

Mémoire présenté devant l'ENSAE Paris
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires
le 08/11/2023

Par : **Antoine Loubeyre**

Titre : **Tarification des traités de réassurance
pour les garanties incapacité et invalidité**

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

Entreprise : MACSF

Nom :

Signature :



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

Directeur du mémoire en entreprise :

Nom : LHARIDON Sebastien

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)**

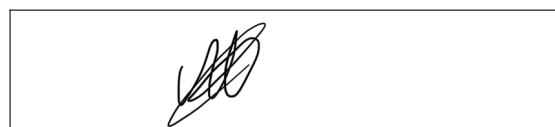
Secrétariat:

Signature du responsable entreprise



Bibliothèque:

Signature du candidat



Résumé

Chaque année, la Mutuelle d'Assurance du Corps de Santé Français (MACSF) doit procéder au renouvellement de ses traités de réassurance pour l'année suivante. Sur le segment prévoyance, qui est l'objet de ce mémoire, ces traités ont pour objectif de limiter la volatilité du résultat et en particulier de limiter la charge pouvant être générée par les expositions élevées sur certaines têtes. Le plan de réassurance est par conséquent composé d'un empilement de traités d'excédent de sinistres par tête.

Afin de négocier au mieux ce renouvellement de réassurance, MACSF modélise les coûts anticipés de son portefeuille d'assurance prévoyance pour l'année à venir afin d'avoir sa propre estimation des tarifs des différents traités. Dans ce cadre, il est particulièrement important de s'intéresser aux sinistres graves susceptibles de dépasser les seuils de réassurance. Ces sinistres sont généralement caractérisés par des individus confrontés à une incapacité de longue durée, détenant plusieurs garanties, ainsi que par les cas où survient une transition vers l'état d'invalidité.

Après avoir introduit de manière générale le produit prévoyance et expliqué les spécificités liées aux risques d'incapacité et d'invalidité, ce mémoire se concentre sur les outils permettant la modélisation des phénomènes de durée et des probabilités de transition.

Pour estimer ces quantités statistiques, nous proposons différentes méthodes statistiques pour estimer les lois d'incidence et de maintien en incapacité, ainsi que la loi de transition de l'état d'incapacité à l'état d'invalidité.

Nous présentons tout d'abord les estimateurs classiques d'Hoem et de Kaplan-Meier. Des méthodes de lissage statistique telles que Whittaker-Henderson ou la méthode à noyau sont ensuite utilisées pour lisser chacune des lois de transition. Les versions univariées et bivariées sont présentées, de même que les méthodes de pondération pour prendre en compte l'exposition. Ensuite, des modèles de régression de type Cox et AFT sont également étudiés afin d'intégrer l'hétérogénéité de notre portefeuille à travers des variables explicatives.

Enfin, des tests statistiques ainsi que des critères d'information sont présentés afin de permettre la validation de nos choix de modèles.

Dans une dernière partie, un rappel sur la réassurance non-proportionnelle est effectué avant de procéder à la tarification des traités proprement dite. Pour cela, les lois calibrées précédemment sont utilisées pour simuler la sinistralité générée par l'exposition de notre portefeuille pour l'année à venir. Pour chaque simulation, nous appliquons nos traités de réassurance non-proportionnels tête par tête à la charge totale simulée (incapacité et invalidité). Grâce à une approche du type Monte-Carlo, il est ainsi possible d'obtenir un tarif de référence pour chaque futur traité de réassurance.

En conclusion, nous mettrons en évidence les différences des résultats de la prime pure obtenue en fonction de la méthode de modélisation sélectionnée. De plus, nous analyserons les impacts d'un changement de traité de réassurance sur les coûts estimés.

Mots clés : Modèles de durée, Kaplan-Meier, modèle de Cox, modèle AFT, lissage Whittaker-Henderson, lissage kernel gaussien, simulation, optimisation, réassurance, traités non-proportionnels, Monte-Carlo.

Abstract

Every year, the French Mutual Insurance Company for Healthcare Professionals (MACSF) must renew its reinsurance treaties for the following year. In the Long Term Care Insurance field, which is the topic of this thesis, these treaties aim to limit result volatility, particularly by mitigating the potential impact of high exposures on certain policyholders. Consequently, the reinsurance plan consists of a stack of Excess of Loss treaties per policyholder.

In order to negotiate this reinsurance renewal optimally, MACSF models the anticipated costs of its long term care insurance portfolio for the upcoming year. This allows them to get their own estimation of the treaties rates. In this context, it is particularly important to focus on severe claims that may exceed the reinsurance thresholds. These claims are typically characterized by individuals facing long-term incapacity period, holding multiple coverages, as well as cases involving a transition to the state of disability.

After providing a general introduction of the long term care insurance product and explaining the features related to incapacity and disability risks, this thesis focus into the tools enabling the modeling of duration phenomena and transition probabilities.

To estimate these statistical quantities, we propose various statistical methods to estimate the incidence and distribution probability of incapacity, as well as the the transition probability from the state of incapacity to the state of disability.

We first present classical estimators Hoem and Kaplan-Meier. Statistical smoothing methods such as Whittaker-Henderson or kernel methods are then used to smooth each of the transition probability. Both univariate and bivariate versions are presented, along with weighting methods to take exposure into account. Subsequently, Cox and AFT regression models are also analysed to include portfolio heterogeneity through explanatory variables.

Finally, statistical tests and information criteria are presented to enable the validation of our model choices.

In a final section, a reminder on non-proportional reinsurance is provided before proceeding with the pricing of the reinsurance treaties. For this purpose, the previously estimated probabilities are used to simulate the claims generated by the exposure of our portfolio for the upcoming year. For each simulation, we apply our non-proportional reinsurance treaties on each policyholder to the total simulated loss (incapacity and disability). Using a Monte Carlo approach, it becomes possible to get a reference rate for each future reinsurance treaty.

In conclusion, we will highlight the differences in the results of the pure premium obtained based on the selected modeling method. Additionally, we will analyze the impacts of a change in reinsurance treaty on the estimated costs.

Keywords : Duration models, Kaplan-Meier, Cox model, AFT model, Whittaker-Henderson smoothing, Gaussian kernel smoothing, simulation, optimization, reinsurance, non-proportional treaties, Monte-Carlo.

Note de synthèse

Tarification des traités de réassurance pour les garanties incapacité et invalidité

Antoine Loubeyre, ENSAE Paris

L'objectif de ce mémoire est de concevoir un modèle de tarification visant à mettre à jour les traités de réassurance de la MACSF pour son produit d'assurance prévoyance. Plus spécifiquement, cette étude se concentre sur les garanties liées à l'incapacité et à l'invalidité. Le processus de construction du modèle comporte deux phases distinctes. Dans un premier temps, nous avons entrepris la modélisation de diverses probabilités de transition d'un état à un autre, ainsi que l'élaboration d'une distribution de durées. Dans un second temps, nous avons procédé à la simulation du portefeuille d'expositions de l'année 2022. Cette dernière étape a permis de générer différents scénarios de sinistralité afin d'obtenir un coût moyen de la charge du portefeuille. Ce montant permet à la direction actuarielle de la MACSF d'ajuster les traités de réassurance pour l'année 2023/2024.

Mots clés : Modèles de durée, Kaplan-Meier, modèle de Cox, modèle Aft, lissage Whittaker-Henderson, lissage kernel gaussien, simulation, optimisation, réassurance, traités non-proportionnels, Monte-Carlo.

Présentation de la construction

Pour établir ce modèle, nous subdivisons plusieurs états distincts. Comme illustré à la figure 1, nous identifions trois issues possibles à partir de l'état initial (état sain). Premièrement, nous avons l'état *Sinistre inférieure à la franchise*¹, représentant des sinistres avec une durée n'excédant pas la période de franchise. Ces événements sont généralement des accidents. Ensuite, l'état *incapacité supérieure à la franchise*² représente la transition due à un sinistre dépassant la franchise, qui peut résulter de maladies ou d'accidents. Enfin, l'état de *grossesse* est également inclus. L'analyse descriptive menée dans ce mémoire révèle la nécessité de cette segmentation, car les probabilités de passage à l'état d'incapacité diffèrent entre ces groupes.

Dans un premier temps, nous cherchons donc à déterminer un taux brut pour le passage vers l'état d'incapacité en fonction de la cause et des diverses variables de segmentation. Comme notre objectif est d'analyser les sinistres ayant une sévérité élevée, nous présentons principalement les résultats des sinistres dépassant leurs franchises. Les sinistres d'une durée inférieure à leurs franchises et ceux causés par des grossesses ne dépassent généralement pas les seuils de la réassurance. La figure montre aussi un état de *rechute* qui sera expliqué par la suite. Enfin, la boucle en bas de l'état *incapacité supérieure à la franchise* indique la durée passée dans cet état. Elle est modélisée par une distribution de probabilité.

Pour finir, nous analysons l'état d'*invalidité* consécutif à l'état d'incapacité. Le calcul de la transition vers cet état est basé sur l'âge auquel survient le premier sinistre et la durée passée en état d'incapacité. Enfin, l'état de *décès* est inclus dans l'illustration à titre indicatif, mais il n'est pas inclus dans la modélisation de ce mémoire.

La construction de la loi de durée du maintien en incapacité est modélisée par deux approches différentes. La première porte sur les **sinistres** et la seconde sur les **arrêts de travail**. La première manière consiste à modéliser la **durée totale du maintien en incapacité**. Pour

¹Sinistres inf FR

²Sinistres sup FR

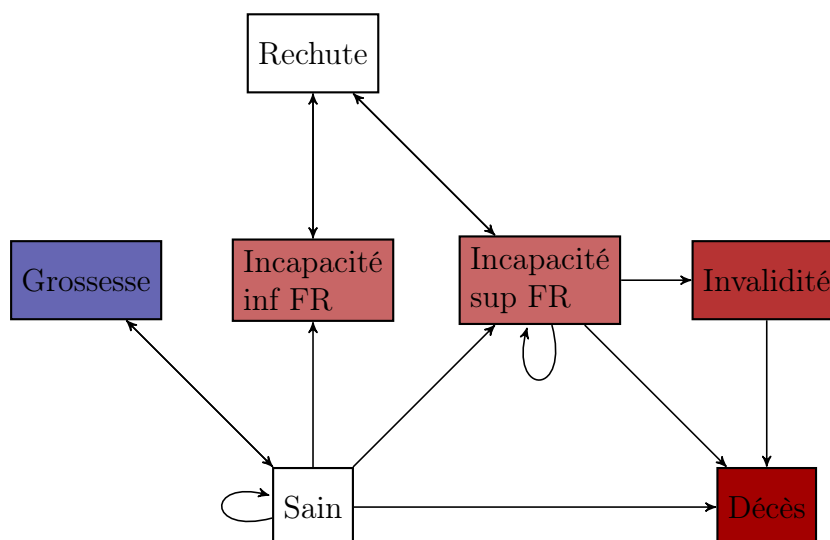


Figure 1: Modèle de tarification complet

cela, nous agrégeons toutes les durées des arrêts par sinistre. La seconde approche concerne la modélisation de la **durée des arrêts de travail** couplée d'une **loi de comptage**. Dans cette partie, nous n'ajoutons aucun traitement sur la durée des arrêts et laissons chaque période d'arrêts d'origine. Afin que les résultats soient cohérents entre les deux méthodes, nous modélisons le **nombre de rechutes** pour la seconde approche. Pour ce faire, nous utilisons un modèle linéaire généralisé de type Binomial Négatif afin d'obtenir une loi de comptage. Cela permet d'obtenir la durée totale d'un sinistre qui est égale à la somme de n durées d'arrêts, où le nombre n provient de la loi de comptage des rechutes. Cette dernière approche est similaire à la construction d'une prime pure à l'aide d'un modèle collectif en Assurance Non-vie.

En ce qui concerne les outils statistiques, nous utilisons des estimateurs unidimensionnels pour la création des taux de passage, à savoir l'estimateur d'Hoem, de Kaplan-Meier et de Nelson-Aalen ainsi que deux modèles économétriques, Cox et AFT. Enfin, des techniques de lissage tel que Whittaker-Henderson et la régression à noyau sont présentes ainsi que des tests de validation de lissage.

Statistiques descriptives et choix de segmentation

Après une analyse approfondie des statistiques descriptives, nous prenons la décision de restreindre notre champ d'étude à la tranche d'âge comprise entre 30 et 60 ans. En-dehors de cet intervalle, les données d'exposition et/ou de passage en incapacité ne sont pas suffisantes pour garantir des estimations significatives. Par ailleurs, notre choix de modélisation se concentre exclusivement sur les individus ayant une période de franchise de **14 jours**, cette option est privilégiée en raison de sa prévalence parmi nos assurés. De plus, nous procédons à la création de catégories socioprofessionnelles, en utilisant un jugement d'expert, aboutissant à la formation de trois groupes professionnels distincts : les *médecins et assimilés*, les *paramédicaux et assimilés*, ainsi que les *infirmiers et assimilés*. Enfin, une variable discriminante basée sur le sexe de l'assuré est intégrée.

Cependant, en ce qui concerne la modélisation du maintien et du passage en invalidité, nous sommes contraints de restreindre le nombre de segments afin de préserver la significativité statistique. Les analyses descriptives relatives à la durée passée en incapacité démontrent que la variable du sexe a moins de pouvoir discriminant que la variable des catégories socioprofessionnelles. Sur cette base, nous choisissons d'écarter la variable du sexe de la modélisation du maintien en incapacité. Par ailleurs, pour estimer le passage en invalidité, nous réduisons

également le nombre de modalités de segmentation. En raison du volume d'environ 700 sinistres d'invalidité, nous prenons la décision de conserver uniquement les variables d'âge et de durée passée en incapacité. De plus, afin d'homogénéiser ces variables, nous optons pour une transformation en classes. Initialement, une tentative d'application d'un algorithme de regroupement (clustering) a été entreprise pour obtenir des catégories homogènes en termes de durées et d'âges, mais les résultats obtenus ne sont pas optimaux. Par conséquent, nous procédons à la création manuelle de classes homogènes.

Passage en incapacité

La figure suivante affiche les résultats des taux d'incidence lissés.

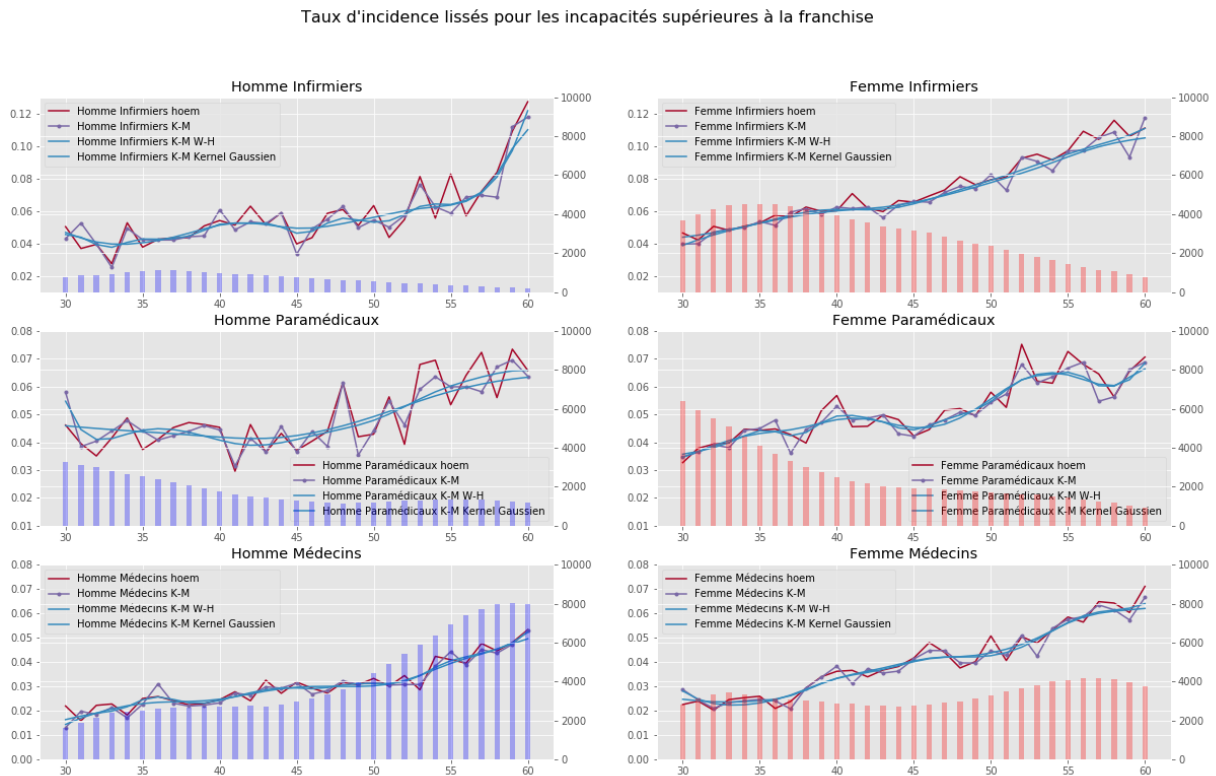


Figure 2: Lissage des taux d'incidence bruts des sinistres supérieurs à la franchise

Nous constatons que l'estimateur de Kaplan-Meier semble démontrer une moindre volatilité par rapport à l'estimateur d'Hoem. Cependant, les deux estimateurs convergent lorsque l'exposition atteint un niveau adéquat (cf les médecins). En ce qui concerne le processus de lissage, nous adoptons deux méthodes distinctes pour déterminer les paramètres optimaux. Pour la régression kernel, nous utilisons la méthode de validation croisée, tandis que pour le lissage de Whittaker-Henderson, nous optons pour une approche graphique, basée sur une détermination itérative des paramètres. Les résultats sont globalement assez similaires. En outre, le test des signes et celui du khi-deux confirment la qualité des lissages. Ceci indique que les procédures de lissage effectuées sur le graphique ci-dessus sont représentatives de la réalité et ajustent de manière cohérente les données brutes. Finalement, la décision est prise de retenir **l'estimateur de Kaplan-Meier avec un lissage de Whittaker-Henderson**. Ces mêmes graphiques ont également été générés pour l'analyse du passage en état de grossesse et en incapacité inférieure à la franchise.

Maintien en incapacité

La prochaine étape consiste à modéliser le maintien en incapacité, pour laquelle nous explorons deux approches distinctes. La première repose sur une méthodologie paramétrique qui peut être réalisée de deux manières différentes. La première consiste à attribuer une loi de durée aux données après avoir préalablement segmenté l'échantillon. La seconde manière nécessite l'utilisation d'un modèle économétrique paramétrique de type "Accelerated Failure Time" (AFT). Dans cette seconde approche, les variables de segmentation jouent le rôle de covariables dans la régression. L'atout du modèle AFT réside dans sa capacité à fournir des informations concernant la signification des variables, leur impact causal sur la variable de durée, ainsi que des évaluations globales de la qualité de la régression.

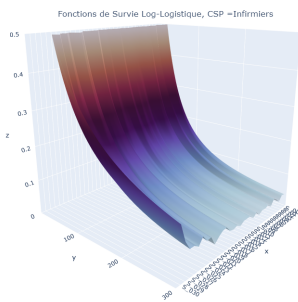


Figure 3: Log Logistique Infirmiers

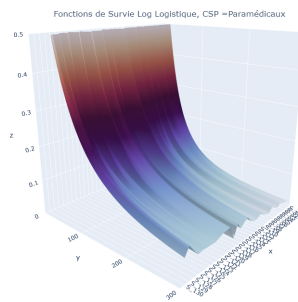


Figure 4: Log Logistique Paramédicaux

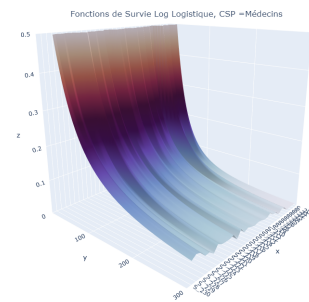


Figure 5: Log Logistique Médecins

La seconde approche, de nature non-paramétrique, implique une estimation à l'aide de l'estimateur de Kaplan-Meier et le modèle semi-paramétrique de type Cox.

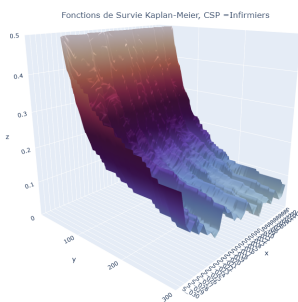


Figure 6: K-M Infirmiers

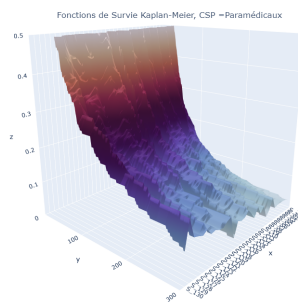


Figure 7: K-M Paramédicaux

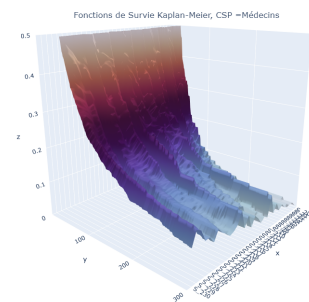


Figure 8: K-M Médecins

Dans les graphiques ci-dessus, la partie supérieure illustre les fonctions de survie de la loi Log-logistique (marginales) pour les diverses catégories professionnelles. Juste en dessous, sont présentées les courbes de survie provenant de l'estimateur de Kaplan-Meier.

En utilisant les critères d'information AIC et BIC, nous concluons que la loi Log-logistique est la mieux adaptée à nos données. Une fois ces lois paramétriques établies, nous les comparons aux modèles "Accelerated Failure Time" (AFT). Cette comparaison révèle des similitudes notables entre les résultats. De plus, toutes les variables explicatives de la régression AFT sont significatives. Cette constatation renforce notre intuition dans la pertinence des variables de segmentation que nous avons choisies.

En ce qui concerne l'approche semi-paramétrique, nous constatons que le modèle de Cox ne respecte pas l'hypothèse de proportionnalité des coefficients. En conséquence, nous décidons de ne pas adopter ce modèle, bien qu'il existe des extensions pour valider cette hypothèse. Pour l'estimateur non-paramétrique, les résultats montrent que les fonctions de survie empirique semblent proches des fonctions de survie théorique. De plus, utiliser une approche non-paramétrique nécessite un temps de calcul bien plus conséquent pour la simulation. Par conséquent, nous optons pour une **approche paramétrique avec la méthode marginale** pour modéliser la durée de maintien en arrêt de travail.

Passage en invalidité

La dernière étape de la modélisation concerne le passage en invalidité. Nous estimons cette probabilité en fonction du temps passé en incapacité et de l'âge de survenance du premier passage. Comme pour les taux d'incidence, nous utilisons l'estimateur de Kaplan-Meier et d'Hoem. La figure suivante présente les résultats des deux estimateurs avant lissage.

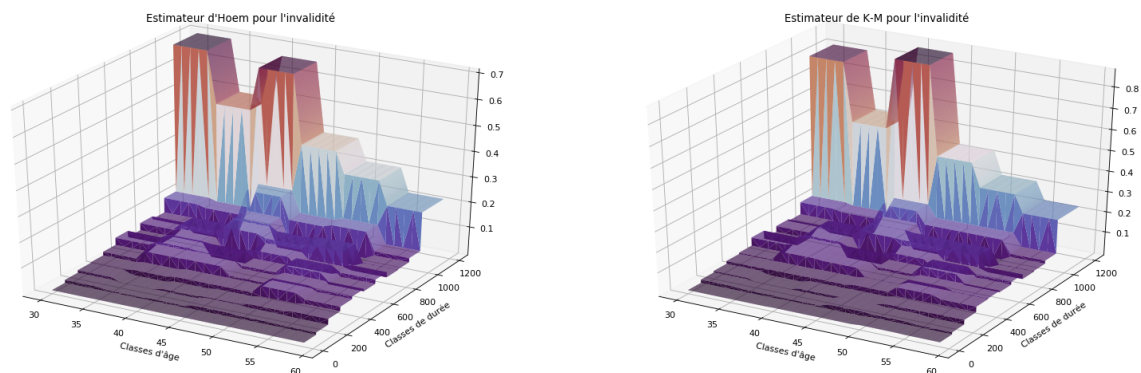


Figure 9: A gauche : estimateur d'Hoem ; A droite : estimateur de Kaplan-Meier

Les résultats montrent une convergence des deux estimateurs. Par la suite, nous appliquons le lissage de Whittaker-Henderson ainsi que la méthode de régression kernel en deux dimensions. À l'issue de l'analyse, nous choisissons de retenir **l'estimateur de Kaplan-Meier avec le lissage de Whittaker-Henderson**.

Modélisation du nombre de rechutes

Pour la deuxième approche concernant la modélisation du maintien en incapacité, nous élaborons une loi de comptage des rechutes. Pour cela, nous utilisons un modèle linéaire généralisé, en supposant une distribution conditionnelle de la loi Binomiale Négative. Ce choix de loi résulte d'une analyse du paramètre de dispersion au sein de notre échantillon. Les résultats montrent la présence d'une sur-dispersion, ce qui nous conduit à privilégier la loi Binomiale Négative à la place de la loi de Poisson. Afin d'obtenir des résultats homogènes, nous lissons les coefficients estimés pour chaque âge. La figure suivante présente les résultats de l'espérance conditionnelle estimée par âge et par CSP.

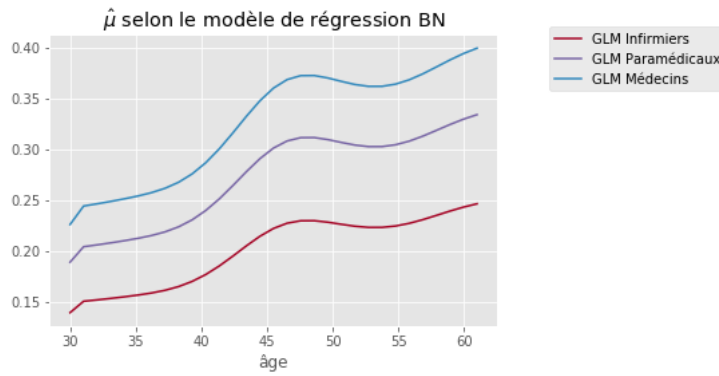


Figure 10: GLM avec une loi Binomiale Négative

La figure montre que la fréquence de rechute est plus élevée pour les médecins, suivie des paramédicaux et enfin les infirmiers.

Simulation et optimisation des traités de réassurance

La dernière partie de ce mémoire concerne la simulation du portefeuille de la MACSF et l'optimisation sous contraintes du traité de réassurance interne.

Une fois chacune des lois de transition et de durée sélectionnées, nous appliquons un algorithme pour simuler notre exposition de l'année 2022. L'objectif de cette étape est, dans un premier temps, de générer autant de scénarios de sinistralité que possibles. Une fois les simulations effectuées, nous calculons la moyenne des simulations afin d'obtenir un coût moyen de sinistralité. Ce montant correspond à la prime pure.

Les résultats de la simulation montrent que la méthode basée sur la modélisation d'une loi de durée totale du sinistre reflète mieux les coûts historiques que l'approche basée sur la durée des arrêts. En raison de ce résultat, nous décidons de conserver la première méthode pour l'optimisation des traités de réassurance.

Optimisation sous contrainte en fonction de la priorité et de l'AAD

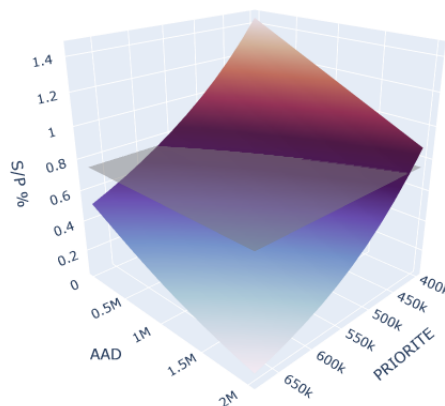


Figure 11: Optimisation du traité de réassurance

La figure 11 représente l'optimisation du traité de réassurance en fonction de la priorité et de l'AAD, sous contrainte d'un S/P fixé. Ce programme nous permet de déterminer les couples de sensibilités (AAD, priorité) optimaux en fonction du S/P .

Conclusion et perspective d'évolution

À travers ce mémoire, nous avons pu élaborer un modèle de tarification pour le produit prévoyance en utilisant différentes techniques statistiques. L'objectif principal était de définir les trois lois de probabilité (taux brut d'incidence, loi de durée de maintien et taux de passage en invalidité) afin de pouvoir simuler le portefeuille de la MACSF et obtenir un coût moyen de sinistralité pour les garanties d'incapacité et d'invalidité par tranche de réassurance.

Pour la construction du taux de passage en incapacité et en invalidité, il a été décidé d'utiliser l'estimateur de Kaplan-Meier avec un lissage de Whittaker-Henderson.

Concernant la modélisation du maintien en incapacité des sinistres et des arrêts, l'analyse des différents critères d'information et des modèles nous a permis de conclure que la loi Log-logistique est la mieux adaptée pour chacune des catégories professionnelles.

Afin de prendre en compte l'effet des rechutes, deux méthodes de modélisation du maintien en incapacité ont été réalisées. L'une portante sur la création d'une loi de durée totale du sinistre et l'autre sur la création d'une loi de durée des arrêts couplée avec une loi de comptage pour les rechutes. Cette dernière permet de ventiler la durée totale d'incapacité en deux composantes. En revanche, elle oblige à faire une hypothèse d'indépendance entre le nombre de rechutes et leur durée. Bien qu'en se basant sur les statistiques descriptives, cette hypothèse semble valide a priori. Les résultats de la simulation montrent que la méthode basée sur la modélisation d'une loi de durée totale du sinistre reflète mieux les coûts historiques que l'approche basée sur la durée des arrêts.

Enfin, le traité de réassurance interne a pu être amélioré grâce à une optimisation sous contraintes. Les résultats ont montré que nous pouvons, dans un premier temps, réduire la priorité afin d'atteindre notre objectif de S/P égal à 0.75. Ensuite, nous avons constaté qu'il est possible de réduire davantage la volatilité en abaissant encore la priorité et en ajoutant une AAD. Pour le traité de réassurance externe, il a été possible de proposer une grille donnant une estimation des chargements de sécurité du réassureur en fonction des différents niveaux de primes proposés. Cette grille pourra être utilisée par la MACSF dans le cadre du renouvellement de ses traités fin 2023.

Limites

Les principales limites auxquelles nous avons fait face concernent la qualité des données. En effet, la reconstitution de la vie de l'ensemble des assurés sur 10 ans a exigé une attention particulière. Enfin, la faible probabilité d'invalidité nécessite un grand nombre de simulations pour obtenir une convergence de la variance dans les tranches de réassurance. Des optimisations du code Python permettraient d'améliorer la convergence de la variance pour les mesures extrêmes.

Perspectives

À travers ce mémoire, nous avons modélisé seulement une partie du modèle présenté en introduction. La perspective finale pour la fonction actuarielle de la MACSF est d'obtenir des lois de durées pour le maintien en invalidité et des taux de transition vers l'état de décès. Ces ajouts permettront d'obtenir une tarification actuarielle complète du produit prévoyance, couvrant les garanties d'incapacité, d'invalidité et de décès.

De plus, un modèle multi-états markovien non-homogène a été modélisé dans le cadre de cette étude. Le modèle n'a pas été présenté dans ce mémoire car, les résultats obtenues n'ont pas été jugé suffisamment robuste. L'avantage de ce modèle serait de mieux prendre en compte la dépendance entre tous les états.

Pour terminer, il pourrait être aussi intéressant d'ajouter à terme la pathologie responsable du passage dans l'état d'incapacité. Avec suffisamment de données, il est probable que cette variable de segmentation apporte davantage d'informations que d'autres variables de segmentation.

Executive Summary

Reinsurance Treaty Pricing for Incapacity and Disability Coverages

Antoine Loubeyre, ENSAE Paris

The aim of this thesis is to design a pricing model to update the reinsurance treaties of MACSF for its long term care insurance product, specifically focusing on incapacity and disability coverages. The construction process of the model consists of two distinct parts. First of all, we undertook the modeling of various transition probabilities from one state to another, as well as the development of a duration probability distribution. Subsequently, we proceeded with the simulation of the portfolio of exposures for the year 2022. This final step generates different claims scenarios in order to get an average portfolio cost. This amount enables the MACSF actuarial team to adjust the reinsurance treaties for the year 2023/2024.

Keywords : Duration models, Kaplan-Meier, Cox model, AFT model, Whittaker-Henderson smoothing, Gaussian kernel smoothing, simulation, optimization, reinsurance, non-proportional treaties, Monte-Carlo.

Construction Overview

To build this model, we divide it into several distinct states. As illustrated in Figure 12, we identify three possible outcomes from the initial state (healthy state). Firstly, we have the state *Claims below the deductible*³, representing claims with a duration not exceeding the deductible period. These events are generally accidents. Next, the state *Incapacity above the deductible*⁴ represents the transition due to a claim exceeding the deductible, which can result from illnesses or accidents. Finally, the state of *pregnancy* is also included. The descriptive analysis conducted in this thesis reveals the necessity of this segmentation, as the transition probabilities to the incapacity state differ among these groups.

Initially, we aim to determine a incidence rate for the transition to the incapacity state based on the cause and various segmentation variables. As our goal is to analyze claims with high severity, we primarily present the results of claims above their deductibles. Claims with durations below their deductibles and those caused by pregnancies generally do not exceed reinsurance thresholds. The figure also shows a state of *relapse*, which will be explained later. Lastly, the loop at the bottom of the *Incapacity above the deductible* state indicates the duration spent in this state, modeled by a probability distribution.

Finally, we analyze the state of *disability* following the incapacity state. The computation of the transition to this state is based on the age at which the first claim occurs and the duration spent in the incapacity state. Lastly, the *death* state is included in the figure for illustrative purposes, but it is not part of the modeling in this thesis.

The construction of the duration distribution for the incapacity period is modeled through two different approaches. The first one concerns **claims**, while the second one focuses on **sick leaves**. The first method involves modeling the **total duration of incapacity period**. To achieve this, we aggregate all sick leave durations per claim. The second approach deals with modeling the **duration of sick leaves** combined with a **counting distribution**. In this part, no adjustments are made to the duration of sick leaves, and the original periods of sick leaves are retained. To ensure consistency between the two methods, we model the **number of relapses** for the second approach. For this purpose, we employ a Negative Binomial generalized

³Claims below DE

⁴Claims above DE

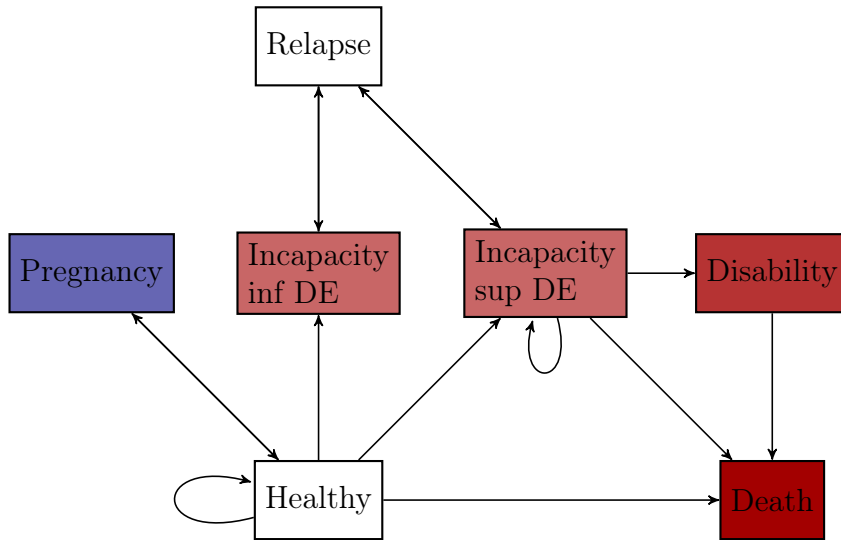


Figure 12: Full pricing model

linear model to obtain a counting distribution. This allows us to calculate the total duration of a claim, which equals the sum of n sick leave durations, where the value of n comes from the relapse counting distribution. This latter approach is similar to compute a pure premium using a collective model in Non-Life Insurance.

Regarding the statistical tools, we use one-dimensional estimators for creating transition rates. These include the Hoem estimator, the Kaplan-Meier estimator, the Nelson-Aalen estimator, as well as two econometric models, Cox and AFT. Finally, smoothing techniques such as Whittaker-Henderson and kernel regression are introduced, along with validation tests for smoothing.

Descriptive Statistics and Segmentation Choices

After an in-depth analysis of the descriptive statistics, we decide to reduce our study to the age range between 30 and 60 years. Outside of this interval, the exposure and/or transition to incapacity data are not sufficient to ensure meaningful estimations. Additionally, our modeling choice exclusively focuses on individuals with a **14-day** deductible period, which is selected due to its prevalence among our policyholders. Furthermore, we create occupation categories using expert judgment, resulting in the formation of three distinct professional groups: *doctors and related professions*, *paramedical and related professions*, as well as *nurses and related professions*. Finally, a discriminant variable based on the gender of the insured is included.

However, regarding the modeling of incapacity period and disability transition, we have to limit the number of segments to preserve statistical significance. Descriptive analyses concerning the duration spent in incapacity show that the gender variable has less discriminatory power compared to the occupation categories variable. Based on this, we choose to exclude the gender variable from the incapacity period modeling. Furthermore, to estimate the disability transition, we also reduce the number of segmentation categories. Given the volume of around 700 disabilities claims, we decide to retain only age and duration spent in incapacity as variables. Additionally, to standardize these variables, we transform them to class. Initially, an attempt to apply clustering algorithms was made to obtain homogeneous categories in terms of durations and ages, but the results obtained were not optimal. Therefore, we proceed with manual creation of homogeneous classes.

Transition to Incapacity

Figure 13 displays the smoothed incidence rates results for the claims database.

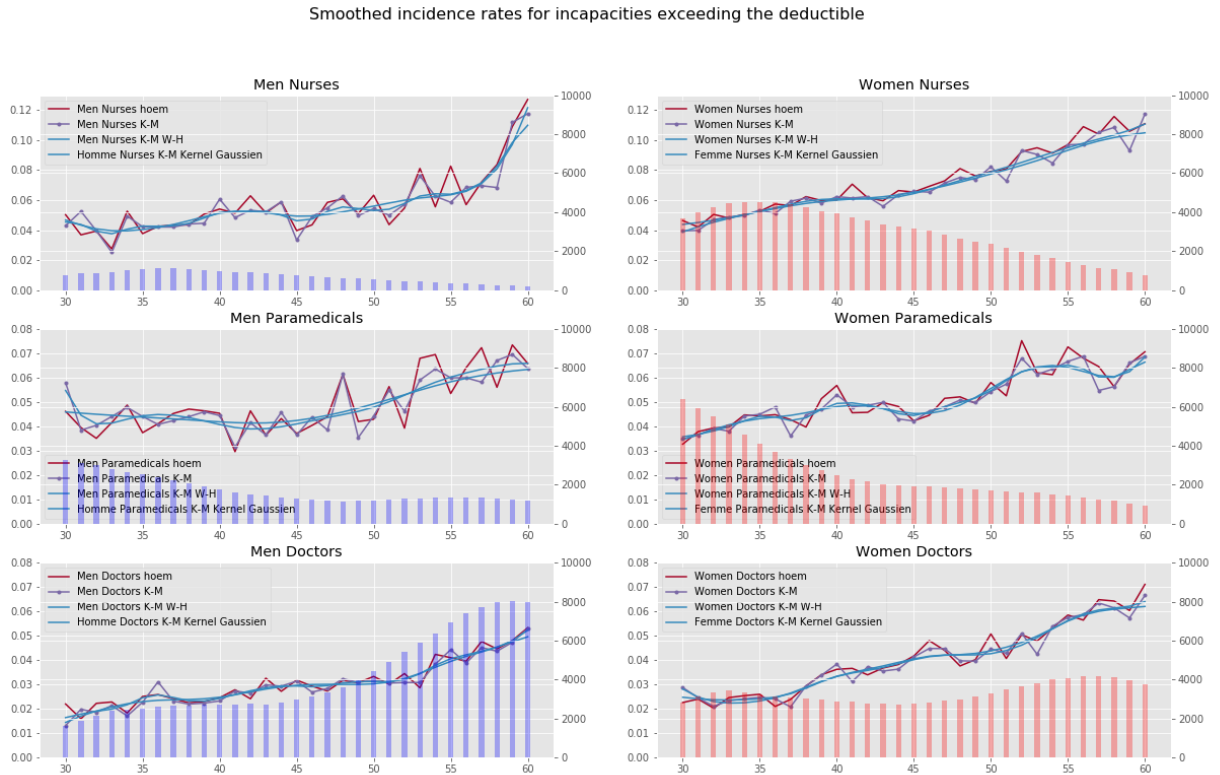


Figure 13: Incidence rate smoothed of claims above their deductible

We observe that the Kaplan-Meier estimator is less volatility compared to the Hoem estimator. However, both estimators converge when the exposure reaches an adequate level (as seen in the case of doctors). Regarding the smoothing process, we employ two distinct methods to determine optimal parameters. For kernel regression, we use cross-validation, while for Whittaker-Henderson smoothing, we opt for a graphical approach based on iterative parameter determination.

The results are generally quite similar. Furthermore, chi-squared and sign tests confirm the quality of the smoothing. This indicates that the smoothing procedures performed on the above graph are representative of reality and consistently adjust the raw data. Ultimately, the decision is made to retain **the Kaplan-Meier estimator with Whittaker-Henderson smoothing**. These same graphs were also generated for the analysis of transition into pregnancy and incapacity for claims below the deductible.

Incapacity Period

The next step involves modeling incapacity period, for which we explore two distinct approaches. The first is based on a parametric methodology that can be executed in two different ways. The first involves assigning a duration distribution to the data after initially segmenting the sample. The second method requires the use of a parametric econometric model of the "Accelerated Failure Time" (AFT) type. In this second approach, the segmentation variables

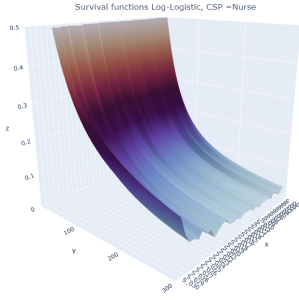


Figure 14: Log Logistique Nurses

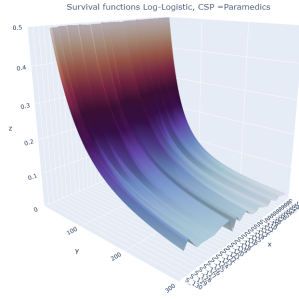


Figure 15: Log Logistique Paramedics

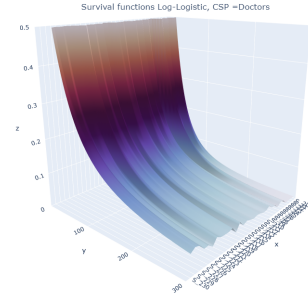


Figure 16: Log Logistique Doctors

is used as covariates in the regression. The advantage of the AFT model is to provide information regarding the significance of variables, their causal impact on the duration variable, as well as overall assessments of the regression quality.

The second approach, of a non-parametric nature, involves estimation using Kaplan-Meier and the semi-parametric Cox model.

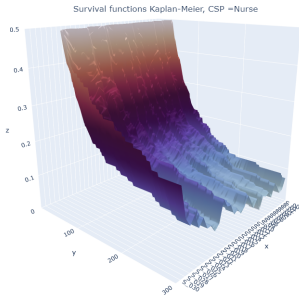


Figure 17: K-M Nurses

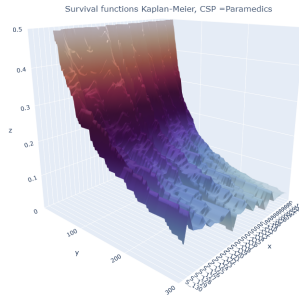


Figure 18: K-M Paramedics

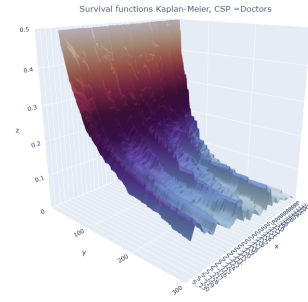


Figure 19: K-M Doctors

In the subsequent graphs, the upper part illustrates the survival functions of the Log-Logistic distribution (marginal) for the various professional categories. Just below, the survival functions from the Kaplan-Meier estimator are presented.

Using the AIC and BIC information criteria, we conclude that the Log-Logistic distribution is the best fit for our data. Once these parametric distributions are established, we compare them to the "Accelerated Failure Time" (AFT) models. This comparison reveals similarities in the results. Moreover, all explanatory variables in the AFT regression are significant. This observation reinforces our confidence in the relevance of the chosen segmentation variables.

Regarding the semi-parametric approach, we find that the Cox model does not satisfy the proportional hazards assumption. As a result, we decide not to adopt this model, even though there are extensions to validate this assumption. For the non-parametric estimator, the results show that the empirical survival functions appear to be close to the theoretical survival functions. Furthermore, employing a non-parametric approach requires significantly more computation time for simulation.

We have therefore opt for a parametric approach using the marginal method to model the incapacity period.

Transition to Disability

The final step of the modeling process concerns the transition to disability. We estimate this probability based on the time spent in incapacity and the age at which the first transition occurs. Similar to the incidence rates, we use both the Kaplan-Meier and Hoem estimators. The following figure presents the results of these two estimators before smoothing.

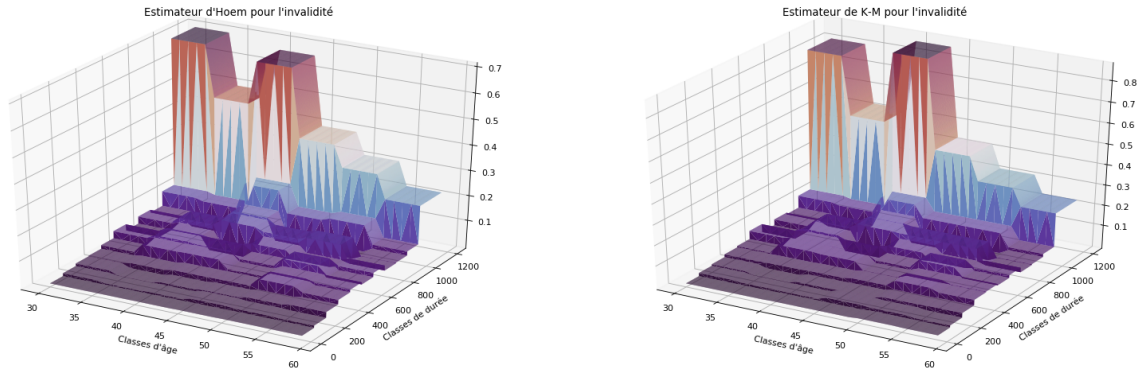


Figure 20: Left : Hoem estimator ; Right : Kaplan-Meier estimator

The results demonstrate a convergence of the two estimators. Subsequently, we apply Whittaker-Henderson smoothing as well as the two-dimensional kernel regression method. Following the analysis, we choose to retain **the Kaplan-Meier estimator with Whittaker-Henderson smoothing**.

Modeling the Number of Relapses

For the second approach regarding the modeling of incapacity period, we develop a counting distribution for relapses. To achieve this, we employ a generalized linear model, assuming a conditional distribution of the Negative Binomial distribution. In order to get homogeneous results, we smooth the estimated coefficients for each age. The following figure presents the results of the estimated conditional mean by age and occupation categories.

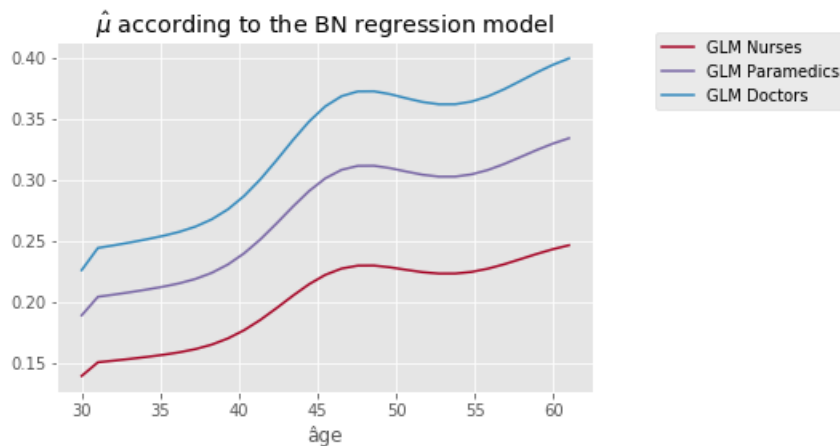


Figure 21: GLM with Negative Binomial distribution

The figure shows that the relapse frequency is higher for doctors, followed by paramedical professionals, and finally nurses.

Simulation and optimisation of reinsurance treaties

The last part of this thesis involves the simulation of MACSF's portfolio.

Once each of the transition and duration distributions is selected, we apply an algorithm to simulate our exposure for the year 2022. The goal of this step is, firstly, to generate as many claims scenarios as possible. After the simulations are performed, we calculate the average of the simulations to obtain an average claims cost. This amount corresponds to the pure premium.

The results of the simulation show that the method based on modeling a total loss duration distribution better reflects historical costs than the approach based on the duration of sick leaves. Due to this result, we decide to retain the first method for the optimization of reinsurance treaties.

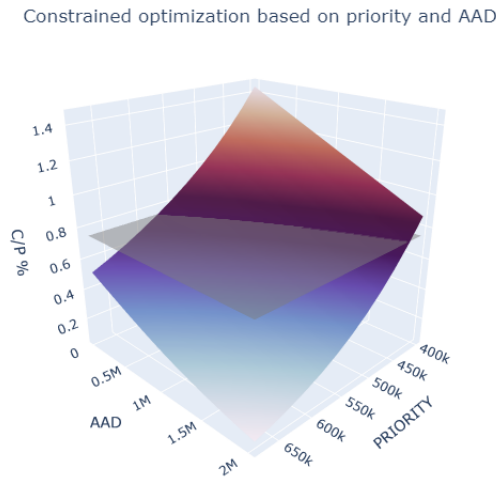


Figure 22: Optimisation of reinsurance treaties

Figure 22 illustrates the optimization of the reinsurance treaty as a function of priority and AAD, under the constraint of a fixed C/P . This program allows us to determine the optimal sensitivity couple (AAD, priority) based on the C/P .

Conclusion and Perspectives

Through this thesis, we have been able to develop a pricing model for the long-term care insurance product using various statistical techniques. The main objective was to define the three probability distributions (incidence rate, duration distribution of incapacity period, and transition rate to disability) in order to simulate MACSF's portfolio and get an average claims cost for the incapacity and disability coverages per reinsurance layer.

For constructing the transition rate to disability and invalidity, it was decided to use the Kaplan-Meier estimator with Whittaker-Henderson smoothing.

Regarding the modeling of the disability claims period and sick leaves, the analysis of various information criteria and models allowed us to conclude that the Log-logistic distribution is the most suitable for each occupational category.

To account for the effect of relapses, two modeling methods for incapacity period were conducted. One focused on creating a total duration distribution for the claim, and the other on creating a duration distribution for sick leaves coupled with a counting distribution for relapses. The latter allows for the breakdown of the total duration of incapacity into two

components. However, it requires assuming independence between the number of relapses and their duration. Although, based on descriptive statistics, this assumption appears valid a priori. The simulation results show that the method based on modeling a total claim duration distribution better reflects historical costs than the approach based on leave duration.

Finally, the internal reinsurance treaty was improved through constraint optimization. The results showed that, initially, we can reduce the priority to achieve our C/P objective of 0.75. Subsequently, we observed that it is possible to further reduce volatility by lowering the priority further and adding an Annual Aggregate Deductible (AAD). For the external reinsurance treaty, we were able to propose a grid providing an estimate of the reinsurer's security loadings based on different premium levels. This grid can be used by MACSF during the renewal of its treaties in late 2023.

Limitations

The main limitations we faced relate to data quality. Indeed, reconstructing the lives of all policyholders over a period of 10 years required special attention. Furthermore, the low probability of disability requires a large number of simulations to achieve variance convergence within the reinsurance tranches. Optimizations of the Python code would help improve variance convergence for extreme measures.

Future Directions

Through this thesis, we have only modeled a part of the introduced model. The ultimate perspective for the actuarial function at MACSF is to obtain duration distribution for disability period and transition rates to the death state. These additions will provide a complete actuarial pricing of the long-term care insurance product, covering disability, incapacity, and death coverages.

Additionally, a non-homogeneous multi-state Markov model was considered in this study. However, the results obtained were not deemed sufficiently robust, so it was not presented in this thesis. The advantage of this model would be to better account for the dependence between all states.

To conclude, it could also be interesting to consider adding the pathology responsible for the transition to the incapacity state in the future. With enough data, it is likely that this segmentation variable would provide more information than other segmentation variables.

Remerciements

Je tiens tout d'abord à remercier chaleureusement mon tuteur Sébastien L'HARIDON, qui m'a permis d'effectuer une alternance enrichissante et passionnante. Grâce à lui, j'ai pu découvrir le fonctionnement de l'assurance prévoyance dans sa globalité ainsi que les aspects actuariels sous-jacents. Cela m'a permis de mettre en application toute la théorie des modèles de durée acquise à l'ENSAE, tout en étant challengé par les contraintes pratiques de l'entreprise. Les conseils de Sébastien et la qualité de son suivi de mes travaux m'ont permis de réaliser ce mémoire en actuariat. Ce fut un réel plaisir d'avoir ce maître de stage, à la fois professionnellement et personnellement.

Je tiens également à remercier le manager de l'équipe, Pierre-François MARCASTEL. Ses conseils et son expertise technique m'ont permis de me challenger et d'améliorer la qualité de mon mémoire.

Je souhaite exprimer ma gratitude envers toute l'équipe de la fonction actuarielle de la MACSF pour les moments partagés en interne et en externe, ainsi que pour leurs conseils tout au long de mon alternance. Ce fut un réel plaisir de réaliser cette expérience professionnelle au sein de cette équipe chaleureuse et motivante.

Sommaire

1	Présentation	1
1.1	Introduction	1
1.2	L'assurance prévoyance	2
1.2.1	L'assurance prévoyance	2
1.2.2	Système d'indemnisation	2
1.2.3	Le marché de la prévoyance en France	5
1.3	Modélisation	6
1.3.1	Illustration du modèle de tarification	6
1.3.2	Sinistres et arrêts de travail	9
2	Théorie des phénomènes de durée	11
2.1	Caractéristiques	11
2.1.1	Principes	11
2.1.2	Censure et Troncature	14
2.2	Statistiques non-paramétriques	17
2.2.1	Estimateurs	17
2.2.2	Techniques de lissage	24
2.3	Statistiques paramétrique	30
2.3.1	Estimateurs	31
2.3.2	Techniques de lissage	34
2.4	Modèle avec variables explicatives	36
2.4.1	Modèle de Cox	36
2.4.2	Modèle AFT	39
2.4.3	Modèle GLM	42
2.5	Synthèse et transition	46
3	Calibration sur le portefeuille MACSF	49
3.0.1	La mutuelle MACSF	49
3.0.2	Présentation des données	51
3.1	Calibration sur des sinistres	56
3.1.1	Statistiques descriptives	56
3.1.2	Passage en incapacité	64
3.1.3	Maintien en incapacité	69
3.1.4	Passage en invalidité	82
3.1.5	Synthèse sur la calibration des sinistres	84
3.2	Calibration sur des arrêts	85
3.2.1	Statistiques descriptives	85
3.2.2	Maintien en incapacité	86
3.2.3	Modélisation des rechutes	91
4	Tarification de traités de réassurance prévoyance	94
4.1	Concepts	94
4.2	Application	99
5	Conclusion	107
6	Annexe	111
6.1	Théorie des phénomènes de durée	111
6.2	Calibration sinistres	114

6.3	Calibration arrêts	123
6.4	Lissage	130
6.5	Réassurance	132

1 Présentation

1.1 Introduction

Chaque année, la Mutuelle d'Assurance du Corps de Santé Français (MACSF) doit procéder au renouvellement de ses traités de réassurance pour l'année suivante. Sur le segment prévoyance, qui est l'objet de ce mémoire, ces traités ont pour objectif de limiter la volatilité du résultat et en particulier de limiter la charge pouvant être générée par les expositions élevées sur certaines têtes.

L'objectif de ce mémoire consiste donc à modéliser les sinistres d'une sévérité élevée, pouvant potentiellement atteindre les différents seuils de réassurance. Pour évaluer les coûts probables qu'ils pourraient engendrer, nous examinons l'historique de leurs sinistralités ainsi que les coûts associés. Ces derniers peuvent découler de deux causes distinctes. La première concerne les paiements d'indemnités journalières pour chaque jour d'arrêt de travail, tandis que la seconde englobe les rentes versées en cas d'invalidité de l'assuré.

Afin d'estimer le coût probable résultant de la somme de ces deux sources de paiement, il est essentiel de déterminer préalablement la probabilité qu'un individu déclenche le paiement d'une indemnité journalière et/ou d'une rente d'invalidité. Par conséquent, plusieurs questions doivent être abordées : Quelles sont les probabilités qu'un individu déclenche le paiement d'indemnités journalières ? Quelle est la durée passée en arrêt de travail ? Quelle est la probabilité qu'un individu déclenche une rente d'invalidité ? et quel est le coût probable résultant de ces paiements cumulés ?

Pour répondre aux premières questions, nous utilisons des méthodes statistiques pour estimer les probabilités qu'un individu déclenche un arrêt de travail suivi d'une invalidité. Pour cela, nous subdivisons le problème en trois parties distinctes. La première porte sur l'estimation de la probabilité qu'un individu entre dans l'état d'incapacité. Une fois qu'il y est, nous modélisons le nombre probable de jours pendant lesquels il reste en arrêt de travail. Enfin, la probabilité de passage en invalidité sachant qu'il est en incapacité est estimée. En pratique, plus un individu reste longtemps en arrêt de travail, plus il a de fortes chances de déclencher une invalidité.

L'objectif est donc d'analyser les sinistres pouvant potentiellement rester longtemps en incapacité, car le cumul des indemnités journalières et de la rente d'invalidité a de fortes chances de dépasser les seuils de la réassurance.

Afin d'estimer les coûts probables des garanties d'incapacité et d'invalidité pour l'année à venir, nous procédons à des simulations de sinistralité. Pour ce faire, nous récupérons notre portefeuille de l'année en cours et générons différentes simulations à l'aide d'un algorithme de Monte-Carlo en utilisant les probabilités calculées précédemment. Sur chacune des simulations individuelle, nous appliquons les traités de réassurance tête par tête. Le résultat obtenu pour l'ensemble du portefeuille représente un coût total probable. En calculant la moyenne des simulations, nous obtenons une estimation du coût pour chacune des tranches de réassurance. Cette moyenne correspond à la prime pure théorique à laquelle nous devons ajouter un coefficient de chargement de sécurité pour obtenir la prime pure commerciale.

Pour des raisons de **confidentialité**, les données présentées dans ce mémoire ont été volontairement modifiées.

1.2 L'assurance prévoyance

Cette section détaille les différents risques de prévoyance et leur fonctionnement.

1.2.1 L'assurance prévoyance

Selon la loi n° 89-1009 du 31 décembre 1989, dite loi EVIN, la prévoyance regroupe "*les opérations ayant pour objet la prévention et la couverture du risque décès, des risques portant atteinte à l'intégrité physique de la personne ou liés à la maternité, des risques d'incapacité de travail ou d'invalidité ou du risque chômage*".

L'assurance prévoyance permet donc de se protéger contre les conséquences financières liées aux accidents de la vie quotidienne. Elle offre plusieurs catégories de garanties :

- La garantie Incapacité Temporaire Totale (ITT) : Elle prévoit une indemnisation en cas d'incapacité de travail d'une durée supérieure à la franchise et inférieure à 1095 jours (3 ans).
- La garantie Invalidité Permanente Total (IPT) : Cette garantie prévoit le versement d'une rente en cas d'invalidité empêchant l'assuré de travailler. Les prestations sont ajustées en fonction du taux d'invalidité défini par la sécurité sociale.
- La garantie emprunteur : Elle concerne la prise en charge partielle ou totale du crédit restant dû de l'assuré en cas de perte de revenus.
- Autres garanties existantes : Dépendance, Décès, Accidents de la vie, Obsèques, ect...

Dans ce mémoire, nous nous concentrerons principalement sur les garanties incapacités, invalidités et emprunteurs.

1.2.2 Système d'indemnisation

La prévoyance joue un rôle de complément à la sécurité sociale. En cas d'accident de la vie, la sécurité sociale ne couvre qu'une partie du salaire de l'assuré, généralement jusqu'à 50 %. Le reste étant à la charge de l'assuré. La partie couverte par la sécurité sociale s'appelle le **régime obligatoire** et la partie restante correspond au **régime complémentaire**.

Depuis le 1er juillet 2021, l'organisme en charge de la sécurité sociale des professions libérales est la Caisse Nationale d'assurance vieillesse des professions libérales (**CNAVPL**). Il s'agit d'un organisme français de droit privé chargé de la gestion du régime d'assurance vieillesse de base des professions libérales et de la gestion des réserves de ce régime. La CNAVPL regroupe la majorité des professionnels de santé⁵.

Cet organisme d'assurance maladie regroupe dix caisses de retraite. Il y a la CARCDSF (Caisse autonome de retraite des chirurgiens-dentistes et des sages-femmes), la CARMF (Caisse autonome de retraite des médecins de France), la CARPIMKO (Caisse autonome de retraite et de prévoyance des infirmiers, masseurs-kinésithérapeutes, pédicures-podologues, orthophonistes et orthoptistes), la CARPV (Caisse autonome de retraites et de prévoyance des vétérinaires), la CAVAMAC (Caisse d'allocation vieillesse des agents généraux et des mandataires non salariés de l'assurance et de la capitalisation), la CAVEC (Caisse d'assurance vieillesse des experts-comptables et des commissaires aux comptes), la CAVOM (Caisse d'assurance vieillesse des officiers ministériels, des officiers publics et des compagnies judiciaires), la CAVP (Caisse

⁵Définition Wikipédia

d'assurance vieillesse des pharmaciens), la CIPAV (Caisse interprofessionnelle de prévoyance et d'assurance vieillesse. Cette caisse regroupe les architectes, les ingénieurs, techniciens, géomètres, experts, conseils consultants) et la CPRN (Caisse de prévoyance et de retraite des notaires).

La figure 23 présente les indemnités des différents organismes en fonction de la durée d'arrêt de travail de l'assuré. Nous remarquons que la CNAVPL indemnise les arrêts de travail du 4e au 90e jour, soit une période de 87 jours d'indemnités journalières. Cette indemnisation est **commune** à tous les professionnels de santé. Durant cette période, la gestion opérationnelle est confiée aux caisses primaires d'assurance maladie (CPAM).

% du salaire indemnisé

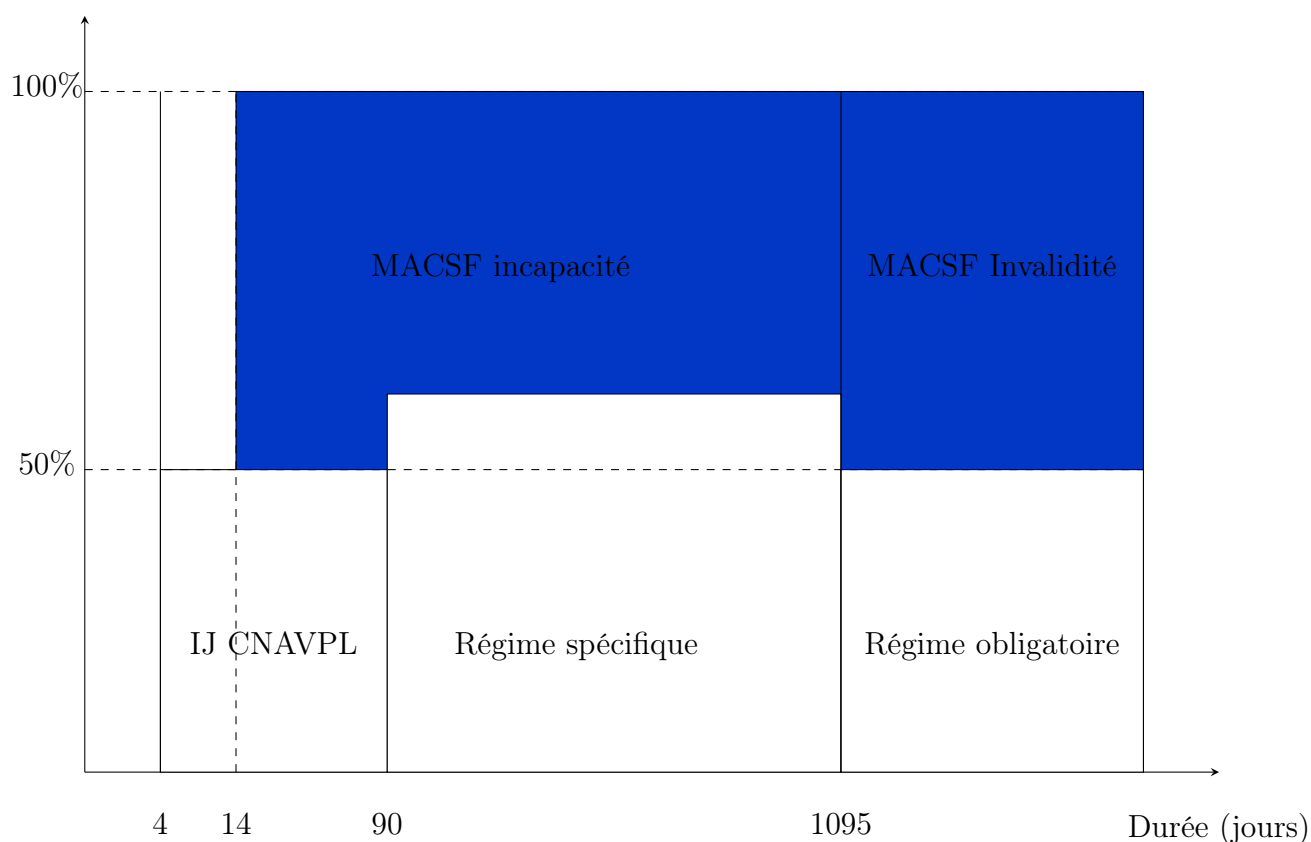


Figure 23: Indemnisation prévoyance

L'indemnisation de l'incapacité du 4ieme au 90ieme jours correspond à 1/730e de la moyenne des revenus annuels des trois dernières années, dans la limite de trois fois le plafond annuel de la Sécurité sociale (PASS). En 2021, cela se traduit par un montant minimum de 22,54€ par jour et un montant maximum de 169€ par jour. Les trois premiers jours correspondent au délai de carence, une période durant laquelle l'assuré n'est pas indemnisé. Ce délai de carence est prévu entre la signature du contrat et la prise d'effet des garanties.

Après 90 jours d'arrêt de travail, le versement des indemnités journalières dépend de la caisse professionnelle à laquelle l'assuré est affilié. Seules la CARMF, la CARPIMKO, la CAVEC et la CARCDSF offrent une indemnisation après le 90ieme jour d'arrêt de travail. Par exemple, la CARMF offre une indemnisation selon la classe de l'assuré. Cette classe dépend du revenu de l'individu. L'indemnisation offerte par la caisse CARPIMKO offre une indemnisation de

55.44 euros/jour. La CAVEC offre une indemnité de 110 euros bruts par jour. Concernant la CARCDSF, l'indemnisation est de 102.58 € pour les chirurgiens-dentistes et de 43.34 € par jour pour les sage-femme.

Il convient également de noter qu'un individu peut rester dans un état d'incapacité pendant une période maximale de 1095 jours (soit 3 ans). Au-delà de cette période, il est considéré comme invalide.

En cas d'invalidité, chaque caisse de retraite indemnise ses cotisants selon différents critères. Par exemple, la CARMF indemnise ses assurés par classe (A, B, C) déterminée en fonction du niveau de revenu. Pour la CARCDSF, la rente d'invalidité est fixe et est fixée à 30 003.90 € pour les chirurgiens-dentistes. Les sages-femmes sont indemnisées de 12 449€/an. Pour la caisse CARPIMKO, les cotisants reçoivent 1 680€/mois.

De plus, la MACSF prend également en charge une partie de l'indemnisation en fonction du degré d'invalidité, et ce, jusqu'à l'âge de départ à la retraite.

Pour chaque caisse d'assurance maladie, la MACSF propose des plans de prévoyance associés à chaque profession de santé. Dans chaque contrat de prévoyance proposé, il existe deux types de garanties, une portant sur la sphère privée et l'autre sur la sphère professionnelle.

La première propose une indemnisation pour la protection des proches de l'assuré. Cela peut être un capital décès, une rente de conjoint ou bien une rente d'éducation. La seconde concerne les indemnisations pour une incapacité ou bien une invalidité.

Chacune des garanties, que ce soit pour la sphère privée ou publique, possède ses propres caractéristiques définies par les variables suivantes :

- Montant de garantie minimal et maximal
- Franchise
- Durée minimale et maximale
- Age limite d'adhésion
- Age limite d'indemnisation

Remarque importante : Lorsque l'arrêt de travail est causé par un **accident**, la MACSF indemnise l'assuré dès le premier jour d'arrêt. En revanche, en cas d'hospitalisation, l'indemnisation débute à partir du 4e jour si la durée de séjour est inférieure ou égale à 21 jours et 7 jours dans le cas contraire. Dans le cas d'une maladie n'entraînant pas d'hospitalisation, la franchise débute au 8e ou 15e jour, selon les garanties du contrat.

La souscription à un contrat de prévoyance diffère selon que l'individu soit salarié d'une entreprise ou travailleur indépendant. En France, les entreprises ont l'obligation de souscrire un contrat de prévoyance pour leurs employés, conformément à la convention collective. Ce contrat comprend généralement, au minimum, la garantie décès. En revanche, les travailleurs indépendants tels que les entrepreneurs ou les professionnels libéraux ne sont pas tenus de souscrire un contrat de prévoyance. La souscription est donc facultative. Ce mémoire porte sur les professionnels libéraux ayant donc souscrit un contrat de prévoyance individuel.

1.2.3 Le marché de la prévoyance en France

Selon les données de France Assureurs, le nombre d'affaires nouvelles de contrats Madelin prévoyance⁶ pour le premier semestre de l'année 2022 a augmenté de 5.9% par rapport à l'année précédente. De plus, le nombre de contrats en cours à la fin de juin 2022 a augmenté de 3.1%, dont 1.3% concerne les garanties incapacité et invalidité.

Les cotisations ont également augmenté de 5.2% par rapport au premier semestre de l'année 2021, avec une hausse de 4.4% pour l'incapacité et l'invalidité. Enfin, les prestations versées pour le premier semestre 2022 sont estimées à 461.1 millions d'euros, soit une baisse de 3.4% par rapport au premier semestre de l'année 2021.

En analysant l'historique des études menées par France Assureurs, nous constatons que le marché de la prévoyance ne cesse de progresser année après année.

	2021		2022	
	Année	Var/2020	1 ^{er} Sem	Var/ 1 ^{er} sem 2021
Nbr affaires nouvelles (milliers)	145.4	+15.5%	85	+5.9%
Nbr contrats en cours (adhésion, milliers)	1 127	+0.7%	1 153.8	+3.1%
<i>dont gar. incapacité-invalidité (milliers)</i>	819	+0.5%	831.1	+1.3%
Cotisations (millions)	1 679.3	+4.8%	1 037.9	+5.2%
<i>dont gar. incapacité-invalidité (milliers)</i>	1 062.1	+5.5%	665.7	+4.4%
Prestations versées (millions)	889.9	-11.2%	461.1	-3.4%
<i>dont gar. incapacité-invalidité (milliers)</i>	633.1	-14.6%	334.0	-4.4%

Table 1: Source : France Assureurs

D'une manière générale, depuis 2018, le nombre de cotisations ne cesse de croître, passant de 1462 milliers d'euros à 1731 milliers d'euros en juin 2022. Cela implique donc une croissance continue du nombre de contrats en cours, passant ainsi de 1111 mille contrats en 2018 à 1154 milliers en juin 2022. À titre de comparaison, le niveau de cotisation pour la garantie dépendance est plutôt stable voire a subi une légère baisse depuis 2019. En revanche, pour la garantie obsèques, nous observons une hausse de 2.2 % des cotisations par rapport à l'année 2021.

La section suivante présente le modèle de tarification qui sera utilisé pour modéliser le risque d'incapacité et d'invalidité.

⁶Contrats de prévoyance destinés aux indépendants ainsi qu'à leur conjoint collaborateur.

1.3 Modélisation

Pour rappel, l'objectif de ce mémoire est de construire un modèle de tarification afin de renouveler les traités de réassurance pour le produit prévoyance. Dans cette section, nous présentons de manière générale les modèles qui seront utilisés dans la partie consacrée à la modélisation.

1.3.1 Illustration du modèle de tarification

Pour illustrer notre modélisation des garanties d'incapacité et d'invalidité, nous allons définir, pour le moment, trois états :

- Etat sain : Un individu est dans l'état sain lorsqu'il n'a subi aucun événement et qu'il est toujours exposé aux risques.
- Etat d'incapacité : Un individu se trouve dans l'état d'incapacité lorsqu'il subit un événement qui le rend incapable de travailler temporairement.
- Etat d'invalidité : Une fois qu'un individu est dans l'état d'incapacité, il peut passer dans l'état d'invalidité soit en raison de l'aggravation de sa maladie, soit parce qu'il a atteint la durée maximale dans l'état d'incapacité. Il est important de rappeler que la durée maximale dans l'état d'incapacité est de 3 ans.

Pour mieux comprendre ces trois états, nous pouvons les représenter graphiquement à l'aide de la figure 24 suivante. Comme nous pouvons le remarquer, un individu dans l'état sain peut passer à l'état d'incapacité et ensuite revenir à l'état sain. Une fois dans l'état d'incapacité, l'individu peut évoluer vers l'état d'invalidité. Cependant, il est rare qu'un individu puisse revenir à l'état sain ou à l'état d'incapacité une fois qu'il est entré dans l'état d'invalidité.

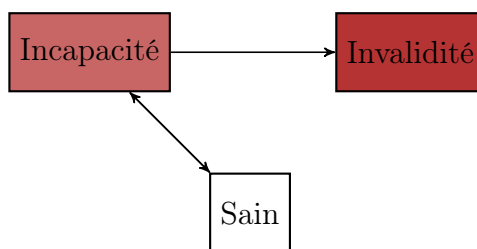


Figure 24: Modèle à 3 états

Le modèle à trois états nous donne une première idée de la modélisation à effectuer, mais il ne représente qu'une partie de la réalité. En pratique, il existe d'autres états possibles. Par exemple, nous pouvons considérer l'état de grossesse, qui représente les individus incapables de travailler en raison d'une grossesse. Les états d'incapacité et d'invalidité peuvent eux, être subdivisés en plusieurs sous-états. L'état d'incapacité peut être divisé en deux catégories : l'incapacité liée à un sinistre d'une durée inférieure à la franchise⁷ et celui supérieure à la franchise⁸. Quant à l'état d'invalidité, il peut être subdivisé en trois sous-états en fonction du type d'invalidité. En France, il existe trois niveaux d'invalidité définis par la sécurité sociale. Chaque niveau représente un certain degré de dépendance.

En outre, nous pouvons ajouter un état de "décès" et un état de "rechute". L'état de rechute permet de distinguer la probabilité de tomber dans l'état d'incapacité pour la première fois à celle d'y retomber à nouveau pour le même sinistre.

Ces différents états et sous-états enrichissent la modélisation en prenant en compte la complexité de la réalité.

⁷Incapacité inf FR

⁸Incapacité sup FR

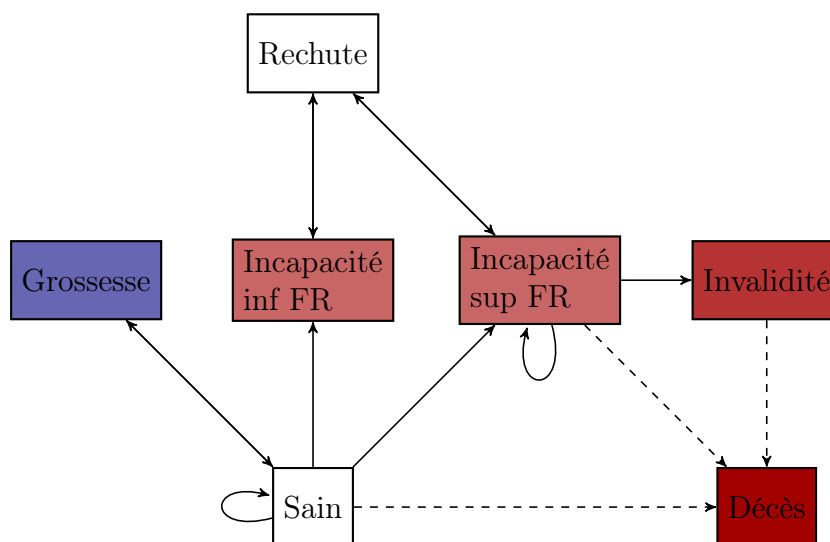


Figure 25: Modèle multi-états complet

Comme le montre la figure 25, il existe de nombreuses probabilités de transition d'un état à un autre. Dans le cadre de ce mémoire, nous n'avons pas subdivisé les états d'invalidité en trois sous-états, mais avons considéré un seul état d'invalidité. De plus, nous n'avons pas modélisé le passage à l'état de décès. Notre étude s'est donc portée sur tous les états hors décès.

En assurance et notamment en actuariat vie et en prévoyance, nous nous intéressons particulièrement aux lois **d'incidence**, de **maintien en incapacité** et de **passage en invalidité**.

- La loi d'incidence concerne la probabilité de passage de l'état sain (ou rechute) à l'état d'incapacité.
- La loi de maintien en incapacité représente la probabilité de rester en état d'incapacité. En général, elle est représentée par une table à deux dimensions (âge d'entrée en incapacité x durée dans l'état) et ayant pour valeur la probabilité de rester dans l'état d'incapacité.
- La loi de passage à l'état d'invalidité représente la probabilité de passer de l'état d'incapacité à l'état d'invalidité.

Ces lois sont principalement utilisées pour le provisionnement et la simulation, comme c'est le cas dans ce mémoire. Elles sont généralement représentées soit en termes de nombre d'individus, soit en termes de probabilité.

En pratique, pour réaliser des modèles de tarifications ou des modèles de provisionnement, il existe deux approches. La première consiste à utiliser les tables du B.C.A.C (Bureau Commun d'Assurances Collectives), définies par une institution agréée. La construction de ces tables est basée sur la population nationale. L'inconvénient donc de ces tables est qu'elles ne représentent pas le risque associé à notre portefeuille d'assurés. En effet, d'une part, le risque de nos assurés n'est pas égal au risque de la population nationale et d'autre part, le risque de tomber en incapacité et ou en invalidité est différent selon les professions de santé. L'utilisation de ces tables peut introduire un biais d'estimation significatif dans l'évaluation du risque de notre portefeuille.

La deuxième approche consiste à créer des tables d'expériences basées sur notre propre portefeuille. C'est cette dernière approche qui sera développée dans ce mémoire.

La construction de la loi de durée du maintien en incapacité est modélisée par deux approches différentes. La première porte sur les **sinistres** et la seconde sur les **arrêts de travail**.

La première manière consiste à modéliser la **durée totale du maintien en incapacité**. Pour cela, nous agrégeons toutes les durées des arrêts par sinistre. La seconde approche concerne la modélisation de la **durée des arrêts de travail** couplée d'une **loi de comptage**. Dans cette partie, nous n'ajoutons aucun traitement sur la durée des arrêts et laissons chaque période d'arrêts d'origine. Afin que les résultats soient cohérents entre les deux méthodes, nous modélisons le **nombre de rechutes** pour la seconde approche. Pour ce faire, nous utilisons un modèle linéaire généralisé de type Binomial Négatif afin d'obtenir une loi de comptage. Cela permet d'obtenir la durée totale d'un sinistre qui est égale à la somme de n durées d'arrêts, où le nombre n provient de la loi de comptage des rechutes. Cette dernière approche est similaire à la construction d'une prime pure à l'aide d'un modèle collectif en Assurance Non-vie.

Pour la construction de chacune des lois, nous avons d'abord segmenté notre échantillon en fonction de différentes modalités avant d'implémenter nos estimateurs. Ensuite, nous avons adopté une approche économétrique. Ce type de modèle permet d'intégrer l'hétérogénéité de notre portefeuille à travers des variables explicatives. De plus, cette approche permet d'estimer l'effet causal d'une variable explicative sur la variable à expliquer et de vérifier si la segmentation créée lors de la première approche est statistiquement significative.

Une fois les lois estimées, nous avons simulé notre portefeuille de l'année 2022. La moyenne de nos simulations a permis d'obtenir une valeur de référence pour ajuster nos traités de réassurance pour l'année 2023/2024. Le processus de tarification général réalisé est récapitulé dans la figure suivante.

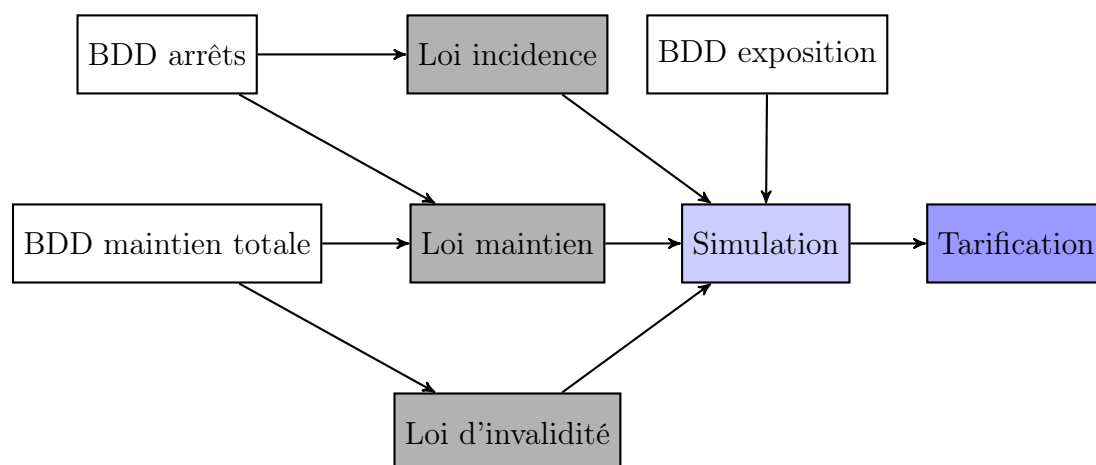


Figure 26: Processus de tarification

Remarque : La base de données des arrêts représente la base de données d'indemnisation initiale. La base de données de maintien totale est issue de la BDD des arrêts où nous utilisons seulement les périodes d'incapacités. De plus, les durées des arrêts ont été agrégées par sinistre.

L'illustration suivante détaille l'algorithme utilisé selon la méthode utilisée pour la construction de la loi du maintien en incapacité. A gauche de la figure, la loi de la durée totale du maintien en incapacité est utilisée alors qu'à droite, se trouve le couplage de la loi des arrêts avec la loi de comptage.

La section suivante explique la différence entre la notion de sinistre et celle des arrêts de travail.

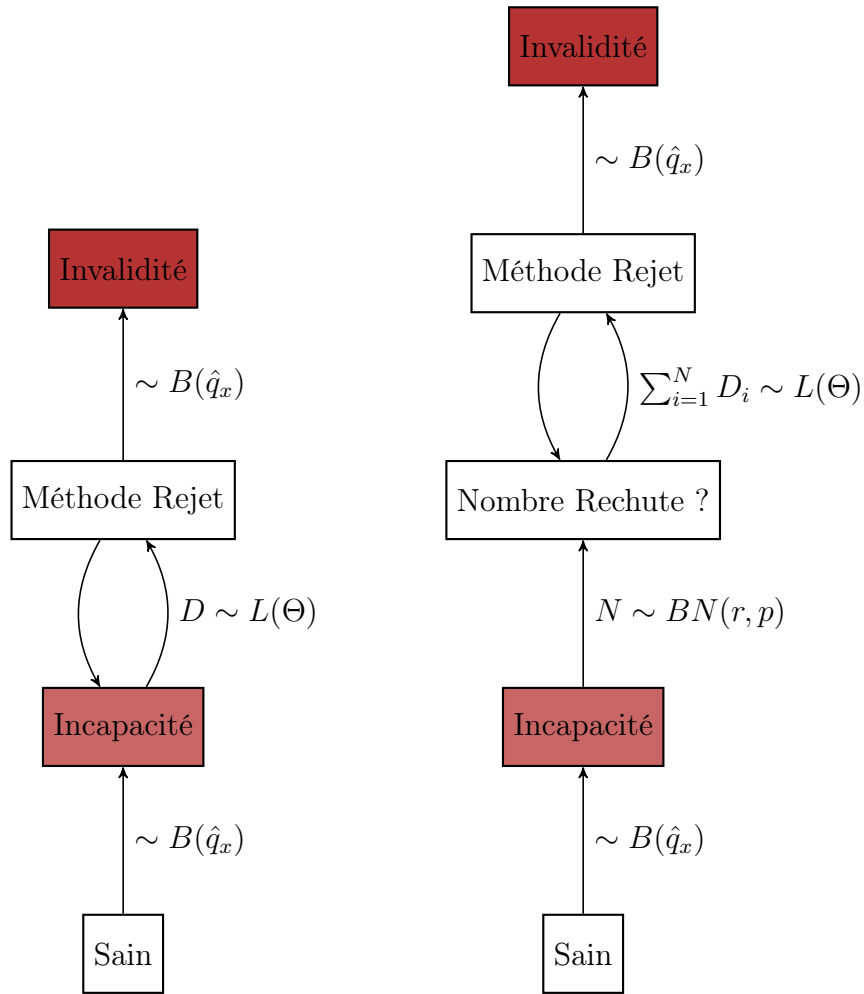


Figure 27: Illustration de l'algorithme selon la loi de maintien utilisée

1.3.2 Sinistres et arrêts de travail

Dans le cadre de ce mémoire, nous avons utilisé deux méthodes pour modéliser le maintien en incapacité. La première consiste à analyser la durée totale d'un sinistre, tandis que la seconde se concentre sur la modélisation des arrêts de travail.

Lorsqu'un individu est en arrêt de travail, il déclenche un sinistre auprès de son assurance. Pour illustrer ce cas, prenons l'exemple d'un individu qui attrape une maladie le rendant incapable de travailler pendant plusieurs semaines. Après une période inférieure à 1095 jours, il peut retrouver son "aptitude", c'est-à-dire être en mesure de travailler comme auparavant. Toutefois, une fois dans l'état "sain", il est courant que l'individu fasse des rechutes, c'est-à-dire qu'il retombe en arrêt de travail pour la même cause que la première incapacité. Ce scénario peut se répéter à plusieurs reprises.

Nous utiliserons donc le terme de **sinistre** pour désigner l'arrêt de travail dans sa globalité, englobant toutes les rechutes éventuelles.

Quant à l'**arrêt de travail**, il se réfère à chaque passage en incapacité de travail associée à un même sinistre. Ainsi, un sinistre comprend au moins un arrêt de travail.

Ces distinctions nous permettront de mieux appréhender les différents aspects et événements liés aux arrêts de travail dans notre analyse.

La figure 44 illustre le parcours potentiel d'un assuré de notre portefeuille.

Au début de la période, l'individu est dans l'état sain, c'est-à-dire qu'il est en état de travailler. Il reste dans cet état pendant une durée D_1 avant de passer à l'état d'incapacité de

travail pour diverses raisons. Après une période de temps D_2 , l'assuré se rétablit et retourne à l'état sain. Malheureusement, les causes de la première période d'arrêt de travail réapparaissent et l'individu subit une rechute en retombant en incapacité. Cela est représenté par une nouvelle période d'incapacité d'une durée de D_4 . Après un certain temps, l'individu retourne à l'état sain. Cependant, il tombe à nouveau malade, mais cette fois-ci de manière plus grave, et passe à l'état d'invalidité à la fin de la troisième période d'incapacité, soit à la fin de la durée D_6 .

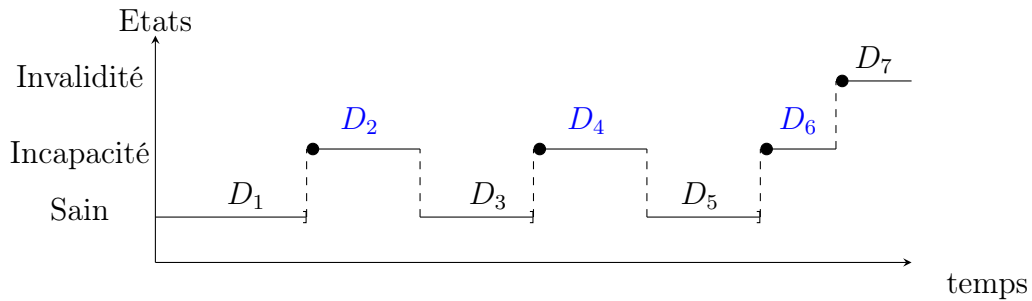


Figure 28: Arrêts de travail

Les constructions des bases de données de sinistres et d'arrêts de travail seront détaillées dans le chapitre dédié à la calibration des lois.

2 Théorie des phénomènes de durée

L'objectif de cette section est de présenter les outils statistiques qui nous permettront de calibrer les différentes lois évoquées en introduction.

2.1 Caractéristiques

Pour construire les trois lois de probabilité (taux d'incidence, loi de maintien, passage en invalidité), il est nécessaire de faire appel à la théorie des phénomènes de durée. Par exemple, pour calculer la durée de maintien en incapacité, nous allons analyser une variable aléatoire T qui est positive ou nulle.

Dans le cas de la construction de la loi d'incidence, nous allons étudier la durée passée dans l'état sain avant le passage dans l'état d'incapacité. Or, il se peut qu'un individu ne passe jamais dans cet état et reste sain pendant toute la durée de l'étude. Ce type de phénomène est appelé censure.

Une autre caractéristique concerne la troncature des données. En prenant le cas du maintien en incapacité, nous allons observer une variable aléatoire T à condition qu'elle soit supérieure à la franchise de l'individu. En effet, nous observons la durée d'incapacité d'un individu à condition que la durée de son arrêt de travail dépasse sa franchise. Avant cette durée, nous n'avons aucune information sur la distribution des données. Cela vient du fait qu'un individu en arrêt de travail pour une durée inférieure à sa franchise ne recevra aucune indemnisation, et par conséquent, il ne signalera pas son sinistre à la MACSF. Cette particularité sera aussi à prendre en compte dans nos données.

Pour récapituler, nous avons deux catégories de caractéristiques à prendre en compte pour construire un modèle de tarification pour les garanties d'incapacité et d'invalidité :

1. $T \geq 0$
2. Censure et troncature

2.1.1 Principes

Comme évoqué précédemment, nous analysons une durée, ce qui se traduit par l'analyse d'une variable aléatoire à valeurs dans l'intervalle des réels positifs \mathbb{R}_+ . Cela nous oblige donc à supposer (dans un cadre paramétrique) des distributions spécifiques comme par exemple la loi de Weibull, la loi Gamma ou encore la loi Log-logistique. De plus, une variable aléatoire est généralement représentée par sa fonction de densité (ou de masse) ou bien sa fonction de répartition. Dans le cas de la modélisation de durée, il est courant d'utiliser la fonction de hasard (ou de risque) ou la fonction de hasard cumulée.

La première fonction peut être interprétée de la manière suivante : si nous considérons que T représente la durée de vie d'un individu, alors la fonction de risque indique la probabilité qu'un individu décède dans la journée donnée, sachant qu'il est en vie le matin. Cette fonction est utilisée pour le calcul de l'estimateur de Kaplan-Meier.

La fonction de risque cumulée représente l'accumulation du taux de risque instantané et est utilisée pour le calcul de l'estimateur de Nelson-Aalen.

Ces quantités statistiques sont représentées de la manière suivante :

$$\forall t \geq 0$$

- Fonction de densité : $f(t) = \lim_{h \rightarrow 0} \frac{\mathbb{P}(t \leq X < T+h)}{h} = F'(t) = -S'(t)$
- Fonction de répartition : $F(t) = \mathbb{P}(T \leq t) = 1 - S(t)$
- Fonction de survie : $S(t) = \mathbb{P}(X > t)$
- Fonction de hasard (taux de risque instantané) : $\lambda(t) = \lim_{h \rightarrow 0} \frac{\mathbb{P}(t \leq X < T+h | X \geq t)}{h} = \frac{f(t)}{S(t)} = -\ln(S(t))'$
- Fonction de hasard cumulée : $\Lambda(t) = \int_0^t \lambda(u) du = -\ln(S(t))$

Ces fonctions possèdent un lien entre elles. Par exemple, la fonction de densité peut être obtenue en dérivant la fonction de répartition, ou inversement. La figure 29 illustre le lien entre chacune des fonctions :

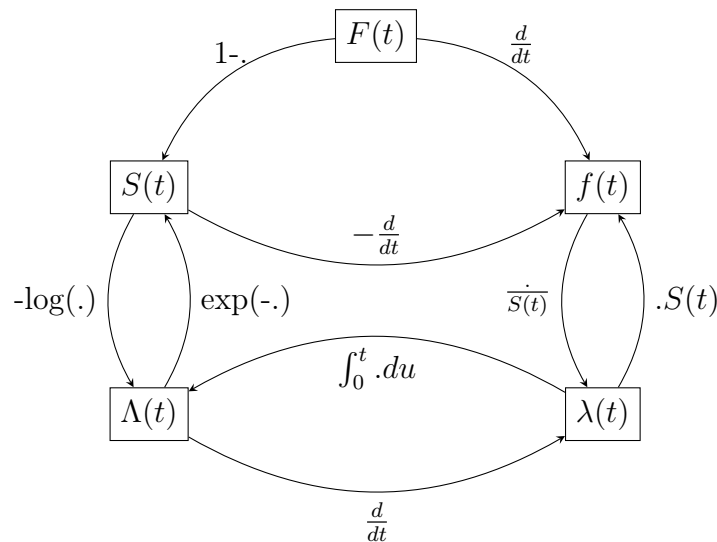


Figure 29: Lien entre les différentes fonctions d'une variable aléatoire

La deuxième particularité de l'analyse des phénomènes de durée concerne l'absence d'observation de l'événement étudié (censure) et l'absence d'information complète sur une partie de l'échantillon (troncature). Par exemple, si nous souhaitons observer la durée de vie d'un individu, nous devons attendre son décès. Toutefois, il arrive souvent que l'étude ou l'observation prenne fin avant que l'individu ne décède. Dans de tels cas, nous savons seulement que l'individu a survécu jusqu'à la fin de l'étude, mais nous ne connaissons pas la durée exacte de sa vie. Il est donc nécessaire de corriger ce biais d'observations incomplètes.

Exemple illustratif d'un individu en arrêt de travail

La figure 30 illustre les principales situations auxquelles un individu peut être confronté dans nos bases de données. Dans ce graphique, l'entrée de l'individu dans l'état d'incapacité est représentée par une étoile bleue, et la durée associée est représentée par une ligne en pointillés de la même couleur. Enfin, le passage à l'état d'invalidité est symbolisé par un point noir.

Pour une meilleure compréhension de l'exemple, il est important de distinguer deux durées distinctes. La première correspond à la durée passée dans l'état sain, notée T_{sain} , et la seconde représente la durée passée en état d'incapacité, notée $T_{incapacite}$.

Lorsque nous analysons la durée T_{sain} , l'événement étudié est le passage en incapacité. Si un individu sort de l'étude sans avoir connu cet événement, ou s'il n'est jamais passé dans l'état

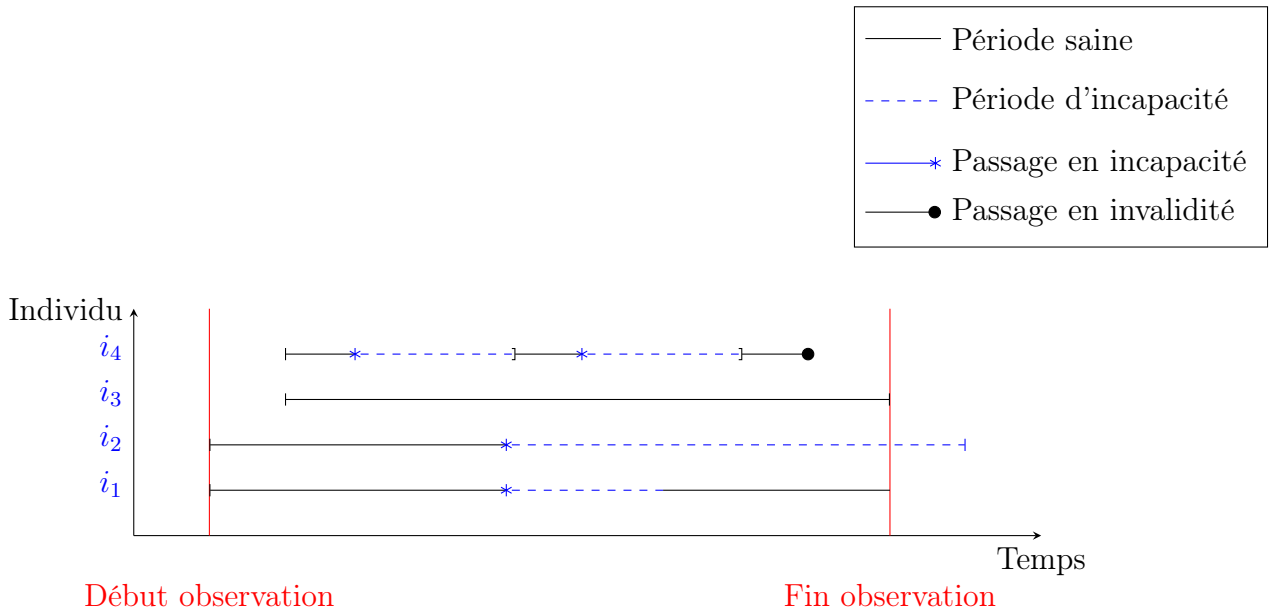


Figure 30: Phénomènes de censure et de troncature

d'incapacité pendant la période d'étude, il sera considéré comme censuré, car nous n'aurons aucune information sur la durée exacte passée dans l'état sain.

En ce qui concerne l'analyse de la durée $T_{incapacite}$, l'évènement étudié est soit le retour à l'état sain et/ou soit le passage à l'état d'invalidité. La table 2 récapitule les différentes causes de sorties et de censures associées à chaque durée.

Lois	Durée	Evènement	Censure
Incidence	T_{sain}	Passage incapacité	Aucun passage en incapacité
Maintien	$T_{incapacit}$	Toutes sorties	Durée = Durée maximale indemnisation
Invalidité	$T_{incapacit}$	Passage invalidité	Aucun passage en invalidité

Table 2: Récapitulatif des causes d'évènements et de censures

Le schéma 30 illustre le parcours de l'individu i_1 à travers trois périodes de durée distinctes. Une première période dans l'état sain, notée T_{sain} , une période en incapacité, $T_{incapacite}$, puis une seconde période dans l'état sain à nouveau.

Pour l'analyse de la première période, T_{sain} , nous n'avons aucune censure, car l'évènement de passage en état d'incapacité s'est bien réalisé. De plus, la période en état d'incapacité a pu être analysée dans son intégralité, car nous observons la transition du départ de l'état d'incapacité vers l'état sain. Cela nous permet donc de connaître exactement la durée passée dans cet état.

En revanche, la deuxième période dans l'état sain pose problème, car nous n'observons pas la durée réelle passée dans cet état. Nous disposons uniquement de la date d'entrée, c'est-à-dire la date à partir de laquelle l'individu a quitté l'état d'incapacité pour redevenir sain. L'information sur la durée de cette période est donc censurée à droite.

Pour l'individu i_2 , nous constatons que sa durée d'incapacité n'est pas entièrement observée. En effet, à la fin de l'observation, l'individu est toujours, en théorie, dans cet état. Dans ce cas, nous ne connaissons pas la durée exacte de son incapacité. Nous disposons uniquement de l'information selon laquelle l'individu est toujours en incapacité à la date de fin d'observation. Nous sommes donc en présence d'une censure à droite.

En ce qui concerne l'individu i_3 , il apparaît dans l'étude à une date ultérieure par rapport

au début de l'observation. Avant cela, nous n'avons aucune information sur cet assuré. Cela correspond à une troncature à gauche, qui est une caractéristique spécifique de l'échantillon de base. Ce phénomène est courant en assurance en raison de la présence de franchises dans les garanties. Par exemple, pour l'analyse du maintien en incapacité, nous observons l'individu conditionnellement au fait que sa durée soit supérieure à sa franchise. La différence entre censure et troncature sera détaillée dans la section suivante.

Pour conclure, l'individu i_4 représente la majorité des cas lorsque nous étudions le passage et le maintien en incapacité. Cet individu entre dans l'échantillon après la date de début d'observation (troncature à gauche), puis passe en arrêt de travail à deux reprises (ou subit une rechute) avant de finalement passer dans l'état d'invalidité. Dans le cas de l'étude des arrêts de travail, cet individu ne présente aucune censure à droite, car nous observons les durées exactes.

La section suivante présente en détail les phénomènes de censure et de troncature dans le contexte des modèles de durée.

2.1.2 Censure et Troncature

Dans l'analyse des phénomènes de durée, il est essentiel de prendre en compte les biais liés aux censures et aux troncatures. Comme nous l'avons vu dans les exemples précédents, la durée d'un événement de durée T n'est généralement pas entièrement observable. Dans notre étude, par exemple, un individu peut résilier son contrat, décéder ou atteindre la durée maximale d'indemnisation pendant la période d'observation. De plus, certains individus peuvent entrer dans l'étude seulement si leur durée d'incapacité dépasse la franchise de leur contrat. Tous ces facteurs limitent l'observation complète de la durée en état sain ou en état d'incapacité.

Mathématiquement, nous observons donc $(Y_1, \delta_1), \dots, (Y_n, \delta_n)$ tel que :

$$\begin{aligned} Y_i &= \inf(T_i, C_i) \\ \delta_i &= \mathbb{1}_{T_i \leq C_i} \\ (Y_i, \delta_i) &\text{ a pour loi } (Y, \delta) \text{ sachant } Y \geq \tau \end{aligned}$$

où T représente la durée réelle du phénomène étudié (période saine ou période d'incapacité), C la durée censurée, c'est-à-dire la durée observée en cas de censure à droite. δ est une indicatrice qui prend la valeur 1 en cas d'observation complète (c'est-à-dire si nous observons T) et 0 sinon. τ représente la troncature à gauche.

Censure

Il existe principalement trois types de censure :

- Censure type 1 : Dans ce cas, la censure est déterministe, c'est-à-dire que la durée censurée C_i est fixée à l'avance. Cela se produit lorsque nous avons préalablement décidé d'analyser la durée sur une période spécifique $[0, C]$. Ainsi, nous n'observons que les durées inférieures à la valeur de censure fixée.
- Censure type 2 : Dans ce cas, les durées (T_1, \dots, T_n) sont triées dans l'ordre croissant et la censure est représentée par $C_i = C = T_{(r)}$. Par exemple, on peut décider d'arrêter l'analyse après avoir observé r événements se produire. Pour les événements restants,

nous ne connaissons que le fait que leur durée sera supérieure à celle du r-ème événement observé.

- Censure de type 3 : Dans ce cas, les variables aléatoires C_i sont indépendantes et identiquement distribuées (i.i.d.). Dans notre étude, nous observons une censure à droite car nous ne faisons aucune hypothèse particulière. Nous n’observons que la censure à droite, ce qui biaise l’information concernant la partie droite de la distribution de T .

Dans le cadre de notre étude, nous constatons des censures de type 3 lorsque nous examinons les périodes saines. En effet, un individu peut sortir de l’étude ou ne pas passer en incapacité et ce, à tout moment de l’étude. En revanche, pour l’analyse du maintien en incapacité, nous observons une censure de type 1 car la durée d’indemnisation maximale est fixée et connue à l’avance.

Remarque : Il est également important de mentionner l’existence de phénomènes de censure par intervalle et de censure à gauche. La censure par intervalle se produit lorsque la durée exacte de l’événement n’est pas connue, mais nous avons des informations sur l’intervalle de temps dans lequel elle se situe. La censure à gauche, quant à elle, survient lorsqu’un événement s’est déjà produit avant que l’observation ne débute, et nous ne disposons que de cette information ainsi que de la date de début d’observation.

Mathématiquement, nous observons donc $(Y_1, \delta_1), \dots, (Y_n, \delta_n)$ tel que :

$$Y_i = \text{Sup}(T_i, C_i)$$

$$\delta_i = \mathbb{1}_{T_i \geq C_i}$$

Troncature

La principale différence entre la censure et la troncature réside dans la façon dont les données sont observées. Dans le cas de la censure, nous avons des informations sur les individus qui ont subi un événement avant la fin de l’observation, ce qui nous permet de compter le nombre d’individus censurés. En revanche, dans le cas de la troncature, nous observons les données uniquement pour les individus dont la durée dépasse un seuil spécifié, ce qui signifie que nous ne pouvons pas compter les individus dont la durée est inférieure à ce seuil.

Prenons l’exemple de notre étude, où nous avons une troncature à gauche. Nous observons les durées de maintien uniquement⁹ pour les individus dont la durée dépasse la franchise. Les durées inférieures à la franchise ne sont, en général, pas observables dans notre étude.

Comme pour la censure, il existe plusieurs catégories de troncature :

- Troncature à gauche : dans ce cas, les durées T_i sont observables seulement si elles sont supérieures à une valeur seuil τ . Cela signifie que nous ne disposons pas d’informations sur les durées inférieures à τ . Dans notre étude, nous observons la durée conditionnellement au fait que la durée dépasse la franchise associée à la garantie du contrat.
- Troncature à droite : inversement, les durées T_i sont observables seulement si elles sont inférieures à une valeur seuil τ . Les durées supérieures à τ ne sont pas observables.

⁹Pour certains individus, nous observons une durée qui est inférieure à leur franchise, car il s’agit probablement d’accidents, et donc leur franchise appliquée est de 0 jour.

- Troncature par intervalle : dans ce cas, la durée est tronquée à la fois à gauche et à droite. Par exemple, nous pourrions souhaiter étudier uniquement les individus dont le salaire se situe entre une certaine valeur minimale et une valeur maximale.

Dans la section suivante, nous présentons les biais générés en cas d'utilisation d'estimateurs ne prenant pas en compte ces deux caractéristiques.

Des méthodes d'estimation spécifique à l'analyse des variables de durée

Comme nous l'avons vu à travers les quatre situations présentées en introduction de cette section, estimer les caractéristiques (moments, fonction de survie, fonction de hasard, etc.) de la variable de durée T , nécessite de prendre en compte les biais causés par les phénomènes de troncature et de censure.

Il est important de noter que l'estimation de l'espérance mathématique par la moyenne empirique conduirait à des résultats biaisés.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n Y_i$$

$$\bar{x} \xrightarrow{n \rightarrow \infty} E[X]$$

En effet, nous ne pouvons observer directement la durée complète T_i de l'individu i , mais plutôt une version tronquée $Y_i = \min(T_i, C_i)$. Cela signifie que si les données ne sont pas censurées, nous observons la durée réelle T_i , sinon nous observons la valeur de censure C_i . Si nous utilisons simplement l'estimateur de la moyenne empirique, nous sous-estimons la vraie valeur de l'espérance de la variable aléatoire T . C'est pourquoi nous avons besoin d'estimateurs prenant en compte ce biais spécifique.

La figure 31 met en évidence les différents biais résultant de l'absence de prise en compte de la censure et/ou de la troncature. Dans chaque illustration, nous avons généré des données suivant une loi Gamma avec des paramètres $X \sim \Gamma(2, 0.1)$, et nous avons également associé cette loi aux données empiriques.

Remarque : Les histogrammes en bleu représentent la vraie densité empirique des observations. Lorsque nous tronquons les données à gauche, la forme de l'histogramme (en rouge) est modifiée, car la hauteur des bâtons varie en fonction du nombre d'observations présentes dans l'échantillon.

La figure située en haut à gauche représente une loi Gamma associée aux données empiriques sans censure ni troncature. Nous remarquons ainsi que la loi théorique correspond parfaitement aux données empiriques.

Pour la simulation présentée en haut à droite, nous avons délibérément tronqué nos données en dessous de 14. Cela permet d'observer les conséquences de ne pas tenir compte d'une **franchise de 14 jours**. Nous constatons que lorsque la troncature n'est pas prise en compte dans l'estimation des paramètres, la loi théorique obtenue commence à partir de la valeur 14. En revanche, lorsque la troncature est prise en compte, la loi obtenue arrive à modéliser les données complètes.

Illustration de la loi Gamma Généralisée avec de la censure et troncature

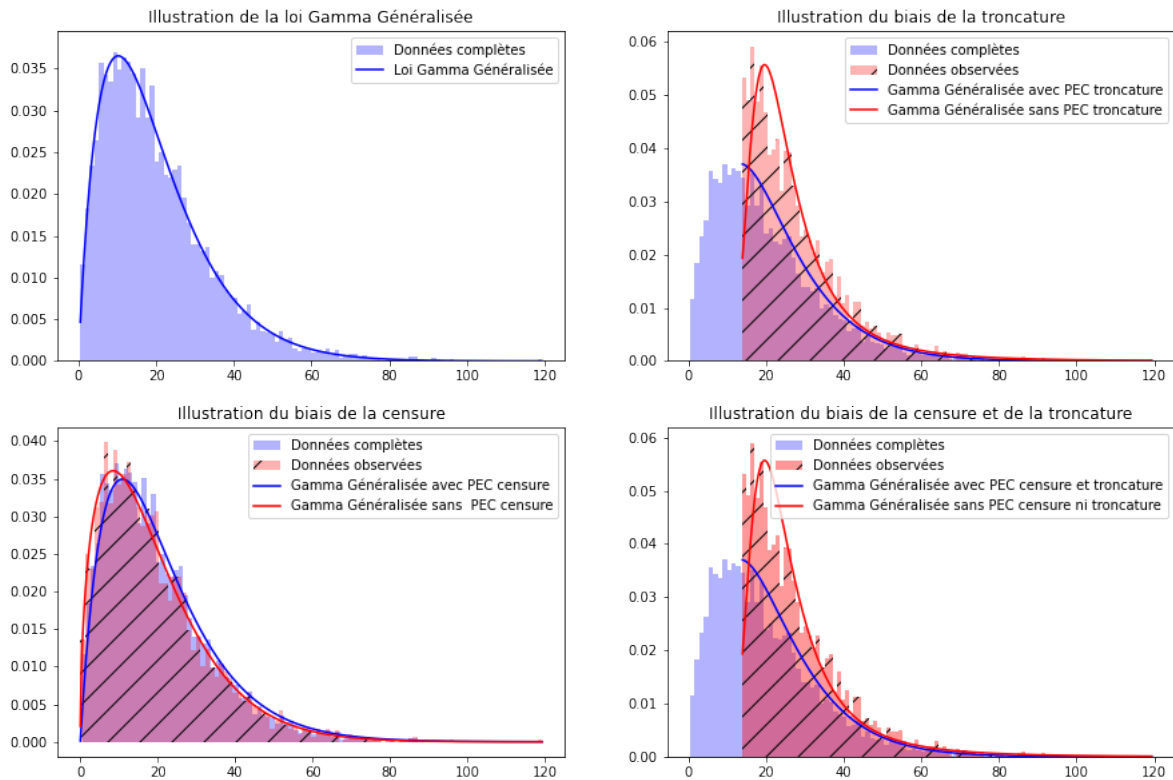


Figure 31: Illustration de la Gamma Généralisée, (PEC=Prise En Compte)

La figure située en bas à gauche illustre le biais qui se produit si nous ne prenons pas en compte les censures à droite dans nos données. De la même manière que précédemment, une loi a été estimée sur des données complètes (sans censure) et une autre avec des valeurs artificiellement censurées. Nous remarquons donc que si les censures ne sont pas spécifiées, la loi théorique obtenue ne reflète pas la véritable densité des observations.

Pour conclure, nous avons illustré le biais créé par la non-prise en compte de la censure et de la troncature. Nous remarquons que la combinaison des deux entraîne un biais important dans l'estimation des paramètres de la loi théorique.

Dans la section suivante, nous allons présenter différents estimateurs couramment utilisés pour modéliser la durée T . Nous expliquerons les méthodes d'implémentation en fonction de la loi à modéliser, à savoir, la loi d'incidence, de maintien ou la loi du passage en invalidité.

2.2 Statistiques non-paramétriques

Cette section se consacre à la présentation des estimateurs non-paramétrique ainsi qu'à des méthodes de lissages.

2.2.1 Estimateurs

Dans cette section, nous allons présenter trois estimateurs couramment utilisés dans l'analyse des données de survie, à savoir les estimateurs d'Hoem, de Kaplan-Meier et de Nelson-Aalen. Bien que l'estimateur d'Hoem soit paramétrique, nous décidons de la présenter dans cette

section, car il est utilisé pour les mêmes objectifs que les estimateurs non-paramétriques. L'avantage de cet estimateur est qu'il est simple à implémenter et permet d'avoir une première approximation des taux bruts. En revanche, il n'est pas le plus robuste en présence de censures et de troncatures. L'avantage des deux autres estimateurs (Kaplan-Meier et Nelson-Aalen) est qu'ils permettent de manière naturelle de prendre en compte ces phénomènes. De plus, ils sont non-paramétriques, il n'est donc pas nécessaire de supposer une hypothèse sur la distribution des données. La mise en place de chacun des estimateurs sera détaillée en fonction de la loi.

Estimateurs des moments

Une méthode simpliste pour estimer les taux bruts de passage en incapacité ou en invalidité consiste à utiliser l'estimateur d'Hoem. Il est en général utilisé pour le calcul des taux de mortalité, mais peut être adapté pour les taux de passage en incapacité. Cet estimateur généralise l'estimateur de la loi Bernoulli en prenant en compte l'exposition de l'individu au sein du portefeuille. Soit :

$$D_x \sim B(q_x) \text{ si sa fonction de masse est définie par :}$$

$$P(D_x = d_x) = q_x^{d_x} (1 - q_x)^{1-d_x} \text{ avec } \forall x \in X(\Omega) = (0, 1)$$

avec

- n_x : nombre d'individus dans l'état sain de l'âge x à l'âge $x+1$
- D_x : la variable aléatoire représentant le nombre de passage en état incapacité observée sur $]x, x + 1]$
- d_x : la réalisation de la variable aléatoire D_x
- q_x : la probabilité de passage en incapacité
- $[\alpha_i, \beta_i]$: représente l'intervalle inclus dans $[x, x + 1]$ pour lequel l'assuré i est sous observation

Nous calculons \hat{q} par maximum de vraisemblance et nous trouvons l'estimateur de la loi de Bernoulli : $\hat{q}_x = \frac{d_x}{n_x}$

Cependant, les individus ne sont pas observés pendant toute la durée de notre étude, qui est de 10 ans. Chaque individu possède sa propre exposition, c'est-à-dire la période pendant laquelle il est suivi dans l'étude. Il est donc nécessaire de remplacer le dénominateur n_x par la somme des poids observés du portefeuille pour prendre en compte cette exposition variable.

L'estimateur Hoem est donc :

$$\hat{q} = \frac{d_x}{\sum_i \beta_i - \alpha_i} \quad (1)$$

Intervalles de confiance asymptotique de l'estimateur d'Hoem

Grâce au théorème central-limite, nous pouvons définir la convergence en loi suivante :

$$Z = n_x \frac{q_x - \hat{q}_x}{\sqrt{q_x(1 - q_x)}} \xrightarrow[n \rightarrow \infty]{L} N(0, 1)$$

L'intervalle de confiance asymptotique obtenu est donc :

$$I_\alpha = \left[\hat{q}_x \pm z_{\frac{\alpha}{2}} \sqrt{\frac{q_x(1-q_x)}{n_x}} \right] \quad (2)$$

Mise en place de l'estimateur d'Hoem

Afin d'appliquer cet estimateur, nous avons découpé notre base de données en années afin d'estimer la probabilité de passage en incapacité par tranche d'âge. Pour chaque année et chaque tranche d'âge, nous avons enregistré le nombre de passages en incapacité ainsi que l'exposition. Par la suite, nous avons calculé le ratio de ces deux quantités par âge, ce qui correspond à l'estimateur d'Hoem.

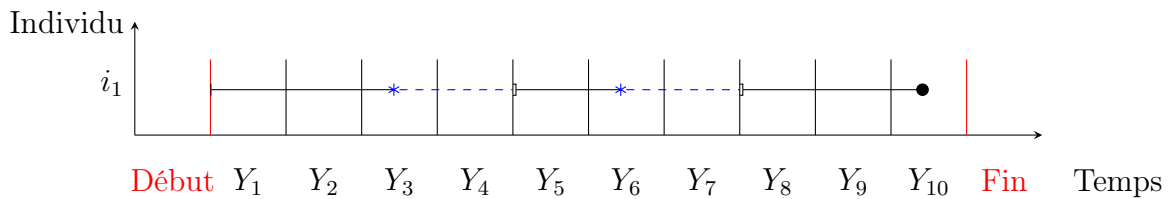


Figure 32: Découpage de notre base de données

La figure 33 représente le découpage de notre base de données, avec un individu ayant connu deux passages en incapacité. Pour le premier passage, nous avons retenu son âge au début de la période (Y_3), et pour le deuxième passage, son âge au début de la période (Y_6).

L'estimateur d'Hoem nous permet d'obtenir une première estimation des probabilités de passage en incapacité ou en invalidité. Les deux estimateurs suivants ne nécessitent pas de découpage de notre échantillon par période annuelle et permettent de prendre en compte naturellement les observations incomplètes (censures et troncatures).

Estimateur de Kaplan-Meier

L'estimateur non-paramétrique de Kaplan-Meier est utilisé pour estimer la fonction de survie (ou la fonction de répartition) d'une variable de durée en présence de censures et de troncatures. Dans une situation où toutes les données sont disponibles, l'estimation de la fonction de survie théorique peut être réalisée à l'aide de la fonction de survie empirique, donnée par :

$$\hat{S}(t) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{T_i > t}$$

Comme nous l'avons expliqué en introduction, la présence de censures nous empêche d'observer pleinement la variable aléatoire T_i . A la place, nous observons les variables aléatoires $Y_i = \min(T_i, C_i)$.

L'idée derrière cet estimateur est la suivante : si nous souhaitons calculer la probabilité qu'un individu né en t_0 survive jusqu'en t_2 , alors cela revient à supposer qu'il a survécu en t_1 . Mathématiquement, nous avons le développement suivant :

$$\begin{aligned}
\mathbb{P}(T > t_2) &= \mathbb{P}(T > t_2, T > t_1) & (3) \\
&= \mathbb{P}(T > t_2 | T > t_1) \mathbb{P}(T > t_1) \\
&= \mathbb{P}(T > t_2 | T > t_1) \mathbb{P}(T > t_1 | T > t_0) \mathbb{P}(T > t_0) \\
&= \left(\frac{S(t_2)}{S(t_1)} \right) \left(\frac{S(t_1)}{S(t_0)} \right) S(t_0) \\
&= \left(1 - \frac{S(t_1) - S(t_2)}{S(t_1)} \right) \left(1 - \frac{S(t_0) - S(t_1)}{S(t_0)} \right) S(t_0) \\
&= \left(1 - \frac{\mathbb{P}(T = t_1)}{S(t_1)} \right) \left(1 - \frac{\mathbb{P}(T = t_0)}{S(t_0)} \right) S(t_0) \\
&= (1 - \lambda(t_1)) (1 - \lambda(t_0)) S(t_0) \\
&= \prod_{t_j < t_2} (1 - \lambda(t_j))
\end{aligned}$$

Dans cette démonstration, nous avons considéré que T est une variable discrète, ce qui signifie que la fonction de survie $S(t)$ est constante par morceaux. Cela explique le passage entre les fonctions de survie $(S(t_1) - S(t_2))$ à la fonction de masse.

Le résultat final de l'estimateur fait apparaître le terme λ , qui représente le taux de risque instantané. Dans le cas discret, λ représente la probabilité qu'un événement se produise au temps t , étant donné que l'individu a survécu jusqu'au temps t .

$$\forall t \geq 0, \lambda(t) = \frac{\mathbb{P}(T = t)}{\mathbb{P}(T \geq t)} = \frac{dF(t)}{1 - F(t-)}$$

Dans le cas continu, le taux de risque instantané est représenté par la formule suivante :

$$\forall t \geq 0, \lambda(t) = \lim_{h \rightarrow 0} \frac{\mathbb{P}(t \leq X < T + h | X \geq t)}{h} = \frac{f(t)}{S(t)} = -\ln(S(t))'$$

Pour construire¹⁰ l'estimateur de Kaplan-Meier, nous devons d'abord estimer le taux de risque instantané puis l'inclure dans le calcul de la fonction de Kaplan-Meier. Le résultat nous donne la fonction suivante :

$$\hat{S}(t) = \prod_{T_i \leq t} \left(1 - \hat{\lambda}(t) \right) = \prod_{Y_i \leq t} \left(1 - \frac{d\hat{F}(t)}{1 - \hat{F}(t)} \right) = \prod_{T_i \leq t} \left(1 - \frac{d\hat{H}_1(t)}{1 - \hat{H}(t)} \right) = \prod_{T_i \leq t} \left(1 - \frac{\delta_i}{Y_i} \right) \quad (4)$$

avec au numérateur, $\delta_i = \begin{cases} 1 & \text{si observation complète, c'est-à-dire si l'évènement se produit} \\ 0 & \text{si observation incomplète} \end{cases}$

et au dénominateur Y_i , la somme des individus en vie juste avant t , soit $Y_t = Y_{t-1} - c_{t-1} - \delta_{t-1} + e_{t-1}$. Avec c correspondant aux individus censurés et e les individus entrant dans l'étude.

La démonstration de l'estimateur est basée sur **l'hypothèse d'identification**, qui suppose l'indépendance entre l'évènement et la censure, c'est-à-dire que T_i et C_i sont indépendants. La formule présentée ci-dessus correspond au cas où il n'y a pas d'ex-aequo, c'est-à-dire qu'un seul évènement peut se produire à un instant donné T_i .

¹⁰Une démonstration est disponible en annexe.

En présence d'ex-aequo, l'estimateur de Kaplan-Meier doit être modifié de la manière suivante :

$$\hat{S}(t) = \prod_{\substack{i=1, \dots, n \\ T_i \leq t}} \left(1 - \frac{d_i}{Y_i}\right) \quad (5)$$

avec d_i , le nombre de d'évènement se produisant en T_i .

Remarque : Dans le cas où il n'y aurait ni censure ni troncature, l'estimateur de Kaplan-Meier est strictement équivalent à la fonction de survie empirique.

Grâce à l'approximation suivante, nous pouvons obtenir l'estimateur de Greenwood de la variance, soit :

$$\widehat{Var}(\log(\hat{S}(t))) \approx \sum_t \frac{d_i}{Y_i(Y_i - d_i)}$$

$$\widehat{Var}(\hat{S}(t)) = \hat{S}(t)^2 \sum_t \frac{d_i}{Y_i(Y_i - d_i)}$$

Comme pour l'estimateur de Kaplan-Meier, nous pouvons construire un intervalle de confiance asymptotique en utilisant le théorème central limite, soit :

$$IC(\alpha) = \left[\hat{S}(t) \pm z_{\frac{\alpha}{2}} \sqrt{\widehat{Var}(\hat{S}(t))} \right]$$

Mise en place de l'estimateur de Kaplan-Meier

Comme mentionné précédemment, nous utiliserons cet estimateur pour estimer nos trois lois de probabilités. Pour mettre en place l'estimateur de Kaplan-Meier, il est important de déterminer les moments où un évènement peut se produire.

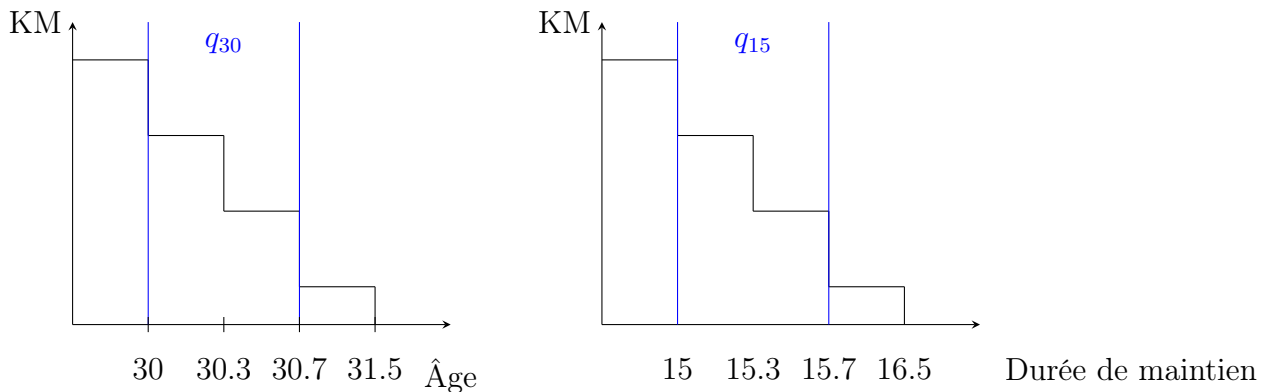


Figure 33: Mise en place de l'estimateur de Kaplan Meier. Incidence à gauche, Maintien et invalidité à droite.

- Loi d'incidence : Pour cette loi, les moments $t_1 < t_2 < \dots < t_n$ représentent les âges auxquels un ou plusieurs événements se produisent. Un événement peut être le passage en incapacité, l'entrée dans l'étude ou la sortie de l'étude. Le temps est discrétisé par les âges d'entrée et de sortie des individus.
- Loi de maintien et loi du passage en invalidité : Dans cette partie, les événements sont représentés par la durée (en jours) passée dans l'état d'incapacité. Chaque jour passé dans cet état peut entraîner soit un passage en invalidité, soit un retour à l'état sain. Ici, le temps est discrétisé par durée (en jours) passée en incapacité.

Construction de l'estimateur des taux bruts à l'aide de Kaplan-Meier

Une fois que nous avons obtenu l'estimateur de Kaplan-Meier, nous pouvons calculer les taux bruts d'incidence.

Pour rappel, le taux brut q_x représente la probabilité de passer en incapacité dans la classe d'âge $]x, x+1]$. Ainsi,

$$\hat{q}_x = 1 - \hat{p}_x = 1 - \frac{\hat{S}(x+1)}{\hat{S}(x)} = 1 - \prod_{t_i > x}^{x+1} \left(1 - \frac{d_{t_i}}{Y_{t_i}}\right) \quad (6)$$

Cependant, comme expliqué dans la section dédiée à la présentation générale du modèle de tarification, nous allons calculer un taux brut par cause de sortie. Ce taux brut est calculé en considérant notre modèle comme un **modèle à risques concurrents**¹¹. Cela nous conduit donc à apporter une légère modification à l'expression de \hat{q}_x .

$$\begin{aligned} q_{0j}(x, x+1) &= \hat{P}(T \leq x+1, X_T = j | T > x) := \sum_{x < t_i < x+1} \frac{\hat{P}(T > t_i)}{\hat{P}(T > x)} \cdot \frac{\Delta N_{0j}(t_i)}{Y_0(t_i)} \\ &= 1 - \prod_{t_i > x}^{x+1} \left(1 - \frac{d_{i,0j}}{Y_{i,0}}\right) \end{aligned}$$

où, $Y_0(t_i)$ correspond au nombre d'individu à risque à l'instant t_i et $N_{0j}(t_i)$ le nombre de transition de l'état sain vers la cause j (incap sup, incap inf ou grossesse).

Estimateur de Nelson-Aalen

L'estimateur de Nelson-Aalen est un autre estimateur non-paramétrique utilisé dans l'analyse des données de survie. Il est construit à partir du taux de risque cumulé, qui est défini comme suit :

$$\Lambda(t) = - \int_0^t \lambda(u) du$$

En utilisant le lien entre la fonction de survie et la fonction de risque instantané, nous pouvons définir l'équation suivante :

$$\forall t \geq 0, \lambda(t) = \frac{\mathbb{P}(T = t)}{\mathbb{P}(T \geq t)} = \frac{dF(t)}{1 - F(t-)} = \frac{dH_1(t)}{1 - H(t)}$$

¹¹Ces modèles sont utilisés lorsqu'il y a un état sain et plusieurs causes de sorties

Alors, on intégrant le taux de risque, nous trouvons la relation suivante :

$$\Lambda(t) = \int_0^t \lambda(u) du = \int_0^t \frac{dH_1(u)}{1 - H(u)} du$$

Nous estimons cette quantité théorique par l'estimateur suivant :

$$\hat{\Lambda}(t) = \sum_{T_i \leq t} \frac{\sum_{j=1}^n \mathbf{1}_{(T_j = T_i, \delta_j = 1)}}{\sum_{j=1}^n \mathbf{1}_{(T_j \geq T_i)}} = \sum_{T_i \leq t} \frac{d_i}{Y_i}$$

Par la relation entre la fonction de survie et le taux de risque instantané cumulé, nous obtenons un estimateur de la fonction de survie, soit :

$$\hat{S}(t) = e^{-\hat{\Lambda}(t)} \quad (7)$$

avec pour variance :

$$\widehat{Var}(\hat{\Lambda}(t)) = \sum_t \frac{d_i}{Y_i^2}$$

Remarque : Il existe une relation entre l'estimateur de Kaplan Meier et celui de Nelson-Aalen :

$$\hat{S}(t) = \prod_{T_i \leq t} (1 - \hat{\lambda}(t)) \approx \prod_{T_i \leq t} (e^{-\hat{\lambda}(t)}) = e^{-\sum_{T_i \leq t} \hat{\lambda}(t)} = e^{-\hat{\Lambda}(t)} := \hat{S}_{NA}(t)$$

Cette démonstration utilise l'approximation $e^x \approx 1 + x$, qui est valable pour de petites valeurs de x . L'approximation entre l'estimateur de Kaplan-Meier et l'estimateur de Nelson-Aalen peut être utilisée à condition que $\hat{\lambda}(t)$ soit faible dans l'intervalle $[t, t + dt)$. La figure 34 confirme que les deux estimateurs sont approximativement équivalents.

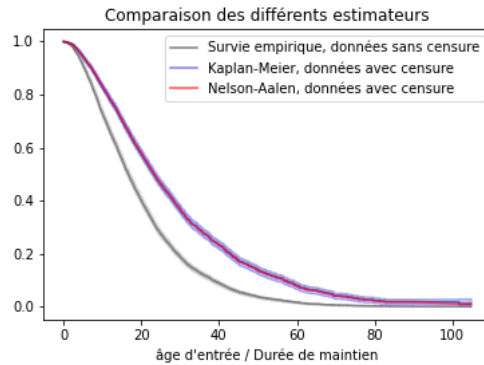


Figure 34: Comparaison des estimateurs non paramétrique

La construction de cet estimateur se fait de manière similaire à l'estimateur de Kaplan-Meier.

2.2.2 Techniques de lissage

Dans cette section, nous présentons différentes méthodes de lissage non-paramétriques. En science actuarielle, il est courant de lisser les données brutes afin d'obtenir des résultats plus homogènes. Le choix de la méthode de lissage dépendra de la distribution sous-jacente des données.

Pour les différentes probabilités de passage, nous utiliserons trois méthodes : la régression par noyau, le lissage de Whittaker-Henderson et la régression LOWESS.

Enfin, la loi de passage en invalidité et celle du maintien en incapacité par une approche non-paramétrique sera lissée avec les versions bidimensionnelles du lissage de Whittaker-Henderson et de la régression kernel.

Whittaker-Henderson en 1 dimension

La méthode Whittaker-Henderson fait partie de la famille des lissages non-paramétrique. Ce lissage permet de répondre à deux critères :

- Critère de fidélité : courbe proche des taux bruts

$$F(q) = \sum_{i=1}^n w_i (q_i - \hat{q}_i)^2 \quad (8)$$

- Critère de régularité : courbe suffisamment régulière

$$S(q) = \sum_{i=1}^{n-z} (\Delta^z q_i)^2 \quad (9)$$

$$\Delta^z q_i = \sum_{j=0}^z \binom{z}{j} (-1)^{z-j} q_{j+i} \quad (10)$$

où w_i représente les poids à l'âge i , q_i le taux brut d'incidence et \hat{q}_i le taux d'incidence lissé.

L'objectif de ce lissage est donc d'avoir $F(q)$ et $S(q)$ petit en fonction d'un paramètre de lissage h .

$$M = F(q) + hS(q) \quad (11)$$

$$q_x(h) = \underset{q}{\operatorname{argmin}} WH_h(q) \quad (12)$$

En dérivant M par rapport à q et en égalisant par rapport à 0, nous obtenons une forme analytique de la solution. Soit :

$$q^* = (w + hK'K_z)^{-1}w\hat{q} \quad (13)$$

Cette solution dépend du paramètre de lissage h .

Lorsque $h \rightarrow 0$: $q_i(h) \rightarrow \hat{q}_i$ et $h \rightarrow \infty$: $q_i(h) = \text{constante}$

Il y a donc 2 paramètres de lissage, \mathbf{h} et \mathbf{z} et un paramètre de poids à renseigner. En général, le paramètre z est égal à 2.

Pour prendre en compte l'exposition, plusieurs choix sont possibles :

- $w_i = 1$: Chaque observation a la même pondération qui est égale à 1

- $w_i = \frac{n_i}{\bar{n}}$ où $\bar{n} = \frac{\sum_{i=i_{min}}^{i=i_{max}} E_i}{i_{max} - i_{min} + 1}$: Les observations sont pondérées par l'effectif à l'âge x rapportées à l'effectif moyen pour tous les âges. Cette version sera utilisée dans le cadre de ce mémoire.
- $w_x = \frac{\sum_{i=1}^n E_x^i}{\hat{q}_x(1-\hat{q}_x)}$: La pondération est égale à l'inverse de la variance de l'estimateur brut.
- $w_x = \sum_{i=1}^n E_x^i$: Cette pondération est utilisée dans le cas où il y a peu de données.

La version du lissage précédente correspond au cas unidimensionnel. Cependant, pour le lissage du maintien en incapacité (non-paramétrique) et pour la probabilité de passage en invalidité, nous aurons besoin de la version bidimensionnelle.

Whittaker-Henderson en 2 dimensions

La méthode de Whittaker-Henderson peut être facilement généralisée pour le cas bidimensionnel. Pour cela, il suffit d'ajouter une somme dans les critères de fidélité et de régularité.

$$F(q) = \sum_{i=1}^p \sum_{j=1}^q w_{ij} (q_{ij} - \hat{q}_{ij})^2 \quad (14)$$

$$S_v(q) = \sum_{j=1}^q \sum_{i=1}^{p-z} (\Delta_v^z q_{ij})^2 \quad (15)$$

$$S_h(q) = \sum_{i=1}^p \sum_{j=1}^{q-z} (\Delta_h^z q_{ij})^2 \quad (16)$$

avec, $1 \leq i \leq p$ et $1 \leq j \leq q$, où i et j les variables correspondent à la dimension 1 et 2 respectivement.

Le programme de minimisation devient ainsi :

$$M = F(q) + \alpha S_v(q) + \beta S_h(q) \quad (17)$$

Ainsi, nous trouvons une forme analytique pour les taux bruts, telle que :

$$q^* = (w + \alpha K^{v'} K^v + \beta K^{h'} K^h)^{-1} w u \quad (18)$$

La section suivante présente une autre famille de lissage.

Régression non-paramétrique

Dans un modèle économétrique classique, notre objectif est d'expliquer la relation entre une variable à expliquer et une ou plusieurs variables explicatives à travers une fonction de lien. Dans un modèle paramétrique, la forme de la fonction de lien est connue et l'objectif consiste à estimer les paramètres. Par exemple, dans le modèle de régression linéaire, nous avons la forme suivante :

$$\mathbb{E}[Y|X = x] = m(x) = \alpha + \beta x$$

En économétrie, il est courant d'inclure les puissances des variables afin de mettre en évidence une relation concave entre les variables explicatives et la variable à expliquer, ou d'introduire des variables croisées pour tenir compte de l'interaction entre deux variables. Les modèles paramétriques présentent l'avantage de fournir des estimations performantes lorsque la forme de la fonction de lien est correctement spécifiée. De plus, ils permettent une interprétation économique des résultats, ce qui peut être plus difficile avec la seconde approche.

Dans le cas d'une régression non-paramétrique, nous ne faisons aucune hypothèse *a priori* sur la spécification de la forme de lien. L'idée consiste donc à estimer cette fonction de lien *a priori* inconnue.

$$\mathbb{E}[Y|X = x] = m(x)$$

Cette méthode présente une robustesse face à une mauvaise spécification du modèle. Cependant, les estimations sont généralement moins précises par rapport à une approche paramétrique correctement spécifiée.

En pratique, l'estimation non-paramétrique est utilisée pour différents objectifs. Elle peut être employée pour mettre en évidence la relation entre deux variables dans un premier temps, puis une fois cette relation détectée, un modèle paramétrique peut être construit en se basant sur les résultats de la relation non-paramétrique.

Certains économistes utilisent ces méthodes pour détecter la présence d'hétéroscédasticité (c'est-à-dire une variance conditionnelle différente pour chaque individu en fonction des variables explicatives).

Dans le domaine de l'actuariat, il est courant de lisser les résultats issus d'un modèle afin de les rendre plus homogènes en vue d'une tarification. Les données à lisser peuvent être des probabilités ou les coefficients d'une régression, par exemple.

La méthode d'estimation non-paramétrique la plus utilisée s'appelle la régression **Nadaraya-Watson**. Elle est basée sur l'utilisation d'un estimateur à noyau.

Lissage de Nadaraya-Watson

Comme mentionné dans l'introduction de cette section, cette méthode économétrique non-paramétrique permet de déterminer la forme de la relation entre y et x sans faire d'hypothèses au préalable sur sa spécification.

Voici le modèle correspondant :

$$\begin{aligned} y &= m(x) + \epsilon \\ \mathbb{E}[y|x] &= m(x) \end{aligned}$$

L'estimateur à noyau est basé sur une généralisation de l'estimateur naïf, qui repose sur le calcul de la moyenne des valeurs de y en fonction des observations voisines. Par exemple, pour lisser le point (y_{ref}, x_{ref}) , nous prenons en compte toutes les observations situées à une distance inférieure à $\frac{h}{2}$ de la référence.

$$\hat{m}(x) = \frac{\sum_{i=1}^n \mathbb{1}_{\left(-\frac{1}{2} < \frac{x_{ref} - x_i}{h} < \frac{1}{2}\right)} y_i}{\sum_{i=1}^n \mathbb{1}_{\left(-\frac{1}{2} < \frac{x_{ref} - x_i}{h} < \frac{1}{2}\right)}} \quad (19)$$

La formule indique qu'il est suffisant de prendre la somme des points proches de la valeur à lisser. Cependant, l'utilisation d'une indicatrice dans l'estimateur naïf, qui attribue la valeur 1 à une observation si elle se trouve dans l'intervalle $[x_{ref} - \frac{h}{2}; x_{ref} + \frac{h}{2}]$ et 0 sinon, entraîne des discontinuités. Pour éviter cela, nous utiliserons l'estimateur à noyau.

L'estimateur à noyau se base sur le principe similaire à celui de l'estimateur naïf, mais avec une différence importante : il introduit une pondération basée sur la distance par rapport à la valeur de référence, plutôt que d'affecter un poids uniforme à chaque observation. Ainsi, au lieu d'utiliser une fonction indicatrice, des fonctions spécifiques appelées noyaux (ou kernel) sont utilisées. Cette approche permet d'obtenir un lissage graduel, où les observations les plus proches de la valeur de référence ont un impact plus fort sur l'estimation finale.

$$\begin{aligned} \hat{m}(x) &= \frac{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right) y_i}{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right)} \\ \hat{m}(x) &= \sum_{i=1}^n w_i y_i \\ w_i(x) &= \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right)} \end{aligned} \quad (20)$$

Dans la littérature, il existe plusieurs types de kernel couramment utilisés.

- Kernel Gaussien : $K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$
- Kernel Epanechnikov : $K(u) = \frac{3}{4} (1 - u^2) \mathbb{1}_{(|u| \leq 1)}$
- Kernel Triangle : $K(u) = (1 - |u|) \mathbb{1}_{|u| \leq 1}$
- Kernel Uniforme : $K(u) = \frac{1}{2} \mathbb{1}_{|u| \leq 1}$

Remarque : L'estimateur uniforme correspond à l'estimateur naïf avec une fonction de poids normalisée.

Comme illustré sur la figure 35, nous pouvons observer que les noyaux Gaussien et Epanechnikov présentent des similitudes dans leurs caractéristiques. La différence réside dans le fait que le noyau Gaussien prend toujours des valeurs positives, tandis que le noyau Epanechnikov prend des valeurs nulles en-dehors d'un intervalle I spécifique.

En pratique, le choix du noyau n'a pas un impact significatif sur le lissage. Le paramètre de lissage h est le critère le plus crucial, car il permet de trouver un équilibre entre le biais et la variance. Une valeur plus élevée de h entraîne une fonction plus lisse, tandis qu'une valeur plus petite de h conduit à une fonction plus irrégulière.

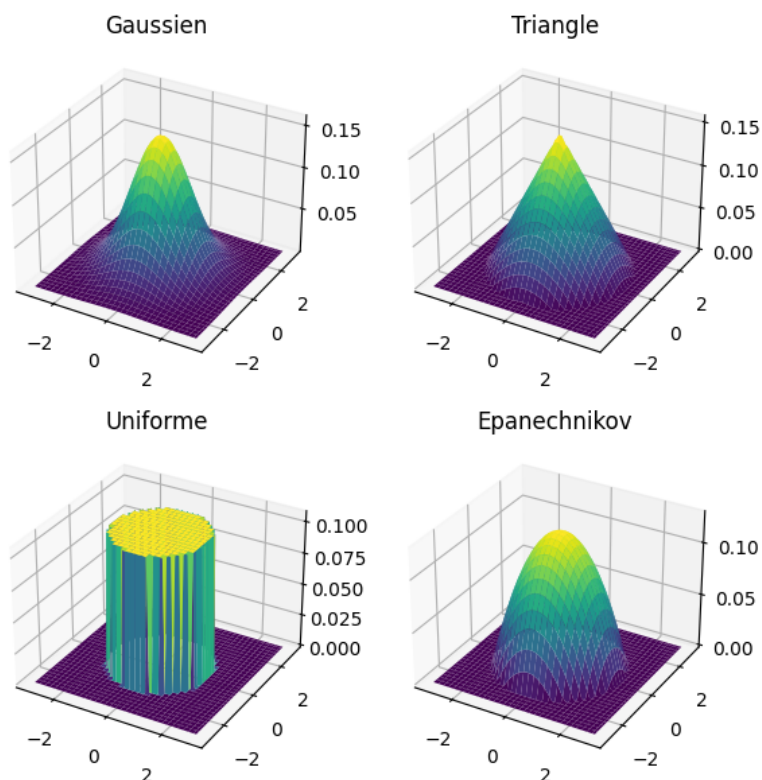


Figure 35: Kernels

Dans le premier cas, lorsque h est élevé, l'estimateur présente une faible variance, mais il peut être biaisé. À l'inverse, pour une valeur réduite de h , l'estimateur est moins biaisé, mais sa variance augmente.

Remarque : Lorsque le paramètre $h = \infty$, nous retrouvons la moyenne empirique des observations et dans le cas où $h = 0$, nous obtenons $\hat{m}(x_i) = y_i$.

Choix du paramètre de lissage

Pour le choix du paramètre de lissage optimal, il existe plusieurs solutions. L'une d'elles consiste à minimiser l'erreur quadratique moyenne intégrée associée au paramètre de lissage h :

$$MISE(h) = E \left[\int [m, h(x) - m(x)]^2 dx \right] \quad (21)$$

Cette quantité théorique représente l'espérance de l'intégrale des biais au carré, mais dans le contexte de la régression, nous ne pouvons pas substituer $f(x)$ par une densité théorique. Par conséquent, des approches asymptotiques sont utilisées pour obtenir des valeurs proches du paramètre optimal, noté h^* .

L'une de ces approches est la méthode de validation croisée (*Cross-validation*), qui permet d'obtenir une estimation optimale du paramètre. Elle consiste à minimiser le critère suivant :

$$CV(h) = \frac{1}{n} \sum_{i=1}^n [y_i - \hat{m}_{-i}(x_i)]^2 \quad (22)$$

où $\hat{m}_{-i}(x_i)$ représente la fonction de régression estimée sur toutes les observations sauf x_i .

On sélectionne ensuite le paramètre optimal qui minimise cette quantité, soit :

$$h_{CV}^* = \text{ArgMin}_{h \in \mathbb{R}^{*+}} CV(h) \quad (23)$$

D'autres critères tels que l'AMISE (Approximate Mean Integrated Squared Error) existent, mais ils ne seront pas abordés dans ce mémoire.

Au lieu d'utiliser une méthode automatique de sélection du paramètre de lissage, nous pouvons opter pour une approche graphique simple. En fonction de notre objectif, nous ajustons manuellement le paramètre de lissage en observant les résultats sur le graphique. Cette méthode nous permet de choisir le paramètre de manière plus subjective et adaptée à nos besoins spécifiques.

Une autre méthode de lissage, appelée **LOWESS** (LOcally WEighted Scatterplot Smoothing), a été appliquée. Ce lissage fait partie de la famille des régressions locales et offre un meilleur ajustement pour les valeurs extrêmes. Cependant, afin de ne pas alourdir davantage la lecture, et étant donné qu'elle n'a finalement pas été retenue dans notre modèle final, nous avons décidé de la présenter en annexe plutôt que dans cette section.

Remarque :

- Dans le cadre des lissages à l'aide de la régression Nadaraya-Watson, nous avons décidé de dupliquer les données en fonction de l'exposition. Cette technique (non-conventionnelle) permet de prendre en compte l'exposition et donc de pondérer le lissage en conséquence.
- Pour le lissage Nadaraya-Watson, nous avons sélectionné le paramètre de lissage à l'aide de la méthode de validation croisée. Quant au lissage de Whittaker-Henderson, nous avons choisi une approche graphique.

Tests de validation de lissage

Cette section traite de deux tests statistiques permettant de vérifier que les taux lissés ne diffèrent pas significativement des taux bruts initiaux. À cet effet, nous présentons deux tests couramment utilisés : le test du Khi-deux et le test des signes.

1. Test d'ajustement du Khi-deux

Ce test permet de s'assurer que les taux lissés ne sont pas trop éloignés des taux bruts initiaux. L'hypothèse H_0 est "Les taux bruts sont proches des taux lissés". Pour réaliser ce test, nous calculons la statistique suivante :

$$T = \sum_{x=x_{min}}^{x_{max}} n_x \frac{(\hat{q}_x - q_x)^2}{q_x(1 - q_x)} \quad (24)$$

Cette statistique de test suit une loi du χ_{p-r-1}^2 . Nous acceptons l'hypothèse H_0 si $T > \chi_k^2$. En cas d'acceptation, cela signifie que les taux lissés sont proches des taux bruts initiaux. Avec $r=0$ dans le cas du lissage de Whittaker-Henderson.

2. Test des signes

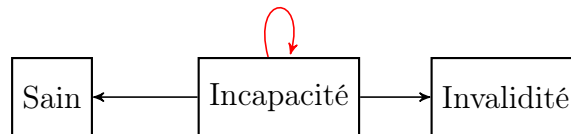
Ce test vérifie que la loi de la différence entre les taux bruts et les taux estimés suit une loi normale, $N(0, 1)$. Si par exemple tous les signes de la série $\delta_x = q_x - \hat{q}_x$ sont positifs, alors cela signifierait que nous sous-estimerions les vrais taux bruts et inversement. Pour réaliser ce test, nous devons calculer la statistique suivante :

$$T = \frac{|n_+ - n_-| - 1}{\sqrt{n}} \quad (25)$$

avec n_+ le nombre de signe positif et n_- le nombre de signes négatifs et $n = n_+ + n_-$.

Sous $H_0 : T \sim N(0, 1)$. Nous rejetons donc l'hypothèse nulle si $|T| > N_{1-\frac{\alpha}{2}}(0, 1)$ avec α un seuil à 5%.

2.3 Statistiques paramétrique



Dans cette section, nous aborderons principalement les estimateurs paramétriques utilisés pour modéliser les distributions de probabilités de la durée passée dans l'état d'incapacité (**loi de maintien**). Étant donné la présence de censures et de troncatures, il est nécessaire d'adapter la méthode de maximisation de vraisemblance pour prendre en compte les données inobservables.

La variable d'intérêt de cette section est la **durée passée dans l'état d'incapacité**. Cependant, nous n'observons que les durées conditionnelles au dépassement de la franchise du contrat de garantie. En raison de la présence de franchise, notre échantillon est tronqué à gauche, ce qui signifie que nous n'observons que les durées d'incapacité supérieures à cette valeur.

De plus, nos données sont également censurées à droite. Certains individus atteignent la limite maximale de leur période d'indemnisation, et une fois cette limite atteinte, la durée passée dans l'état d'incapacité n'est plus observable. Nous savons simplement que l'individu est toujours en incapacité lorsqu'il atteint cette limite, sans avoir d'information sur la durée exacte passée dans cet état.

Les phénomènes de censure et de troncature nous obligent à modifier la fonction de vraisemblance afin d'estimer la distribution de probabilité des données incomplètes.

Pour illustrer cette problématique, la figure 36 présente un exemple avec une distribution de Weibull de paramètres $X \sim W(26, 0.5)$.

La partie en bleu foncé représente la densité de probabilité correspondant aux données observées, qui peuvent être censurées à certains points (par exemple, à 90 jours) en raison des garanties des contrats. La partie en bleu clair représente la partie inconnue que nous cherchons à estimer à partir des données observées (troncature à gauche).

Notre objectif est donc de développer des méthodes permettant de tirer le meilleur parti des données disponibles malgré la présence de censures et de troncatures, afin d'estimer au mieux

Densité d'une distribution de Weibull : $f(x; \beta, \alpha) = \frac{\beta}{\alpha^\beta} x^{\beta-1} e^{-(\frac{x}{\alpha})^\beta}$

- β : paramètre de forme
- α : paramètre d'échelle

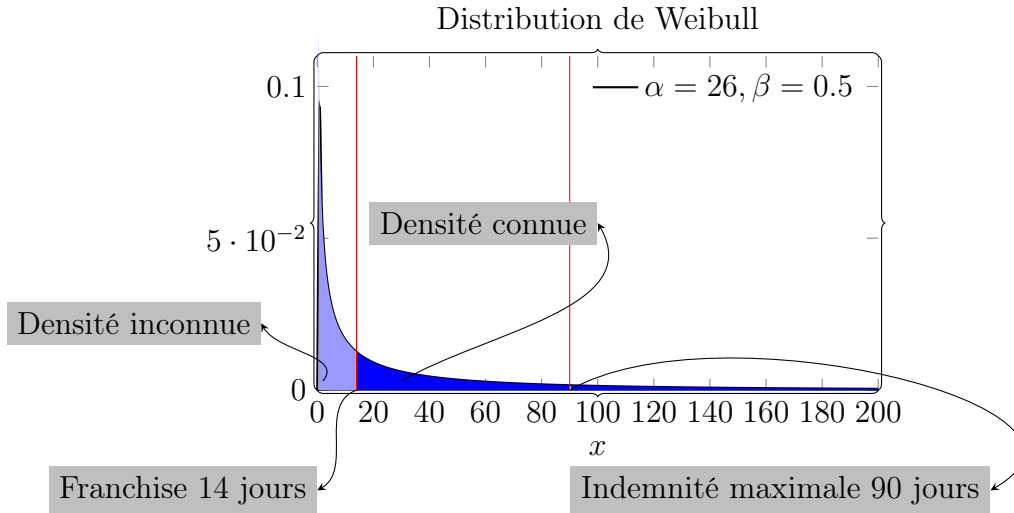


Figure 36: Illustration d'une distribution tronquée et censurée

la distribution de probabilité des durées passées dans l'état d'incapacité.

2.3.1 Estimateurs

Maintenant, que nous avons exposé les principales caractéristiques de notre échantillon, nous allons aborder la méthode du **maximum de vraisemblance**.

La méthode du maximum de vraisemblance est l'une des approches couramment utilisées pour estimer les paramètres d'une distribution théorique. Elle est préférée à la méthode des moments en raison de sa robustesse et de son efficacité.

Pour estimer les paramètres d'une loi de durée en présence de données censurées et/ou tronquées, nous devons construire une fonction de vraisemblance spécifique qui tienne compte de ces caractéristiques¹². La fonction de vraisemblance complète est représentée par la formule suivante :

$$L = \prod_{i \in D} f(T_i) \cdot \prod_{i \in R} S(C^{(r)}) \cdot \prod_{i \in L} (1 - S(C^{(r)})) \cdot \prod_{i \in I} (S(L_i) - S(R_i)) \quad (26)$$

avec,

- D : Représente les données complètes
- R : Représente les données censurées à droite
- L : Représente les données censurées à gauche
- I : Représente les données censurées par intervalle

¹²Voir le cours de P.Breheny [22]

Pour détailler ce résultat, nous faisons l'**hypothèse d'identification** comme pour la construction de l'estimateur de Kaplan-Meier.

Lorsque nous observons la durée exacte de l'individu i , nous maximisons la densité $f(T_i)$. Dans le cas où nous observerions seulement la durée jusqu'au temps C_i (données censurées à droite), nous maximisons la fonction de survie $S(C^{(r)})$. Inversement, dans le cas des données censurées à gauche, nous maximisons la fonction de répartition $(1 - S(C^{(r)}))$. Pour les données censurées par intervalles, la vraisemblance maximise la différence entre la fonction de survie des données censurées à gauche et celle des données censurées à droite. Les deux derniers cas ne correspondent pas à nos données. Pour avoir une vraisemblance adaptée à nos données, nous devons ajouter un facteur prenant en compte la troncature à gauche. Ainsi, la densité $f(T_i)$ est remplacée par $f(T_i)/S(\tau)$ où τ représente la valeur de la franchise, et la fonction de survie $S(C^{(r)})$ est remplacée par $S(C^{(r)})/S(\tau)$.

La vraisemblance finale utilisée est donc la suivante :

$$L = \prod_{i=1}^n \left(\frac{f(t_i)}{S(\tau)} \right)^{\delta_i} \cdot \left(\frac{S(t_i)}{S(\tau)} \right)^{1-\delta_i} \quad (27)$$

La démonstration de cette formule se trouve dans l'annexe consacrée à la théorie des phénomènes de durées.

Lois de distributions candidates

Pour modéliser la durée de maintien, nous avons choisi trois lois candidates :

- Weibull
- Gamma généralisée
- Log Logistique

Cette décision provient du fait que le risque instantané de ces lois est propice à l'analyse de la durée du maintien en incapacité d'un individu. En effet, si nous prenons l'exemple de la **loi Weibull**, $W(\lambda, \rho)$, où λ est le paramètre d'échelle et ρ le paramètre de forme, nous avons :

$$f_T(t) = \frac{\rho}{\lambda} \left(\frac{t}{\lambda} \right)^{\rho-1} \exp \left(- \left(\frac{t}{\lambda} \right)^\rho \right) \quad (28)$$

$$S_T(t) = \exp \left(- \left(\frac{t}{\lambda} \right)^\rho \right) \quad (29)$$

$$\lambda(t) = \frac{\rho}{\lambda} \left(\frac{t}{\lambda} \right)^{\rho-1} \quad (30)$$

Si $\rho = 1$, nous retrouvons la loi exponentielle. Si $0 < \rho < 1$, alors nous avons une décroissance monotone du risque instantané et si $\rho > 1$, le risque est monotone croissant. La figure 37 illustre les différents cas possibles.

Pour la **loi Log-Logistique**, $LL(\lambda, \frac{1}{\sigma})$, où λ est le paramètre d'échelle et $\frac{1}{\sigma}$ le paramètre de forme, nous avons :

$$f(t) = \frac{1}{1 + \left(\frac{t}{\lambda}\right)^{-\frac{1}{\sigma}}} \quad (31)$$

$$S(t) = \frac{1}{1 + \left(\frac{t}{\lambda}\right)^{\frac{1}{\sigma}}} \quad (32)$$

$$\lambda(t) = \frac{\frac{1}{\sigma\lambda} \left(\frac{t}{\lambda}\right)^{\frac{1}{\sigma}-1}}{1 + \left(\frac{t}{\lambda}\right)^{\frac{1}{\sigma}}} \quad (33)$$

Comme pour la loi de Weibull, le taux de risque instantané dépend du paramètre de forme. Si $\frac{1}{\sigma}$ est inférieur à 1, nous avons une fonction croissante au dénominateur et décroissante au numérateur, ce qui entraîne un taux de risque décroissant. Dans le cas où $\frac{1}{\sigma} > 1$, nous avons une fonction de hasard unimodale.

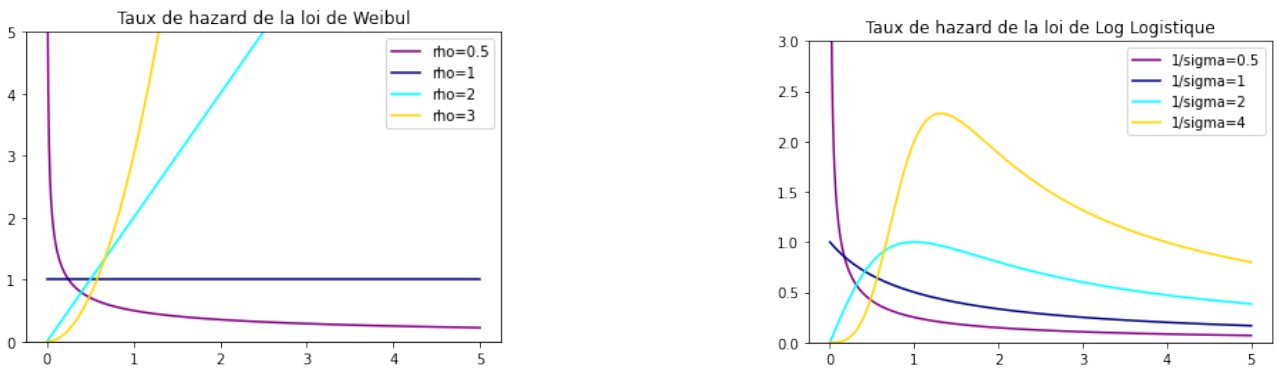


Figure 37: Gauche : Taux hazard Weibull | Droite : Taux hazard Log Logistique

De la même manière que pour la loi de Weibull et la loi Log-logistique, nous pouvons définir la **distribution de Gamma Généralisée**¹³, $GG(a, d, p)$ où a est le paramètre d'échelle et $p, d > 0$.

$$f(t) = \frac{\frac{p}{a^d} t^{d-1} e^{-\left(\frac{t}{a}\right)^p}}{\Gamma\left(\frac{d}{p}\right)} \quad (34)$$

$$F(t) = \frac{\gamma\left(\frac{d}{p}, \left(\frac{t}{a}\right)^p\right)}{\Gamma\left(\frac{d}{p}\right)} = P\left(\frac{d}{p}, \left(\frac{t}{a}\right)^p\right) \quad (35)$$

où P est la fonction gamma incomplète faible et si,

- $p = d$, alors nous retrouvons la distribution de Weibull
- $p = 1$, nous obtenons la distribution Gamma
- $p = d = 1$, nous retrouvons la distribution Exponentiel

Maintenant, que nous avons présenté les distributions de probabilités couramment utilisées en analyse des phénomènes de durée, nous allons à présent détailler les méthodes de sélection.

¹³Il existe plusieurs façons de paramétrer la distribution de Gamma Généralisée.

Sélection de la loi optimale

Pour sélectionner la loi optimale, il existe différents critères ou tests à utiliser :

- AIC
- BIC
- Kolmogorov-Smirnov
- Cramer-Von-Mises

Dans le cas de notre étude, nous avons principalement utilisé les critères d'AIC et de BIC.

$$AIC = 2k - 2\ln(L) \quad (36)$$

$$BIC = -2\ln(L) + k.\ln(N) \quad (37)$$

Avec k comme le nombre de paramètres, N comme le nombre d'observations et $\ln(L)$ le logarithme de la vraisemblance.

L'AIC représente un compromis entre le biais (diminue en fonction du nombre de paramètres) et la parcimonie du modèle (représentation des données avec le moins de paramètres possibles).

Le critère du BIC pénalise les modèles avec un grand nombre de paramètres.

2.3.2 Techniques de lissage

Pour les lois paramétriques liées à la durée du maintien en incapacité, nous avons élaboré une approche visant à lisser les paramètres des distributions tout en prenant en compte leur interdépendance. Les distributions de durées utilisées comportent deux paramètres, et chacun de ces paramètres dépend de l'autre.

Afin de lisser les distributions, nous avons utilisé les moments d'ordre 1 et 2. Cette approche permet de considérer conjointement les deux paramètres et d'obtenir un ajustement plus précis en prenant en compte leur interdépendance. L'inconvénient de cette méthode est qu'elle ne peut être appliquée pour la loi Log-logistique car le moment d'ordre 2 n'existe que si le paramètre de forme est supérieur à 2. Cette méthode peut donc s'appliquer seulement pour la loi de **Weibull**.

Etape 1

Dans un premier temps, nous avons récupéré les moments 1 et 2 des lois de probabilités. Soit :

$$m_k = E[X^k] = \int_{-\infty}^{+\infty} x^k dF_X(x) \quad (38)$$

$$m_1 = E[X] = \int_{-\infty}^{+\infty} x dF_X(x) \quad (39)$$

$$m_2 = E[X^2] = \int_{-\infty}^{+\infty} x^2 dF_X(x) \quad (40)$$

$$m_1 = \mu \quad (41)$$

$$m_2 - (m_1)^2 = \sigma^2 \quad (42)$$

Etape 2

Nous avons par la suite lissé les moments à l'aide de la méthode de Whittaker-Henderson et la régression kernel avec un noyau gaussien (voir section sur les statistiques non-paramétrique).

Etape 3

Une fois les moments 1 et 2 lissés, \tilde{m}_1 et \tilde{m}_2 , nous avons utilisé la méthode des moments pour récupérer les nouveaux paramètres de distribution.

$$\mu = \alpha \Gamma\left(1 + \frac{1}{\beta}\right) = \tilde{m}_1 \quad (43)$$

$$\sigma^2 = \alpha^2 \Gamma\left(1 + \frac{2}{\beta}\right) - \mu^2 = \tilde{m}_2 - (\tilde{m}_1)^2 \quad (44)$$

Nous avons remplacé μ et σ^2 par leur version empirique que nous avons lissé (à travers les moments) à savoir $\hat{\mu}$ et la $\hat{\sigma}^2$. En appliquant le log de chaque côté de l'équation, nous obtenons la forme suivante :

$$\log(\hat{\mu}) = \log(\alpha) + \log \Gamma\left(1 + \frac{1}{\beta}\right) \quad (45)$$

$$\log(\hat{\sigma}^2 + \hat{\mu}^2) = 2 \log(\alpha) + \log \Gamma\left(1 + \frac{2}{\beta}\right) \quad (46)$$

En multipliant par 2 la première équation et en insérant la deuxième équation dans la première, nous trouvons le résultat suivant :

$$\log(\hat{\sigma}^2 + \hat{\mu}^2) - \log \Gamma\left(1 + \frac{2}{\beta}\right) = 2 \log(\hat{\mu}) - 2 \log \Gamma\left(1 + \frac{1}{\beta}\right) \quad (47)$$

$$\log \Gamma\left(1 + \frac{2}{\beta}\right) - 2 \log \Gamma\left(1 + \frac{1}{\beta}\right) - \log(\hat{\sigma}^2 + \hat{\mu}^2) + 2 \log(\hat{\mu}) = 0 \quad (48)$$

La dernière étape consiste à résoudre l'équation précédente par un algorithme d'optimisation. Nous avons initialisé le paramètre $\beta_0 = 0.5$.

Une fois le nouveau paramètre $\hat{\beta}$ déterminé, nous pouvons à son tour, recalculer le paramètre $\tilde{\alpha}$ par la formule suivante :

$$\tilde{\alpha} = \frac{\hat{\mu}}{\Gamma\left(1 + \frac{1}{\hat{\beta}}\right)} \quad (49)$$

Pour la loi **Log-logistique**, nous décidons d'utiliser une méthode simplifiée en ne lissant que le moment d'ordre 1. Cette décision est due à la condition sur le moment d'ordre 2 qui ne nous permet pas d'utiliser l'approche ci-dessus.

1. Lisser le paramètre $\beta \rightarrow \tilde{\beta}$
2. Récupérer le moment d'ordre 1: $m_1 = \alpha \frac{\pi}{\sin(\frac{\pi}{\tilde{\beta}})}$

3. Lisser le moment d'ordre 1, $m_1 \rightarrow \tilde{m}_1$

4. Récupérer le paramètre α avec l'équation en 2, soit : $\tilde{m}_1 = \alpha \frac{\frac{\pi}{\beta}}{\sin(\frac{\pi}{\beta})}$

Jusqu'à présent, nous avons présenté des méthodes d'estimations unidimensionnelles. La section suivante porte sur la modélisation multidimensionnelle.

2.4 Modèle avec variables explicatives

Depuis le début de ce mémoire, nous avons utilisé des estimateurs unidimensionnels spécifiques à l'analyse des modèles de durée sur des sous-groupes de notre échantillon.

L'objectif de cette section est donc d'examiner les différentes alternatives permettant d'inclure directement les variables explicatives (les variables de segmentations) dans un modèle statistique. L'un des avantages de l'utilisation d'un modèle est qu'il permet d'évaluer l'effet des variables explicatives sur la variable d'intérêt. Nous présentons donc différents modèles pour estimer nos lois de probabilité ou les taux bruts.

2.4.1 Modèle de Cox

Le modèle le plus connu dans l'analyse des phénomènes de durée est le modèle de Cox. Ce modèle permet de prendre en compte l'effet des variables explicatives sur le taux de risque instantané. Il est considéré comme semi-paramétrique car, il ne suppose aucune distribution particulière sur la variable d'intérêt, mais il impose une forme spécifique à la fonction de lien.

Habituellement, un modèle économétrique s'écrit de la manière suivante :

$$E[Y|X] = \alpha + \beta X \quad (50)$$

Cependant, avec le modèle de Cox, la relation entre les variables explicatives et la variable de durée se fait à travers le taux de risque instantané. C'est ce taux qui capture l'effet des variables explicatives sur la probabilité que l'événement survienne à un instant donné.

Le modèle à hasards proportionnels s'écrit de la manière suivante :

$$\lambda(t|X) = \lambda_0(t)h(\beta X) \quad (51)$$

où λ_0 représente la fonction de risque de base (ou de référence), X les variables explicatives et β les coefficients associés.

Comme illustré par l'équation, ci-dessus, nous pouvons observer que le taux de risque de base ne dépend pas de l'individu i , mais uniquement du temps t . Cette observation implique que la dépendance temporelle du risque que l'événement se réalise (par exemple, le risque de tomber en incapacité ou en invalidité) est la même pour tous les individus. De plus, la fonction h est une fonction positive.

Le modèle est dit de **Cox** lorsque la fonction $h()$ est la fonction exponentielle, soit:

$$\lambda(t|X) = \lambda_0(t)\exp(\beta X) \quad (52)$$

De plus, c'est un modèle dit proportionnel, car, pour 2 individus i et j , le rapport des fonctions de hasard est constant dans le temps, soit :

$$\frac{\lambda(t|X_i)}{\lambda(t|X_j)} = \frac{\lambda_0(t)h(\beta X_i)}{\lambda_0(t)h(\beta X_j)} = \frac{h(\beta X_i)}{h(\beta X_j)} = \frac{\exp(\beta' X_i)}{\exp(\beta' X_j)} = \exp(\beta'(X_i - X_j))$$

Ce ratio permet donc de mesurer le risque que l'individu i subisse l'événement par rapport au risque de l'individu j .

Par exemple, considérons un modèle avec une seule variable explicative représentant le sexe, où la valeur 1 indique un homme et 0 indique une femme. Dans ce cas, le rapport de taux, égal à $\exp(\beta' * 1)$, s'interprète comme la probabilité (constante dans le temps) de subir l'événement en étant un homme par rapport à une femme. Si nous considérons une variable explicative continue, l'interprétation sera liée à l'augmentation d'une unité de cette variable par rapport à la référence.

Dans le modèle de Cox, il est nécessaire de définir des intervalles de temps $[t_i, t_i + \Delta t]$ où des événements tels que le passage en incapacité ou en invalidité, l'entrée ou la sortie de l'étude se produisent. À partir de cette discrétisation du temps, nous construisons la vraisemblance du modèle de Cox qui nous permet d'estimer les coefficients.

Dans un premier temps, nous pouvons définir la probabilité de subir un événement dans l'intervalle $I = [t_i, t_i + \Delta t]$:

$$\sum_{j \in I} \lambda_0(t_i) \exp(\beta X_j)$$

De la même manière, nous pouvons définir la probabilité qu'un individu subisse un événement en t_i :

$$\frac{\exp(\beta' X_i)}{\sum_{j \in I} \exp(\beta X_j)}$$

A l'aide de cette probabilité, nous pouvons construire la vraisemblance partielle qui représente le produit sur les temps d'événements. Soit:

$$L_{Cox}(\beta) = \prod_{i=1}^D \frac{\exp(\beta X_i)}{\sum_j \exp(\beta X_j)}$$

Comme nous pouvons le remarquer, la vraisemblance ne dépend pas de la fonction de risque de base $\lambda_0(t)$. Il n'est donc pas nécessaire de la connaître pour estimer les paramètres du modèle. La suite du modèle consiste à maximiser la log-vraisemblance partielle à l'aide d'un algorithme d'optimisation tel que Newton-Raphson.

Le score est défini comme suit :

$$U(\beta) = \frac{\partial \ell(\beta)}{\partial \beta} = \left(\frac{\partial \ell(\beta)}{\partial \beta} \right) = \left(\frac{\partial \ell(\beta)}{\partial \beta_1} \right) = \dots = \left(\frac{\partial \ell(\beta)}{\partial \beta_p} \right) = 0$$

Mise en place

Dans un premier temps, nous pouvons présenter une méthode pour estimer les taux bruts du passage en incapacité en utilisant le modèle de Cox. En intégrant l'équation entre la fonction de risque et les variables explicatives au taux brut d'incidence, nous obtenons la formule suivante:

$$q_x = P(x \leq T \leq x + 1 | T \geq x) = 1 - \frac{S(x+1)}{S(x)} = 1 - \frac{e^{\int_{\infty}^x \lambda(u) du}}{e^{\int_{\infty}^{x+1} \lambda(u) du}} \quad (53)$$

$$q_x(x|X, \beta) = 1 - \frac{e^{\int_{\infty}^x \lambda_0(u) \exp(\beta' X) du}}{e^{\int_{\infty}^{x+1} \lambda_0(u) \exp(\beta' X) du}} = 1 - e^{\int_x^{x+1} \lambda_0(u) \exp(\beta' X) du} = 1 - \left(\frac{S_0(x+1)}{S_0(x)} \right)^{\exp(\beta' X)}$$

$$q_x(x|X, \beta) = 1 - (1 - q_0(x))^{\exp(\beta' X)}$$

Pour calculer cette quantité, il est nécessaire de définir un risque de référence. Pour ce faire, nous utilisons les taux bruts du risque de référence, $q_0(x)$, calculés à l'aide de la méthode de Kaplan-Meier. Une fois ce taux brut de référence défini, nous pouvons calculer les autres taux bruts à l'aide des coefficients estimés à partir de la régression du modèle de Cox.

Dans notre contexte, la population de référence correspond à celle où les variables explicatives sont toutes égales à 0, c'est-à-dire lorsque $X_1 = \dots = X_p = 0$.

Dans un second temps, nous pouvons utiliser ce modèle pour estimer la fonction de survie du maintien en incapacité. La relation entre le taux de risque instantané et la fonction de survie nous permet d'obtenir une estimation de la forme suivante :

$$\hat{S}(t|X) = e^{-\int_0^t \lambda(u|X) du} = e^{(-\hat{\Lambda}_0(t) \exp(\beta' X))} \quad (54)$$

Vérification des hypothèses

Pour tester les hypothèses du modèle, plusieurs méthodes sont envisageables. La première méthode consiste à tracer un graphique des résidus de Schoenfeld pour chaque variable explicative. Cependant, cette méthode n'est malheureusement pas suffisamment robuste. C'est pourquoi nous préférons opter pour un test numérique.

La seconde méthode consiste à effectuer un test de proportionnalité des coefficients.

Résidus de Schoenfeld

L'idée du test est de vérifier l'hypothèse de proportionnalité des coefficients. Pour ce faire, nous utilisons le test de Grambsch et Therneau (1994), qui se fonde sur les résidus de Schoenfeld. À cet effet, nous formulons le modèle suivant :

$$\lambda_i(t) = \lambda_0(t) e^{X_i(t)^T (\beta + G(t)\theta)} \quad (55)$$

avec θ en tant que scalaire et G une fonction pouvant être une transformation de Kaplan-Meier, la fonction identité ou la fonction logarithme.

L'objectif est de tester l'hypothèse nulle suivante :

$$H_0 : \theta = 0_{p*1} \text{ vs } H_1 : \theta \neq 0_{p*1}$$

où $p * 1$ représente une matrice avec p lignes et 1 colonne.

Si l'hypothèse nulle est acceptée, alors dans ce cas, les coefficients sont invariants en fonction du temps.

Pour réaliser ce test, il faut, dans un premier temps, calculer les résidus de Schoenfeld.

$$\hat{r}_{ij} = x_{ij} - \sum_j x_{j,k} \hat{\rho}_{j,t_i} \quad (56)$$

$$\hat{\rho}_{j,t_i} = \frac{e^{\beta \hat{x}_j}}{\sum_j e^{\beta x_j}}$$

où $\hat{\rho}_{j,t_i}$ est la vraisemblance estimée que l'individu (j) à risque, connaisse l'événement en t_i .

L'idée derrière ces résidus est de calculer, pour chaque individu et à chaque événement qui se produit pour lui, la différence entre la valeur de la j -ième variable explicative de cet individu et une moyenne pondérée des valeurs de cette variable explicative sur l'ensemble des autres individus.

Enfin, la seconde étape consiste à calculer la statistique de test suivante :

$$T(G) = \left(\sum_{l=1}^d G_l \hat{r}_l \right)^T H^{-1} \left(\sum_{l=1}^d G_l \hat{r}_l \right) \quad (57)$$

$$H = \sum_{l=1}^d G_l \hat{V}_l G_l - \left(\sum_{l=1}^d G_l \hat{V}_l \right) \left(\sum_{l=1}^d G_l \hat{V}_l \right)^{-1} \left(\sum_{l=1}^d G_l \hat{V}_l \right)^T \quad (58)$$

Cette statistique de test suit une loi asymptotique du χ_p^2 .

Pour plus de détails sur ce test, vous pouvez consulter en annexe l'article d'origine, *Proportional hazards tests and diagnostics based on weighted residuals, 1994*, de Grambsch, P. M. et Therneau [14].

2.4.2 Modèle AFT

Le modèle AFT (Accelerated Failure Time) fait partie des modèles de régression paramétrique dans l'analyse des phénomènes de survie. Ce modèle de régression paramétrique représente une alternative au modèle de Cox et permet une plus grande précision lorsque la loi est en accord avec les données. Un autre avantage des modèles paramétriques est la possibilité d'obtenir une loi de probabilité qui peut être utilisée pour simuler l'exposition lors de la dernière partie de ce mémoire. De plus, du fait de son caractère paramétrique, les résultats obtenus sont facilement interprétables.

En général, les modèles paramétriques peuvent être représentés de deux manières différentes. La première consiste à utiliser un **modèle de durée de vie accélérée**, tandis que la deuxième utilise un **modèle linéaire**.

La première représentation stipule que la fonction de survie d'un individu avec les variables explicatives X au temps t est la même que celle d'un individu avec la fonction de survie de base au temps $te^{\beta X}$.

Représentation sous forme d'AFT

$$S(t|X) = S_0(te^{\theta^T X}) \quad (59)$$

$$\lambda(t|X) = e^{\theta^T X} \lambda_0(te^{\theta^T X}) \quad (60)$$

Le terme $e^{\theta^T X}$ représente un facteur d'accélération de temps.

Représentation sous forme de modèle linéaire

La seconde forme consiste à modéliser le logarithme de la v.a T en fonction de covariables X, soit :

$$\log(T) = \beta_0 + \beta^T X + \sigma \epsilon \quad (61)$$

$$T = e^{\beta_0} \cdot e^{\beta^T X} \cdot e^{\sigma \epsilon} \quad (62)$$

avec β_0 l'ordonnée à l'origine, σ un paramètre d'échelle, ϵ un terme d'erreur et $\theta = -\beta$. Dans cette représentation, les covariables ont un effet multiplicatif sur le temps de survie T ou de manière équivalente, nous pouvons dire que les covariables ont un effet additif sur le logarithme du temps, $\log(T)$.

Les deux représentations sont relativement proches. En effet, si nous supposons que la durée de vie de base est $T_0 = e^{\beta_0 + \sigma \epsilon}$ ou de manière équivalente que $S_0(t) = \mathbb{P}(e^{\beta_0 + \sigma \epsilon} > t)$, alors dans ce cas, les deux formes sont équivalentes.

En reprenant l'équation ci-dessus, nous pouvons obtenir l'équivalence suivante :

$$T = e^{\beta^T X} \cdot e^{\beta_0 + \sigma \epsilon} = e^{\beta^T X} T_0 = e^{-\theta^T X} T_0 \quad (63)$$

$$T_0 = e^{-\beta^T X} T \quad (64)$$

où T_0 représente la durée de vie "de base", c'est-à-dire la durée de vie qui ne dépend pas des variables explicatives, contrairement à T .

Cette formulation permet d'interpréter les modèles AFT de manière intuitive. Par exemple, si nous considérons une variable explicative représentant le sexe, où 1 correspond à une femme et 0 à un homme, l'équation nous permet de dire que le temps s'accélère ou ralentit pour les femmes en fonction de la valeur du coefficient β . Si le coefficient est positif, la durée s'accélère, et inversement. Le facteur $e^{-\beta^T X}$ modifie le temps par rapport à la durée de base.

La fonction de survie de la variable aléatoire T peut se réécrire en fonction d'un terme d'erreur ϵ . La démonstration complète est disponible en annexe. Soit :

$$S(T|X) = \mathbb{P}(T > t|X) \quad (65)$$

$$= S_\epsilon \left(\log \left(\frac{t}{e^{\beta^T X + \beta_0}} \right)^{\frac{1}{\sigma}} \right) \quad (66)$$

$$= S_\epsilon(\psi) \quad (67)$$

avec $\psi = \log \left(\frac{t}{e^{\beta^T X + \beta_0}} \right)^{\frac{1}{\sigma}}$ et $S_\epsilon(\psi) = \mathbb{P}(\epsilon > \psi)$.

Après avoir présenté les différentes formes du modèle AFT, nous allons à présent exposer les lois couramment utilisées dans ces modèles de régression paramétrique. Dans les modèles AFT, il est nécessaire de spécifier une distribution pour la durée de la variable $T|X$ (ou, de manière équivalente, de supposer une distribution pour le terme d'erreur ϵ). Établir un lien entre la loi de distribution de $T|X$ et celle de ϵ nous permet de déterminer les paramètres de la première loi en fonction du vecteur de variables aléatoires et des estimations.

Le tableau suivant 3 récapitule le lien entre les deux lois, soit :

Distribution de $T X$	Distribution de ξ
Exponentielle	Valeur extrême
Weibull	Gumbel
Log Logistique	Logistique
Log Normale	Normale
Gamma	Log Gamma

Table 3: Distributions pour le modèle AFT

Dans notre cas, nous allons présenter les modèles avec les trois mêmes lois candidates utilisées lors de la modélisation marginale de la durée du maintien en incapacité, à savoir :

- La loi de Weibull
- La loi Log logistique
- La loi Gamma généralisée

Le fait de spécifier ces lois nous permettra de comparer les fonctions de survie obtenues du modèle AFT à celles obtenus sans modèle de régression.

AFT paramétrique avec la loi Weibull

Pour démontrer le modèle AFT avec une loi de Weibull, nous allons partir de la loi du terme d'erreur, à savoir une Gumbel. A l'aide de cette loi, nous allons pouvoir retrouver la loi de durée de $T|X$, soit:

$$f_{\epsilon}(t) = e^t e^{-e^t} \quad (68)$$

$$\epsilon = g^{-1}(T) = \frac{\log(T) - \beta X}{\sigma} \quad (69)$$

$$f_T(t) = f_{\epsilon}(g^{-1}(T)) = e^{\frac{\log(t) - \beta X}{\sigma}} e^{-e^{\frac{\log(t) - \beta X}{\sigma}}} \quad (70)$$

$$= \frac{1}{e^{\beta X}} \left(\frac{t}{e^{\beta X}} \right)^{\frac{1}{\sigma} - 1} \exp \left(- \left(\frac{t}{e^{\beta X}} \right)^{\frac{1}{\sigma}} \right) \quad (71)$$

En comparant cette densité à celle de la fonction de densité classique d'une Weibull $W(\rho, \lambda)$, qui est pour rappel (λ le paramètre d'échelle et ρ le paramètre de forme) :

$$f_T(t) = \frac{\rho}{\lambda} \left(\frac{t}{\lambda} \right)^{\rho-1} \exp \left(- \left(\frac{t}{\lambda} \right)^{\rho} \right) \quad (72)$$

$$S_T(t) = \exp \left(- \left(\frac{t}{\lambda} \right)^{\rho} \right) \quad (73)$$

Nous pouvons retrouver la densité d'une loi de weibull, $W(\rho, \lambda)$ avec comme paramètre, $\rho = \frac{1}{\sigma}$ et $\lambda = \exp(\beta X)$. Nous avons à présent la loi de Weibull suivante : $T|X \sim W(\lambda = \exp(\beta X), \rho = \frac{1}{\sigma})$. La fonction de survie est de la forme suivante.

$$S(t|X) = \exp\left(-\left(\frac{t}{e^{\beta T X}}\right)^{\frac{1}{\sigma}}\right) = \exp\left(-\left(\frac{t}{\lambda}\right)^{\rho}\right) \quad (74)$$

Pour estimer les paramètres du modèle AFT Weibull, nous utilisons la méthode du maximum du vraisemblance de la même manière que pour une loi sans variable explicative. Soit :

$$Y = \log(T) \quad (75)$$

$$L(\beta, \sigma, y_i) = \prod_{i=1}^n \left(\frac{f_Y(y_i)}{S(\tau)}\right)^{\delta_i} \cdot \left(\frac{S_Y(y_i)}{S(\tau)}\right)^{1-\delta_i} \quad (76)$$

En reproduisant la même démarche, nous pouvons obtenir la loi conditionnelle Log-logistique.

AFT paramétrique avec la loi log-Logistique

On suppose que le terme d'erreur suit la logistique, soit $\epsilon \sim \text{Logistique}$, alors dans ce cas nous obtenons la fonction de survie suivante :

$$S_{\epsilon}(\psi) = \mathbb{P}(\epsilon > \psi) = \frac{1}{1 + e^{\psi}} = \frac{1}{1 + \left(\frac{t}{e^{\beta T X + \beta_0}}\right)^{\frac{1}{\sigma}}} \quad (77)$$

$$= \frac{1}{1 + \left(\frac{t}{\lambda}\right)^{\frac{1}{\sigma}}} = S(t|X) \quad (78)$$

donc $T|X \sim \text{loglogistique}(\lambda = e^{\beta T X + \beta_0}, \frac{1}{\sigma})$.

Il existe plusieurs méthodes pour sélectionner la meilleure loi candidate. Nous allons principalement utiliser les critères d'information AIC/BIC.

2.4.3 Modèle GLM

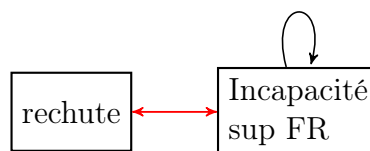


Figure 38: Modélisation des rechutes

Dans le cadre de ce mémoire, nous avons décidé d'utiliser le modèle linéaire généralisé avec une loi de comptage pour modéliser la fréquence des rechutes à partir de la base de données des arrêts.

Pour rappel, cette base de données nous permet de modéliser une **durée de maintien en arrêt**. Cependant, l'objectif visé est de modéliser la durée d'incapacité totale du sinistre. Afin d'estimer cette durée, il est essentiel de connaître le nombre de rechutes qu'un individu peut subir.

Pour cela, nous allons utiliser le modèle linéaire généralisé avec une loi de comptage. Ce modèle nous permettra d'obtenir l'espérance du nombre de rechutes en fonction des caractéristiques des individus. Avec les résultats de la régression, nous allons pouvoir récupérer les paramètres de la loi conditionnelle sous-jacente. Cette distribution sera utilisée pour simuler le nombre de rechutes lors de la tarification des traités de réassurance. Les réalisations obtenues à partir de la loi représenteront le **nombre de tirages** à effectuer dans la loi de durée (des arrêts).

Modèle mathématique

Le modèle GLM (*Generalized Linear Models* en anglais) est couramment utilisé en assurance non-vie pour modéliser la fréquence ou la sévérité des sinistres attritionnels¹⁴. Ces modèles économétriques sont une extension des modèles linéaires dans le sens où plusieurs hypothèses sont relâchées. Ou inversement, le modèle linéaire est un cas particulier des modèles linéaires généralisés avec plusieurs hypothèses restrictives. Avant d'introduire les GLM, nous introduisons les modèles linéaires classiques.

Modèle Linéaire

Hypothèses¹⁵

- H1 Linéarité : $Y = X\beta + \epsilon$ La relation en y et les variables explicatives est linéaire
- H2 Plein rang : La matrice X de variables explicatives est de rang k , cela signifie que les p colonnes sont des vecteurs linéairement indépendants.
- H3 Exogénéité des variables explicatives : $E[\epsilon] = 0 \Leftrightarrow E[Y|X] = X\beta$ La perturbation est d'espérance nulle
- H4 Homoscédasticité et absence d'autocorrélation : $E[\epsilon\epsilon'] = \sigma^2 I_N \Leftrightarrow \begin{cases} E[\epsilon_i^2] = V(Y|X) = \sigma^2 \forall i \\ E[\epsilon_i \epsilon_{i'}] = 0 \forall i \neq i' \end{cases}$
- H5 Génération des données : Les observations peuvent être des constantes ou des variables aléatoires
- H6 Distribution normale : $\epsilon \sim N(0, \sigma^2 I_N) \Leftrightarrow Y|X \sim N(X\beta, \sigma^2 I_N)$
Cette hypothèse permet d'obtenir l'équivalence des MCO avec l'estimateur du maximum de vraisemblance.

Théorème de Gauss Markov

Sous les hypothèses H1 à H4, l'estimateur des moindres carrés ordinaires $\hat{\beta}$ est un estimateur BLUE (Best Linear Unbiased Estimator *en anglais*) de β .

$$\hat{\beta} = (X'X)^{-1}X'y$$

Limites du modèle linéaire

Le modèle linéaire est mal adapté pour les variables à expliquer binaire ou pour les données de comptages. En effet, le terme à gauche appartient à $(0, 1)$ tandis que le terme à droite appartient à \mathbb{R} .

¹⁴Sinistres avec une fréquence élevée et une sévérité faible

¹⁵Voir livre économétrie de W.Green

$$\underbrace{E[Y|X]}_{\in(0,1)} \neq \underbrace{X\beta}_{\in\mathbb{R}}$$

De plus l'hypothèse d'homoscédasticité ne permet pas de traiter des données avec des variances conditionnelles différentes. Il convient donc d'utiliser les modèles linéaires généralisés afin de palier ces limites. Ci-dessous, un graphique¹⁶ permettant de visualiser les différences :

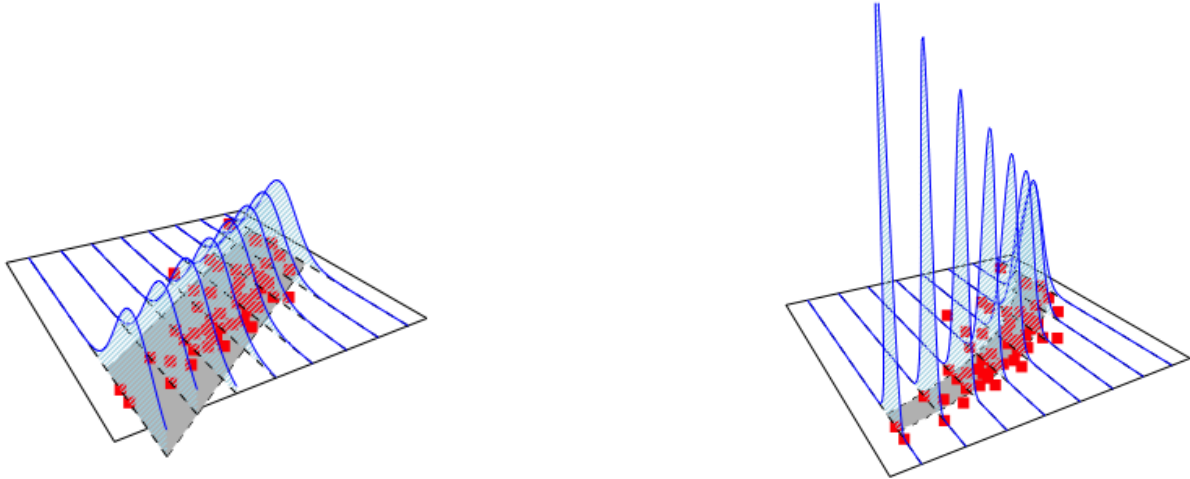


Figure 39: Gauche : modèle linéaire | Droite : Modèle GLM log Poisson

Modèle linéaire généralisé

Hypothèses

- $Y|X \sim \mathcal{L}(\theta, \psi)$: La loi conditionnelle de la variable à expliquer appartient à la famille de loi Exponentielle.
- Un ensemble de variables explicatives (x_1, x_2, \dots, x_p) de rang p (donc inversible) et les paramètres associés $(\beta_1, \beta_2, \dots, \beta_p)$
- Une fonction g , monotone et dérivable (appelé fonction de lien) tel que :

$$\mu = E[Y|X] = g^{-1}(X\beta) = b'(\theta)$$

Famille exponentielle

La famille exponentielle regroupe les lois de probabilité dont la fonction de masse ou de densité est représentée par : $f(y, \theta, \psi) = \exp\left(\frac{y\theta - b(\theta)}{a(\psi)} + c(y, \psi)\right)$

Selon les paramètres¹⁷, nous retrouvons chacune des lois usuelles. Ci-dessous, le tableau regroupant les différentes fonctions de lien.

$$\begin{aligned} \mu &= E[Y|X] = b'(\theta) = g^{-1}(X\beta) \\ \theta &= g(\mu) = g(E[Y|X]) = g(b'(\theta)) \end{aligned}$$

Loi	$g(\mu) = \theta$	$\mu = g^{-1}(X\beta)$	Var
$N(\mu, \sigma^2)$	μ	$X\beta$	1
$Ber(\mu)$	$\log(\frac{\mu}{1-\mu})$	$\frac{1}{1+e^{-X\beta}}$	$\mu(1-\mu)$
$Pois(\mu)$	$\log(\mu)$	$e^{X\beta}$	μ
$Gam(\alpha, \beta)$	$\frac{1}{\mu}$	$(X\beta)^{-1}$	μ^2
$Tweedie(\mu, p, \phi)$	$\log(\mu)$	$e^{X\beta_{Tweedie}}$	$\frac{\phi+2}{\phi+1}$

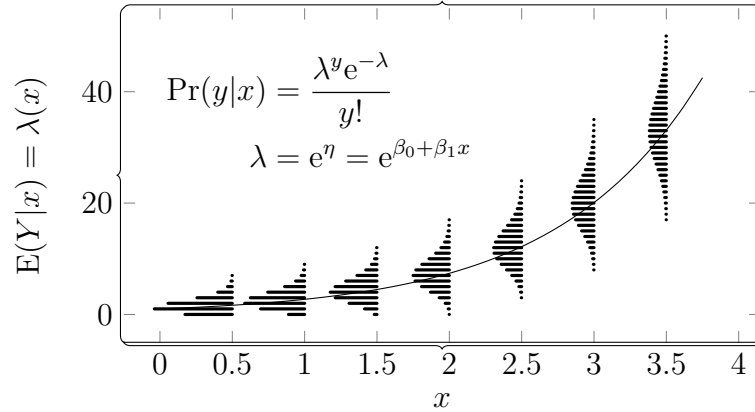


Figure 40: Représentation d'un modèle GLM Poisson (1 variable)

La figure 40 représente une illustration d'un modèle GLM de type poisson avec une variable explicative et la fonction de lien logarithme. Nous remarquons que la loi de poisson conditionnelle varie en fonction des modalités de la variable explicative.

La loi que nous utiliserons dépendra du paramètre de dispersion.

- Si $E[N] > V(N)$: Loi Géométrique
- Si $E[N] = V(N)$: Loi de Poisson
- Si $E[N] < V(N)$: Loi Binomiale Négative

Le paramètre de dispersion sera estimé par la formule suivante :

$$\phi = \frac{S_N^2}{m_N} = \frac{\frac{\sum_{i=1}^n (Y_i - m_N \cdot E_i)^2}{\sum_{i=1}^n E_i}}{\frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n E_i}} \quad (79)$$

Dans le cas d'une régression avec des variables explicatives, nous pouvons calculer le paramètre de dispersion par modalité. Soit:

$$\phi = \frac{\frac{\sum_{i=1, X_i=x}^n (Y_i - m_N \cdot E_i)^2}{\sum_{i=1, X_i=x}^n E_i}}{\frac{\sum_{i=1, X_i=x}^n Y_i}{\sum_{i=1, X_i=x}^n E_i}} \quad (80)$$

La section suivante résume les modèles que nous utiliserons par la suite.

¹⁶Graphique issu du site d'A.Charpentier

¹⁷Voir tableau en annexe

2.5 Synthèse et transition

L'objectif de cette section est de récapituler tous les estimateurs et modèles présentés ci-dessus. Le tableau 4 récapitule les estimateurs et les modèles utilisés selon la loi modélisée.

Modèles/Lois	Incapacité	Maintien	Invalidité
Non-Paramétrique	K-M/N-A	K-M	K-M
Paramétrique	Hoem	Lois de distribution et modèle AFT	
Semi-Paramétrique	Modèle Cox	Modèle Cox	Modèle Cox

Table 4: Récapitulatif des modèles/estimateurs utilisés selon la loi (K-M : Kaplan-Meier, N-A : Nelson-Aalen)

Remarque : Étant donné les résultats non concluants du modèle de Cox pour le passage et le maintien en incapacité, ce dernier ne sera pas modélisé pour le passage en invalidité.

Pour chacun de ces modèles, la durée considérée diffère en fonction de la loi. Le tableau 5 récapitule les différentes durées et variables d'intérêt en fonction de la loi étudiée.

Pour la loi d'incidence, la variable d'intérêt est la durée passée dans l'état sain avant la transition vers l'état d'incapacité. Pour les autres lois (maintien et passage en invalidité), la variable d'intérêt est la durée passée dans l'état d'incapacité.

Variables/Lois	Incapacité	Maintien	Invalidité
Temps étudié	T_{sain}	$T_{incapacite}$	$T_{incapacite}$
Variable de discrétisation	âge	durée incapacité	durée incapacité
Variable d'intérêt	q_x	$f(t)$	q_t

Table 5: Récapitulatif des variables d'intérêts selon la loi

Pour utiliser un estimateur non-paramétrique, il est nécessaire de discrétiser le temps. La discrétisation du temps se fait en fonction de la loi modélisée. Pour la loi d'incidence, le temps est discrétisé en fonction de chaque âge auquel un individu entre ou sort de l'étude. Pour les autres lois, le temps est discrétisé en fonction de chaque durée de maintien à laquelle un individu retourne à l'état sain ou passe en invalidité.

La variable à estimer pour le passage dans l'état d'incapacité ou d'invalidité est le **taux d'incidence brut**. Il représente la probabilité qu'un individu passe dans l'état d'incapacité ou d'invalidité entre l'âge x et $x + 1$, sachant qu'il a l'âge x . Mathématiquement, il s'exprime comme suit :

$$q_x = \mathbb{P}(x \leq T < x + 1 | T \geq x) = \frac{\mathbb{P}(x \leq T < x + 1)}{\mathbb{P}(T \geq x)} = 1 - \frac{S(x + 1)}{S(x)} = 1 - \frac{\frac{l_{x+1}}{l_0}}{\frac{l_x}{l_0}} = 1 - \frac{l_{x+1}}{l_x} \quad (81)$$

Remarque : Les probabilités de maintien en incapacité sont utilisées pour construire les tables de provisionnement. Elles sont en général représentées en terme de nombre de survivants. Le calcul de ces tables repose sur une cohorte initiale d'individus, généralement de 10 000 (ou 100 000) individus. À chaque pas de temps, le nombre d'individus est multiplié par la probabilité de survie $p_x = 1 - q_x$. Chaque ligne de la table représente le nombre de survivants restants, noté l_x .

$$l_0 = 10000$$

$$l_x = l_0 S_X(x)$$

La densité de probabilité $f(t)$ est la variable d'intérêt dans le contexte du maintien en incapacité.

Le graphique suivant illustre les différentes étapes de création des lois d'incidence, de maintien et de passage en invalidité.

La partie suivante traite de l'application sur le portefeuille de la MACSF. Elle est divisée en trois chapitres. Le premier porte sur la présentation du portefeuille de la MACSF et la construction de la base de données. Le chapitre suivant concerne la calibration sur la base de données des sinistres. Enfin, le dernier chapitre aborde la calibration sur la base de données des arrêts.

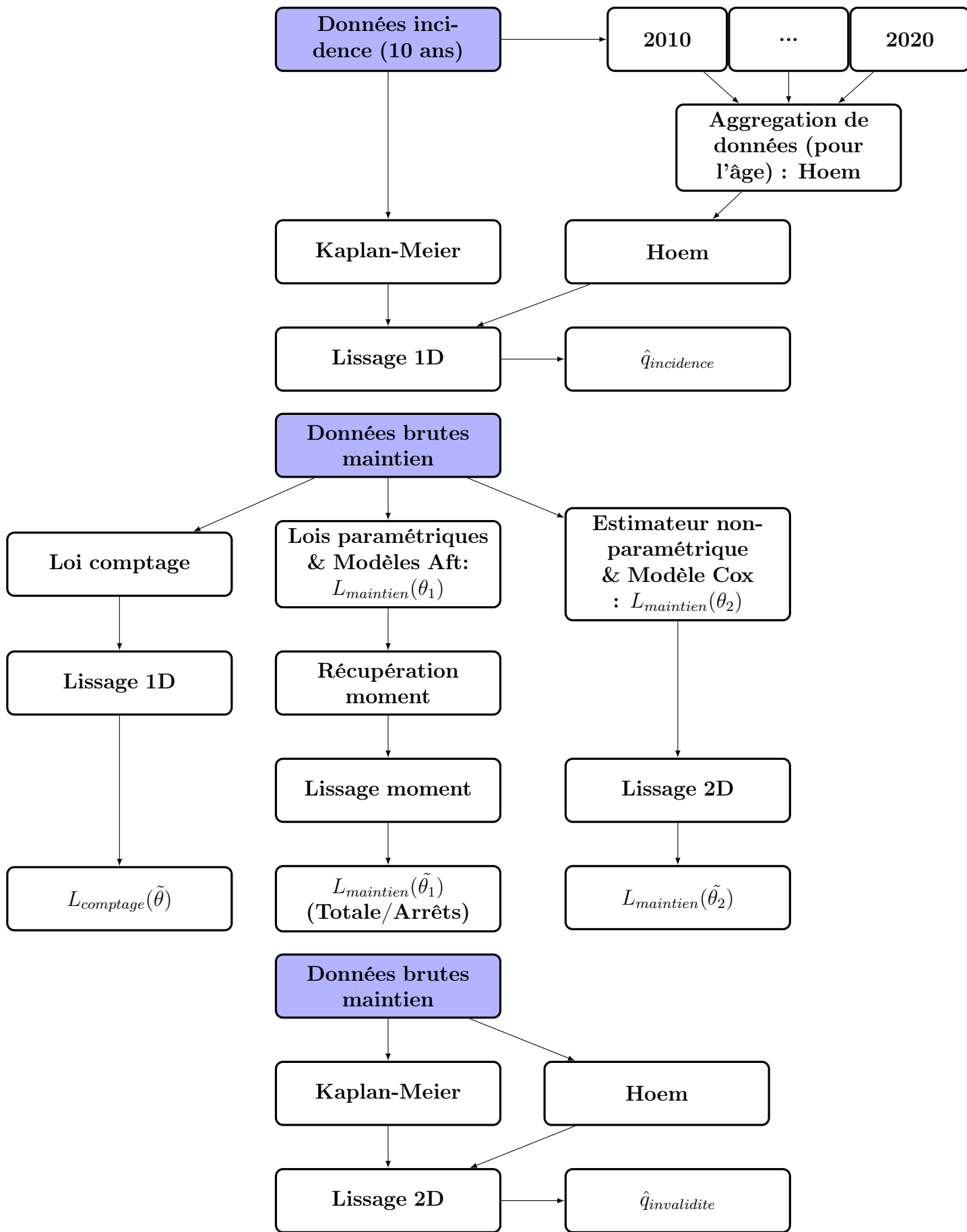


Figure 41: Récapitulatif des modèles utilisés

3 Calibration sur le portefeuille MACSF

3.0.1 La mutuelle MACSF

La MACSF (Mutuelle d'Assurance du Corps de Santé Français) est une société d'assurance mutuelle française fondée en 1935. Le groupe est présidé et gouverné depuis sa création par des administrateurs issus également des professions médicales.

Les sociétés constituant le périmètre de consolidation du groupe MACSF sont les suivantes:

- MACSF SGAM, Société de groupe d'Assurance Mutuelle, a pour objectif de mettre en place et de gérer des liens de solidarité financière durables entre les sociétés affiliées ; elle fédère les sociétés d'assurance mutuelles en y exerçant une influence dominante, afin de veiller et de garantir leur bonne santé et solvabilité.
- MACSF assurances, société d'assurance non vie, pratique les risques accidents corporels et maladie, ainsi que tous les risques dommages.
- MACSF prévoyance, société d'assurance vie.
- MACSF épargne retraite, société d'assurance vie.
- MACSF Libéa, société d'assurance non vie, pratique les risques accidents corporels et maladie, ainsi que tous les risques dommages y compris l'assistance.
- MACSF RÉ S.A, société de réassurance de droit luxembourgeois.

Le graphique suivant illustre l'organigramme des sociétés constituant le groupe MACSF.

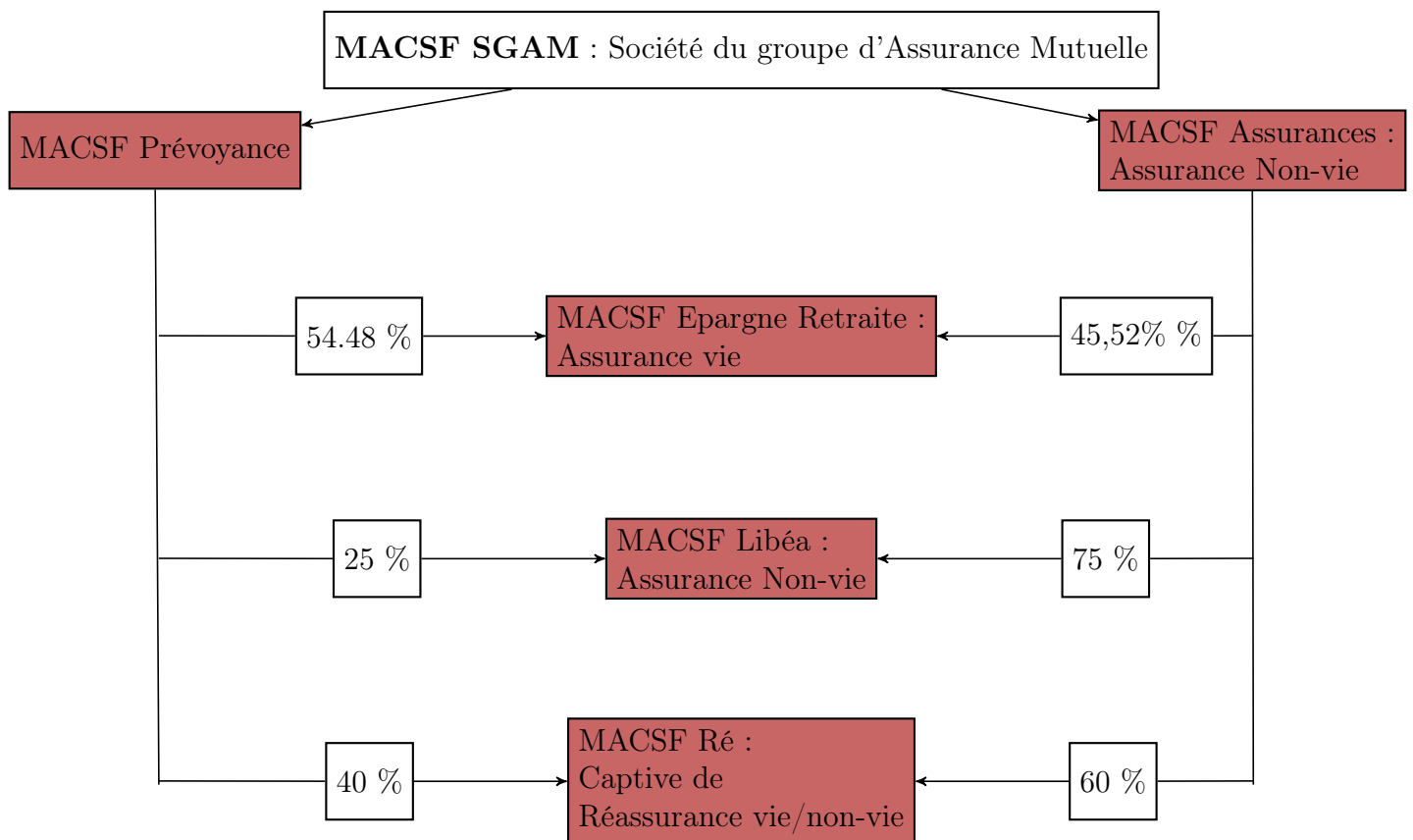


Figure 42: Organigramme MACSF

Dans ce mémoire, nous avons principalement travaillé sur MACSF Prévoyance et MACSF Ré.

Le groupe MACSF génère un chiffre d'affaires de 2 228 millions d'euros, dont 1 429 millions d'euros en lien avec l'activité d'épargne et de retraite, 752 millions d'euros sur les activités d'assurance IARD, RCP et assurance de personnes et 47 millions d'euros sur les autres activités.

Le résultat consolidé porté par toutes les activités du groupe est de 247 millions d'euros. De plus, le groupe possède un portefeuille de 2,32 millions contrats.

Enfin, le groupe propose une gamme variée de produits d'assurance destinés aux professionnels de la santé ainsi qu'à leurs proches. Le tableau suivant récapitule les produits en fonction des différentes entités juridiques.

	SGAM	Epargne Retraite	Assurances	Prévoyance	Libéa	Ré
Assurance automobile			X		X	X
Bateaux de plaisance			X			
Assurance dommages			X		X	X
Garantie des accidents de la vie			X			
Catastrophes naturelles			X		X	X
Responsabilité civile générale			X		X	X
Responsabilité civile professionnelle			X		X	X
Protection juridique			X		X	
Pertes pécuniaires diverses			X			
Assistance			X		X	
Plan de prévoyance			X	X	X	X
Assurance emprunteur			X	X		X
Santé individuelle			X		X	
Santé collective			X		X	
Epargne, retraite et PERP		X	X			

Figure 43: Produits d'assurance du groupe MACSF

Le portefeuille de la MACSF est atypique dans le sens où les sociétaires sont des professionnels de santé. Les produits d'assurance visent principalement les métiers suivants :

- Internes à l'hôpital
- Docteurs
- Jeunes diplômés
- Infirmiers
- Proches des professionnels de santé

Pour chacune de ces catégories de métiers, les produits sont définis en fonction du lieu d'exercice qui peut être à l'hôpital, dans un établissement privé ou bien dans un cabinet.

Les produits principalement vendus par la MACSF sont la Responsabilité Civile Professionnelle et Protection Juridique (RCP-PJ), les produits de prévoyance et la complémentaire santé. La RCP-PJ permet de se protéger en cas de mise en cause par un patient (RCP), tandis que

la PJ permet d'obtenir une assistance pour tout autre litige d'ordre privé ou professionnel. La complémentaire santé permet d'obtenir des remboursements des frais de santé.

La section suivante porte sur l'analyse des données du portefeuille concernant les garanties incapacité et invalidité.

3.0.2 Présentation des données

Dans cette section, nous allons présenter en détail la création des différentes bases de données qui seront utilisées pour la modélisation. Pour rappel, nous allons créer deux bases de données pour la modélisation du maintien en incapacité : la première portant sur la **durée du sinistre totale** et une autre portant sur les **arrêts**. Pour une meilleure compréhension de la construction, nous reprenons la figure utilisée en introduction.

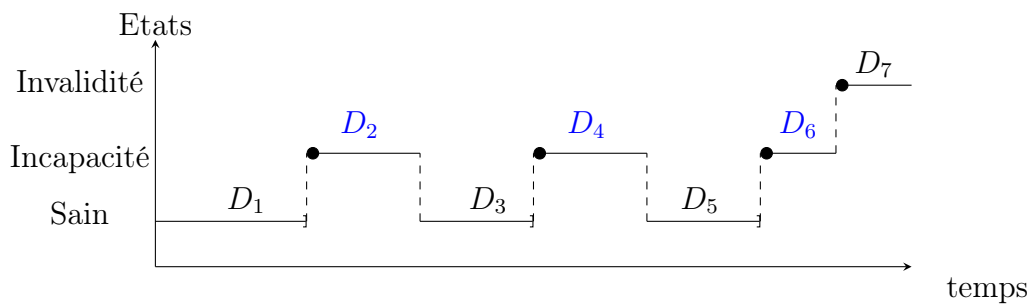


Figure 44: Arrêts de travail

La modélisation des sinistres diffère de la modélisation des arrêts de travail. En effet, comme nous pouvons le voir à travers la figure 44, un individu peut connaître plusieurs périodes d'arrêt de travail pour un seul et même sinistre. Par conséquent, la durée totale d'un sinistre sera différente de celle des arrêts de travail.

Notre base de données initiale (**BDD des arrêts**) comporte la base de données d'exposition et la base de données **d'indemnisation**. Cela signifie que tous les sinistres que nous observons sont des coûts que nous avons payés. Dans cette base de données, nous observons les individus par période. C'est-à-dire que pour chaque individu, nous avons la date à laquelle il est entré en incapacité et le nombre de jours associé.

Remarque : Pour rappel, la base de données de la **durée de maintien totale** représente la base de données des **arrêts** filtrée sur les périodes d'incapacité. De plus, les durées des arrêts ont été agrégées par sinistre.

Pour le **passage en incapacité**, nous avons considéré que le **premier passage associé à un sinistre** dans l'état d'incapacité représentait le seul passage dans cet état, tandis que les rechutes associées au même sinistre étaient considérées comme des censures. En considérant que toutes les rechutes sont des censures, nous conservons l'exposition de l'individu avant son passage dans l'état d'incapacité et n'ajoutons pas de nouveaux passages dans cet état. La table suivante représente une illustration de la base de données utilisée.

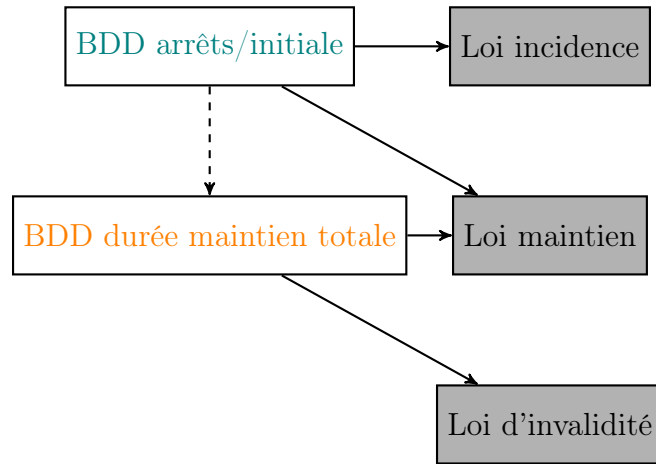


Figure 45: Bases de données

Id	Age_{Entree}	Age_{Sortie}	$Incap_{sinistre}$	$Duree$
1	t_1	t_2	1	D_1
1	t_3	t_4	"cens"	D_3
1	t_5	t_6	"cens"	D_5

Table 6: Construction de la table pour le passage en incapacité, **BDD arrêts**

Pour la **loi du maintien en incapacité**, c'est-à-dire la loi de probabilité qui modélise la durée du temps passé dans l'état d'incapacité, nous utiliserons deux approches différentes et donc deux bases de données (BDD des **arrêts** filtrée sur les périodes d'incapacité et la BDD de la **durée de maintien totale**).

Pour la modélisation de la **durée totale du sinistre**, nous allons sommer toutes les durées associées au même sinistre. En reprenant l'exemple présenté ci-dessus sur la figure 44, la durée de maintien du sinistre sera donc égale à $D_{totale} = D_2 + D_4 + D_6$. La table 7 représente une illustration de la base de données de la **durée de maintien totale**. Nous pouvons remarquer dans cette table que nous avons une ligne avec comme âge d'entrée, l'âge de l'individu lors de la survenance de la première incapacité, et une colonne "Inval" représentant une indicatrice du passage dans l'état d'invalidité. Cette indicatrice sera utilisée pour définir l'événement de sortie pour la construction des probabilités du passage en invalidité. L'âge de sortie n'est pas spécifié en raison de l'agrégation des durées dans l'état d'incapacité. Cette manipulation des données ne permet pas d'indiquer l'âge auquel l'individu est passé en invalidité. En effet, entre l'âge de survenance du sinistre et l'âge au moment du passage en invalidité, les périodes saines n'ont pas été prises en compte dans le calcul.

Id	Age_{Entree}	Age_{Sortie}	Inval	Durée
1	t_1	NA	1	$D_2 + D_4 + D_6$

Table 7: Base de données de la **durée de maintien totale**

La table 8 représente la base de données des **arrêts** pour le maintien en incapacité. Pour cette modélisation, la durée sera celle associée à chaque arrêt de travail et ne sera donc pas agrégée.

Id	Age Entrée	Age Sortie	Inval	Durée
1	t_2	t_3	0	D_2
1	t_4	t_5	0	D_4
1	t_6	t_7	1	D_6

Table 8: BDD des **arrêts** filtrée sur les périodes d'incapacité

Nous allons donc modéliser une **loi de durée totale du sinistre** d'un côté et une **loi de durée des arrêts** de l'autre.

Pour récapituler, nous avons créé deux bases de données distinctes : l'une portant sur les sinistres (incluant le passage en incapacité du sinistre, la modélisation de la durée totale du sinistre et le passage en invalidité du sinistre), et l'autre sur les arrêts de travail.

En fonction de la statistique à calculer, nous avons utilisé les bases de données suivantes :

- Taux d'incidence : BDD des **arrêts** filtrée sur les périodes saines
- Loi du maintien :
 - Loi de durée totale : BDD de la durée de **maintien totale**
 - Loi de durée des arrêts : BDD des **arrêts** filtrée sur les périodes d'incapacité
- Taux de passage en invalidité : BDD de la durée de **maintien totale**

La section qui porte sur la **calibration des arrêts** porte seulement sur la modélisation de la durée des arrêts et des rechutes à l'aide de la base de données des **arrêts** filtrée sur les périodes d'incapacité. Pour le reste, nous parlerons de **calibration des sinistres**.

En général, la base de données des arrêts de travail est la plus couramment utilisée, mais nous avons souhaité créer ces deux bases pour une raison bien spécifique. L'idée étant qu'avec la base de données de la durée de maintien totale, nous pouvons générer une durée de maintien représentant la durée totale du sinistre (donc la somme de chaque rechute). Cette méthode permet de générer des simulations sans trop de complexité. En revanche, en ce qui concerne les **rechutes**, cette méthode ne permet pas d'analyser l'effet des rechutes sur la durée totale d'incapacité. C'est dans ce cas-là qu'intervient la seconde base de données. La seconde base de données nous permet de générer dans un premier temps un nombre de rechutes et une durée associée à chaque arrêt de travail. Ce sont deux approches distinctes. L'idée est donc d'utiliser ces deux approches et voir si les résultats finaux convergent.

Pour construire ces base, nous avons dans un premier temps extrait les données de la MACSF à l'aide du logiciel SAS. Ensuite, les données ont été retravaillées afin d'avoir une structure permettant d'utiliser les estimateurs présentés dans la partie sur la théorie des phénomènes de durées. Au final, après ces manipulations, nous avons obtenu une base de données initiale de 307 124 lignes et 19 variables.

Parmi ces variables, nous avons la variable représentant le sexe de l'individu, *qualite*, qui est égale à 1 pour un homme et 0 pour une femme.

La variable *Grp_{calibration}* a été encodé de la manière suivante :

- 1 si Salarié d'un hôpital (non modélisé)
- 2 si Infirmiers et assimilés

- 3 si Paramédicaux et assimilés
- 4 si Médecins et assimilés

La variable *etat* représente l'état dans lequel l'individu se trouve actuellement, elle est égale à *sain* ou à *incap*.

Dans notre base de données, nous avons aussi la variable *sin_{num}* et *arret_{num}*. La première représente le numéro de sinistre associé à l'individu et la seconde correspond au numéro d'arrêt de travail pour le sinistre associé. Il existe une indicatrice *isRechute* valant 1 en cas de rechute et 0 sinon. Dans le cas où la variable *arret_{num}* est égale à 2 ou plus, alors c'est obligatoirement une rechute.

Pour identifier les sinistres relatifs à une grossesse, une indicatrice *isGross* a été créée. De même que pour identifier les sinistres finissant en invalidité, la variable *isInval* est égale à 1 en cas de passage en invalidité et 0 sinon.

Pour terminer, chaque ligne possède un numéro d'assuré et une clé de sinistre. Cette clé correspond à la concaténation du numéro d'assuré et du numéro de sinistre.

A l'aide de toutes ces variables, nous avons pu créer nos deux bases de données distinctes.

Pour la base de données des **arrêts**, nous avons pour chaque numéro de sinistre, récupéré les périodes en état sain pour la modélisation de l'incidence. Pour chacune de ces périodes saines, nous avons créé une variable *isInffran* indiquant si la période saine est associée à un sinistre inférieur à la franchise. Pour chaque période, nous avons construit la variable *isIncap* égale à 1 en cas de passage en incapacité (toutes causes) ou 0 dans le cas contraire.

Pour le calcul des taux d'incidence bruts, 3 indicatrices ont été construites, soit :

- $isIncap_{sup} = 1$ si :

- $isIncap = 1$
- $isGross = 0$
- $isInffran = 0$

- $isIncap_{inf} = 1$ si :

- $isIncap = 1$
- $isGross = 0$
- $isInffran = 1$

- $isIncap_{Gross} = 1$ si :

- $isIncap = 1$
- $isGross = 1$

Une fois chaque période saine extraite, nous avons calculé l'âge d'entrée et de sortie de l'état à l'aide de la variable associée à l'âge de naissance, $date_{naissance}$. Ces variables sont notées age_{entree} et age_{sortie} . Ces variables seront nécessaires pour discrétiser le temps pour l'utilisation de l'estimateur de Kaplan-Meier.

Pour la modélisation des arrêts, nous avons récupéré les périodes d'incapacité de la BDD des **arrêts**.

Pour créer la base de données correspondant à la durée totale et au passage en invalidité nous avons, pour chaque *clé de sinistre*, sommé la durée passée en état d'incapacité et nous

l'avons stocké dans la variable $nbr_{jours_{sum}}$.

La figure suivante récapitule le nombre de lignes restantes pour chacune des bases de données:

	Nombre de lignes	Nombre d'assurés
Base de données des arrêts (arrêts/initiale)	307 121	129 506
Base de données des arrêts (incidence)	227 894	129 506
Base de données des arrêts (maintien)	79 230	37 731
Base de données de maintien (durée totale)	55 397	37 731

Table 9: Nombres de ligne des bases de données

Maintenant, que nous disposons de ces bases de données, nous pouvons passer à la présentation des caractéristiques de chacune d'entre elles. La section suivante fournit des statistiques descriptives permettant d'analyser le contenu et les caractéristiques de ces bases.

3.1 Calibration sur des sinistres

Cette section est consacrée à l'analyse descriptive de la base de données des **arrêts filtrée sur les périodes saines** et celle qui porte sur la **durée totale du maintien en incapacité**. Nous présentons dans un premier temps les données utilisées pour modéliser l'incidence, puis les données associées au maintien en incapacité et à l'invalidité.

3.1.1 Statistiques descriptives

Base de données de l'incidence en incapacité

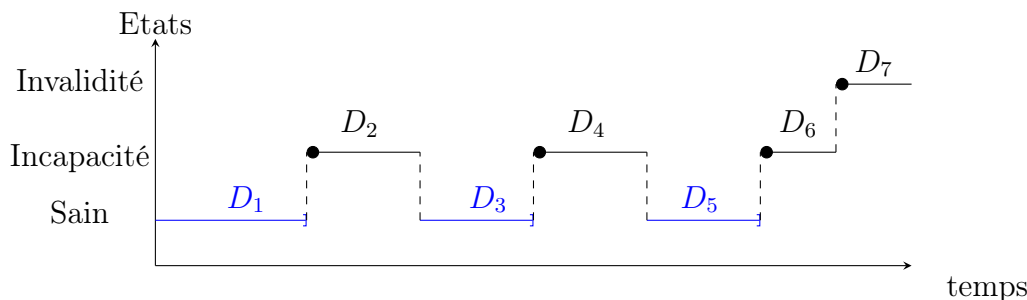


Figure 46: BDD des arrêts filtrée sur les périodes saines

Dans cette sous-base de données, se trouvent donc toutes les durées en **état sain** des individus. Sur la figure 46, la durée analysée est représentée par la couleur bleue. Dans le but d'obtenir des résultats significatifs pour la modélisation, nous avons choisi de prendre en compte une période de **10 ans**, de 2010 à 2019. L'année 2020 et 2021 n'ont pas été incluses en raison de leur caractère exceptionnel lié à la pandémie de Covid-19.

La base de données des arrêts utilisée pour l'incidence a été segmentée par année afin de détecter d'éventuelles tendances temporelles. La figure 47 présente l'exposition du portefeuille.

La partie de gauche du graphique concerne l'exposition totale du portefeuille. Un filtre a été appliqué sur la base de données représentée. Ce dernier concerne le statut des assurés. Nous avons gardé uniquement les praticiens libéraux qui correspondent au périmètre de l'étude demandée. Ce choix sera expliqué en détail dans la description du prochain graphique.

Nous pouvons observer une légère tendance à la hausse de l'exposition du portefeuille de la MACSF au fil des années. En effet, l'exposition de la base est passée d'environ 60 000 années d'assurance à près de 70 000 années d'assurance. La partie de droite du graphique met en évidence une forte augmentation du nombre de femmes de 2010 à 2017. Après cette année, il y a une légère stagnation jusqu'en 2018, suivie d'une augmentation l'année suivante. En revanche, l'exposition des hommes n'a pas connu d'évolution significative, voire a légèrement diminué.

Les individus peuvent être regroupés en trois catégories de professions :

- Les infirmiers et assimilés
- Les paramédicaux et assimilés
- Les médecins et assimilés

Ces catégories de professions ont été créées à l'aide de l'expertise et le jugement d'experts. Il a été conclu que ces catégories possèdent en général les mêmes risques de sinistralité.

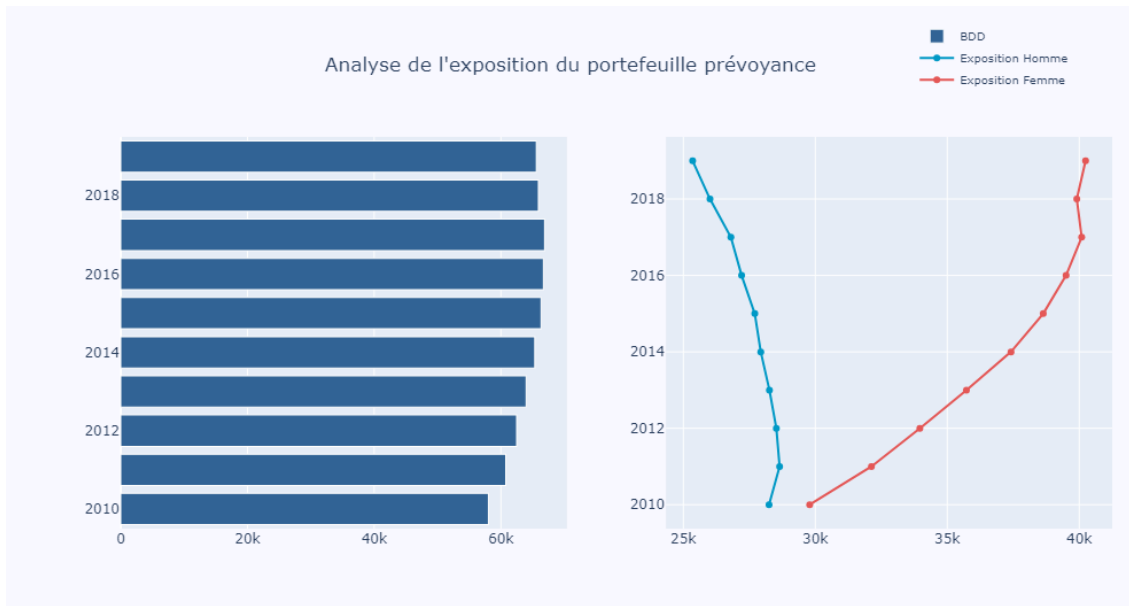


Figure 47: BDD sinsitres : Exposition totale et exposition par franchise

La Table 10 présente un échantillon partiel des différentes professions incluses dans chaque groupe.

Métier	Groupe Csp
Infirmiers	Infirmiers et assimilés
Aide soignant	
Agent hospitalier	
Assistant dentaire et médical	Paramédicaux et assimilés
Kinésithérapeute	
Sage femme	
Opticien	
Autres	
Médecin	Médecins et assimilés
Chirurgien dentiste	
Pharmacien	
Vétérinaire	

Table 10: Classification CSP

Les catégories présentées regroupent les professionnels de santé exerçant en tant que libéraux, ayant souscrit une assurance prévoyance individuelle. Les médecins sont principalement regroupés avec les chirurgiens, les dentistes, les pharmaciens et les vétérinaires. La catégorie des paramédicaux a été associée aux professions de kinésithérapeutes, sages-femmes, opticiens, etc.

Enfin, la dernière catégorie, *infirmiers et assimilés*, comprend les métiers d'assistants médicaux. On y trouve les infirmiers eux-mêmes, les aides-soignants et les agents hospitaliers.

Le graphique 48 illustre la répartition de ces catégories de professions médicales ainsi que les différentes franchises proposées par la MACSF.

Sur la partie gauche du graphique, nous observons que la catégorie des médecins représente la plus grande part de l'exposition du portefeuille, suivie par les paramédicaux, puis par les infirmiers. En termes de tendance, nous constatons une croissance significative de l'exposition

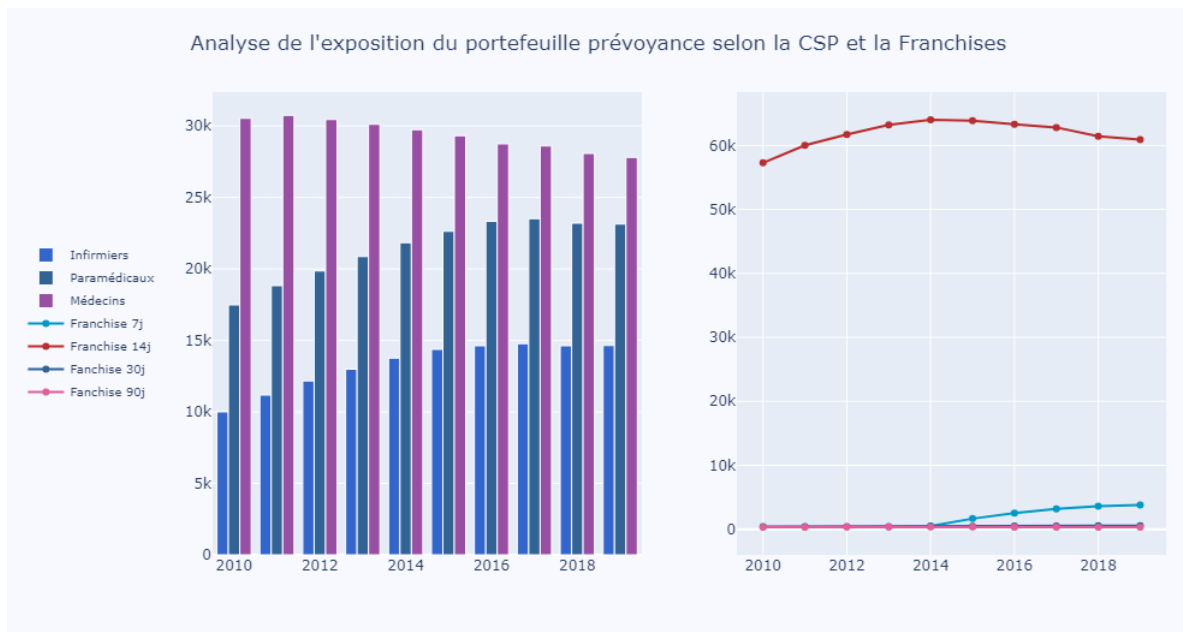


Figure 48: Exposition selon la CSP et la franchise

des infirmiers et paramédicaux de 2010 à 2016, suivie d'une légère baisse. Pour chacune des catégories, nous observons un nombre d'individus significatif, ce qui nous permettra d'obtenir des résultats convergents.

La partie droite présente l'exposition du portefeuille en fonction de la franchise des garanties. Les contrats d'incapacité proposent généralement quatre franchises différentes : 7, 14, 30 et 90 jours. La franchise représente le délai à partir duquel l'assuré peut commencer à percevoir une indemnisation. Nous observons que la majorité des contrats de prévoyance souscrits concernent une franchise de **14 jours**. Sur la base de cette constatation, nous avons décidé de ne conserver que les individus ayant une franchise de 14 jours pour la modélisation.

Le graphique suivant présente le nombre de sinistres. L'objectif de ce graphique est d'analyser la significativité du nombre de sinistres pour le calcul des taux d'incidence présentés dans la section suivante.

Dans la partie supérieure gauche du graphique, nous pouvons observer le nombre total de sinistres pour toutes les causes confondues. Nous remarquons un pic de sinistralité lié aux grossesses autour de l'âge de 30 ans. Passé 40 ans, le nombre de sinistres pour cause de grossesse tend vers 0. De plus, nous constatons que le nombre de sinistres dépassant le seuil de franchise (peut-être des maladies ou des accidents) est supérieur aux nombres de sinistres inférieurs à la franchise. Cet écart se creuse à partir de 55 ans. Ce premier graphique nous permet de conclure que le nombre de sinistres diffère en fonction de la cause du sinistre. Les taux d'incidence seront donc calculés en fonction de chaque cause de sinistre afin de prendre en compte l'hétérogénéité des risques.

Les autres graphiques dissocient les causes de sinistres en fonction du sexe. En ce qui concerne les sinistres au dessus du seuil de la franchise (partie inférieure gauche), nous constatons que les femmes ont un nombre total de sinistres d'environ 400 à partir de 35 ans, suivi d'une légère diminution jusqu'à l'âge de 58 ans. Ce nombre est deux fois supérieur à celui des hommes. En effet, pour ces derniers, le nombre de sinistres varie autour de 180-200 sinistres entre les âges de 30 et 50 ans.

Cependant, à partir de la cinquantaine, la sinistralité des hommes augmente fortement,

Nombre de sinistres selon la cause

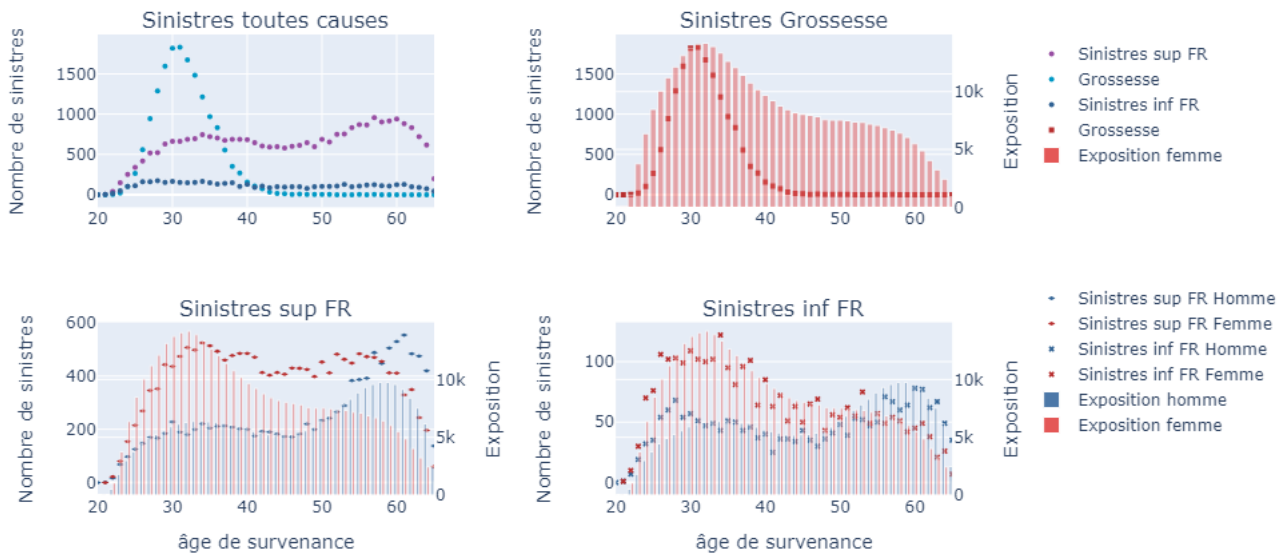


Figure 49: Nombre de sinistres de la BDD sinistre

dépassant les 400 sinistres au total. À l'inverse, le nombre de sinistres des femmes chute brutalement à partir de 58 ans. Il est important de noter que cette différence s'explique par le fait que l'exposition des femmes est nettement supérieure à celle des hommes. Lorsque nous calculerons les taux d'incidence bruts, la différence ne sera pas aussi marquée, car il s'agit d'un ratio entre le nombre de sinistres et l'exposition (dans le cas de l'estimateur d'Hoem).

La partie inférieure droite du graphique concerne le nombre de sinistres avec une durée d'incapacité inférieure au seuil de franchise. En moyenne, nous remarquons une sinistralité comprise entre 30 et 120 sinistres pour chaque sexe. Entre 30 et 40 ans, nous constatons une différence entre les deux sexes qui tend à diminuer avec l'âge. Dans cette tranche d'âge, la sinistralité des femmes se situe aux alentours de 80 sinistres, tandis que celle des hommes se situe autour de 40 sinistres. Nous constatons par ailleurs que cet écart tend à diminuer au fil des années jusqu'à s'équilibrer aux alentours de la cinquantaine. À partir de cet âge, la différence entre les deux sexes n'est plus significative.

Pour conclure, ces graphiques nous ont permis d'analyser la différence de sinistralité selon la cause du sinistre pour le passage en incapacité, mais aussi selon le sexe de l'individu. Afin d'obtenir des taux d'incidence significatifs et propres à chaque risque, il convient donc de segmenter notre base de données selon le sexe, mais aussi de calculer les taux d'incidence selon la cause du sinistre.

Le graphique suivant (50), représente le nombre de passages en incapacité causés par un sinistre d'une durée supérieure à la franchise en fonction de chaque catégorie professionnelle.

Ce graphique permet d'analyser si l'exposition et le nombre de sinistres sont suffisamment significatifs pour chaque âge. Pour les sinistres des *infirmiers et assimilés*, nous remarquons que l'exposition des hommes avant 30 ans est inférieure à 1000. De plus, le nombre de sinistres est assez faible. Nous constatons également une faible exposition des hommes après 60 ans.

Chez les *paramédicaux et assimilés*, nous constatons que l'exposition après 60 ans est seulement de 1000 années d'assurance et que le nombre de sinistres avoisine les 50. Le même constat

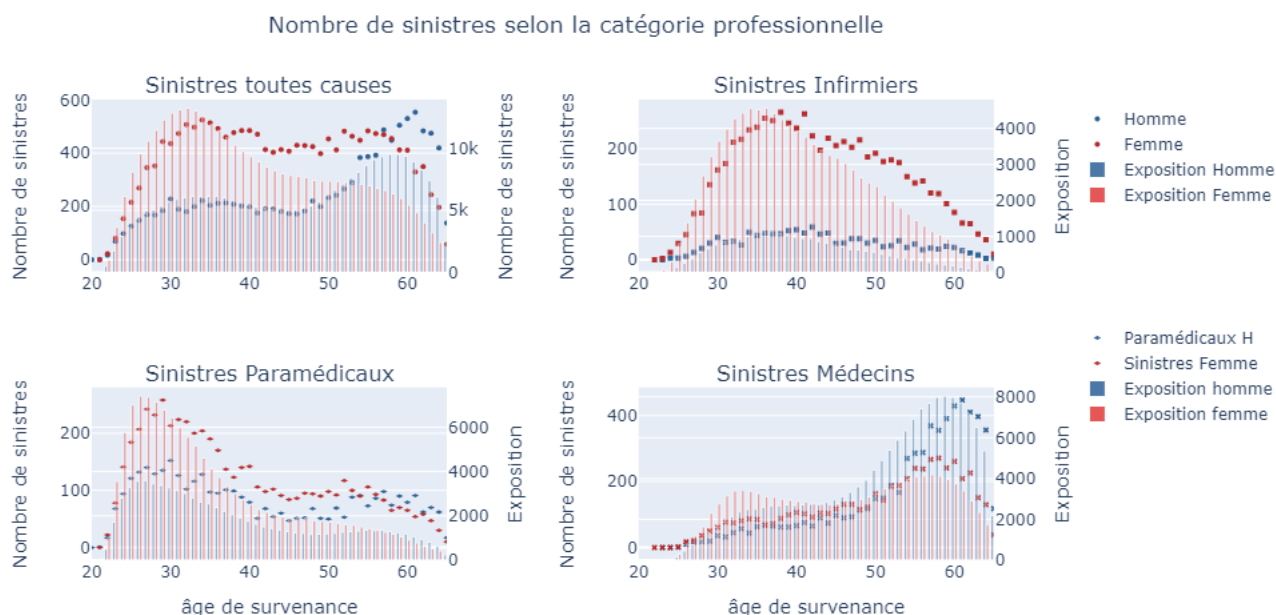


Figure 50: Nombre de sinistre de la BDD sinistre en fonction de la CSP

s'applique aux *médecins et assimilés* (figure en bas à droite), où l'exposition et le nombre de sinistres sont assez faibles avant 30 ans. Ces graphiques mettent en lumière les limites de nos bases de données en termes d'exposition et de sinistralité. Pour des raisons de significativité, nous avons donc décidé de ne travailler que sur la plage d'âge **30-60** ans.

Base de données du maintien en incapacité

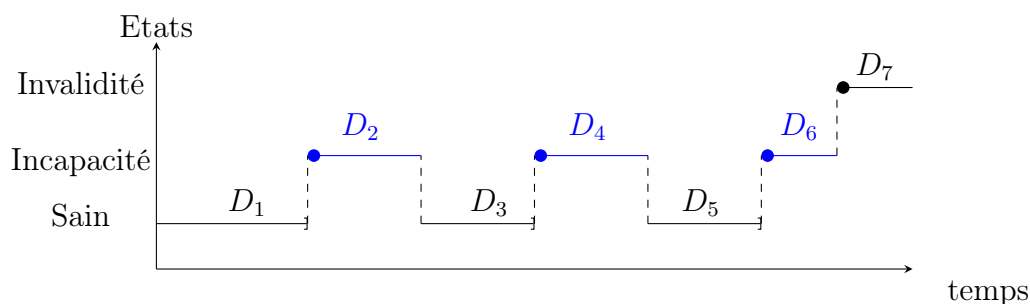


Figure 51: Durées analysées pour la BDD du maintien

Comme indiqué précédemment, une base de données spécifique a été construite pour la modélisation du maintien en incapacité. Cette base de données ne contient que **la somme des périodes d'incapacité** des assurés de la base de données des arrêts. Les durées analysées sont représentées en bleu sur la figure 51.

Étant donné que nous disposons de moins de données que dans la base d'incidence, il n'était pas possible de conserver la segmentation par sexe et par CSP. Nous avons donc effectué des statistiques descriptives pour décider quelle variable de segmentation conserver.

La partie supérieure gauche du graphique 52 présente les quantiles de la durée d'incapacité en fonction du sexe. Pour des raisons de clarté, nous avons choisi d'afficher uniquement la médiane, le premier quantile (Q1) et le troisième quantile (Q3).

En moyenne, nous constatons que la médiane est assez similaire entre les deux sexes jusqu'à l'âge de 55 ans. En effet, nous retrouvons une tendance à la hausse similaire, allant de 38 jours d'incapacité pour les individus de 25 ans à 55-58 jours pour les individus de 55 ans. À partir de cet âge, nous observons une légère augmentation de la durée pour les femmes jusqu'à 58 ans puis une baisse jusqu'à 62 ans. Pour les hommes, nous constatons une stagnation de la durée d'incapacité autour de 44 jours.

En ce qui concerne la durée du quantile 1, les valeurs sont similaires pour les deux sexes, et ce pour tous les âges. Pour le quantile d'ordre 3, nous remarquons qu'à partir de 50 ans, la différence entre les sexes s'accroît. Avant cet âge, la tendance haussière est la même pour les deux sexes.

Ce premier graphique nous permet de voir que la variable de segmentation du sexe n'est pas discriminante pour tous les âges. En effet, nous retrouvons des différences seulement pour les âges après 50 ans.

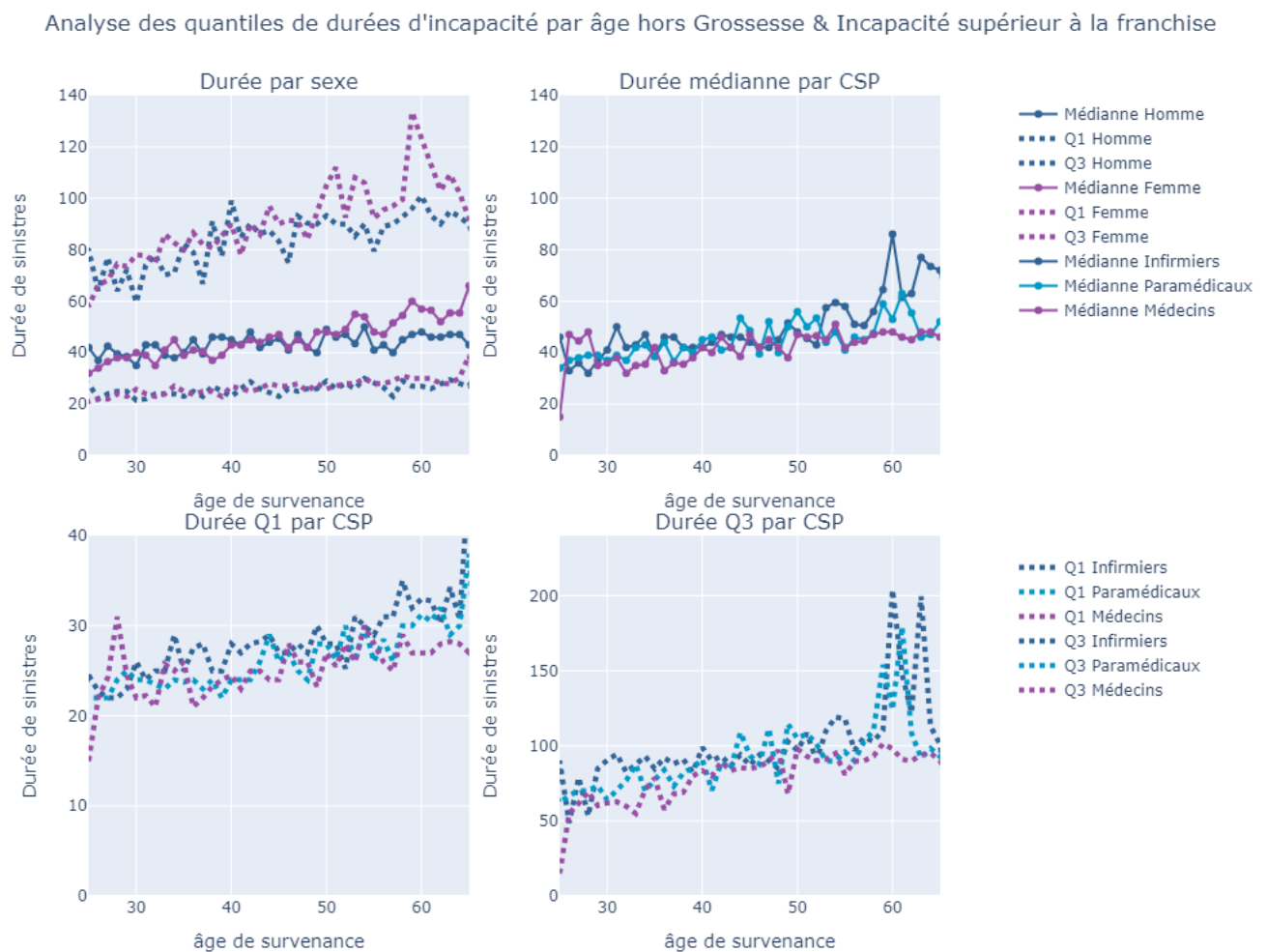


Figure 52: Nombre de sinistres par variable de segmentation

Pour la variable des professions, nous avons décidé de diviser le graphique en trois parties afin de rendre plus visible l'analyse. Un graphique par quantile a donc été réalisé.

Dans la partie supérieure droite, nous présentons les médianes pour les trois catégories de CSP. Dans un premier temps, nous constatons que la volatilité est plus élevée par rapport au

graphique précédent. Nous pouvons identifier deux clusters de volatilité distincts. Le premier groupe concerne les tranches d'âge de 25 à 40 ans, où la durée d'incapacité varie entre 30 et 50 jours. Le deuxième groupe de volatilité concerne les tranches d'âge de 40 à 65 ans, avec des durées oscillant entre 35 et 60 jours.

Nous constatons aussi qu'à partir de l'âge de 55 ans, l'hétérogénéité s'accroît entre les classes socio-professionnelles. En effet, nous observons un pic pour les infirmiers avec une valeur dépassant les 80 jours d'incapacité à l'âge de 60 ans. Après cet âge, la durée des CSP des médecins et des paramédicaux diminue, contrairement aux infirmiers. Cependant, il faut noter qu'après cet âge, l'exposition de notre portefeuille est beaucoup plus faible.

Les graphiques inférieurs présentent les quantiles associés aux durées pour chaque catégorie de CSP. Le graphique portant sur le quantile Q1 montre une tendance haussière similaire entre les catégories professionnelles. Les durées commencent aux alentours de 25 jours d'arrêt pour finir à 25-30 jours. Avec ce graphique, nous pouvons conclure que le quantile Q1 des durées passées dans l'état d'incapacité est plus volatile entre les CSP qu'entre les sexes.

Le dernier graphique (en bas à droite) porte sur le quantile Q3. De manière similaire au graphique précédent, nous constatons une même tendance entre les professions. Cependant, à partir de 58 ans, nous remarquons une forte hausse des durées d'incapacité pour les paramédicaux et les infirmiers.

Cette analyse nous conduit donc à ne garder que la variable de la **catégorie socioprofessionnelle** pour l'analyse du maintien en incapacité.

Le graphique (53) suivant représente le nombre de passages en invalidité sous forme d'un histogramme en deux dimensions. En abscisse, nous avons le nombre de jours passés en incapacité, et en ordonnée, nous avons l'âge de survenance du sinistre.

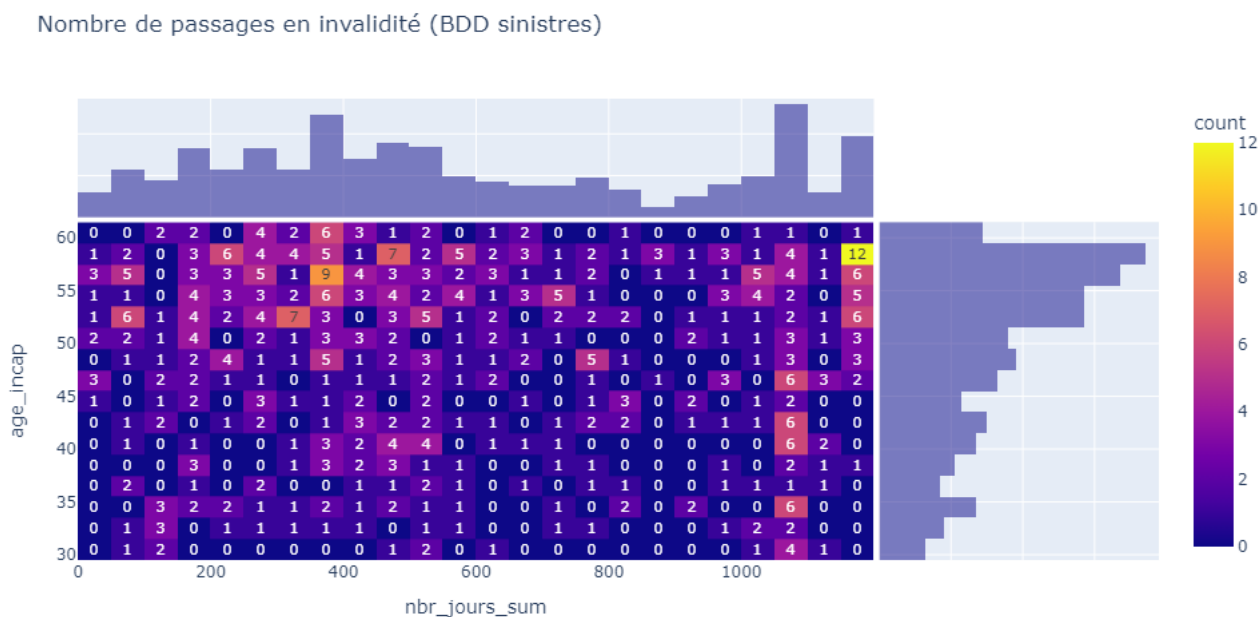


Figure 53: Nombre de passage en invalidité, BDD maintien

Nous constatons que le nombre de transitions vers l'état d'invalidité est très faible voire nul pour certains intervalles d'âges et de durées. En effet, lorsque nous sélectionnons les sinistres

supérieurs à la franchise, nous obtenons une base de données de 30 000 lignes pour **718 passages en invalidité**. Cela nous conduit donc à observer des zones du graphique sans aucun passage en invalidité. Néanmoins, nous remarquons tout de même des zones de passages en invalidité condensées. Après 1100 jours de passage en incapacité, il y a un léger cluster de passage en invalidité pour les âges entre 50 et 55 ans. Cela s'explique par le fait qu'au bout de 1095 jours, les individus toujours en incapacité ont une forte probabilité de passer en invalidité. Le fait de disposer d'un faible nombre de passage en invalidité, nous oblige donc à réaliser des **classes d'âges** et des **classes de durées**. Dans la partie modélisation consacré à l'invalidité, nous détaillerons une méthode statistiques permettant de créer des classes.

Synthèse des statistiques descriptives

L'analyse descriptive nous a permis d'évaluer l'importance des variables de segmentation. Pour l'incidence, nous avons conclu que le sexe et la catégorie des CSP étaient des variables déterminantes pour la segmentation. L'âge est également considéré comme variable discriminante, mais nous l'utilisons comme variable de discrétisation du temps. De plus, l'analyse de l'exposition de ces bases de données nous a permis de conclure que nous devons sélectionner un certain intervalle d'âges. De ce fait, nous avons décidé de ne garder que les âges entre **30 et 60 ans**. En-dehors de cet intervalle, le nombre de sinistralités et d'expositions n'était pas suffisamment élevé.

Concernant la base de données du maintien, nous avons conclu que la variable du sexe n'était pas aussi déterminante comparée à la variable des catégories professionnelles. Étant donné la diminution du nombre d'individus dans cette base de données, nous avons décidé de supprimer cette variable de segmentation et de ne conserver que la segmentation portant sur la CSP (catégorie socio-professionnelle).

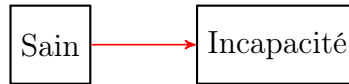
Enfin, la statistique descriptive portant sur le nombre de passages en invalidité a montré qu'il n'était pas possible de conserver la segmentation par âge et par durée d'incapacité valeur par valeur. Cela nous a donc conduit à la conclusion qu'il est nécessaire de créer des classes pour ces deux variables de segmentation.

La table suivante récapitule les différentes caractéristiques qui seront réalisées pour chacune des lois de transitions et sur la loi de durée du maintien en incapacité.

	Temps discrétisé par	q_x calculé par	Segmentation
Passage en incapacité	âge	âge	Cause, Sexe, Professions
Maintien en incapacité	Durée		Cause, Professions
Passage en invalidité	Durée	Classe de durées	Cause, Classe d'âges

Table 11: Récapitulatif des segmentations des différentes lois

3.1.2 Passage en incapacité



Dans cette section, nous allons utiliser les trois estimateurs, à savoir Hoem, Kaplan-Meier et Nelson-Aalen, pour estimer les taux d'incidence bruts pour le passage en incapacité. Comme introduit précédemment, nous allons présenter les résultats selon le sexe et la catégorie socio-professionnelle (CSP). Concernant la cause de sortie, nous avons décidé de ne présenter que le passage en incapacité pour les sinistres supérieurs à la franchise (Sinistres sup), les autres résultats étant disponibles en annexe. Ce choix est lié à notre objectif final, qui est d'analyser les sinistres pouvant potentiellement passer en invalidité et donc atteindre la priorité de la réassurance. Les sinistres inférieurs à la franchise et ceux causés par des grossesses ne font généralement pas l'objet d'un passage en invalidité et n'arrivent donc pas à déclencher la réassurance¹⁸.

Concernant les estimateurs de Kaplan-Meier et de Nelson-Aalen, il est nécessaire d'effectuer une discrétisation du temps. Pour ce faire, nous discrétisons le temps en fonction des âges auxquels un événement se réalise. Ainsi, les pas de temps $t_1 < t_2 < \dots < t_n$ correspondent aux âges où un individu peut :

- Passer en incapacité (événement étudié)
- Entrer dans l'étude
- Etre censuré (sortie de l'étude, changement administratif, passage en incapacité pour cause de grossesse, passage en incapacité pour cause de sinistre inf, décès, etc.)

La figure 54 représente l'estimation des taux d'incidence bruts pour le passage en incapacité avec un sinistre d'une durée supérieure à la franchise (sinistres sup). En arrière-plan de chaque graphique, se trouve l'exposition associée. Nous analysons donc la durée T_{sain} et l'évènement $Incap_{sinistre_{sup}}$ égal à 1 si le passage dû à un sinistre d'une durée supérieur à la franchise et 0 si le passage est lié à une autre cause ou à une censure.

Nous pouvons observer que les estimateurs de Nelson-Aalen et de Kaplan-Meier sont quasiment identiques pour chaque segmentation. En effet, pour chaque segmentation, les estimations des taux bruts se superposent.

De plus, nous remarquons que l'estimateur d'Hoem converge vers les estimateurs de Kaplan-Meier et de Nelson-Aalen à mesure que l'exposition augmente. En effet, en prenant l'exemple de la catégorie des *médecins et assimilés* pour les sexes hommes et femmes, les estimateurs sont presque équivalents. En revanche, dès lors que l'exposition diminue, les écarts entre l'estimateur paramétrique et ceux non-paramétriques se creusent. Nous pouvons constater ce phénomène pour les infirmiers. Les taux bruts entre les âges 30 à 50 ans reste assez similaire puis passé cet âge, la volatilité de l'estimateur d'Hoem augmente fortement.

Chez les médecins, les estimations des taux d'incidence bruts varient de 1% à 5%. En général, les taux sont légèrement plus élevés (d'environ 1% en niveau) chez les médecins femmes

¹⁸Réassurance en XS

Taux d'incidence bruts pour les incapacités supérieures à la franchise

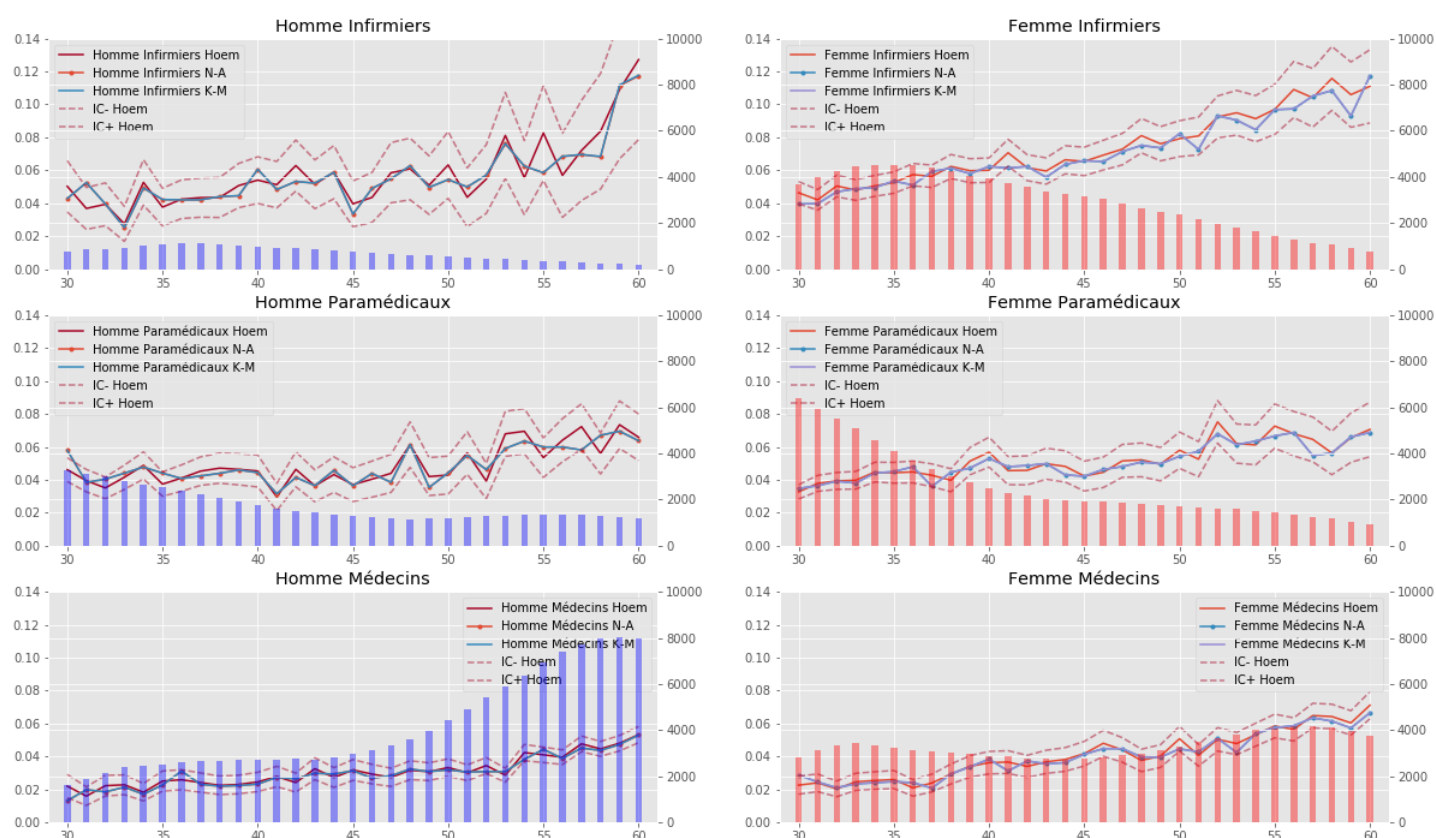


Figure 54: Taux d'incidence bruts pour les sinistres supérieurs à la franchise

par rapport aux médecins hommes. Nous observons une légère baisse du passage en incapacité chez les femmes vers l'âge de 37 ans. Cela peut être dû au fait qu'à cet âge-là, la majorité des incapacités des femmes est liée à une grossesse. De manière générale, pour les deux sexes, nous observons une tendance à la hausse. Nous remarquons aussi que l'exposition la plus élevée est celle des hommes médecins âgés de plus de 50 ans. En effet, l'exposition atteint la valeur de 8000 années d'assurance alors qu'à l'âge de 30 ans, la valeur est de 2000 années d'assurance. Chez les femmes, l'exposition reste constante aux alentours de 4000 années d'assurances.

Chez les paramédicaux hommes, les taux d'incidence bruts se situent autour de 2% entre 30 et 50 ans. À partir de cet âge, les taux augmentent pour atteindre environ 6%. Nous constatons que les taux des hommes sont beaucoup plus volatils que ceux des femmes. En effet, pour certains âges comme 49 ans par exemple, la valeur du taux oscille de 1% à 2%. Chez les paramédicaux femmes, nous retrouvons des oscillations beaucoup moins marquées. Après 50 ans, les taux varient seulement de 1% en moyenne. De manière générale, chez les hommes, nous constatons une légère tendance haussière jusqu'à 50 ans. À partir de cet âge, nous observons une hausse plus élevée, passant de 3% à 5%. Même constat pour les femmes concernant la tendance jusqu'à 50 ans. En revanche, après cet âge-là, la hausse est moins prononcée que celle des hommes.

Enfin, les estimations des taux d'incidence pour les infirmiers sont généralement plus élevées

et plus volatiles que ceux des autres catégories de professions médicales. Chez les hommes, nous observons un creux de 2% du taux d'incidence à l'âge de 45 ans. Nous constatons aussi des pics vers l'âge de 54 ans avec des taux d'incidence d'environ 8%. Au vue de l'exposition affichée en arrière plan, nous pouvons conclure que la forte hétérogénéité des taux est due à la faible exposition associée. Afin de rendre ces taux homogènes, il sera nécessaire d'appliquer un lissage.

Chez les femmes, les taux d'incidence bruts sont croissants. Ils passent de 4% à 9% en 30 ans. Même constat que pour les hommes, vers les âges 55-60 ans, une volatilité apparaît. Cela est dû une à la décroissance de l'exposition au sein de notre portefeuille.

Ces graphiques montrent que l'estimateur d'Hoem semble assez volatile dès lors que l'exposition s'affaiblit. Pour des raisons de robustesse, nous faisons le choix de continuer avec l'estimateur de **Kaplan-Meier** qui nous paraît plus stable. Néanmoins, afin d'avoir des résultats homogènes, il est nécessaire d'appliquer un lissage sur les estimations obtenues. La section suivante illustre l'application des méthodes de lissage évoquées dans la partie théorique.

Application du lissage pour le passage en incapacité

Dans cette partie, nous allons appliquer les méthodes de lissage de **Whittaker-Henderson** et de la **régression à noyau**. Comme expliqué dans la partie théorique sur les lissages, deux méthodes sont utilisées pour déterminer le paramètre optimal de lissage. On peut soit utiliser une méthode statistiques, soit trouver le paramètre optimal graphiquement.

Pour le lissage avec la régression à noyau, nous avons décidé d'utiliser la méthode de cross-validation, tandis que pour la méthode de Whittaker-Henderson, une sélection graphique du paramètre optimal a été utilisée. Pour chaque valeur du paramètre z , nous avons testé différentes valeurs du paramètre h jusqu'à obtenir une courbe de lissage satisfaisante.

Les résultats des paramètres sélectionnés sont affichés dans le tableau 12.

Sexe	Catégories	$h_{Gaussien}^*$	$h_{Epanech}^*$	$h_{Uniforme}^*$	z^*	h^*
Homme	Infirmiers	1.1	1.5	1.5	4	15
	Paramédicaux	3.5	7	5.4	4	15
	Médecins	1.2	2.3	1.5	4	20
Femme	Infirmiers	2.4	5.3	3.5	4	10
	Paramédicaux	1.1	1.8	1.5	4	15
	Médecins	1.8	3.6	2.5	4	10

Table 12: Résultats de $h_{optimal}$

Comme pour la présentation des taux bruts d'incidence, nous présentons les résultats seulement pour les incapacités supérieurs à la franchise. Les lissages appliqués pour les autres causes de sortie sont en annexe. De plus, afin de ne pas alourdir la présentation et l'analyse des résultats, nous avons décidé de ne pas afficher le lissage avec un kernel uniforme et le lissage Lowess.

Interprétation

Nous remarquons que globalement, la différence entre la méthode de Whittaker-Henderson et la régression kernel est assez minime. De plus, comme expliqué dans la partie consacrée à la théorie des lissages, le choix du kernel a peu d'impact sur le résultat. En effet, pour chacun des résultats, le lissage avec un kernel Gaussien et un kernel Epanechkinov se superposent.

Taux d'incidence lissés pour les incapacités supérieures à la franchise

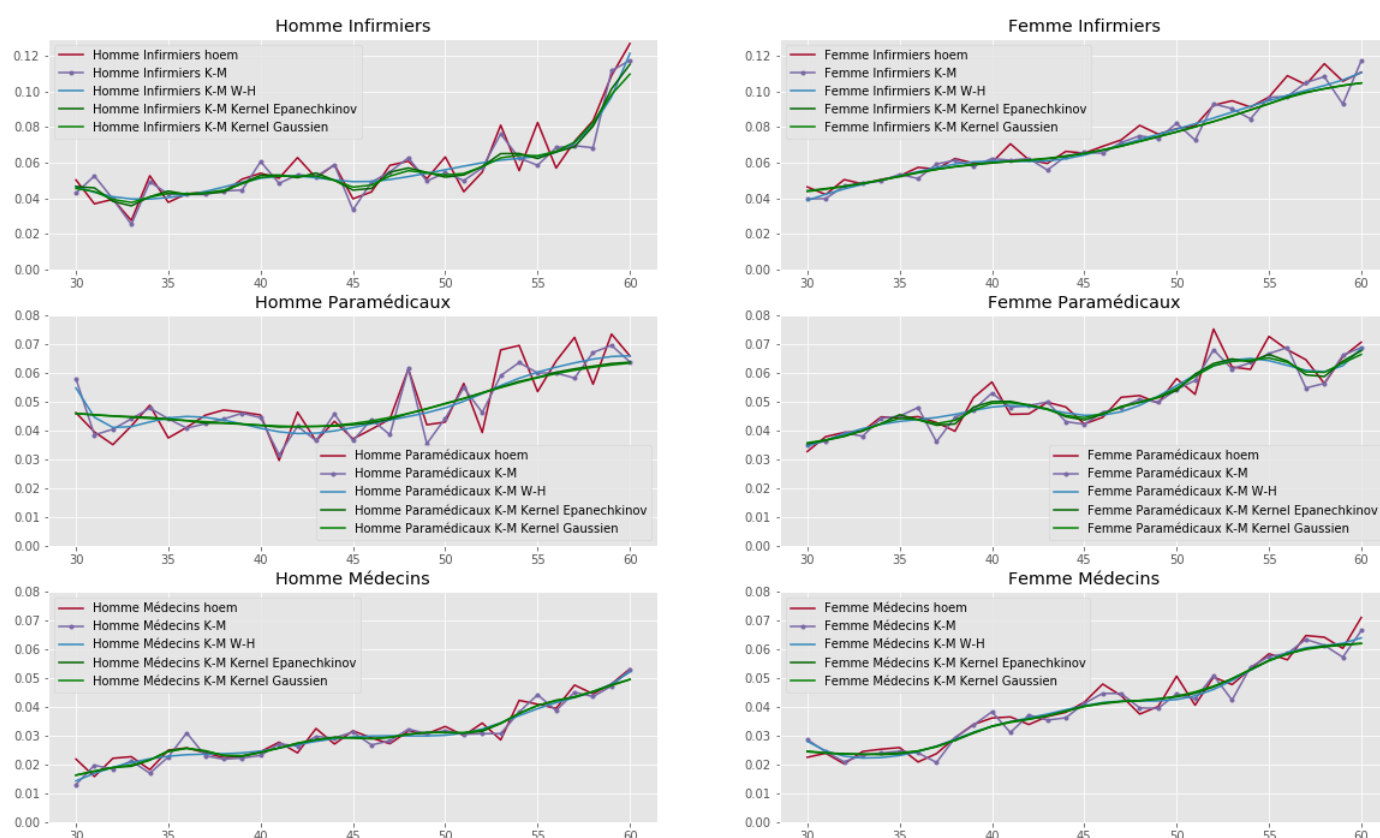


Figure 55: Lissage des taux d'incidence bruts des sinistres supérieurs à la franchise

Dans le cas de la catégorie des *infirmiers et assimilés*, nous observons certains écarts dès lors que les taux bruts varient brusquement. Nous faisons le même constat pour les paramédicaux. Concernant les médecins, les taux bruts étaient à l'origine assez lisses. Le fait d'appliquer un lissage ne change pas grand-chose mais permet de corriger les quelques irrégularités.

La table 13 représente les résultats des deux tests statistiques permettant de vérifier si les nouvelles valeurs lissées ne sont pas trop éloignées des taux bruts initiaux.

Lissage/Test	Test du χ^2			Test signes		
	T	C	Décision	T	C	Décision
W-H	H : [24.4 ; 25.9 ; 23.6] F : [15 ; 16 ; 22]	42.55	NR	H : [0.35 ; 0.54 ; 0.35] F : [1.4 ; 0.35 ; 0.7]	1.96	NR
Noyau Gaus	H : [18.2 ; 32 ; 19.9] F : [19.4 ; 10.8 ; 24.4]			H : [0.35 ; 0.71 ; 1.07] F : [1.43 ; 0.18 ; 1.07]		
Noyau Epa	H : [14.4 ; 31.2 ; 20.4] F : [20.2 ; 9.4 ; 25.4]			H : [0.18 ; 0.35 ; 1.07] F : [1.07 ; 0.35 ; 1.07]		

Table 13: Résultats des tests de lissage

Au vu des résultats des deux tests, nous en concluons que la méthode de lissage semble être adaptée et représentative des taux.

Au final, nous décidons de garder la méthode de Whittaker-Henderson, car elle permet de mieux prendre en compte l'exposition des données.

Modèle de Cox appliqué au passage en incapacité

Dans cette partie, nous présentons le modèle de Cox appliqué au passage en incapacité. La table 14 présente les résultats de la régression. Le risque de référence est celui des hommes médecins, car le taux brut est le moins volatile selon l'estimateur de Kaplan-Meier. Nous modélisons donc l'équation suivante :

$$q_x = (1 - q_{x,ref}) \exp(\beta_1 * CSP_{infirmiers} + \beta_2 * CSP_{para} + \beta_3 * Sexe)$$

Modèle de Cox vs K-M pour les sinistres > franchise

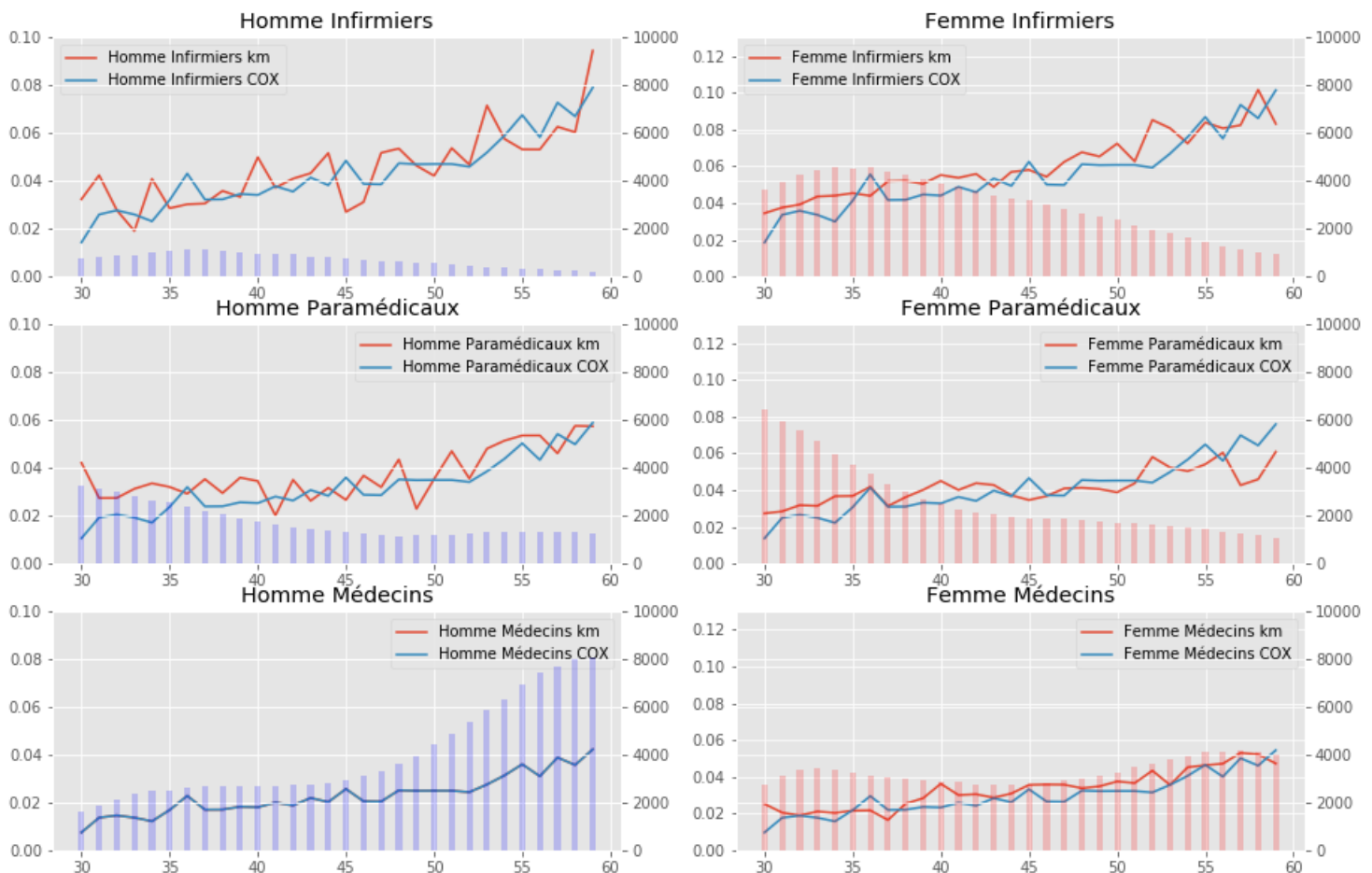


Figure 56: Comparaison entre le modèle Cox et l'estimateur de Kaplan-Meier

Comme illustré dans le tableau 14, nous pouvons remarquer que tous les coefficients sont significatifs au seuil de 5%. De plus, chacun des coefficients est positif, ce qui entraîne une

Variabes	Coef	e^{coef}	IC_-	IC_+	z	pvalue
$CSP_{infirmiers}$	0.61	1.84	1.77	1.90	35.11	<0.005
CSP_{para}	0.36	1.43	1.38	1.48	21.16	<0.005
Sexe	0.20	1.22	1.18	1.25	13.76	<0.005

Table 14: Résultats régression Cox

augmentation du risque instantané par rapport à la variable de référence. Par exemple, le fait d'être une femme entraîne légèrement une augmentation du risque de tomber en incapacité par rapport à être un homme. De même, pour les autres variables, le fait d'être un professionnel de santé associé à la classe des paramédicaux entraîne une augmentation du risque instantané de 43%.

La table suivante représente le test de proportionnalité des résidus de Schoenfeld. Nous constatons que pour chacune des fonctions utilisées, le test rejette, au seuil de 5%, l'hypothèse de proportionnalité des coefficients pour au moins une variable explicative. Ce test nous conduit donc à la conclusion de ne pas continuer avec le modèle de Cox.

Variables	Identité		Log		Rank	
	T	p	T	p	T	p
$CSP_{infirmiers}$	2.16	0.14	2.88	0.08	3.04	0.08
CSP_{para}	35.64	<0.05	38.34	<0.05	38.63	<0.05
Sexe	1.91	0.17	0.84	0.24	0.7	0.41

Table 15: Test de proportionnalité (Résidus de Schoenfeld)

La section suivante porte sur la modélisation du maintien en incapacité.

3.1.3 Maintien en incapacité

Dans cette section, nous présentons les lois paramétriques ainsi que la loi non-paramétrique obtenue à l'aide de l'estimateur de Kaplan-Meier pour estimer le maintien en incapacité.

Estimateur non paramétrique pour le maintien en incapacité

Dans le cas du maintien en incapacité, la discrétisation ne se fait plus en fonction de l'âge, comme c'était le cas pour l'incidence, mais plutôt en fonction du temps passé en incapacité. Ainsi, les instants t_1, t_2, \dots, t_n représentent chaque jour de maintien où un événement, tel qu'une sortie ou une censure, peut potentiellement se produire. De plus, en raison de la franchise à 14 jours, la fonction de survie ne commencera qu'à partir de cette durée.

Les graphiques 60, 61 et 62 représentent les estimateurs de Kaplan-Meier pour chacune des catégories de CSP.

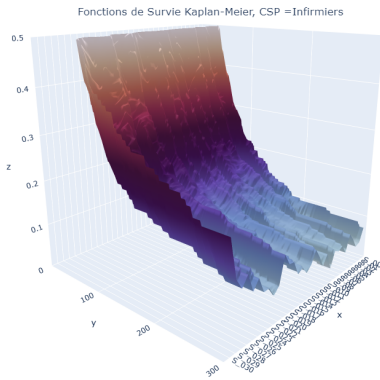


Figure 57: K-M Infirmiers

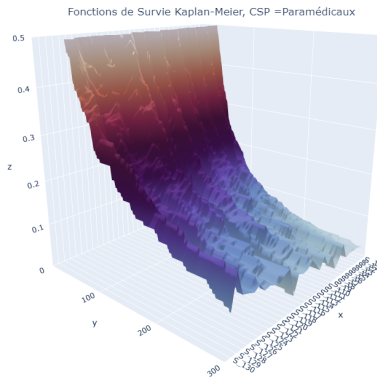


Figure 58: K-M Paramédicaux

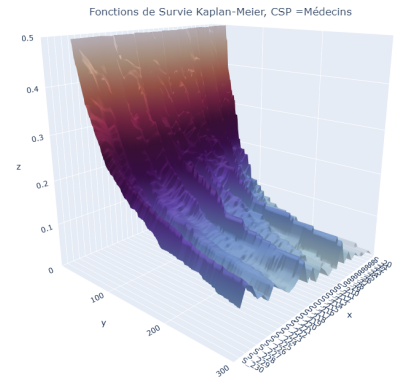


Figure 59: K-M Médecins

Interprétation

Nous observons à travers les figures 60, 61 et 62 le caractère non-paramétrique de l'estimateur de Kaplan-Meier. Cette approche permet de refléter au mieux les données en calculant une nouvelle valeur de la fonction de survie à chaque pas de temps où un événement se réalise. La section suivante présente le lissage des résultats obtenus

Application du lissage pour les lois non-paramétriques du maintien en incapacité

Pour lisser les différentes fonctions de survie non-paramétriques, nous avons décidé d'appliquer un lissage en deux dimensions avec la régression kernel. Les figures suivantes présentent les différents lissages.

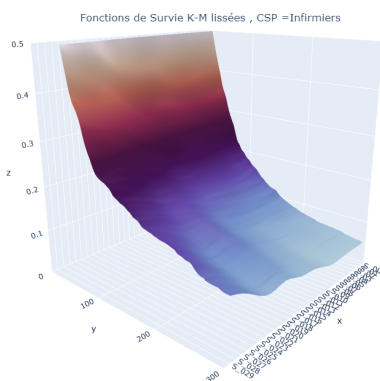


Figure 60: K-M Infirmiers

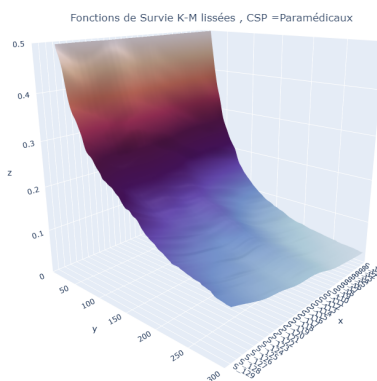


Figure 61: K-M Paramédicaux

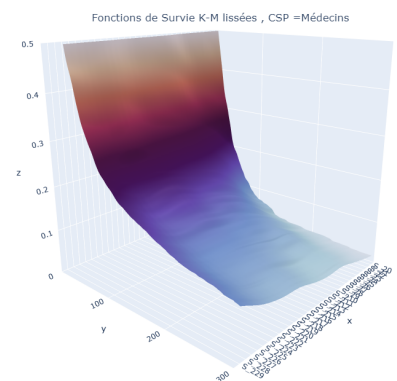


Figure 62: K-M Médecins

Les résultats montrent que le lissage a permis d'homogénéiser les fonctions de survie empiriques. L'avantage du modèle de Kaplan-Meier est qu'il permet d'obtenir des résultats plus précis qu'une fonction de survie théorique. Cependant, l'inconvénient de cette approche réside dans la phase de simulation. Pour simuler notre portefeuille, l'utilisation de lois non-paramétriques nécessite un temps de calcul bien plus élevé que celui d'une simulation basée sur

une loi paramétrique. En effet, il est nécessaire d'effectuer une simulation à chaque intervalle de temps ou d'utiliser la méthode de l'inverse.

De plus, à travers les résultats obtenus, nous pouvons également remarquer que les fonctions de survie empirique sont plutôt stables et se rapprochent d'une fonction de survie théorique. L'avantage d'utiliser une fonction de survie théorique est d'obtenir une meilleure estimation dans les queues de distribution, là où l'estimateur de Kaplan-Meier risque de donner des résultats instables. Pour toutes ces raisons, nous décidons de ne pas utiliser l'estimateur de Kaplan-Meier.

Estimation paramétrique pour le maintien en incapacité

Pour rappel, les distributions présentées dans cette section sont calibrées en tenant compte de la troncature à gauche des données. Cela signifie qu'avant 14 jours, la distribution n'est pas aussi fiable qu'après cette valeur.

Les figures ci-dessous représentent les fonctions de survie des lois de **Weibull**, de **Log-logistique** et de **Gamma généralisée** pour chaque catégorie socio-professionnelle. Afin d'observer clairement les différences, nous avons volontairement tronqué l'axe des ordonnées à 0,5.

Concernant les abscisses, le côté droit indique l'âge associé à la fonction de survie, qui commence à 30 ans (au fond du graphique) et se termine à 60 ans (au premier plan). L'axe des abscisses à gauche indique la durée passée dans l'état d'incapacité. Tout comme l'axe des ordonnées, nous avons volontairement tronqué cet axe à la valeur 300 pour rendre les différences entre les lois et les CSP plus visibles.

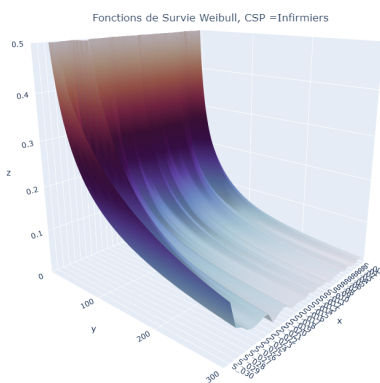


Figure 63: Weibull Infirmiers

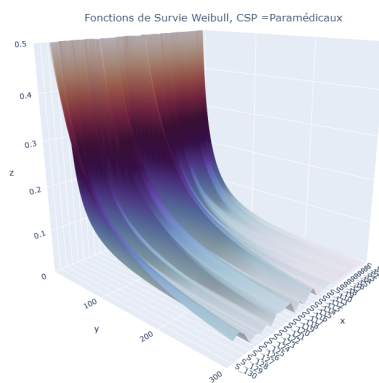


Figure 64: Weibull Paramédicaux

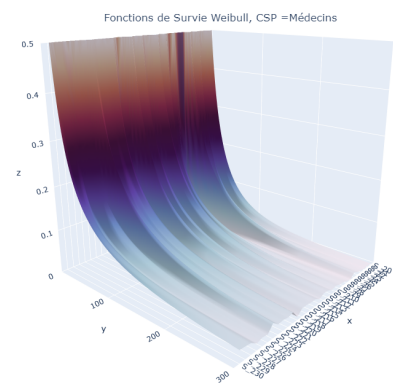


Figure 65: Weibull Médecins

En analysant les fonctions de survie de la loi de Weibull (partie supérieure), nous remarquons des variations entre les différentes catégories de CSP. Les infirmiers ont une probabilité de survie plus élevée pour des durées courtes par rapport aux autres catégories. Les fonctions de survie des infirmiers présentent une forme plus convexe, ce qui indique une probabilité de survie plus faible de rester en incapacité pour de longues périodes. Les paramédicaux, quant à eux, ont des fonctions de survie avec des formes plus creuses, traduisant une probabilité de rester en incapacité plus élevée pour de longues périodes. Les fonctions de survie des médecins se situent entre celles des infirmiers et des paramédicaux.

Les fonctions de survie de la loi Log-logistique (graphiques suivants) présentent des valeurs plus élevées que la loi de Weibull pour des longues durées.

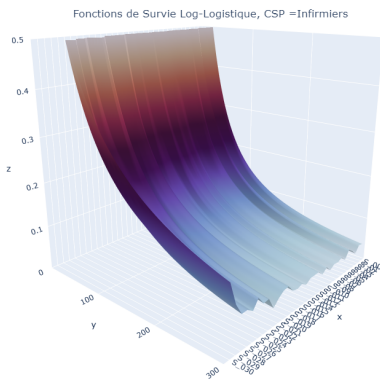


Figure 66: Log logistique Infirmiers

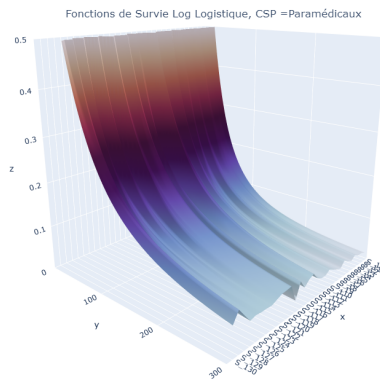


Figure 67: Log logistique Paramédicaux

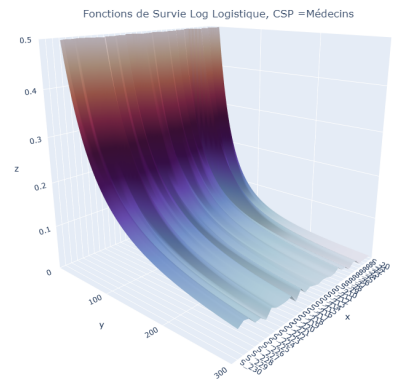


Figure 68: Log logistique Médecins

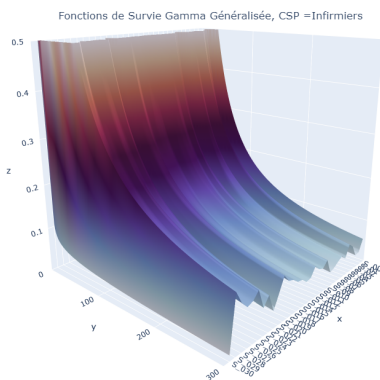


Figure 69: Gamma G Infirmiers

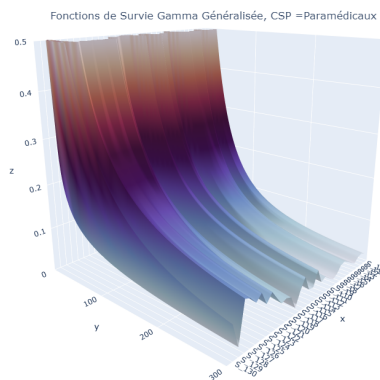


Figure 70: Gamma G Paramédicaux

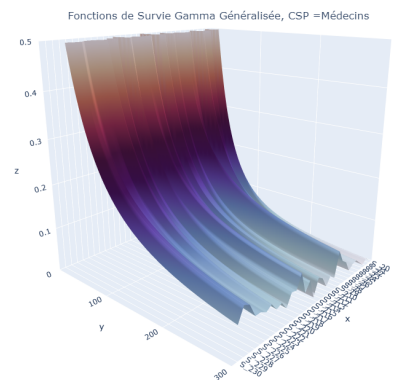


Figure 71: Gamma G Médecins

En ce qui concerne la loi de Gamma Généralisée (figure 69,70,71), nous observons une forte hétérogénéité entre les différentes fonctions de survie. Cela est dû à la nature de la loi de Gamma Généralisée, qui englobe plusieurs lois. Pour rappel, la loi de Gamma Généralisée comprend les lois suivantes :

- Exponentielle
- Weibull
- Gamma
- Log Normale
- Weibull inversée
- Gamme Inversée

Afin de déterminer laquelle des lois est la plus optimale pour la modélisation de la durée en incapacité, nous avons calculé les critères d'information AIC et BIC.

La figure 72 présente les critères d'information pour les infirmiers. Une observation importante est qu'il n'existe pas une loi de probabilité unique qui minimise tous les critères d'information. Cependant, nous pouvons remarquer que la loi Log-logistique est celle qui minimise le plus grand nombre de critères du BIC selon les différentes tranches d'âge.



Figure 72: AIC et BIC pour les infirmiers

La figure 73 met en évidence les critères d'informations pour les paramédicaux. Nous remarquons que la loi de Weibull minimise le plus grand nombre de critères d'information, que ce soit le critère de l'AIC ou celui du BIC.

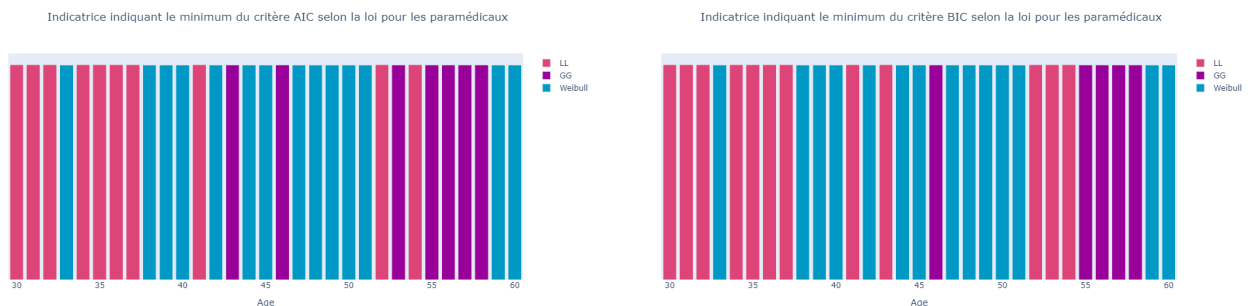


Figure 73: AIC et BIC pour les paramédicaux

Enfin, en ce qui concerne les médecins (figure 74), nous observons que la Loi log-logistique minimise le plus le critère BIC selon les différentes tranches d'âge.

Il est important de noter que l'analyse des critères d'AIC et de BIC est une approche intéressante qui permet de déterminer la loi la plus adéquate pour chacune des CSP. Comme nous l'avons constaté pour chaque catégorie socio-professionnelle et chaque tranche d'âge, il n'existe pas de loi qui minimise de manière absolue tous les critères d'information. Néanmoins, nous pouvons éliminer dès à présent le choix de la loi Gamma Généralisée car elle ne minimise que très peu de critères d'information. Nous pouvons donc établir à présent une loi a priori optimale pour chaque catégorie de CSP, à savoir :

- Infirmiers : Log-Logistique
- Paramédicaux : Weibull
- Médecins : Log-Logistique

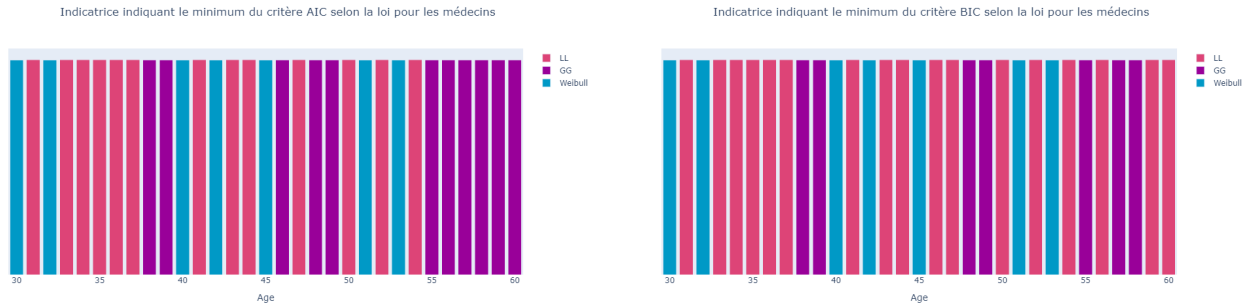


Figure 74: AIC et BIC pour les médecins

La section suivante s'intéresse au lissage des deux lois paramétriques sélectionnées.

Application du lissage pour les lois paramétriques du maintien en incapacité

Cette partie illustre la méthode de lissage des moments de la de **Weibull** et la loi **Log-logistique**.

La figure 75 illustre l'application des lissages de Whittaker-Henderson, de la régression kernel avec un noyau gaussien ainsi que le lissage Loess sur les moments 1 et 2 de la loi de **Weibull**. La troisième colonne représente l'exposition associée à chaque âge et à chaque catégorie de CSP.

Concernant le moment d'ordre 1 des infirmiers, nous pouvons constater qu'il croît jusqu'à l'âge de 40 ans, puis augmente jusqu'à dépasser la valeur 120. Nous pouvons donc dire que la moyenne théorique (l'espérance) de la loi de Weibull augmente avec l'âge. Autrement dit, plus un individu est âgé, plus il a de chances de rester longtemps en état d'incapacité (en moyenne). Nous faisons la même constatation pour le moment d'ordre 2.

Chez les paramédicaux, nous remarquons une très forte hétérogénéité du moment d'ordre 1, notamment à partir de 45 ans. Le lissage permet d'obtenir une courbe lisse croissante. Nous constatons après lissage, que les individus âgés jusqu'à 50 ans ont tendance à rester plus longtemps en incapacité en moyenne. Après cette valeur, nous remarquons une volatilité qui est due à la faible exposition. Il faut donc rester prudent après cet âge.

Pour les médecins, nous faisons le même constat que pour les paramédicaux, c'est-à-dire qu'il y a une tendance à la hausse pour le moment d'ordre 1.

Les figures suivantes présentent les fonctions de survie de la loi de Weibull avant et après le lissage des moments. Nous observons que le lissage a permis d'homogénéiser les résultats en fonction de l'âge tout en préservant la forme originelle des fonctions de survie. De plus, cette méthode de lissage a permis de préserver la dépendance entre les deux paramètres de la loi de Weibull.

Lissage des lois de Weibull

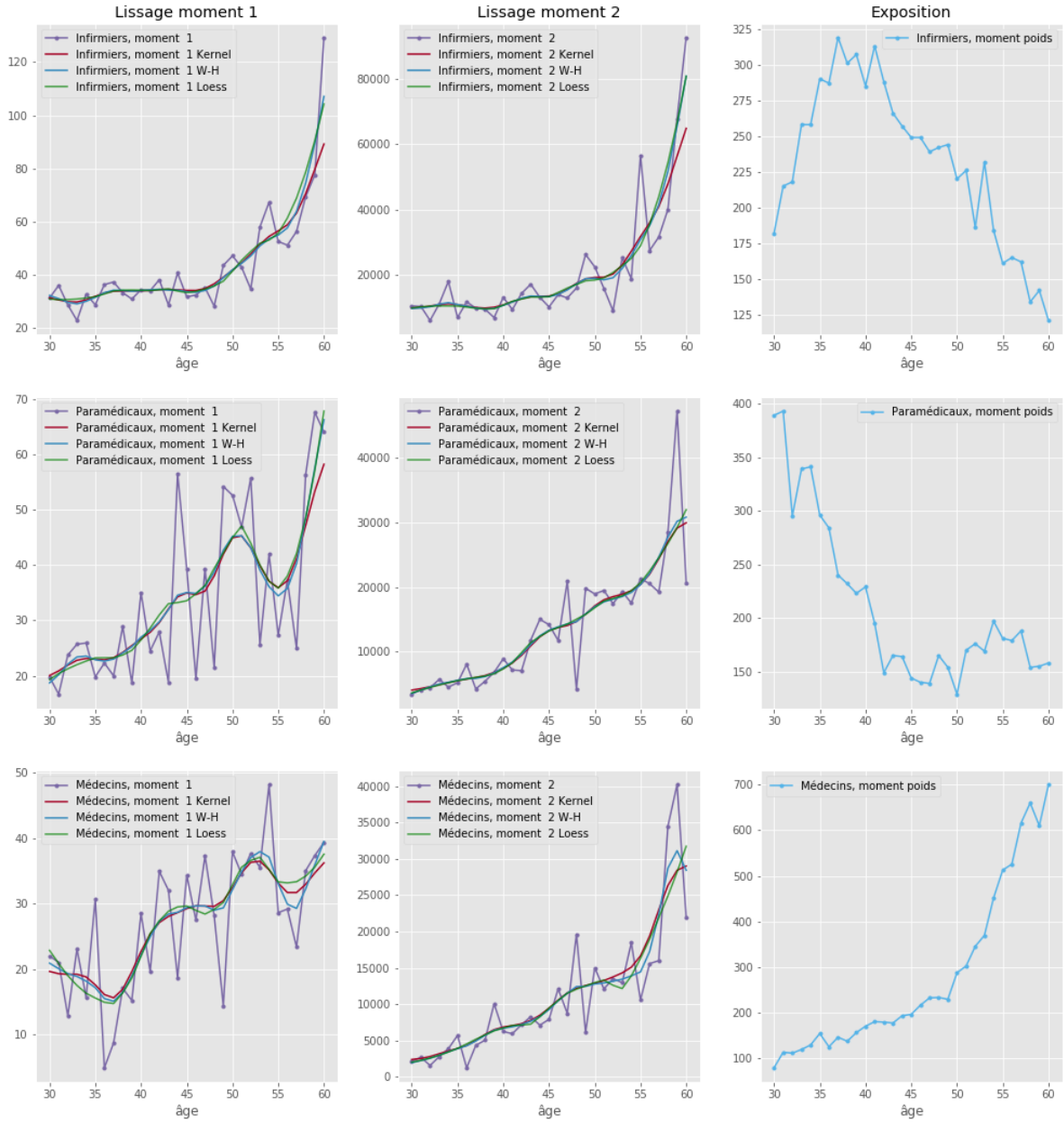


Figure 75: Lissage des moments de la loi de Weibull

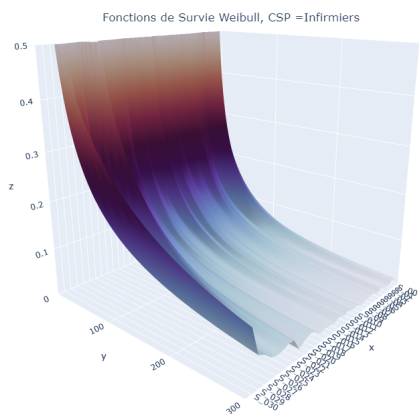


Figure 76: Weibull Infirmiers

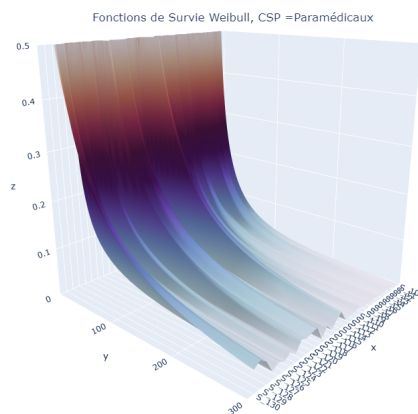


Figure 78: Weibull Paramédicaux

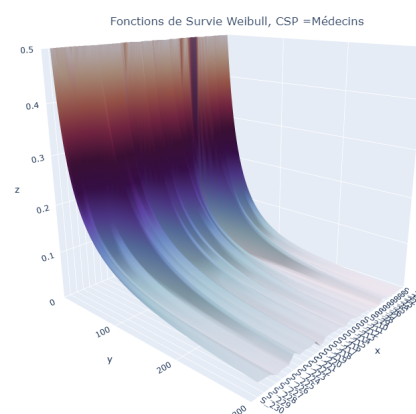


Figure 80: Weibull Médecins

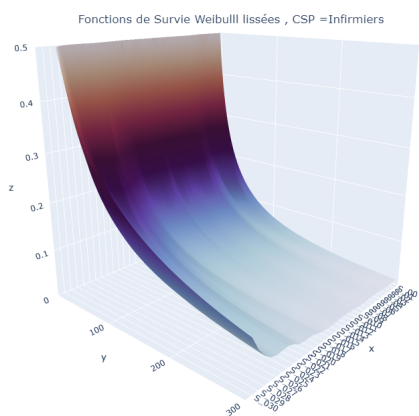


Figure 77: Weibull lissées Infirmiers

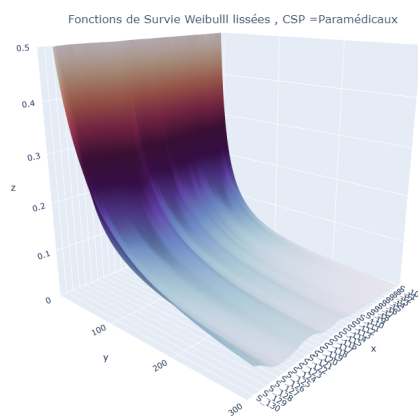


Figure 79: Weibull lissées Paramédicaux

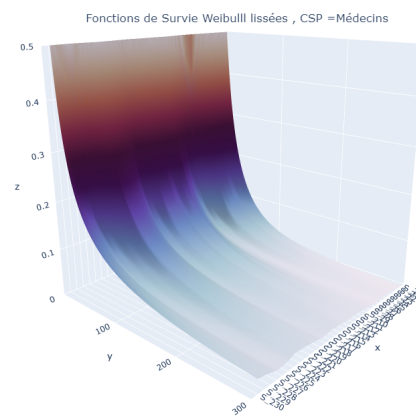


Figure 81: Weibull lissées Médecins

Les figures suivantes représentent le lissage (simplifié) pour la loi Log-logistique. Tout comme la loi de Weibull, nous remarquons que le lissage a permis de lisser les irrégularités entre les âges.

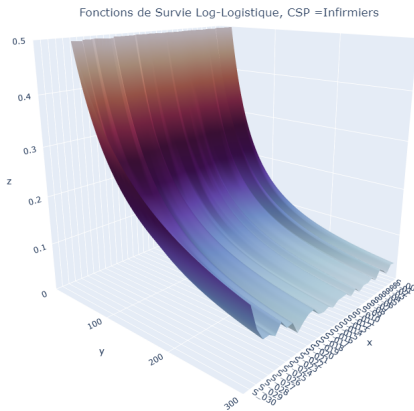


Figure 82: Log Logistique Infirmiers

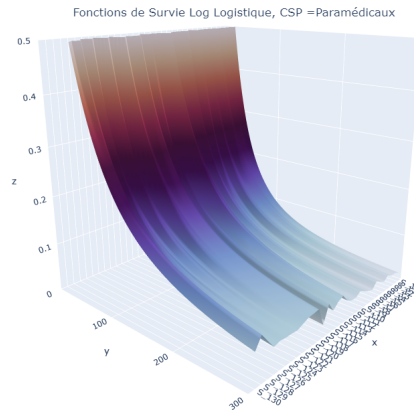


Figure 84: Log Logistique Paramédicaux

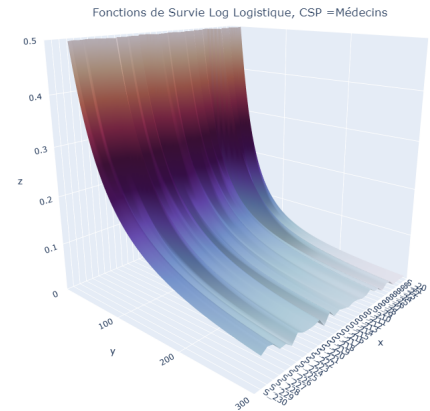


Figure 86: Log Logistique Médecins

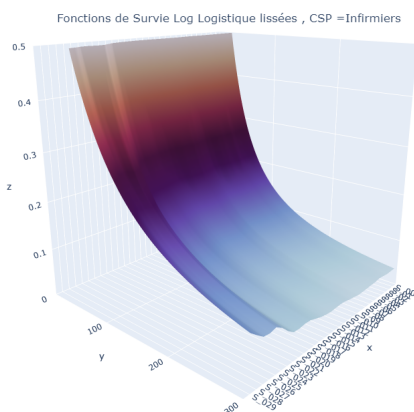


Figure 83: Log Logistique lissées Infirmiers

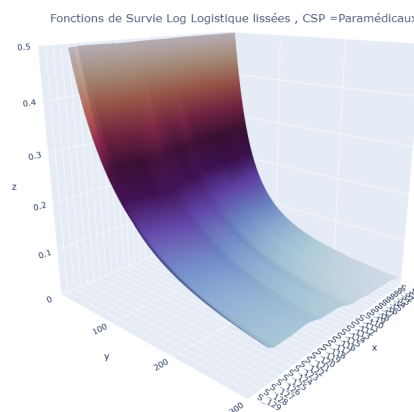


Figure 85: Log Logistique lissées Paramédicaux

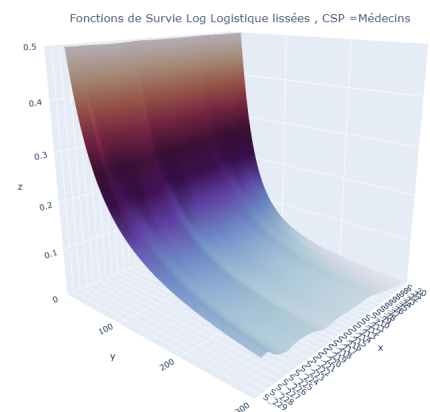


Figure 87: Log Logistique lissées Médecins

Au vu des résultats des critères d'informations, nous décidons de garder pour la suite de l'étude la loi de **Weibull** et la loi **Log-logistique**.

La section suivante porte sur l'application des modèles économétrique de type Cox et Aft.

Modèle de Cox pour le maintien en incapacité

Cette partie est consacrée à l'application du modèle de Cox pour le maintien en incapacité. Ce modèle permet de modéliser l'impact des variables explicatives sur le taux de risque instantané. La relation modélisée est la suivante :

$$\lambda(t|X) = \lambda_0(t) \exp(\beta_1 * CSP_{para} + \beta_2 * CSP_{medecins} + \beta_3 * Age)$$

Les résultats du modèle de Cox sont présentés dans le tableau 16. Tout d’abord, les résultats indiquent que toutes les variables explicatives sont significatives au seuil de 5 %. Concernant les valeurs des coefficients estimées, nous pouvons dire que TCEPA¹⁹, le fait d’appartenir à la catégorie des paramédicaux par rapport à être dans la catégorie des infirmiers, multiplie la probabilité instantanée de sortir de l’état incapacité de 1.09. Nous faisons le même constat pour les médecins. Enfin, nous remarquons que l’âge n’a pas d’effet significatif sur la probabilité de rester en incapacité.

Variabiles	Coef	e^{coef}	IC_-	IC_+	z	pvalue
CSP_{para}	0.09	1.09	1.06	1.12	5.51	<0.005
$CSP_{medecins}$	0.15	1.16	1.13	1.20	9.53	<0.005
Age	-0.01	0.99	0.99	0.99	-20.41	<0.005

Table 16: Résultats régression Cox BDD maintien

Les intervalles de confiance calculés sont les suivants :

$$IC = [e^{\hat{\beta}_j \pm 1.96 \sqrt{\hat{V}(\hat{\beta}_j)}}] \quad (82)$$

La table suivante présente le test de proportionnalité des coefficients. Nous remarquons que lorsque nous prenons la fonction identité, nous rejetons l’hypothèse de coefficients constants au cours du temps pour la variable âge. Pour la fonction logarithme ou la fonction de rank, nous constatons cette fois-ci que la variable $csp_{medecins}$ rejette l’hypothèse nulle de proportionnalité au seuil de 5 %.

Variables	Identité		Log		Rank	
	T	p	T	p	T	p
CSP_{para}	0.31	0.57	1.05	0.31	1.13	0.29
$CSP_{medecins}$	0.39	0.53	6.36	0.01	6.54	0.01
Age	7.53	0.00	0	0.98	0.00	0.99

Table 17: Test de proportionnalité (Résidus de Schoenfeld)

Ce test nous conduit à la conclusion que nous ne pouvons pas utiliser le modèle de Cox pour la suite de l’analyse, car il ne satisfait pas l’hypothèse de proportionnalité des coefficients. Bien qu’il existe des extensions du modèle de Cox pour prendre en compte la variation temporelle des coefficients, compte tenu du nombre de modèles et d’estimateurs déjà présentés, nous décidons de ne pas les appliquer afin de ne pas alourdir le mémoire. Malgré l’invalidité de l’hypothèse de proportionnalité, nous avons présenté en annexe une comparaison des résultats entre le modèle de Cox et l’estimateur de Kaplan-Meier.

La section suivante présente le modèle AFT.

Application des modèles AFT pour le maintien en incapacité

Cette section porte sur l’application du modèle AFT pour le maintien en incapacité. Nous avons appliqué le modèle AFT à la fois pour la loi de Weibull et la loi Log-logistique. Pour réaliser ce modèle, nous avons encodé la variable CSP de la manière suivante :

- CSP=0 : correspond aux infirmiers

¹⁹Toutes choses égales par ailleurs

- CSP=1 : correspond aux paramédicaux
- CSP=2 : correspond aux médecins

Afin de déterminer le modèle optimal, nous avons tout d'abord calculé les critères d'information AIC et BIC. La table 18 présente les résultats des différents critères d'information ainsi que le logarithme de la vraisemblance selon les hypothèses de loi utilisées.

Nous pouvons remarquer que le modèle avec la loi Log-logistique minimise à la fois le critère AIC et le critère BIC. De plus, la valeur de la vraisemblance est plus élevée que celle du modèle AFT Weibull. Cependant, bien que le modèle AFT avec la Log-logistique donne de meilleurs résultats que l'hypothèse de la loi de Weibull, la différence reste relativement faible.

Table 18: Critères d'information pour les modèles AFT

Modèle	Log Likelihood	AIC	BIC
Weibull	-146353.52	292717.03	292727.57
Log Logistique	-146206.74	292423.47	292434.01

La table 19 présente l'estimation des paramètres du modèle AFT sous la forme d'un modèle log-linéaire avec l'hypothèse d'une loi log-logistique.

Les coefficients estimés représentent donc l'accélération ou le ralentissement du temps passé en incapacité. Pour rappel,

- Si $\beta_k > 0$: TCEPA²⁰, une augmentation de X_k a un impact négatif sur le taux de hasard et par conséquent un impact positif sur la durée T.
- Si $\beta_k < 0$: TCEPA, une augmentation de X_k a un impact positif sur le taux de hasard et par conséquent un impact négatif sur la durée T.

Comme introduit précédemment, le modèle AFT est un modèle de type **log-level** sur T. Cela signifie que si une variable explicative k augmente d'une unité, cela entraîne une augmentation (ou une diminution) de la durée T de $(100 \times \beta_k)$ %.

Table 19: Paramètres AFT Log Logistique

Paramètres	Variable	Coef	e^{coef}	IC ₋	IC ₊	Z	pvalue
Lambda	CSP_{para}	-0.13	0.88	0.83	0.92	-4.91	<0.005
	$CSP_{medecins}$	-0.26	0.77	0.73	0.81	-9.78	<0.005
Age	Age	0.02	1.02	1.02	1.02	18.15	<0.005
	Constante	2.72	15.16	13.71	16.76	53.10	<0.005
	Rho	Constante	0.31	1.36	1.34	1.38	37.78

²⁰Toutes choses égales par ailleurs

Table 20: Paramètres AFT Weibull

Paramètres	Variable	Coef	e^{coef}	IC ₋	IC ₊	z	p
Lambda	CSP_{para}	-0.22	0.81	0.75	0.87	-5.67	<0.005
	$CSP_{medecins}$	-0.37	0.69	0.64	0.75	-9.61	<0.005
	Age	0.03	1.03	1.03	1.03	20.43	<0.005
	Intercept	1.18	3.26	2.68	3.97	11.82	<0.005
Rho	Intercept	-0.90	0.41	0.40	0.42	-59.36	<0.005

Interprétation

Concernant les résultats du modèle AFT avec la loi Log-logistique, nous remarquons que toutes les variables explicatives sont statistiquement significatives au seuil de 5%. En ce qui concerne les coefficients, nous observons que la variable explicative représentant la CSP des paramédicaux ($CSP_{paramedicaux}$) a un coefficient négatif. Cela indique que le fait d'appartenir à cette CSP est associé à une durée plus courte en état d'incapacité par rapport aux infirmiers. Ce résultat confirme les observations des statistiques descriptives selon lesquelles, en moyenne, les infirmiers et infirmières restent plus longtemps en incapacité. Nous faisons le même constat pour les médecins. $\beta_{csp_{medecins}} = -0.26$, ce qui signifie que le fait d'être médecin par rapport à être infirmier diminue la durée en état d'incapacité de 26%. La durée d'incapacité passe donc plus vite pour les médecins et les paramédicaux. De plus, la variable âge a un effet positif sur la durée passée en incapacité. Toutes choses égales par ailleurs, chaque année supplémentaire d'âge est associée à un risque de rester en incapacité plus longtemps.

La table 20 illustre l'estimation des paramètres du modèle AFT sous la forme d'un modèle log-linéaire avec l'hypothèse d'une loi Weibull.

De la même manière que le modèle Log-logistique, nous pouvons constater que tous les coefficients sont significatifs au seuil de 5%. Nous remarquons également que les coefficients (à l'exception des constantes) ont tous le même signe que le modèle AFT avec une loi Log-logistique. Nous avons donc les mêmes effets des covariables sur le temps passé en incapacité. Cependant, avec ce modèle, les valeurs des paramètres diffèrent légèrement. En effet, les coefficients des variables de CSP sont légèrement plus élevés que ceux du précédent modèle. Concernant l'âge, le coefficient est supérieur seulement de 1%.

L'avantage d'utiliser les modèles AFT est de déterminer si les variables de segmentation ont un effet causal (ou non) sur la durée passée en incapacité. D'après les résultats des deux modèles de régression, nous pouvons confirmer que la variable des catégories de professions ainsi que la variable âge ont un effet causal sur la durée passée en incapacité. Un autre avantage des modèles AFT est de pouvoir récupérer une loi paramétrique dont les paramètres dépendent des coefficients estimés. Cela facilite la partie simulatoire par rapport à une approche non-paramétrique.

La section suivante compare les résultats entre les lois marginales (avec segmentation préalable sur l'échantillon) et les modèles AFT.

Comparaison entre les modèles AFT et les lois marginales

$$T \sim W(\lambda, \rho) \text{ vs } T|X \sim W(\lambda = e^{\beta X}, \rho = \frac{1}{\sigma})$$

$$T \sim LL(\lambda, \alpha) \text{ vs } T|X \sim LL(\lambda = e^{\beta^T X - \beta_0}, \alpha = \frac{1}{\sigma}).$$

La figure 88 illustre la comparaison entre les lois marginales estimées lors de la première

section et les résultats des modèles AFT. Afin de mieux analyser les différences entre les deux méthodes, nous avons sélectionné les âges de 40 ans et 50 ans pour chacune des catégories de professions. Les résultats pour d'autres âges sont disponibles en annexe. De plus, les fonctions de survie ont été mises à la même échelle que l'estimateur de Kaplan-Meier afin d'avoir une comparaison par rapport à une estimation non-paramétrique. Cela signifie que nous avons divisé la fonction de survie par la fonction de survie à 14 jours. Soit :

$$S(x|x > 14) = \frac{S(x)}{S(14)} \quad (83)$$

En réalisant ce changement, nous créons une fonction de survie tronquée.

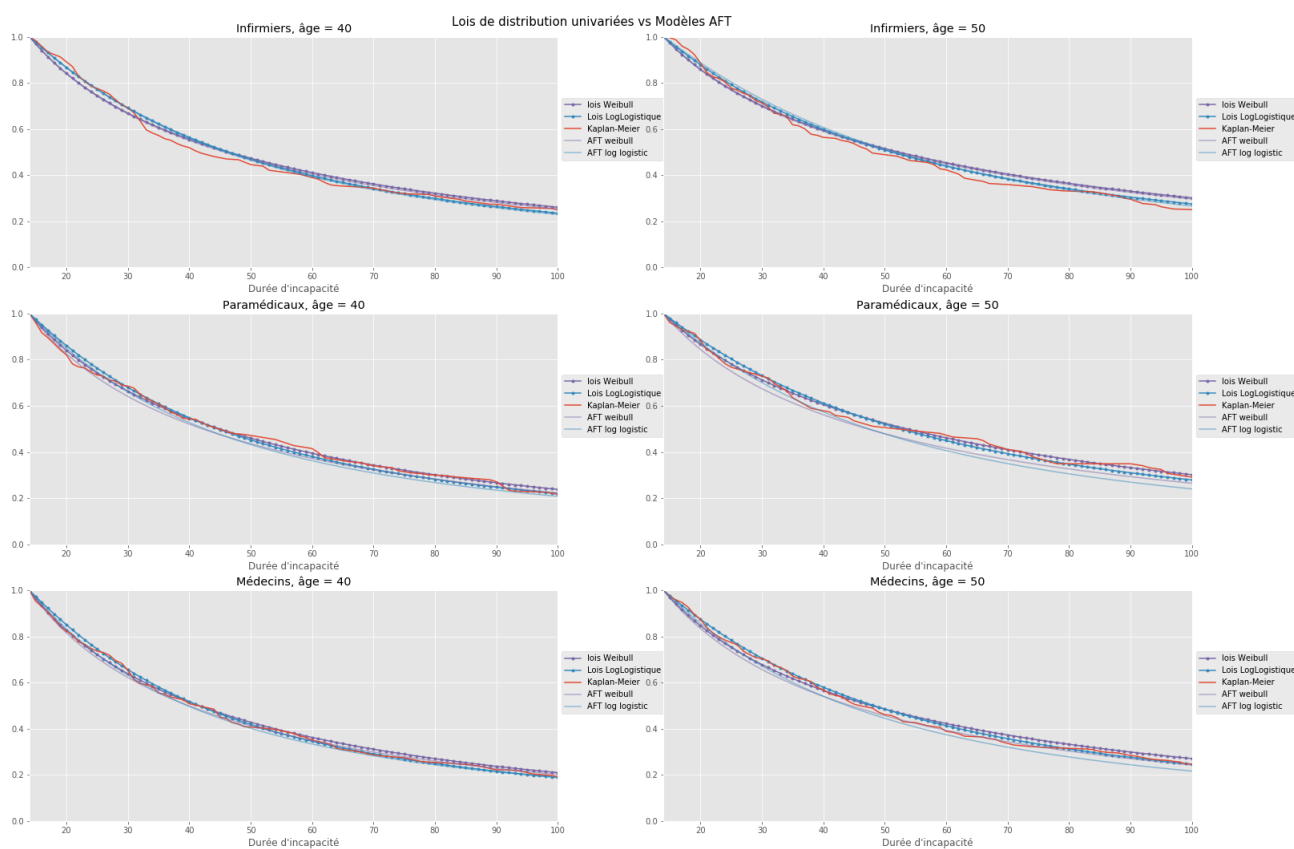


Figure 88: Comparaison entre les lois de distribution et les modèles AFT

D'une manière générale, nous remarquons que les résultats sont globalement assez similaires. Pour les *infirmiers et assimilés* âgés de 40 ans, les fonctions de survie théoriques semblent donner des valeurs légèrement plus élevées pour les durées entre 40 et 50 jours d'arrêts. Pour les âges de 50 ans, nous constatons un léger écart à 65 jours.

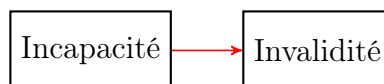
Chez les paramédicaux, nous remarquons peu d'écart pour l'âge de 40 ans entre l'approche non-paramétrique et les approches paramétriques. Cependant, à l'âge de 50 ans, nous commençons à observer des différences entre les lois marginales et les modèles AFT. Ces derniers semblent donner des valeurs légèrement inférieures.

Enfin, chez les médecins, nous constatons le même phénomène que dans la catégorie professionnelle précédente. Pour l'âge de 40 ans, il y a peu de différence, mais pour l'âge de 50 ans, des différences se manifestent entre les méthodes marginales et les modèles économétriques.

En examinant les résultats des critères d'information des modèles AFT, nous faisons le choix de conserver la loi **Log Logistique** pour toutes les catégories de CSP. En ce qui concerne le choix entre l'utilisation d'une loi de Log Logistique marginale ou issue d'un modèle AFT, nous décidons de maintenir la première approche, car les fonctions de survie sont similaires entre les deux approches.

La section suivante traite de la dernière partie de la modélisation, à savoir l'estimation des taux bruts de passage en invalidité.

3.1.4 Passage en invalidité



Dans cette section, nous présentons l'estimation des taux de passage en invalidité. Nous avons utilisé les estimateurs unidimensionnels **d'Hoem** et de **Kaplan-Meier** pour estimer les taux d'incidence bruts en fonction des classes d'âge et des classes de durée.

Comme expliqué dans la section des statistiques descriptives, nos deux bases de données ne contiennent qu'un faible nombre de sinistres (environ 700). Par conséquent, il est impossible de calculer un taux d'incidence pour chaque âge et chaque durée passée en incapacité.

Pour pallier cette limite, nous avons choisi de créer des classes pour agréger les données. Pour ce faire, nous avons exploré une méthode pour déterminer des classes de durées et d'âges homogènes. Cette méthode repose sur l'utilisation d'un modèle d'apprentissage automatique non-supervisé, à savoir l'algorithme des K-means.

La figure 89 illustre l'algorithme sur notre base de données de maintien. Nous avons inséré comme input les trois variables suivantes :

- Indicatrice de passage en invalidité
- Age à la survenance du sinistre
- Nombre de jours en incapacité

Comme nous pouvons le remarquer, l'algorithme parvient à détecter 5 classes d'âge pour une durée inférieure à 200 jours. Ensuite, il ne détecte plus que deux classes. Ces résultats ne nous permettent pas de créer des classes d'âge homogènes pour toutes les durées d'incapacité. Cela nous conduit donc à prendre la décision de ne pas utiliser d'algorithme de classification, mais plutôt à créer des classes manuellement. Pour des raisons de simplicité, nous avons donc créé des classes d'âge contenant 5 années et des classes de durée de 100 jours, comme suit :

- Classes d'âge : [30-35],[36-40],[40-45],[46-50],[51-55],[56-60]

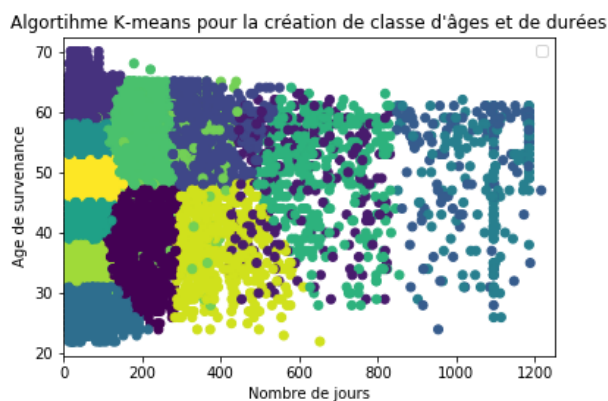


Figure 89: Algorithme de Kmeans pour la création de classes d'âge et de durée

- Classes de durée : $[0-100],[101-200],\dots,[801-900],[901-1000],[1001-1100]$

Maintenant, que nous avons défini les classes d'âge et de durée, nous pouvons passer à l'estimation des taux bruts.

La figure 90 présente les taux bruts de passage en invalidité obtenus à l'aide de l'estimateur d'Hoem et de Kaplan-Meier.

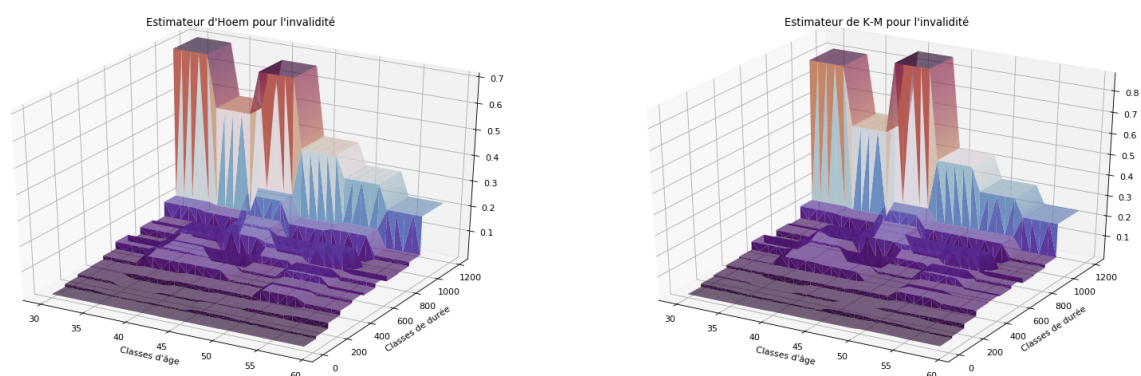


Figure 90: A gauche : estimateur d'Hoem ; A droite : estimateur de Kaplan-Meier

Les deux estimateurs fournissent des résultats similaires. En effet, nous retrouvons généralement les mêmes bosses. Après 1000 jours en état d'incapacité, des divergences commencent à se manifester entre les deux estimateurs. Pour l'estimateur d'Hoem, la probabilité de passer en incapacité monte jusqu'à 0.7 pour les classes d'âge $[30-35]$ et $[40-45]$ ans, alors que cette valeur est plutôt aux alentours de 0.8 pour l'estimateur de Kaplan-Meier. Cependant, il faut prendre en compte qu'à partir de cette durée, l'exposition est très faible. Il est donc nécessaire de rester prudent quant aux valeurs des probabilités après 1000 jours d'incapacité.

En résumé, les deux estimateurs se révèlent assez similaires en pratique. Par conséquent, nous avons décidé d'adopter l'estimateur de Kaplan-Meier pour notre simulation. La section suivante est consacrée à l'application du lissage.

Application du lissage pour les taux bruts du passage en invalidité

Cette section porte sur l'application du lissage de Whittaker-Henderson et la régression kernel en deux dimensions pour l'estimation des taux bruts du passage en invalidité via l'estimateur de Kaplan-Meier.

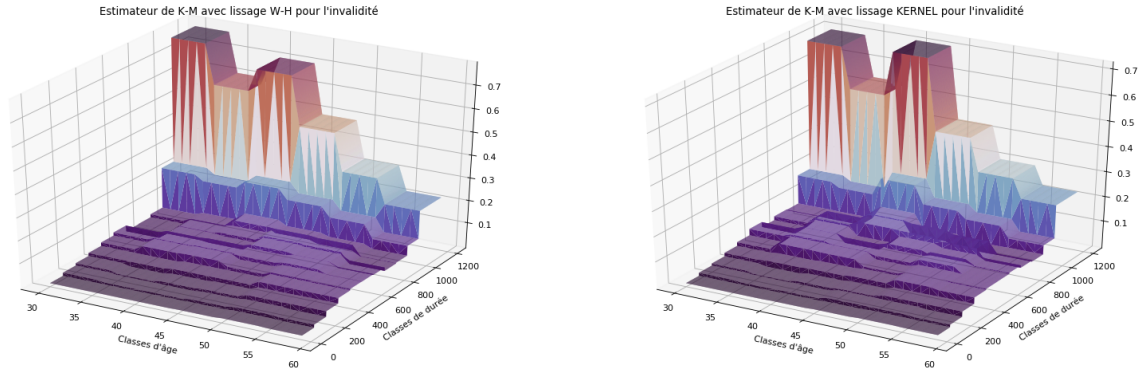


Figure 91: A gauche : Lissage Whittaker-Henderson ; A droite : Lissage kernel

Nous remarquons globalement que les deux méthodes de lissage donnent des résultats similaires. Par conséquent, nous décidons de sélectionner l'estimateur de Kaplan-Meier avec un lissage de Whittaker-Henderson.

3.1.5 Synthèse sur la calibration des sinistres

Pour conclure, l'application des différentes méthodes statistiques sur la base de données des sinistres nous a permis de comparer et de déterminer le modèle/l'estimateur le plus adéquat en fonction de la loi à modéliser.

Pour les taux d'incidence, nous avons décidé de garder l'estimateur de Kaplan-Meier avec un lissage de Whittaker-Henderson. Concernant le maintien en incapacité, nous avons opté pour une approche paramétrique marginale avec la loi Log-logistique pour chacune des catégories professionnelles.

Enfin, pour l'estimation du passage en invalidité, nous avons décidé de conserver l'estimateur de Kaplan-Meier avec un lissage de Whittaker-Henderson.

À présent, nous allons passer à la calibration sur la base de données des arrêts. Pour rappel, l'objectif de cette calibration sur une seconde base de données est de pouvoir comparer les résultats de la simulation. En théorie, les résultats devraient être assez proches.

3.2 Calibration sur des arrêts

Cette section traite de la calibration des méthodes statistiques en utilisant la base de données des **arrêts** pour le maintien en incapacité. Les estimateurs et les modèles économétriques demeurent identiques à ceux présentés dans la base de données de la durée totale du maintien. Afin de faciliter la lecture, les interprétations ont été simplifiées, car l'objectif est de présenter les résultats de manière générale plutôt que de les détailler à nouveau.

3.2.1 Statistiques descriptives

Dans cette section, nous procédons à une nouvelle analyse descriptive de la base de données des arrêts de travail.

Base de données des arrêts : Analyse des périodes d'incapacité

Le graphique 92 illustre la durée d'incapacité en fonction des différentes modalités des variables de segmentation.

En ce qui concerne le graphique relatif à la segmentation par sexe (partie supérieure gauche), la durée est cette fois-ci moins élevée que dans la précédente base de données. Cette observation est logique, car dans la base de données précédente, les durées étaient agrégées par sinistre. En revanche, dans cette nouvelle base de données, les durées représentent les périodes passées en incapacité pour chaque rechute, et non la durée globale du sinistre. Globalement, nous observons une faible volatilité de la médiane des durées des arrêts.

Pour les autres graphiques, nous constatons une volatilité plus importante parmi les catégories de professions par rapport au graphique précédent. En moyenne, la médiane varie entre 30 et 45 jours, alors qu'avec la base de données des sinistres, la médiane était d'environ 40 à 60 jours d'incapacité.

Les graphiques correspondant aux quantiles 1 et 3 demeurent similaires à ceux de l'ancienne base de données.

Synthèse

Cette analyse nous a permis d'examiner les caractéristiques spécifiques de la base de données des arrêts et de mettre en évidence les différences des durées par rapport à la base de données des sinistres.

Nous en concluons qu'en général, nous retrouvons les mêmes tendances et caractéristiques que dans la première base de données. Cela nous amène à maintenir les mêmes variables de segmentation pour chaque loi de durée à modéliser. Cette démarche d'homogénéisation nous permet de garantir la cohérence des résultats et de faciliter la comparaison des analyses entre les deux bases de données.

Dans la section suivante, nous nous intéressons à la modélisation statistique de la loi de durée des arrêts de travail.

Analyse des quantiles de durées d'incapacité par âge hors Grossesse & Incapacité supérieur à la franchise

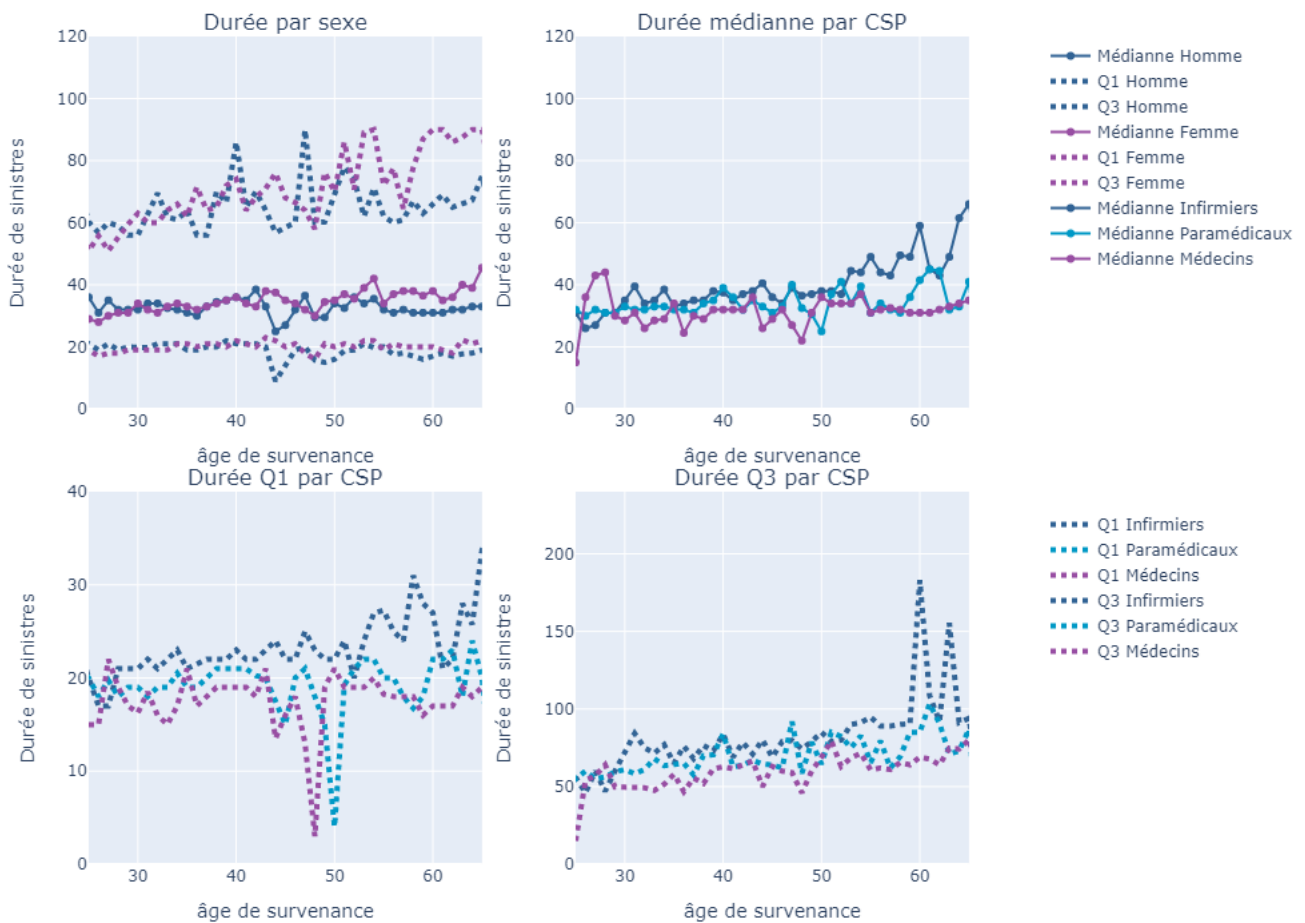


Figure 92: Durée des sinistres BDD arrêts

3.2.2 Maintien en incapacité

Cette section traite de la modélisation de la durée des arrêts. Tout comme pour la base de données des durées de sinistres, nous prenons en compte les trois mêmes lois candidates (Weibull, Log-logistique et Gamma généralisée) pour la modélisation paramétrique. Concernant la modélisation non-paramétrique, nous utilisons à nouveau l'estimateur de Kaplan-Meier. Contrairement à la base de données précédente, ici, nous n'avons pas de troncature à gauche. Nous analysons des durées de rechutes (qui peuvent être inférieures à 14 jours), mais la somme des durées d'arrêts pour le sinistre est supérieure à 14 jours.

Estimateur non-paramétrique pour le maintien en incapacité

Cette section traite de l'application de l'estimateur de Kaplan-Meier pour estimer le maintien en incapacité.

Les figures 93 94 95 présentent les résultats de l'estimateur de Kaplan-Meier pour chaque catégorie.

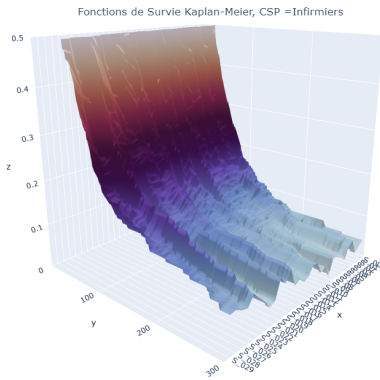


Figure 93: Infirmiers

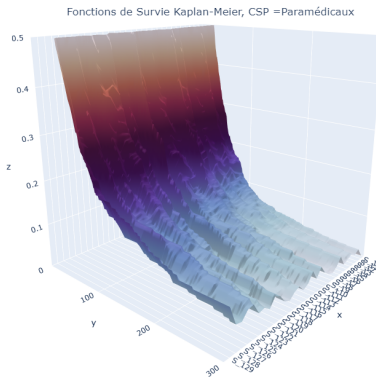


Figure 94: Weibull
Paramédicaux

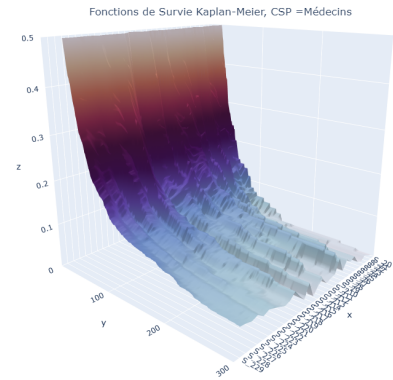


Figure 95: Médecins

Interprétation

Nous retrouvons à nouveau le caractère non-paramétrique de l'estimateur de Kaplan-Meier. Cependant, au vue des résultats plutôt homogène et similaire à des fonctions de survie théorique, nous décidons comme pour la base de données des sinistres de ne pas poursuivre avec cette méthode.

Estimation paramétrique pour le maintien en incapacité des arrêts de travail

Le tableau suivant présente les critères d'information AIC et BIC optimaux pour chaque CSP. En annexe, se trouve le détail du critère optimal pour chaque âge. En ce qui concerne les infirmiers, nous constatons qu'aucune loi ne minimise majoritairement chaque critère AIC. Le choix se situe entre la loi Gamma généralisée et la loi Log-Logistique. Cependant, en ce qui concerne le critère BIC, c'est la loi Log-logistique qui minimise la plupart des âges.

Pour les paramédicaux, la loi optimale est soit la loi de Weibull soit la loi Log-logistique, et ce, pour les deux critères d'information. Enfin, pour les médecins, la loi optimale semble être la loi Log-logistique si l'on prend en compte les critères AIC et BIC. Nous en concluons que finalement, nous retrouvons les mêmes lois qu'avec la précédente base de données.

	AIC	BIC
Infirmiers	GG/LL	LL
Paramédiccaux	LL	LL
Médecins	GG/LL	LL

Table 21: Récapitulatif des critères d'informations

Ensuite, nous avons effectué le lissage des lois candidates à l'aide de la méthode des moments. Les résultats sont disponibles en annexe. Cette modélisation nous permet de conclure que, pour le moment, les lois de Weibull et Log-Logistique semblent les plus appropriées pour la modélisation paramétrique de la durée de maintien des arrêts de travail. Dans le but de challenger nos lois marginales, nous avons réalisé une estimation semi-paramétrique dans la section suivante.

Application du modèle Cox pour le maintien en incapacité

Cette section se concentre sur l'application du modèle de Cox pour modéliser la durée des arrêts de travail. Pour rappel, l'équation modélisée est la suivante :

$$\lambda(t|X) = \lambda_0(t) \exp(\beta_1 * CSP_{para} + \beta_2 * CSP_{medecins} + \beta_3 * Age)$$

Les résultats de la régression se trouvent dans la table ci-dessous.

Variabes	Coef	e^{coef}	IC_-	IC_+	z	pvalue
CSP_{para}	0.13	1.14	1.11	1.17	9.46	< 0.005
$CSP_{medecins}$	0.25	1.28	1.25	1.32	17.63	< 0.005
Age	-0.01	0.99	0.99	0.99	-16.04	< 0.005

Table 22: Résultats régression modèle COX, BDD arrêts

Tout d'abord, nous constatons que toutes les variables explicatives sont significatives au seuil de 5%. De plus, les variables csp_{para} et $csp_{medecins}$ ont un signe positif. Cela indique que, pour les individus appartenant à l'une de ces deux catégories par rapport aux infirmiers, le taux instantané augmente de 1.14 pour la CSP des paramédicaux et de 1.20 pour la CSP des médecins. Par conséquent, cela a un effet négatif sur la durée d'incapacité, ce qui est en cohérence avec les statistiques descriptives.

Nous avons réalisé le test des résidus de Schoenfeld afin de vérifier si le modèle de Cox valide l'hypothèse de proportionnalité des coefficients. La table 23 présente les résultats du test. Ce dernier indique que, pour chaque fonction $g()$ utilisée, il y a à chaque fois une variable explicative qui va à l'encontre de l'hypothèse de proportionnalité. Par conséquent, cela nous amène à rejeter le modèle de Cox pour la modélisation du maintien. En annexe, une comparaison des fonctions de survie entre le modèle de Cox et l'estimateur de Kaplan-Meier sont disponible.

Variables	Identité		Log		Rank	
	T	p	T	p	T	p
CSP_{para}	1.81	0.17	4.77	0.02	3.53	0.06
$CSP_{medecins}$	5.07	0.02	18.46	0.00	22.35	0.00
Age	1.75	0.18	2.55	0.11	8.41	0.00

Table 23: Test des résidus de Schoenfeld pour les arrêts

La prochaine section porte sur les modèles AFT.

Application des modèles AFT pour le maintien en incapacité

Cette section traite de l'application des modèles AFT en supposant la loi de Weibull et la loi Log-logistique.

La table 24 présente les valeurs des critères d'information AIC, BIC ainsi que la log-vraisemblance pour les deux modèles.

Tout comme avec la première base de données, nous constatons que le modèle AFT en supposant la loi Log-Logistique minimise les deux critères d'information et présente la log-vraisemblance la plus élevée. Cependant, la différence entre les deux modèles reste également assez minime dans cette nouvelle base de données.

Table 24: Critères d'information pour les modèles AFT de la base de données des arrêts

Modèle	Log Likelihood	AIC	BIC
Weibull	-187000.05	374010.10	374021.09
Log Logistique	-182475.55	364961.09	364972.08

Les résultats des coefficients de la régression sont présentés dans la table 25. Les signes des coefficients demeurent les mêmes que précédemment. Les variables csp_{para} et $csp_{medecins}$ conservent un signe négatif, ce qui indique que la durée passée en incapacité ralentit si l'individu appartient à l'une de ces deux catégories professionnelles par rapport aux infirmiers. La valeur du coefficient pour l'âge est quasiment identique et est égale à 0.01.

Table 25: Paramètres AFT Log logistique BDD arrêts

Paramètres	Variabes	Coef	e^{coef}	IC ₋	IC ₊	z	p
Lambda	CSP_{para}	-0.15	0.86	0.84	0.89	-10.27	<0.005
	$CSP_{medecins}$	-0.30	0.74	0.72	0.77	-20.26	<0.005
	Age	0.01	1.01	1	1.01	10.29	<0.005
	Intercept	3.50	33.08	31.47	34.77	137. of 47	<0.005
Rho	Intercept	0.5	1.64	1.63	1.66	110.67	<0.005

Les résultats de la régression avec la loi de Weibull, présentés dans la table 26, affichent à nouveau les mêmes signes pour les coefficients. Seules les valeurs des coefficients diffèrent. Les interprétations demeurent identiques à celles précédemment évoquées.

Table 26: Paramètres AFT Weibull BDD arrêts

Paramètres	Variable	Coef	e^{coef}	IC ₋	IC ₊	z	p
Lambda	CSP_{para}	-0.19	0.83	0.80	0.85	-11.75	<0.005
	$CSP_{medecins}$	-0.33	0.72	0.70	0.75	-19.94	<0.005
	Age	0.01	1.01	1.01		21.22	<0.005
	Intercept	3.76	42.89	40.52	45.40	129.58	<0.005
Rho	Intercept	-0.15	0.86	0.85	0.87	-41.06	<0.005

En conclusion, l'analyse des critères d'information semble montrer que le modèle AFT avec la loi Log-logistique surpasse légèrement le modèle AFT avec la loi de Weibull. La section suivante procède à la comparaison entre les lois marginales et les lois obtenues à l'aide du modèle AFT.

Comparaison entre les modèles AFT et les lois marginales

La figure 96 illustre la différence entre les fonctions de survie issues des lois marginales et celles obtenues avec le modèle AFT pour les âges 40 ans et 50 ans. Comme évoqué en introduction, nous ne travaillons pas sur des données tronquées individuellement, mais plutôt sur une troncature à gauche de l'ensemble des arrêts du sinistre. Pour cette raison, nous n'avons donc pas modifié l'échelle en divisant par la fonction de survie à 14 jours.

En ce qui concerne les infirmiers, nous remarquons que la fonction de Kaplan-Meier est très proche de la fonction de survie Log-Logistique. Entre l'approche marginale et le modèle

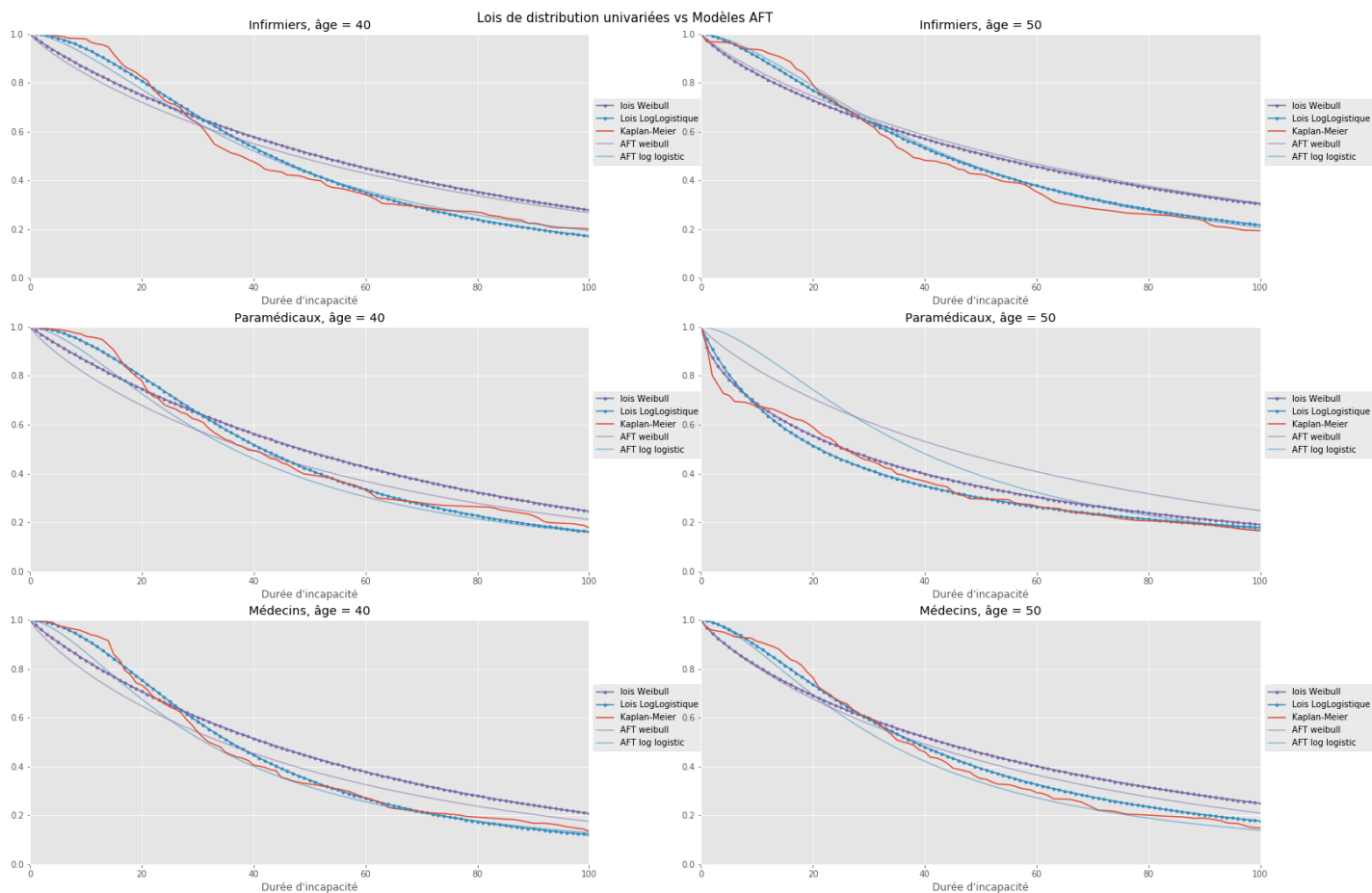


Figure 96: Comparaison entre les lois de survie marginales et les modèles AFT

économétrique, il n'y a pas de différence.

Pour les paramédicaux, nous constatons le même phénomène que précédemment entre l'approche non-paramétrique et celle paramétrique. En effet, la fonction de survie de Kaplan-Meier semble similaire à la fonction de survie de la loi Log-logistique marginale. En revanche, cette fois-ci, nous remarquons des divergences entre la méthode marginale et la méthode AFT.

Enfin, pour les médecins de 40 ans, des différences apparaissent pour les deux approches. Pour les médecins de 50 ans, nous observons une différence avec la loi de Log-logistique, mais l'écart est moins prononcé que celui observé chez les paramédicaux du même âge. D'autres comparaisons avec différents âges sont disponibles en annexe.

En conclusion, nous sélectionnons la loi **Log-Logistique** pour modéliser la durée du maintien en incapacité pour chaque CSP. La modélisation avec les modèles AFT a confirmé notre a priori sur la loi. Concernant le choix de la méthode, nous décidons à nouveau de garder l'approche marginale.

La section suivante porte cette fois-ci sur la modélisation des rechutes.

3.2.3 Modélisation des rechutes

Dans cette section, nous présentons la modélisation du nombre de rechutes. Auparavant, la modélisation des rechutes n'était pas requise, car la durée englobait l'ensemble des arrêts du sinistre, c'est-à-dire la durée totale du sinistre. Cependant, dans ce contexte, étant donné que nous traitons des arrêts spécifiques, il est nécessaire de modéliser le nombre de rechutes. Tout comme en assurance Non-Vie, la loi optimale utilisée pour modéliser la fréquence de rechute (ou des sinistres) est déterminée en fonction de la dispersion des données.

Pour rappel :

- Sous-dispersion : Loi Géométrique
- Equi-dispersion : Loi Poisson
- Sur-dispersion : Loi Binomiale Négative

La figure suivante présente les résultats du paramètre de dispersion pour chaque modalité de la variable de la catégorie socioprofessionnelle.

CSP	ϕ
Infirmiers	2.9
Paramédicaux	8.2
Médecins	13.6

Table 27: Résultats du paramètre de dispersion pour chaque CSP

Les résultats montrent qu'il y a de la sur-dispersion pour chacune des modalités de la variable explicative de la catégorie socioprofessionnelle. Nous faisons donc le choix de prendre une loi Binomiale Négative. La fonction de masse associée à la distribution est :

$$P(Y = y|x) = \frac{\Gamma(\theta + y)}{\Gamma(y + 1)\Gamma(\theta)} r^y (1 - r)^\theta \quad (84)$$

$$\mu = E[Y|X] = b'(\theta_i) = g^{-1}(\eta_i) \quad (85)$$

$$= e^{x^T \beta} = e^{\beta_0 + \beta_1 CSP_{para} + \beta_2 CSP_{medecins} + \sum_{i=1}^p \beta_i age_{i+29}} \quad (86)$$

$$r = \frac{\mu}{\theta + \mu} \quad (87)$$

En posant, $r = \theta$ et $p = r$, nous obtenons la loi suivante, $\forall k \in N$:

$$f(y) = \binom{y}{y + r - 1} p^r (1 - p)^y \quad (88)$$

avec l'objectif final, d'obtenir une loi de comptage pour simuler notre portefeuille. La loi obtenue sera la suivante :

$$Y|X \sim BN(r, p) \quad (89)$$

Les résultats de la régression sont présentés dans le tableau suivant. Pour des raisons de clarté, seuls certains coefficients par âge ont été inclus dans ce tableau. La version complète de la régression est disponible en annexe pour consultation.

Variables	Estimation	Std.Error	z	value
(Intercept)	-1.96680	0.12664	-15.531	<2e-16
factor(Grp_calibration) Paramédicaux	0.30390	0.05323	5.710	1.13e-08
factor(Grp_calibration) Médecins	0.48248	0.05086	9.486	<2e-16
factor(age_maintien)31	0.48248	0.05086	9.486	<2e-16
factor(age_maintien)35	0.10003	0.16628	0.602	0.54747
factor(age_maintien)40	0.11077	0.16979	0.652	0.51415
factor(age_maintien)45	0.31380	0.17381	1.805	0.07102
factor(age_maintien)50	0.24828	0.17098	1.452	0.14648
factor(age_maintien)55	0.28832	0.16028	1.799	0.07203
factor(age_maintien)60	0.61892	0.15368	4.027	5.64e-05

Table 28: Résultats totaux de la régression Binomiale Négative

Critères	valeur
Theta	0.15611
2xLog likelihood	-30919.95
AIC	30988

Table 29: Critères d'information de la régression

Interprétation

Nous remarquons que tous les coefficients estimés sont positifs. Par exemple, pour un individu appartenant à la catégorie des paramédicaux et assimilés, la fréquence de rechute est **plus élevée de 30.3 %** par rapport à un assuré appartenant à la catégorie des infirmiers et assimilés. De même, pour les médecins, la fréquence de rechute est **supérieure de 48.24 %** par rapport aux infirmiers. Nous pouvons également observer une tendance similaire en ce qui concerne les différents groupes d'âge, où une augmentation de l'âge est associée à une augmentation de la fréquence de rechute.

Dans le but d'avoir des paramètres de loi homogène, nous avons décidé d'appliquer de lisser les paramètres de la variable âge avec une régression kernel de type gaussien.

Pour déterminer le paramètre optimal de lissage dans notre régression kernel, nous avons utilisé la méthode de validation croisée. Celle-ci nous a indiqué que $h_{optimal} = 3.2$ était le paramètre de lissage optimal. La figure 98 illustre le résultat du lissage pour la variable *âge*.

Nous remarquons que le lissage avec une valeur de paramètre de $h_{optimal} = 3.2$ permet d'homogénéiser la valeur des coefficients estimés.

La figure suivante expose le paramètre μ issu du modèle GLM. D'une manière générale, nous pouvons observer que la fréquence de rechute est plus élevée pour les médecins, suivis des paramédicaux et assimilés, puis des infirmiers.

Cette section ferme le chapitre concernant la modélisation des probabilités de transitions et de la loi de durée. En conclusion, nous décidons de garder le même choix que sur la base de données des sinistres, à savoir, utiliser la loi marginale Log-logistique pour modéliser la durée des arrêts de travail. A cela, nous décidons d'utiliser une loi Binomiale Négative pour la modélisation des rechutes. Le chapitre suivant se consacre à la tarification des traités de réassurance.

Lissage des $\hat{\beta}_{\text{âge}}$ estimés avec une régression kernel

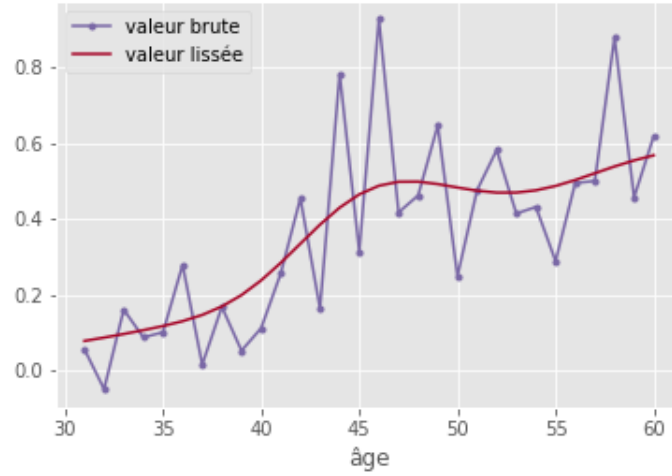


Figure 97: Lissage des coefficients des âges

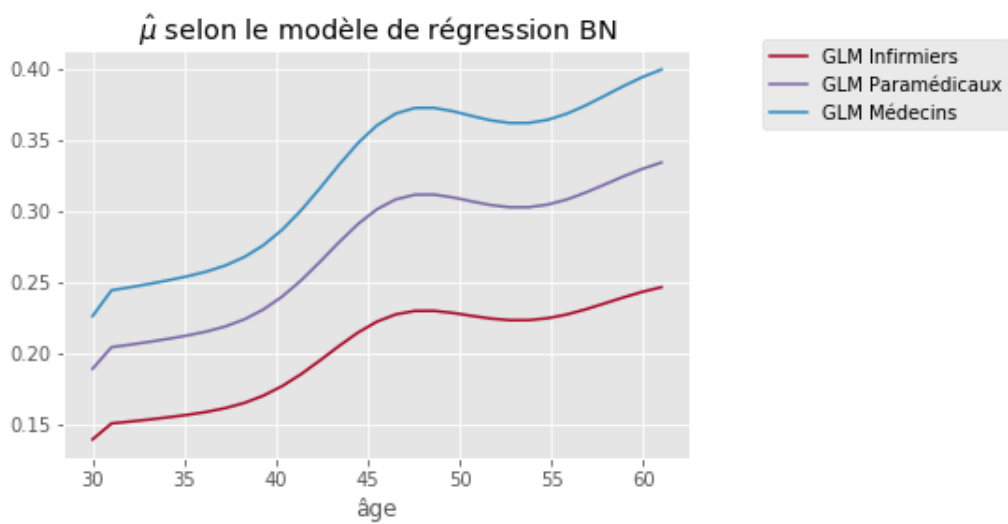


Figure 98: GLM avec une loi Binomiale Négative

4 Tarification de traités de réassurance prévoyance

Cette section vise, tout d'abord, à expliquer les principales caractéristiques de la réassurance, en mettant en évidence notamment les traités non-proportionnels. Ensuite, nous présenterons les résultats de la simulation réalisée sur notre portefeuille d'exposition pour l'année 2022. Enfin, le programme d'optimisation du traité de réassurance interne sera détaillé.

4.1 Concepts

En réassurance, deux grandes catégories de traités sont principalement utilisées : les traités proportionnels et non-proportionnels. Chacune de ces catégories comprend des sous-catégories de traités, à savoir :

- Les Traités proportionnel
 - Quote-Part (QS) : Le réassureur s'engage à payer les sinistres proportionnellement aux primes qu'il reçoit de la cédante²¹.
 - Excédent Plein (XP) : Le réassureur s'engage à payer un certain taux de cession de sinistralités à la cédante. Ce taux est calculé sur la base du montant de capital à assurer.
- Les Traités non-proportionnel
 - Excédent de sinistre (Excess of Loss XL): Le réassureur s'engage à payer police par police le montant qui excède la priorité et dans la limite de la portée.
 - Excédent de perte annuelle (Stop Loss, SL) : C'est un traité similaire à l'excédent de sinistre. La différence vient du fait que la variable d'intérêt représente le résultat plutôt que les montants des sinistres.

La principale distinction entre un traité proportionnel et un traité non-proportionnel réside dans la variable d'intérêt utilisée. Le traité proportionnel est basé sur le **montant du capital assuré**, tandis que le traité non-proportionnel se concentre sur la **sinistralité**.

En ce qui concerne le produit prévoyance, la MACSF a opté pour des traités en excédent de sinistre appliqués individuellement à chaque assuré, autrement dit, nous appliquerons un traité de réassurant en excédent de sinistre tête par tête.

Ce type de traité de réassurance présente deux caractéristiques principales : la portée (ou limite) notée l et la franchise notée d .

La franchise représente le seuil à partir duquel le réassureur prend en charge les sinistres de la cédante. En dessous de ce seuil, la cédante est responsable du paiement des sinistres.

La portée correspond au montant entre la franchise et le montant maximum que le réassureur s'engage à couvrir.

En notation générale, ce type de traité est représenté par $l \text{ XS } d$, où " l " représente la portée et " d " représente la franchise.

Il est important de noter que chaque cédante a la possibilité d'établir plusieurs traités de réassurance avec différentes contreparties, en fonction de ses besoins spécifiques.

La figure 99 présente un schéma illustrant une cédante (la MACSF) et deux réassureurs. La cédante prend en charge les sinistres jusqu'à la franchise d_1 . Ensuite, le réassureur numéro

²¹L'entité qui cède sa sinistralité

1 prend en charge les sinistres au-delà de cette franchise, dans la limite de la portée l_1 . Enfin, le réassureur numéro 2 intervient pour les sinistres dont le montant dépasse la franchise d_2 , et dans la limite du plafond.

Pour illustrer notre propos, nous pouvons analyser la flèche en bleu qui représente un sinistre. Nous constatons que le montant du sinistre en bleu a dépassé la franchise d_1 , mais il est inférieur à la priorité du réassureur numéro 2. Dans ce cas, la MACSF prend en charge le montant jusqu'à la franchise d_1 , tandis que le réassureur numéro 1 prend en charge la partie du montant qui dépasse sa franchise.

Pour le deuxième sinistre en rouge, la MACSF ainsi que les deux réassureurs devront payer la totalité de la portée de leurs traités. La partie supérieure au traité souscrit avec le réassureur 2 sera à la charge de la MACSF.

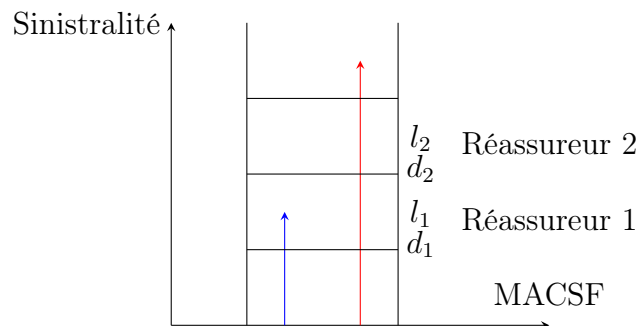


Figure 99: Exemple d'un traité en excédent de sinistre

La partie suivante explique la construction de nos traités de réassurance à l'aide des modèles construits dans ce mémoire.

Création du modèle simulatoire pour le produit prévoyance

Tout d'abord, nous avons extrait les données d'exposition de notre portefeuille pour l'année 2022. Ces données d'exposition contiennent les mêmes variables de segmentation utilisées pour la création des différentes lois de probabilité, ainsi que toutes les caractéristiques des garanties de chaque assuré.

Il est important de noter que chaque individu peut détenir plusieurs contrats. Chaque contrat possède ses propres garanties spécifiques.

Chaque garantie est caractérisée par différentes variables, nous trouvons principalement les modalités suivantes :

- Franchise
- Durée d'indemnité maximale
- Montant d'indemnité journalière (en cas d'incapacité)
- Valeur de la rente (en cas d'invalidité)

En général, en cas d'incapacité, le revenu de l'assuré est généralement couvert par le biais d'indemnités journalières, tandis que pour l'invalidité, les assurés perçoivent généralement des rentes. Les rentes sont payées mensuellement et doivent donc être provisionnées dans l'algorithme.

Une deuxième base de données, connue sous le nom de "base de simulation", est créée à partir de cette première base. La seule différence est que chaque individu est représenté par

une seule et même ligne tout en conservant ses caractéristiques personnelles. Cette base nous permet de simuler les individus un par un en fonction de leurs caractéristiques.

Pour conclure, nous présentons l'algorithme 1 utilisé pour la simulation de notre portefeuille. Dans cet algorithme, nous avons exposé les deux méthodes de simulation, en fonction de l'utilisation des lois de maintien modélisées à partir de la base de données de la durée totale du maintien en incapacité ou de celles issues de la base de données des arrêts.

De plus, étant donné que nous avons travaillé avec des individus ayant une franchise de **14 jours** depuis le début de l'étude, nous avons ajouté un filtre après la simulation des durées. Ce filtre correspond à la méthode de rejet. Comme nous simulons des passages en incapacité avec un q_x basé uniquement sur les incapacités d'une durée supérieure à la franchise de 14 jours, il est nécessaire de simuler des durées d'incapacité d'au moins 14 jours.

Pour la méthode basée sur la durée totale du sinistre, si la réalisation générée est inférieure à 14 jours, alors dans ce cas, nous tirons à nouveau une durée. Pour la méthode des arrêts, si la somme des durées des arrêts est inférieure à 14 jours, alors nous tirons à nouveau des réalisations de durées d'arrêt.

Algorithm 1 Algorithme de simulation

```

1: Initialisation: Set number of simulations NS
2: Input :  $f(j)$  = Contracts of individual j
3: for i in NS do
4:   for each individual (j) do
5:      $\Delta_{i,j}^{Incap} \sim B(\tilde{q}_x)$ 
6:     if  $\Delta_{i,j}^{Incap} == 1$  then
7:
8:       if Relapse distribution selected then
9:         while  $\Delta_{i,j}^{time_{total}} \leq 14$  do
10:           $\Delta_{i,j}^{relapse} \sim BN(r, p)$ 
11:           $\Delta_{i,j}^{time_{total}} \sim L(\theta)$ 
12:          for each (k) in  $\Delta_{i,j}^{relapse}$  do
13:             $\Delta_{i,j,k}^{time_{relapse}} \sim L(\theta)$ 
14:          end for
15:           $\Delta_{i,j}^{time_{total}} = \Delta_{i,j}^{time_{total}} + \Delta_{i,j,k}^{time_{relapse}}$ 
16:        end while
17:       else
18:         while  $\Delta_{i,j}^{time_{total}} \leq 14$  do
19:           $\Delta_{i,j}^{time_{total}} \sim L(\theta)$ 
20:        end while
21:       end if
22:
23:      $\Delta_{i,j}^{Inval} \sim B(q_x^{Inval})$ 
24:
25:      $C_{i,j}^{incap} = \Delta_{i,j}^{time_{total}} * f(j)$ 
26:      $C_{i,j}^{inval} = \Delta_{i,j}^{Inval} * f(j)$ 
27:      $C_{i,j}^{total} = C_{i,j}^{incap} + C_{i,j}^{inval}$ 
28:   end if
29: end for
30: end for

```

Pour un descriptif numérique plus détaillé de l’algorithme, le lecteur pourra trouver en annexe une version illustrée de l’algorithme.

Le résultat final sera donc un coût total par simulation et par individu. Ce coût regroupe les indemnités journalières de chaque individu (agrégés de tous leurs contrats) lié à leur passage dans l’état d’incapacité, ainsi que la rente obtenue en cas de passage en invalidité.

La suite consiste à calculer la somme des coûts par simulation afin d’obtenir un histogramme des coûts pour l’ensemble du portefeuille. Par la suite, nous pouvons calculer la moyenne de toutes ces simulations. Cette moyenne permet d’obtenir une **référence de coût** ou autrement dit, une **prime pure** pour le portefeuille de la MACSF en vue du renouvellement des prochains traités de réassurance.

Programme d’optimisation pour le traité de réassurance interne

La seconde partie de ce chapitre vise à optimiser le traité de réassurance interne (tranche 1) en déterminant les paramètres optimaux du traité. Pour ce faire, nous utilisons un programme d’optimisation sous contraintes. L’objectif est de maximiser la volatilité des sinistres de la tranche 1 tout en respectant un plafond de réassurance et un ratio $\frac{S}{P}$ fixe. En d’autres termes, nous cherchons à minimiser la volatilité des coûts de sinistralité de la MACSF.

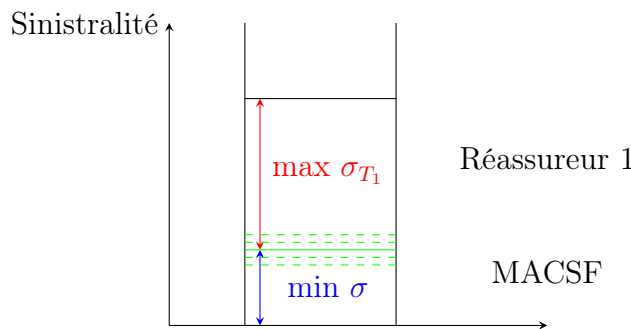


Figure 100: Objectif du programme d’optimisation

Mathématiquement, nous formulons le programme d’optimisation suivant :

$$\begin{aligned}
 & \max_{AAD, \text{priorité}} \quad \sigma_{T_1} \\
 & \text{s.t.} \quad d+l=x_1 \\
 & \quad \quad \frac{S}{P} = x_2 \\
 & \quad \quad b_-^1 < d \leq b_+^1 \\
 & \quad \quad b_-^2 < AAD \leq b_+^2
 \end{aligned} \tag{90}$$

Avec x_1 et x_2 des valeurs fixes, d la priorité, l la portée et b les bornes.

Dans un premier temps, nous allons donc faire varier les valeurs de la priorité et de l’AAD²². Ensuite, nous allons sélectionner les couples (priorité, AAD) qui permettent d’obtenir un ratio $\frac{S}{P}$ égal au montant x_2 . Pour chacun de ces couples, nous analyserons l’écart-type correspondant et sélectionnerons celui ayant la valeur la plus élevée.

²²Annual Aggregate Deductible

		Priorité	
		b_-^1	b_+^1
AAD	b_-^2		x_2
	b_+^2	x_2	

		Priorité	
		b_-^1	b_+^1
AAD	b_-^2		σ_1
	b_+^2	σ_4	σ_2

Table 30: Exemple du programme d'optimisation sous contraintes

Traités de réassurance externe

Dans le cas des traités de réassurance externes (tranche 2), nous allons utiliser différents principes de calcul de la prime pure afin d'obtenir le coefficient de chargement de sécurité. Le coefficient ainsi déterminé nous servira de référence pour évaluer la marge de sécurité appliquée par le réassureur lors du processus de négociation des traités de réassurance.

Dans la pratique, la prime demandée par l'assureur (ou le réassureur à un assureur) ne correspond pas à l'espérance des sinistres (prime pure théorique). En réalité, elle se compose de cette prime pure théorique et d'un chargement de sécurité visant à se prémunir contre le risque de ruine.

Il existe plusieurs principes de calcul de la prime pure. En voici les principaux²³ :

- Principe de l'espérance mathématique : $\Pi(X) = (1 + \rho)E(X)$
- Principe de la variance : $\Pi(X) = E(X) + \rho V(X)$
- Principe de l'écart-type : $\Pi(X) = E(X) + \rho\sqrt{V(X)}$
- Principe de quantile : $\Pi(X) = F_X^{-1}(\alpha)$
- Principe exponentiel : $\Pi(X) = \ln E[e^{\rho X}]$
- Principe d'Esscher : $\Pi(X) = \frac{E[Xe^{\rho X}]}{E[e^{\rho X}]}$

Dans notre étude, nous nous intéressons aux quatre premiers principes.

Dans le cas de notre étude, $E(X)$ et $V(X)$ correspondent respectivement à la moyenne et à la variance des coûts des simulations. $\Pi(X)$ correspond à la prime actuellement payée par MACSF. Avec ces trois mesures statistiques, nous pouvons estimer le coefficient de chargement de sécurité selon les différentes approches.

²³Voir Mathématiques de l'assurance Non-Vie d'A.Charpentier, [20]

En plus de calculer le coefficient de chargement de sécurité pour le traité actuel, nous allons le calculer pour différents niveaux de primes. Cela nous permettra d'obtenir une référence sur le chargement de sécurité demandé par le réassureur. Cette grille servira de référence dans le cadre de la négociation du prochain traité.

La section suivante présente les résultats des simulations.

4.2 Application

Remarque : Pour rappel, les résultats ont été modifiés pour des raisons de confidentialités.

Cette section traite de l'application de l'algorithme présenté dans la section précédente, des résultats de l'optimisation du traité de réassurance interne, ainsi que le calcul du coefficient de chargement de sécurité pour la réassurance externe.

Nombres de simulation

Pour obtenir des résultats statistiquement significatifs, il est nécessaire d'obtenir la convergence de la moyenne et de la variance des simulations. La figure 101 présente deux graphiques de simulation selon les deux approches de simulation utilisées (loi totale et loi des arrêts).

Le premier graphique illustre la convergence de la moyenne des simulations. La formule appliquée est la suivante :

$$\bar{X} = \frac{1}{B} \sum_{j=1}^B C_j \quad (91)$$

$$C_j = \sum_{i=1}^N c_{j,i} \quad (92)$$

Avec B représentant le nombre de simulations, C_j le coût total du portefeuille de la simulation j , et c_i le coût total de l'individu i .

Nous en concluons que pour obtenir une convergence moyenne des coûts totaux, il est nécessaire de réaliser au moins 200 simulations par individu.

Le second graphique illustre la convergence en termes de variance des simulations. La formule appliquée est la suivante :

$$\sigma^2 = \frac{1}{B} \sum_{j=1}^B (C_j - \bar{C}_j)^2 \quad (93)$$

$$\bar{C}_j = \frac{1}{n} \sum_{i=1}^N c_{j,i} \quad (94)$$

Nous concluons qu'à partir de 300 simulations, la variance commence à se stabiliser. Finalement, pour des raisons de convergence et de puissance de calcul, nous décidons de maintenir ce nombre de **300 simulations** pour chaque individu.

Cela revient donc à un total de $300 * 80\,000 = 24\,000\,000$ de simulations.

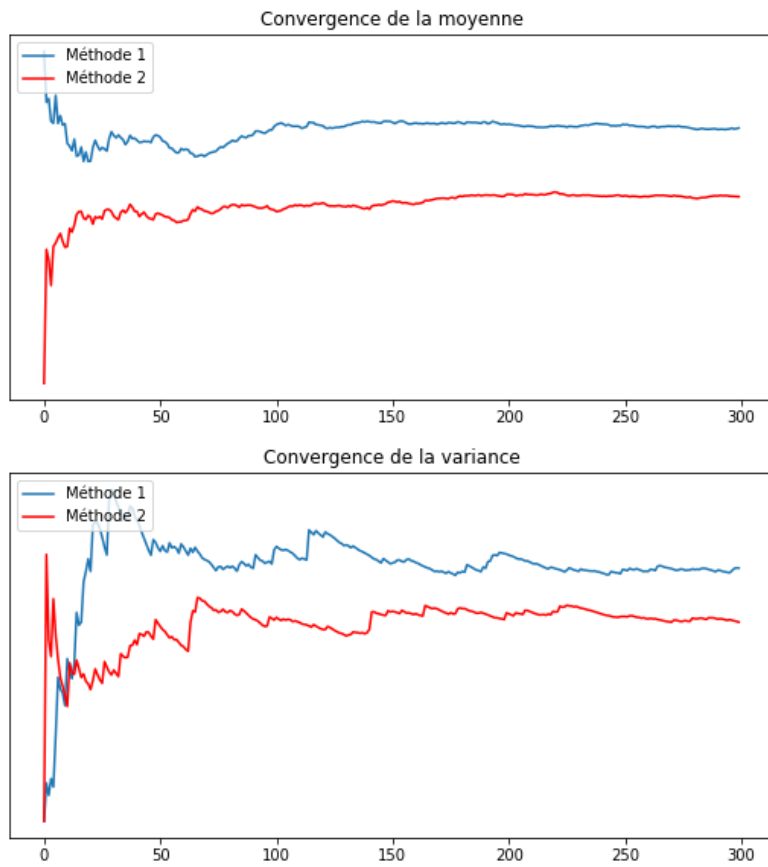


Figure 101: Convergence des simulations

Coûts des simulations par tranche

Le graphique 102 représente les coûts totaux du portefeuille par simulation.

En première ligne, nous avons les coûts bruts (hors réassurance) ventilés en fonction des garanties, ainsi que le total des coûts d'incapacité et d'invalidité à droite.

En deuxième ligne, à gauche, nous avons les coûts en deçà de la première franchise, c'est-à-dire les coûts à la charge de la MACSF. Au milieu, se trouvent les coûts associés à la première tranche de réassurance (interne). Enfin, le graphique de droite représente les coûts de la tranche du second réassureur (externe).

En **bleu**, nous avons la loi totale des sinistres et en **rouge**, nous avons la durée des arrêts couplée à une loi de comptage.

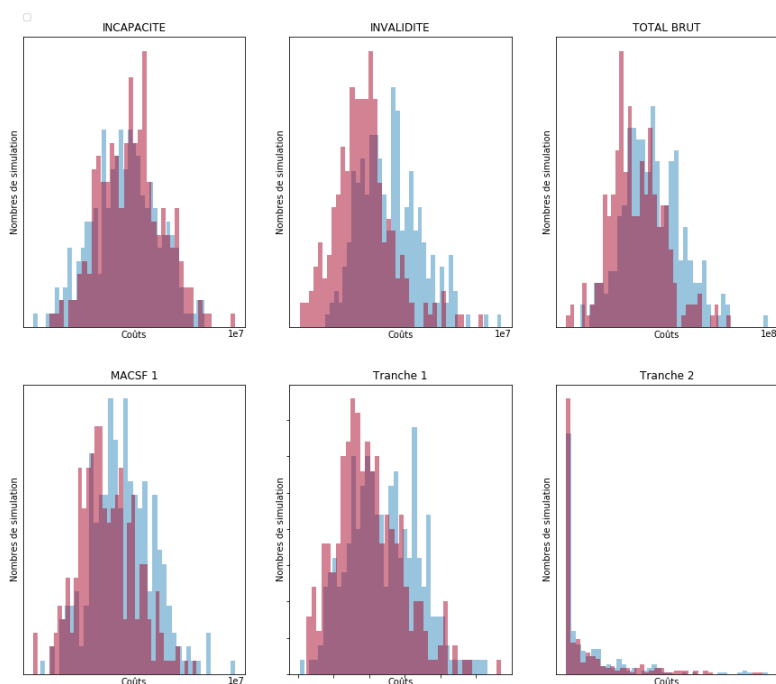


Figure 102: Résultats des simulations en fonction des tranches de réassurance

Sur la sinistralité brute de réassurance, nous remarquons des coûts d'incapacité similaires entre les deux méthodes. En revanche, pour l'invalidité, la méthode utilisant la durée totale des sinistres montre un coût plus élevé que la seconde méthode.

Afin de départager les deux méthodes, nous avons analysé l'écart entre le coût moyen simulé (stochastique) et le coût réellement payé (déterministe) pour l'incapacité et l'invalidité. Conformément aux résultats de ce backtest (non affichés pour des raisons de confidentialité), nous avons choisi de conserver la méthode de simulation avec la loi totale du sinistre. En effet, les résultats pour l'invalidité sont plus proches de la réalité que ceux issus de la seconde approche avec la loi des arrêts couplée à une loi de comptage. Cet écart est également visible dans les graphiques de convergence présentés en figure 101. L'hypothèse d'indépendance entre le nombre de rechutes et leur durée, ainsi que la dépendance entre les durées de rechute, doit être remise en question.

Nous constatons que cet écart de modélisation de l'invalidité se traduit par une différence relative de 2.6% entre les variances des deux méthodes. Pour la tranche 1, les écarts sont plus significatifs. Nous observons une différence relative de 15% pour la moyenne et de 15% pour la

Tranche	Méthode	Q1	Médiane	Moyenne	Q3	Variance
MACSF 1	Sinistres	q_1	ME	m	q_3	Var
	Arrêts	-1.5 %	-1.6%	-1.2%	-1.3%	-2.6%
Tranche 1	Sinistres	q_1	ME	m	q_3	Var
	Arrêts	-13%	-21%	-15%	-11.5%	-15%
Tranche 2	Sinistres	q_1	ME	m	q_3	Var
	Arrêts	0	-62%	-17%	-2.5	-32%

variance. Enfin, pour la dernière tranche de réassurance, la différence relative entre la moyenne et la variance augmente, notamment pour la variance.

Optimisation sous contrainte de la réassurance

Dans cette partie, nous présentons l'application du programme d'optimisation sous contraintes pour la tranche de réassurance interne. Pour rappel, notre objectif est de déterminer le couple (priorité/AAD) optimal pour un ratio S/P fixé et égal à 0,75. Le tableau suivant présente les S/P obtenus en fonction des différents couples de priorité et d'AAD.

AAD/Priorité	400000	425000	450000	475000	500000	525000	550000	575000	600000	625000	650000	675000
0	1.44	1.31	1.19	1.08	0.99	0.90	0.82	0.75	0.68	0.62	0.56	0.51
100000	1.41	1.28	1.16	1.05	0.96	0.87	0.79	0.72	0.65	0.59	0.53	0.48
200000	1.38	1.25	1.13	1.03	0.93	0.84	0.76	0.69	0.62	0.56	0.51	0.46
300000	1.36	1.22	1.10	1.00	0.90	0.81	0.73	0.66	0.59	0.53	0.48	0.43
400000	1.33	1.19	1.08	0.97	0.87	0.78	0.71	0.63	0.57	0.50	0.45	0.40
500000	1.30	1.17	1.05	0.94	0.84	0.76	0.68	0.60	0.54	0.48	0.42	0.37
600000	1.27	1.14	1.02	0.91	0.81	0.73	0.65	0.58	0.51	0.45	0.39	0.35
700000	1.24	1.11	0.99	0.88	0.79	0.70	0.62	0.55	0.48	0.42	0.37	0.32
800000	1.21	1.08	0.96	0.85	0.76	0.67	0.59	0.52	0.45	0.39	0.34	0.29
900000	1.18	1.05	0.93	0.83	0.73	0.64	0.56	0.49	0.43	0.37	0.32	0.27
1000000	1.16	1.02	0.90	0.80	0.70	0.61	0.54	0.46	0.40	0.34	0.29	0.25
1100000	1.13	0.99	0.88	0.77	0.67	0.59	0.51	0.44	0.37	0.32	0.27	0.22
1200000	1.10	0.97	0.85	0.74	0.64	0.56	0.48	0.41	0.35	0.29	0.25	0.20
1300000	1.07	0.94	0.82	0.71	0.62	0.53	0.45	0.38	0.32	0.27	0.22	0.18
1400000	1.04	0.91	0.79	0.68	0.59	0.50	0.43	0.36	0.30	0.25	0.20	0.17
1500000	1.01	0.88	0.76	0.66	0.56	0.48	0.40	0.33	0.28	0.23	0.18	0.15
1600000	0.98	0.85	0.73	0.63	0.53	0.45	0.38	0.31	0.26	0.21	0.17	0.13
1700000	0.96	0.82	0.71	0.60	0.51	0.42	0.35	0.29	0.24	0.19	0.15	0.12
1800000	0.93	0.80	0.68	0.57	0.48	0.40	0.33	0.27	0.22	0.17	0.13	0.10
1900000	0.90	0.77	0.65	0.55	0.45	0.37	0.31	0.25	0.20	0.15	0.12	0.09
2000000	0.87	0.74	0.62	0.52	0.43	0.35	0.28	0.23	0.18	0.14	0.10	0.07

Figure 103: S/P obtenus en fonction du couple (AAD/Priorité)

Nous remarquons que la priorité par rapport au S/P est plus sensible que l'AAD. Cela est logique car la priorité est appliquée au coût total par tête, tandis que l'AAD est appliquée à la charge totale du portefeuille. Actuellement, nous avons un programme de réassurance avec une priorité de 600 000 euros et une AAD fixée à 0. Les résultats du programme nous montrent qu'avec le traité de réassurance interne actuel, nous avons un ratio S/P égal à 0,68. Or, notre objectif est d'atteindre un S/P de 0,75. Pour ce faire, nous pouvons dans un premier temps baisser la priorité de 25 000 euros sans toucher à l'AAD. Ensuite, nous avons deux possibilités. La première consiste à garder une AAD nulle et une priorité à 575 000 euros. Ou bien, nous pouvons encore réduire la priorité, mais cette fois-ci en mettant en place une AAD. Pour décider quelle méthodologie est la mieux adaptée, nous allons analyser l'écart-type associé à ces couples de sensibilité. Le couple ayant l'écart-type le plus élevé sera la stratégie optimale de réassurance (en théorie).

Le graphique suivant illustre les résultats du programme d'optimisation sous forme d'un graphique en trois dimensions.

Optimisation sous contrainte en fonction de la priorité et de l'AAD

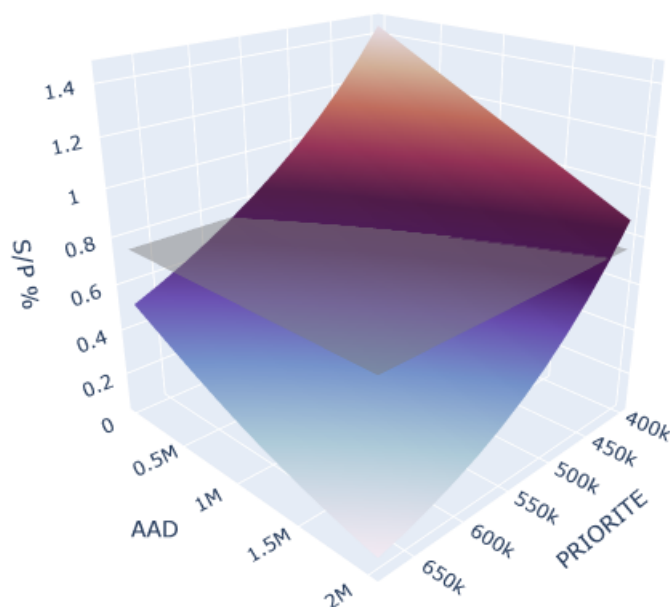


Figure 104: Optimisation de la réassurance

Le plan horizontal représente la contrainte de S/P fixée au préalable. Les valeurs au croisement de la surface et du plan de S/P représentent les couples (AAD, priorité) qui permettent d'obtenir un S/P égal à 0,75. Le graphique illustre notamment la différence entre la pente liée à la priorité et celle de l'AAD. La première est plus raide que la seconde. La suite consiste donc à prendre ces couples candidats et à examiner la volatilité de la charge cédée à la tranche 1.

AAD (k€)	Priorité (k€)
0	575
200	550
500	525
800	500
1200	475
1500	450
1900	425

Table 31: Couples de sensibilité pour un $\frac{S}{P} = 0.75$

Le graphique 105 représente l'écart-type en fonction des couples sélectionnés. Nous remarquons que plus la priorité diminue, plus l'écart-type augmente. De plus, nous constatons une légère bosse à partir d'un AAD de 1 200 000. Cela est dû au fait que l'AAD varie à partir du moment où elle atteint le minimum des charges simulées. Ce phénomène s'observe également sur le graphique 106.

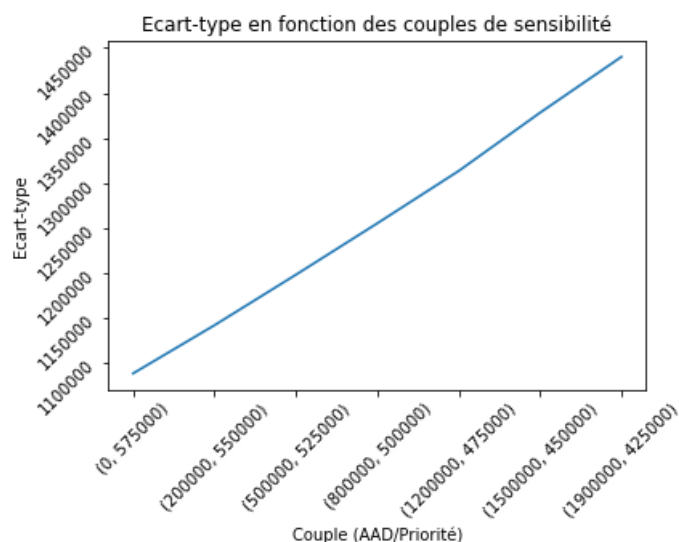


Figure 105: Résultats simulation Tranche 1 par âge et par CSP

Volatilité en fonction de la priorité et de l'AAD

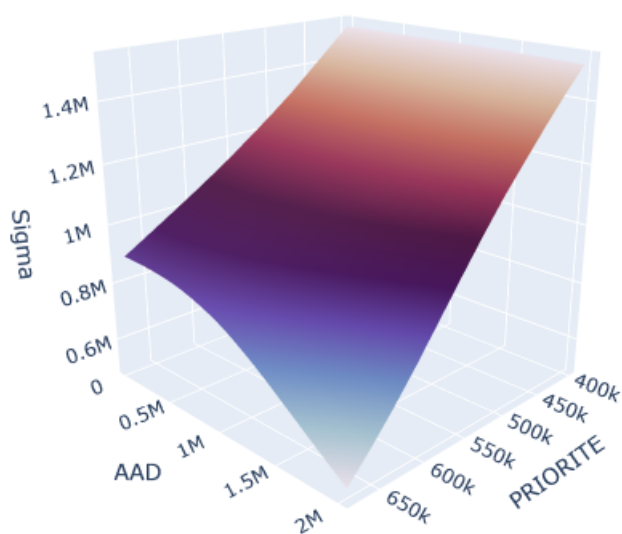


Figure 106: Résultats simulation de la volatilité

En conclusion, pour MACSF, le programme de réassurance le plus adapté en théorie est celui avec la priorité la plus faible couplée à une AAD élevée. Dans la pratique, il n'est pas possible de diminuer la priorité en deçà d'un certain niveau pour tout transférer à une AAD. Afin d'optimiser le traité, il convient dans un premier temps de seulement réduire la priorité afin de maintenir notre objectif d'un ratio S/P égal à 0,75. Dans un second temps, après avoir analysé les effets de ce changement, nous pourrions progressivement réduire la priorité en augmentant l'AAD.

Traité de réassurance externes

Cette partie concerne le traité de réassurance externe. Actuellement, le ratio S/P constaté pour le réassureur est d'environ 25%. Ce montant est relativement bas et signifie que la cédante (la MACSF) paie une prime élevée. L'objectif de cette section est donc de calculer le coefficient de chargement de sécurité appliqué par le réassureur en fonction des différents principes de prime pure afin de s'assurer que le niveau de facturation du coût du risque semble raisonnable.

La figure suivante présente les valeurs des coefficients de chargement de sécurité (ρ) selon le principe de l'espérance, de l'écart-type, ainsi que celui de la variance.

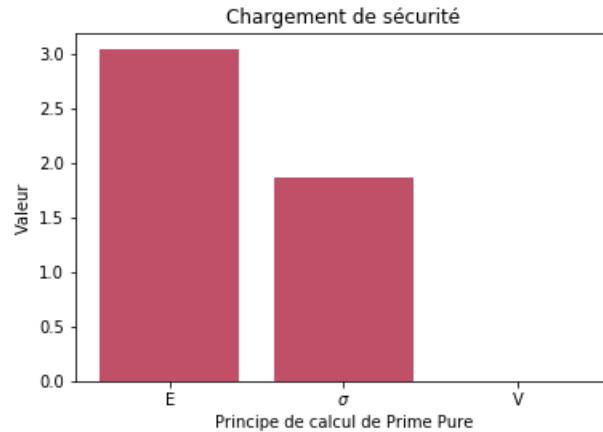


Figure 107: Résultats des coefficients de chargement de sécurité

Concernant le principe du quantile, il est nécessaire de calculer la fonction de répartition empirique des coûts de la charge cédée (voir figure 108). En inversant la formule suivante, $\Pi(X) = F_X^{-1}(\alpha)$, nous trouvons la valeur de α . Cette méthode nous permet donc de déterminer à quel quantile α le réassureur fixe sa tarification du traité de réassurance. Pour le ratio S/P estimé de 25 %, nous estimons un $\alpha = 0.95$. Cela signifie que la prime de la cédante est tarifée par le réassureur au 95ème centile de la distribution de la sinistralité cédée.

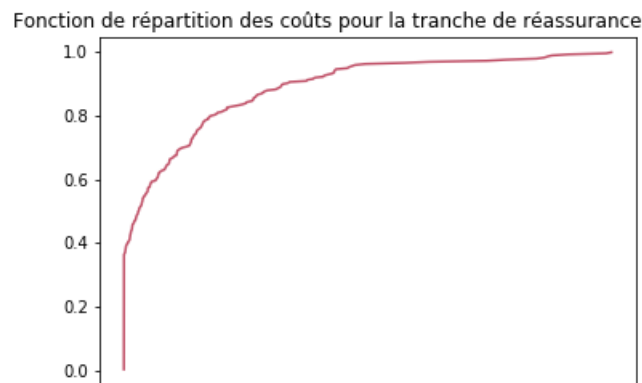


Figure 108: Fonction de répartition de la chargé cédée de la tranche 2

Le coefficient de chargement de sécurité semble donc élevé.

En vue des prochaines négociations du traité de réassurance, nous pouvons estimer la variation des coefficients de chargement de sécurité en fonction de différents niveaux de prime selon

les différentes approches. Le tableau suivant présente les coefficients de chargement de sécurité selon les différentes approches.

S/P	E	σ	Quantile
41%	1,43	0,88	0,87
38%	1,63	1	0,88
35%	1,84	1,13	0,89
33%	2,04	1,25	0,91
31%	2,24	1,37	0,91
29%	2,44	1,5	0,92
27%	2,65	1,62	0,93
26%	2,85	1,75	0,95
25%	3,05	1,87	0,95
24%	3,25	1,99	0,96
22%	3,46	2,12	0,96
21%	3,66	2,24	0,96
21%	3,86	2,37	0,96

Table 32: Coefficient de chargement de sécurité en fonction de l'approche retenue

À titre d'exemple, pour un ratio $\frac{S}{P}$ estimé à 29 %, nous obtenons un coefficient de chargement de sécurité de 2.44 fois l'espérance, soit 1.5 fois l'écart-type, et un quantile de l'ordre de 0.92.

Pour conclure, cette partie nous a permis d'optimiser notre traité de réassurance interne et d'estimer les coefficients de chargement de sécurité associés à différents niveaux de prime pour le traité externe.

Enfin, nous avons fait le choix de maintenir un plafond de réassurance interne fixe dans notre modèle d'optimisation. Une perspective d'évolution serait donc de relâcher cette hypothèse et ainsi d'ajuster la charge cédée entre notre réassureur interne et externe.

5 Conclusion

À travers ce mémoire, nous avons pu élaborer un modèle de tarification pour le produit prévoyance en utilisant différentes techniques statistiques. L'objectif principal était de définir les trois lois de probabilité (taux brut d'incidence, loi de durée de maintien et taux de passage en invalidité) afin de pouvoir simuler le portefeuille de la MACSF et obtenir un coût moyen de sinistralité pour les garanties d'incapacité et d'invalidité par tranche de réassurance.

Pour la construction du taux de passage en incapacité et en invalidité, il a été décidé d'utiliser l'estimateur de Kaplan-Meier avec un lissage de Whittaker-Henderson.

Concernant la modélisation du maintien en incapacité des sinistres et des arrêts, l'analyse des différents critères d'information et des modèles nous a permis de conclure que la loi Log-logistique est la mieux adaptée pour chacune des catégories professionnelles.

Afin de prendre en compte l'effet des rechutes, deux méthodes de modélisation du maintien en incapacité ont été réalisées. L'une portant sur la création d'une loi de durée totale du sinistre et l'autre sur la création d'une loi de durée des arrêts couplée avec une loi de comptage pour les rechutes. Cette dernière permet de ventiler la durée totale d'incapacité en deux composantes. En revanche, elle oblige à faire une hypothèse d'indépendance entre le nombre de rechutes et leur durée. Bien qu'en se basant sur les statistiques descriptives, cette hypothèse semble valide a priori. Les résultats de la simulation montrent que la méthode basée sur la modélisation d'une loi de durée totale du sinistre reflète mieux les coûts historiques que l'approche basée sur la durée des arrêts.

Enfin, le traité de réassurance interne a pu être amélioré grâce à une optimisation sous contraintes. Les résultats ont montré que nous pouvons, dans un premier temps, réduire la priorité afin d'atteindre notre objectif de S/P égal à 0.75. Ensuite, nous avons constaté qu'il est possible de réduire davantage la volatilité en abaissant encore la priorité et en ajoutant une AAD. Pour le traité de réassurance externe, il a été possible de proposer une grille donnant une estimation des chargements de sécurité du réassureur en fonction des différents niveaux de primes proposés. Cette grille pourra être utilisée par la MACSF dans le cadre du renouvellement de ses traités fin 2023.

Limites

Les principales limites auxquelles nous avons fait face concernent la qualité des données. En effet, la reconstitution de la vie de l'ensemble des assurés sur 10 ans a exigé une attention particulière. Enfin, la faible probabilité d'invalidité nécessite un grand nombre de simulations pour obtenir une convergence de la variance dans les tranches de réassurance. Des optimisations du code Python permettraient d'améliorer la convergence de la variance pour les mesures extrêmes.

Perspectives

À travers ce mémoire, nous avons modélisé seulement une partie du modèle présenté en introduction. La perspective finale pour la fonction actuarielle de la MACSF est d'obtenir des lois de durées pour le maintien en invalidité et des taux de transition vers l'état de décès. Ces ajouts permettront d'obtenir une tarification actuarielle complète du produit prévoyance, couvrant les garanties d'incapacité, d'invalidité et de décès.

De plus, un modèle multi-états markovien non-homogène a été modélisé dans le cadre de cette étude. Le modèle n'a pas été présenté dans ce mémoire car, les résultats obtenus n'ont pas été jugés suffisamment robuste. L'avantage de ce modèle serait de mieux prendre en compte

la dépendance entre tous les états.

Pour terminer, il pourrait être aussi intéressant d'ajouter à terme la pathologie responsable du passage dans l'état d'incapacité. Avec suffisamment de données, il est probable que cette variable de segmentation apporte davantage d'informations que d'autres variables de segmentation.

References

- [1] F.Planchet, *Modèles de Durée : Applications actuarielles*, 2006.
- [2] J P.Klein, M L. Moeschberger, *Survival analysis techniques for censored and truncated data*, 2005.
- [3] G.Colletaz, *Econométrie des durées de survie*, 2020.
- [4] F.Helms, *Estimating LTC Premiums using GEEs for Pseudo-Values*, 2003.
- [5] F.Gillaizeau, *Développement et validation de modèles multi-états semi-Markoviens pour le pronostic de patients atteints de maladies chroniques*, 2016.
- [6] K.Lim, E.Liu, *Using the Weibull accelerated failure time regression model to predict time to health events*, 2018.
- [7] A-F Majeed, *Accelerated Failure Time Models : An Application in Insurance Attrition*, 2020.
- [8] A.Allignol, J.Beyersmann, M.Schumacher, *Competing Risks and Multistate Models with R* 2011.
- [9] F.Planchet, *Lignes directrices de la construction des lois de maintien en incapacité et en invalidité*, 2010.
- [10] Q.Guibert, *Sur l'utilisation des modèles multi-états pour la mesure et la gestion des risques d'un contrat d'assurance*, 2016.
- [11] P.Saint-Pierre, *Analyse de survie, modèles multi-états et processus de comptage*, 2021.
- [12] H.Wang, Y.Xue, *An online updating approach for testing the proportional hazards assumption with streams of survival data*, [lien](#) 2021.
- [13] F.Planchet, J.Tomas, *Critères de validation : aspects méthodologiques*, [lien](#)
- [14] P.Grambsch, T.Therneau, *Proportional hazards tests and diagnostics based on weighted residuals*, 1994.
- [15] O.Lopez, *Survival Analysis*, 2023.
- [16] F.Emmanuel, A.Ibrahim, *Econométrie non-paramétrique*, 2008.
- [17] C.Hurlin, *Econométrie et Statistique Non Paramétrique*, 2008.
- [18] E.Duguet, *Statistique et économétrie appliquées des variables de durée*, 2014.
- [19] J.Friedman, T.Hastie, R.Tibshirani, *The Elements of Statistical Learning*, 2009.
- [20] A.Charpentier, *Mathématiques de l'assurance non-vie*, 2004.
- [21] France Assureurs, *Les contrats de prévoyance : tendance à fin juin 2022*, 2022.
- [22] P.Breheny, *Likelihood construction*, 2019. [lien](#)
- [23] PM.Mbow, *Mémoire actuariat : Calibration de lois d'incidence en incapacité*, 2020.

- [24] A.Gaumet, *Mémoire actuariat : Construction de tables d'expérience pour l'entrée et le maintien en incapacité*, 2001.
- [25] A.De La Morinerie, *Mémoire actuariat : Conceptualisation d'un modèle multi-états en arrêt de travail et application à une loi d'incidence en incapacité*, 2016.
- [26] A.Luzon, *Mémoire actuariat : Mise à jour de la loi d'incidence et de maintien d'un contrat dépendance et Impacts sur le provisionnement*, 2019.
- [27] G.Biessy, *Mémoire actuariat : Construction d'un modèle multi-états semi-markovien dans le contexte de l'assurance dépendance*, 2013.

6 Annexe

6.1 Théorie des phénomènes de durée

Démonstration Kaplan-Meier

Pour construire l'estimateur de Kaplan-Meier, nous définissons dans un premier temps les deux quantités suivantes :

$$\begin{aligned}H(t) &= \mathbb{P}(Y \leq t) \\H_1(t) &= \mathbb{P}(Y \leq t, \delta = 1)\end{aligned}$$

que l'on peut réécrire sous la forme suivante :

$$\begin{aligned}1 - H(t) &= \mathbb{P}(Y > t) \\&= \mathbb{P}(\inf(T, C) > t) \\&= \mathbb{P}(T > t, C > t) \\&= \mathbb{P}(T > t)\mathbb{P}(C > t) \\&= (1 - G(t))(1 - F(t))\end{aligned}$$

$$\begin{aligned}H_1(t) &= \mathbb{E}[\delta \mathbf{1}_{T \leq t}] \\&= \mathbb{E}[\mathbb{E}[\mathbf{1}_{T \leq C} \mathbf{1}_{T \leq t} | T]] \\&= \mathbb{E}[(1 - G(t)) \mathbf{1}_{T \leq t}] \\&= \int_{-\infty}^t (1 - G(u)) dF(u)\end{aligned}$$

L'étape suivante consiste à calculer le ratio des deux quantités ci-dessus ce qui nous permet d'obtenir l'estimateur de Kaplan-Meier. Soit :

$$\hat{S}(t) = \prod_{T_i \leq t} \left(1 - \hat{\lambda}(t)\right) = \prod_{Y_i \leq t} \left(1 - \frac{d\hat{F}(t)}{1 - \hat{F}(t)}\right) = \prod_{T_i \leq t} \left(1 - \frac{d\hat{H}_1(t)}{1 - \hat{H}(t)}\right) = \prod_{T_i \leq t} \left(1 - \frac{\delta_i}{Y_i}\right)$$

Démonstration de la vraisemblance

Nous allons faire une brève démonstration de la fonction de vraisemblance. Nous rappelons par ailleurs que nous observons $(Y_1, \delta_1), \dots, (Y_n, \delta_n)$ tel que :

$$\begin{aligned}Y_i &= \inf(T_i, C_i) \\ \delta_i &= \mathbf{1}_{T_i \leq C_i} \\ (Y_i, \delta_i) &\text{ a pour loi } (Y, \delta) \text{ sachant } Y \geq \tau\end{aligned}$$

Pour le cas non censuré (données complètes) ($\delta_i = 1$):

$$\begin{aligned}
f(t_i) &= \frac{f(t_i)}{(1 - S(C^{(r)}))} \cdot (1 - S(C^{(r)})) \\
&= \mathbb{P}(C^{(r)} = Y_i | T_i \leq C^{(r)}) \cdot \mathbb{P}(T_i \leq C^{(r)}) \\
&= \mathbb{P}(Y_i = C^{(r)} | \delta_i = 1) \cdot \mathbb{P}(\delta_i = 1) \\
&= \mathbb{P}(Y_i, \delta_i = 1)
\end{aligned}$$

Pour le cas censuré (données incomplètes) ($\delta_i = 0$) :

$$\begin{aligned}
S(C^{(r)}) &= \mathbb{P}(T_i > C^{(r)}) = \mathbb{P}(\delta_i = 0) \\
&= \mathbb{P}(\delta_i = 0) \cdot \mathbb{P}(Y_i = C^{(r)} | \delta_i = 0) \\
&= \mathbb{P}(Y_i, \delta_i = 0)
\end{aligned}$$

En combinant les deux termes, nous obtenons l'expression suivante :

$$\mathbb{P}(Y_i, \delta_i) = f(t_i)^{\delta_i} \cdot S(t_i)^{1-\delta_i}$$

Ce qui conduit à trouver la vraisemblance de la forme suivante pour des données censurées à droite et tronquées à gauche :

$$L = \prod_{i=1}^n \frac{\mathbb{P}(t_i, \delta_i)}{S(\tau)} = \prod_{i=1}^n \left(\frac{f(t_i)}{S(\tau)} \right)^{\delta_i} \cdot \left(\frac{S(t_i)}{S(\tau)} \right)^{1-\delta_i}$$

Démonstration modèle AFT²⁴

$$\begin{aligned}
S(T|X) &= \mathbb{P}(T > t|X) \\
&= \mathbb{P}(e^{\beta^T X} T_0 > t|X) \\
&= \mathbb{P}(\log(T_0) > \log(t) - \beta^T X|X)
\end{aligned}$$

A partir de là, nous pouvons en déduire deux résultats.

$$\begin{aligned}
&= \mathbb{P}(\log(T_0) > \log(t) - \beta^T X|X) \\
&= \mathbb{P}(T_0 > te^{-\beta^T X}|X) \\
&= S_{T_0}(te^{-\beta^T X})
\end{aligned}$$

et

²⁴Cette démonstration s'appuie sur le papier de Abdul-Fatawu Majeed 2020 [7]

$$\begin{aligned}
&= \mathbb{P}(\log(T_0) > \log(t) - \beta^T X | X) \\
&= \mathbb{P}(\beta_0 + \sigma\epsilon > \log(t) - \beta^T X | X) \\
&= \mathbb{P}(\epsilon > \frac{1}{\sigma}(\log(t) - (\beta^T X + \beta_0)) | X) \\
&= \mathbb{P}(\epsilon > \log(\frac{t}{e^{\beta^T X + \beta_0}})^{\frac{1}{\sigma}} | X) \\
&= S_\epsilon(\log(\frac{t}{e^{\beta^T X + \beta_0}})^{\frac{1}{\sigma}}) \\
&= S_\epsilon(\psi)
\end{aligned}$$

avec $\psi = \log(\frac{t}{e^{\beta^T X + \beta_0}})^{\frac{1}{\sigma}}$

Cette démonstration nous permet d'obtenir la densité et la fonction de survie de $T|X$, soit

:

$$\begin{aligned}
f(t|X) &= f_0(e^{\beta^T X} t) e^{\beta^T X} \\
S(t|X) &= S_0(e^{\beta^T X} t)
\end{aligned}$$

6.2 Calibration sinistres

Passage en incapacité

1 : Sinistre supérieur à la franchise

Taux d'incidence lissés pour les incapacités supérieures à la franchise

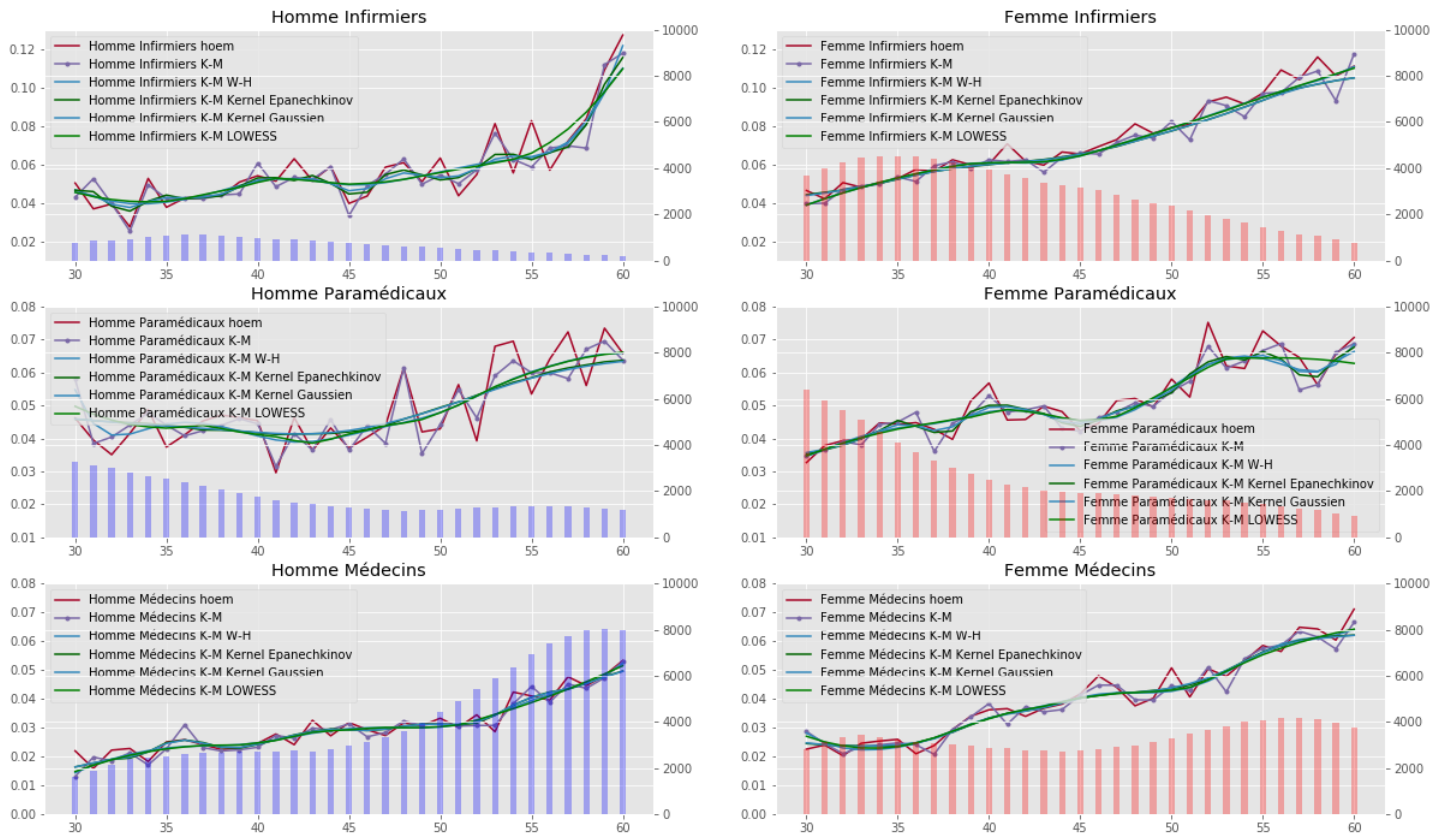


Figure 109: Taux d'incidence bruts des sinistres supérieurs à la franchise avec le lissage LOWESS

1 : Sinistre inférieur à la franchise

Taux d'incidence bruts pour les incapacités inférieures à la franchise

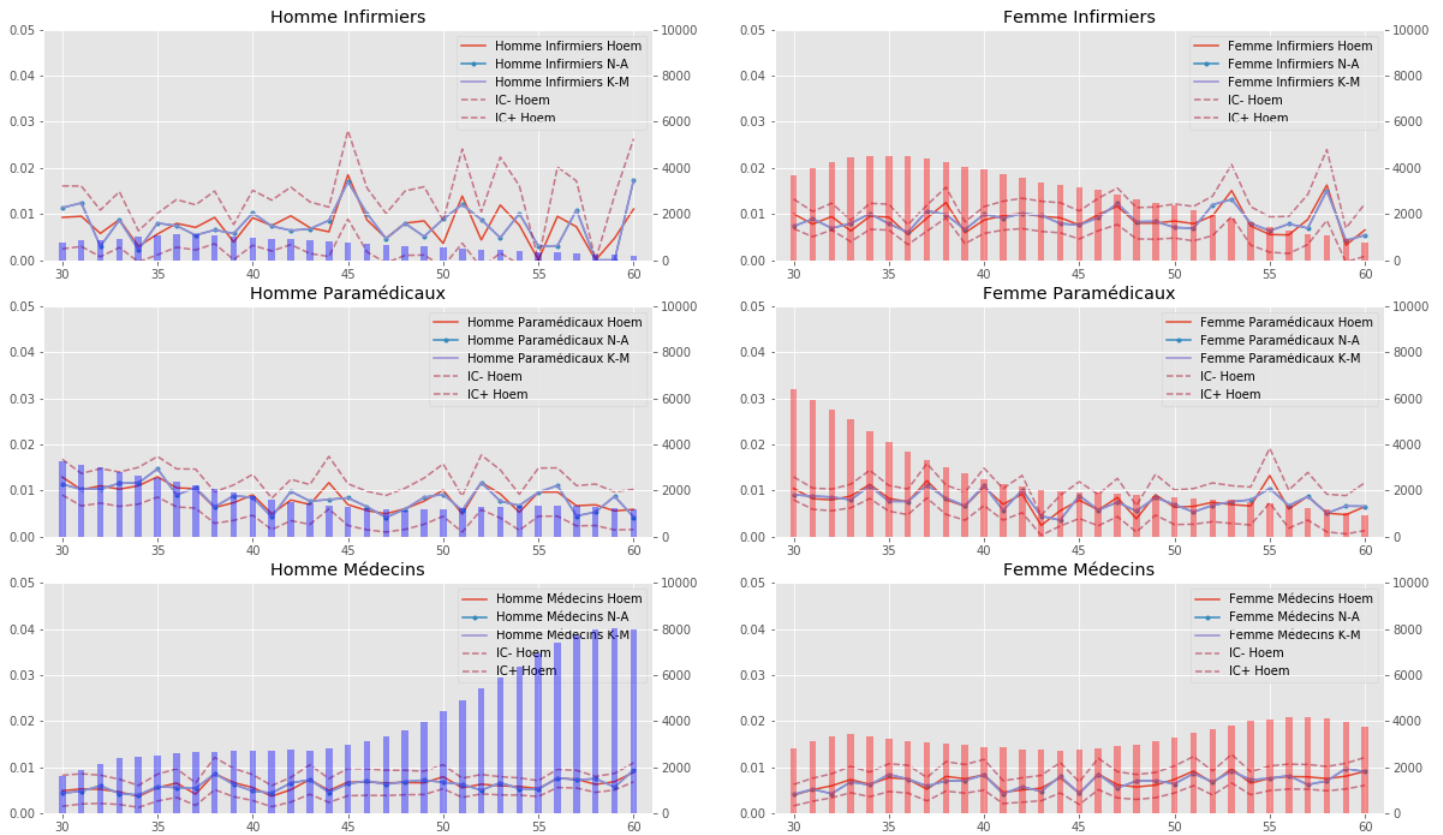


Figure 110: Taux d'incidence bruts des sinistres inférieurs à la franchise

Sexe	Catégories	$h_{Gaussien}^*$	$h_{Epanech}^*$	$h_{Uniforme}^*$	z^*	h^*
Homme	Infirmiers	7	7	6.5	4	15
	Paramédicaux	3.1	4.7	3.5	4	10
	Médecins	5.3	5.3	4.5	4	20
Femme	Infirmiers	7	7	6.4	4	15
	Paramédicaux	4.9	7	7	4	10
	Médecins	2.5	4.9	4.5	4	15

Table 33: Résultats $h_{optimal}$ pour les sinistres inférieurs à la franchise

Taux d'incidence lissés pour les incapacités inférieures à la franchise

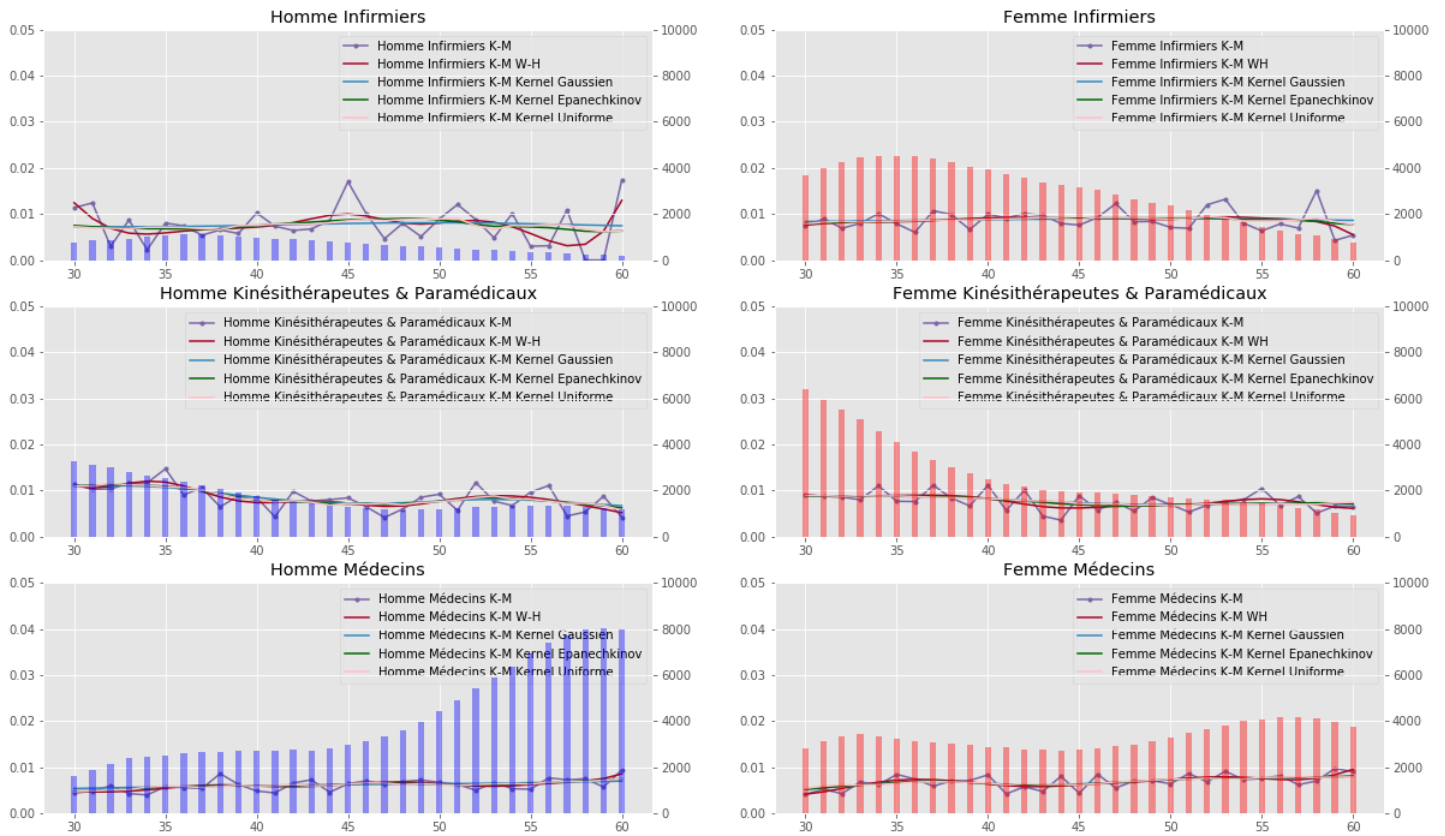


Figure 111: Taux d'incidence bruts lissés des sinistres inférieurs à la franchise

2. Sinistre pour cause de grossesse

Sexe	Catégories	$h_{Gaussien}^*$	$h_{Epanech}^*$	$h_{Uniforme}^*$	z^*	h^*
Homme	Infirmiers	2.6	7	6	4	15
	Paramédicaux	1.1	1.2	4.3	4	10
	Médecins	1.1	2.3	1.4	4	20
Femme	Infirmiers	2	4.7	3.2	4	15
	Paramédicaux	3.2	5.9	4.3	4	10
	Médecins	1.2	2.5	1.3	4	15

Table 34: Résultats $h_{optimal}$ pour les sinistres grossesse

Taux d'incidence bruts pour les incapacités dues à une grossesse

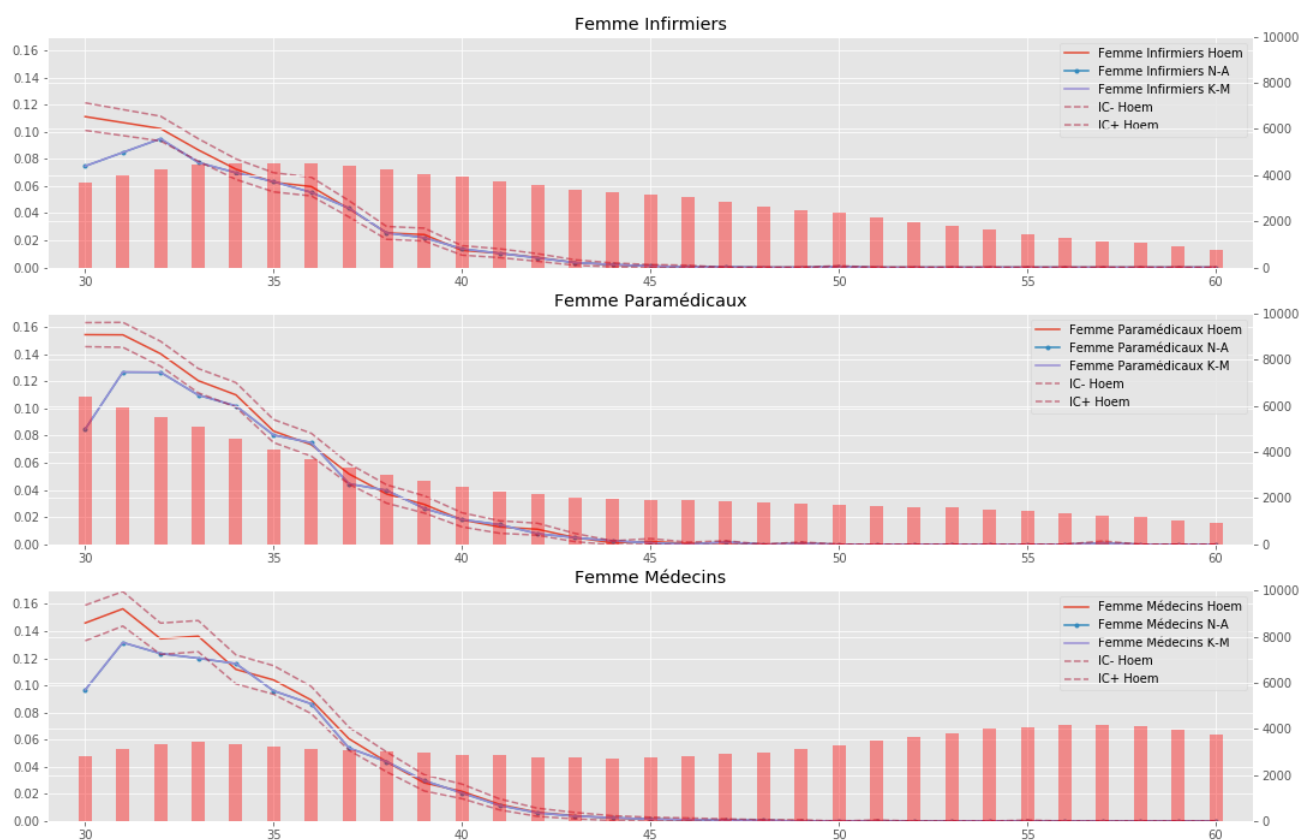


Figure 112: Taux d'incidence bruts des sinistres inférieurs à la franchise

Taux d'incidence lissés pour les incapacités dues à une grossesse

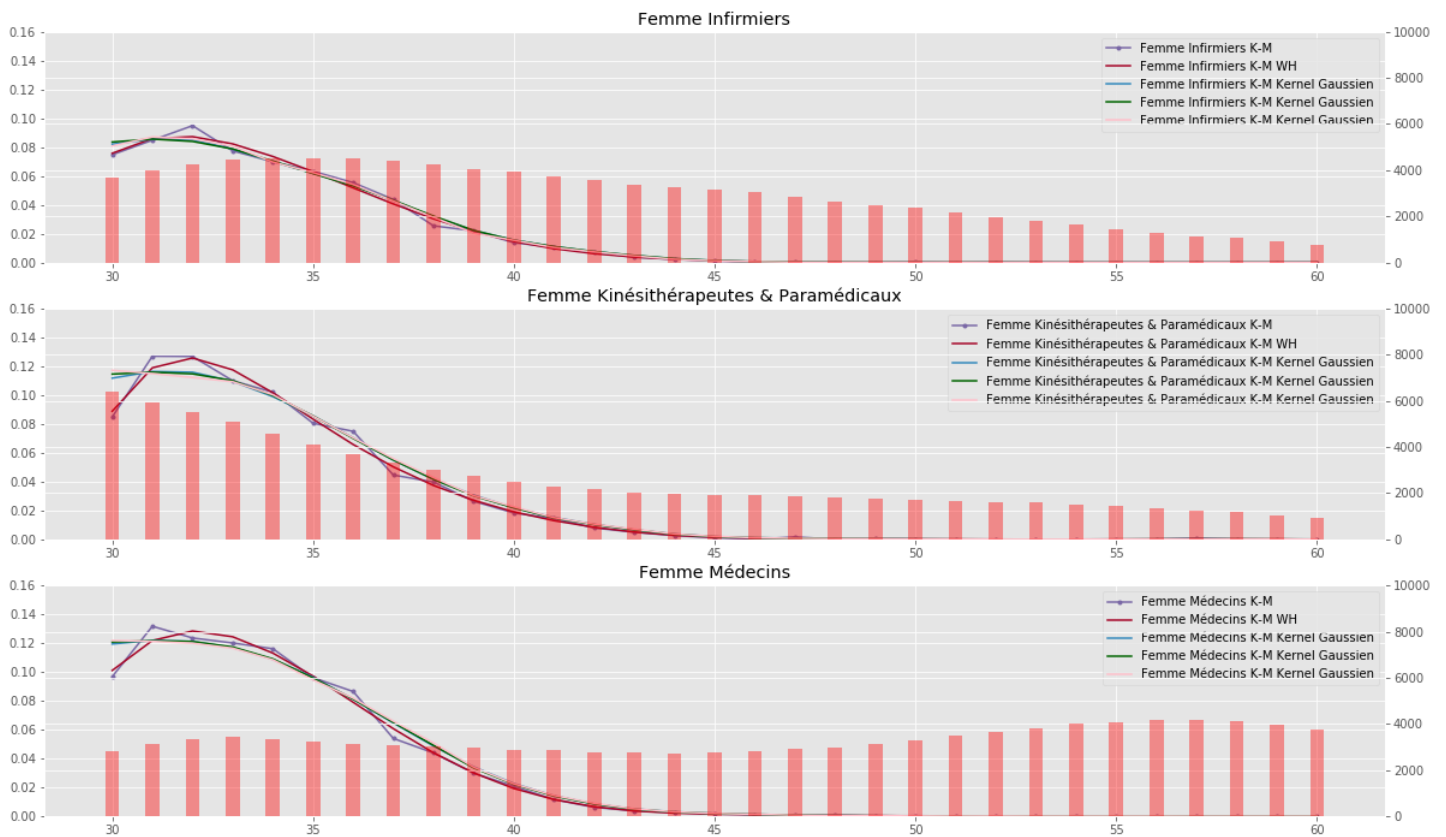


Figure 113: Taux d'incidence bruts lissés des sinistres grossesse

3. : Comparaison entre le modèle Cox et l'estimateur de Kaplan-Meier

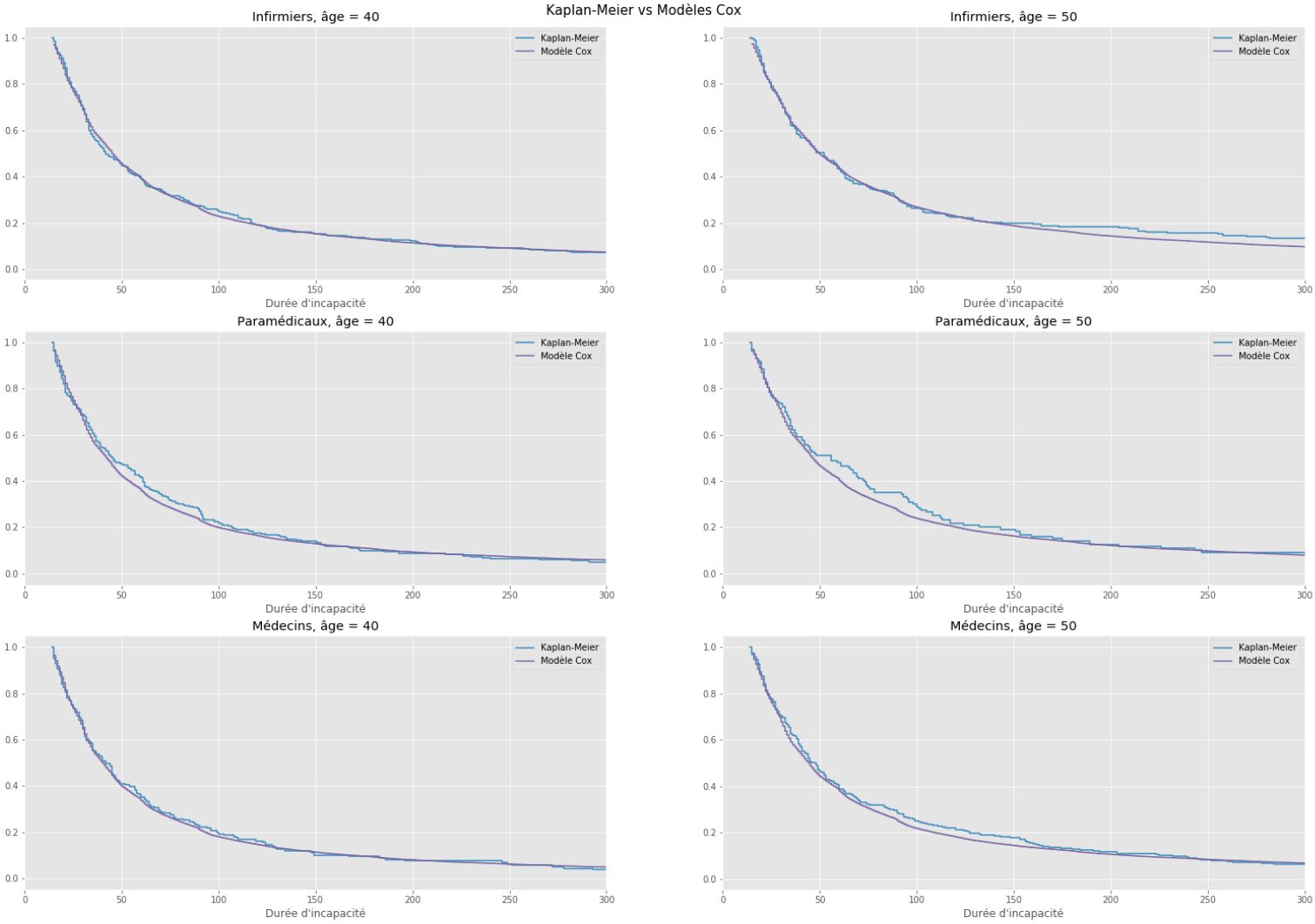


Figure 114: Comparaison des fonctions de survie de maintien par Kaplan-Meier et par le modèle Cox

3. : Comparaison entre les lois marginales et le modèle AFT

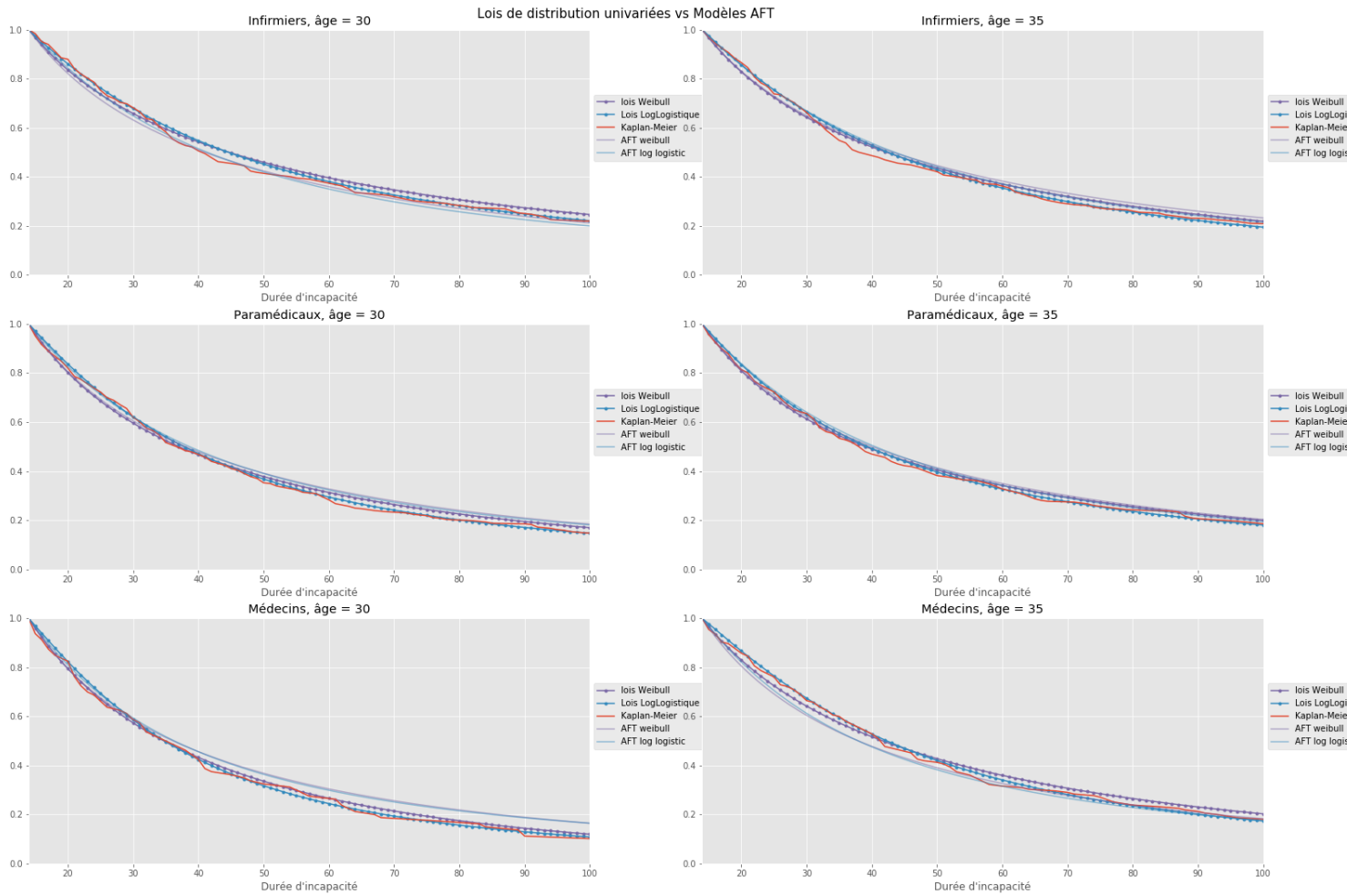


Figure 115: Comparaison lois marginales et modèle AFT, âge 30 ans vs 35 ans

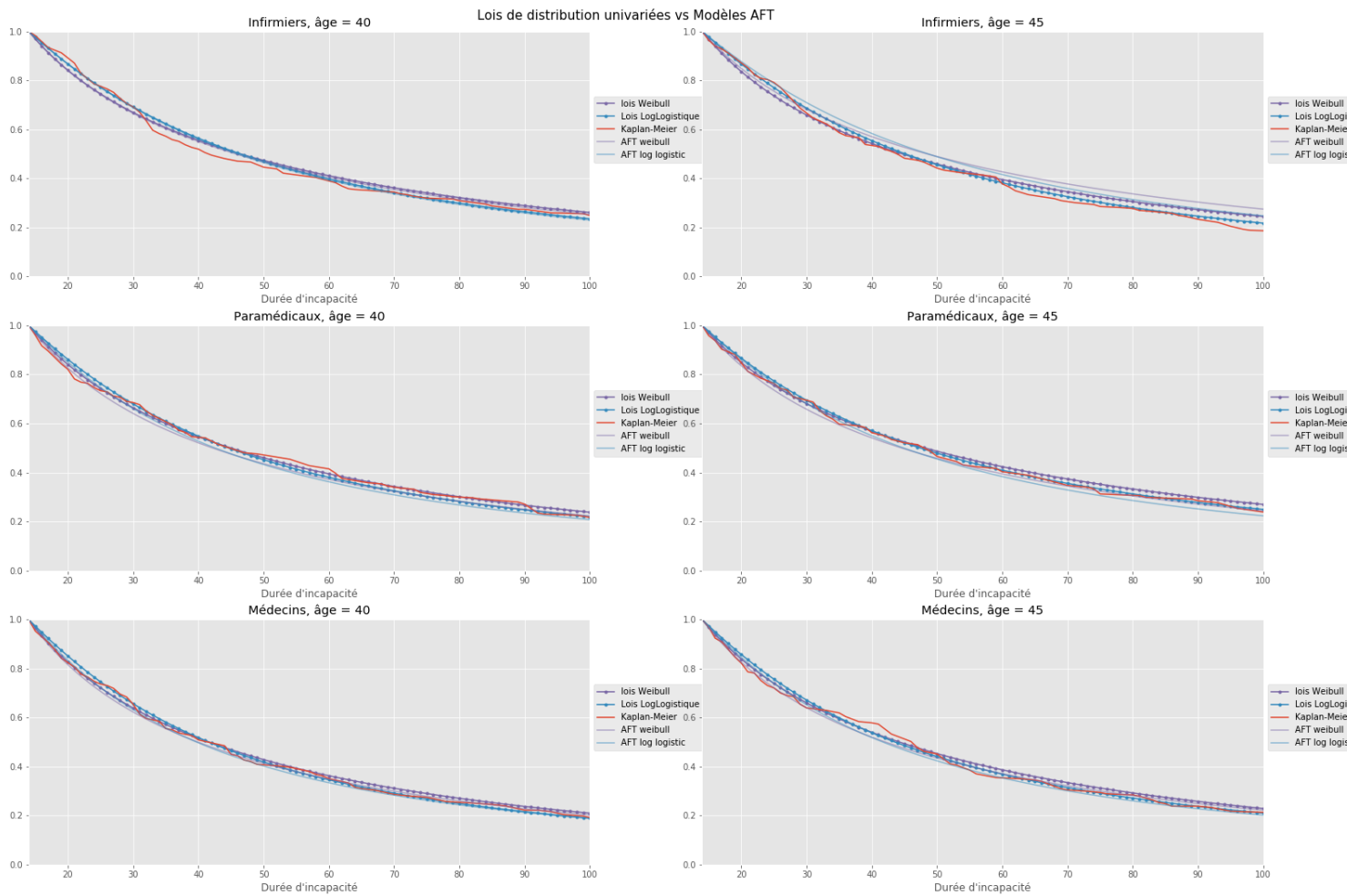


Figure 116: Comparaison lois marginales et modèle AFT, âge 40 ans vs 45 ans

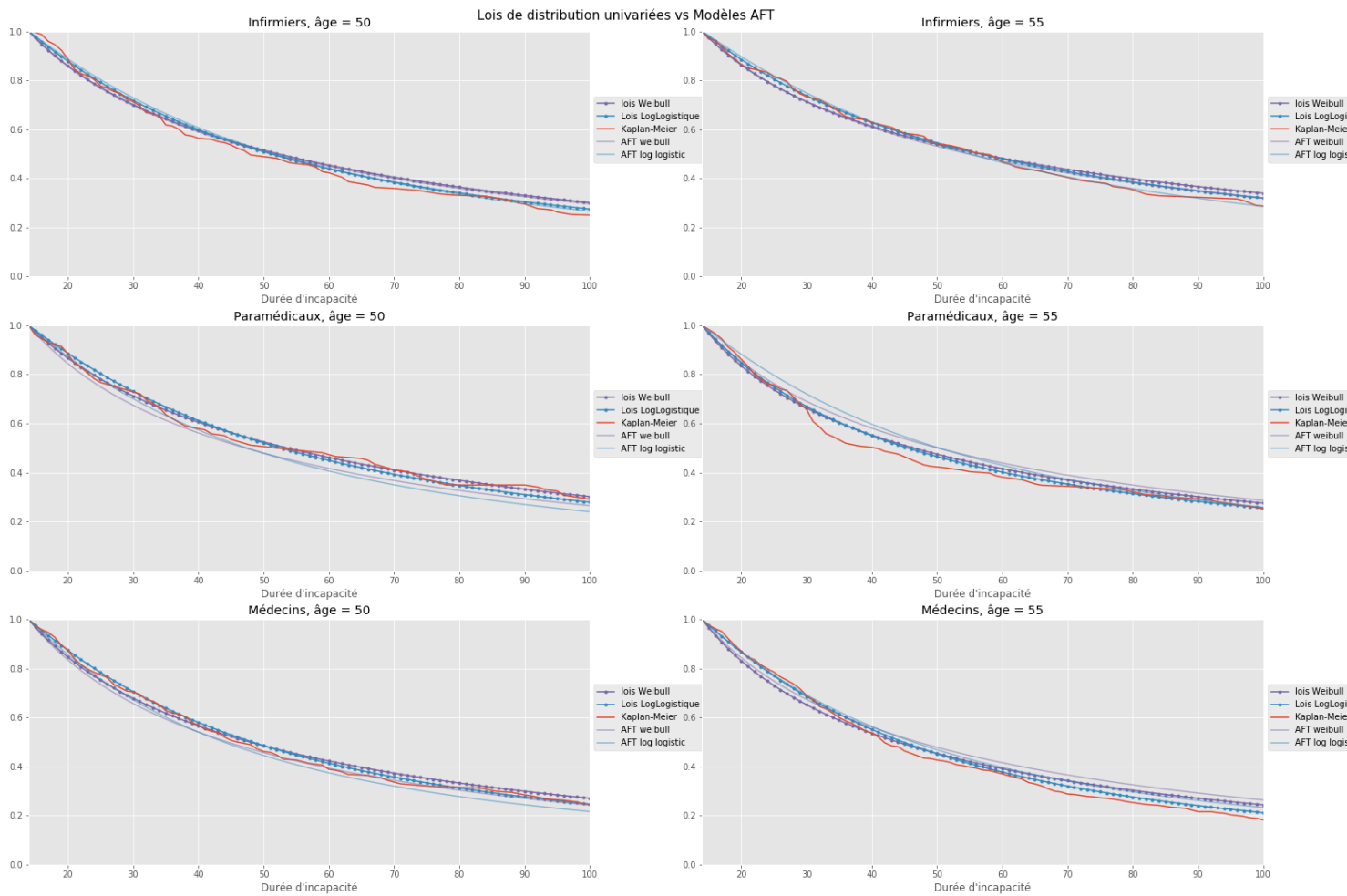


Figure 117: Comparaison lois marginales et modèle AFT, âge 50 ans vs 55 ans

6.3 Calibration arrêts

Maintien en incapacité incapacité

3 : Lois paramétrique

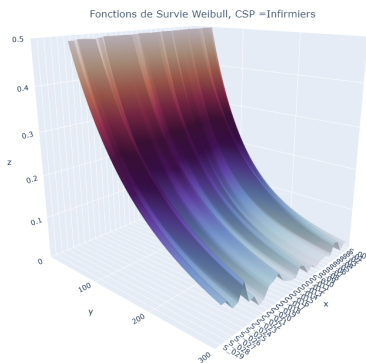


Figure 118: Weibull Infirmiers

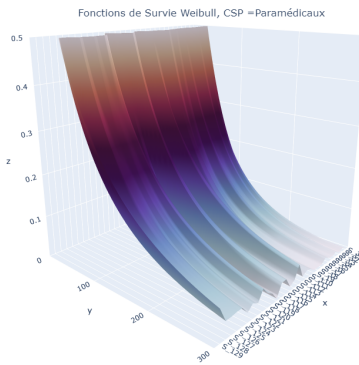


Figure 121: Weibull Kinésithérapeutes

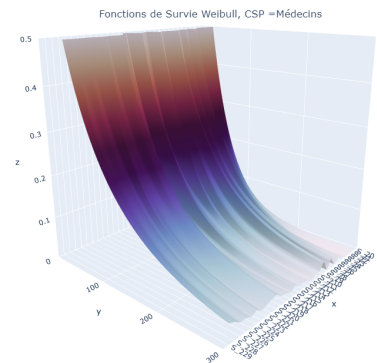


Figure 124: Weibull Médecins

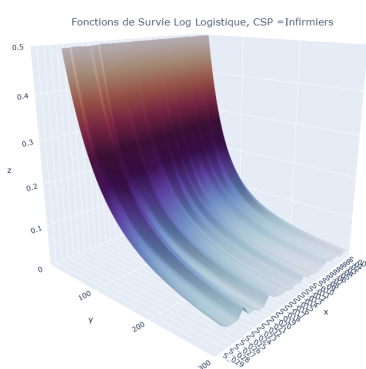


Figure 119: Log Logistique Infirmiers

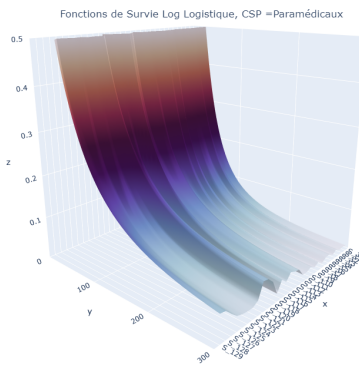


Figure 122: Log Logistique Kinésithérapeutes

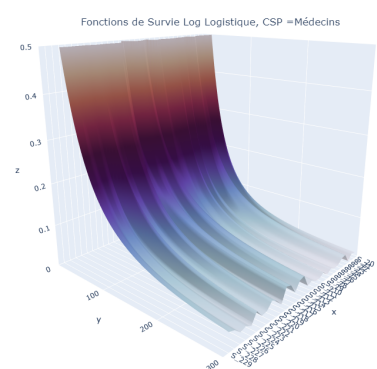


Figure 125: Log Logistique Médecins

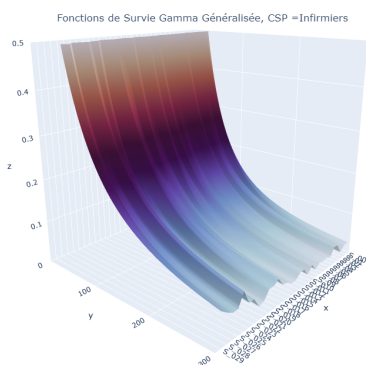


Figure 120: Gamma Généralisée Infirmiers

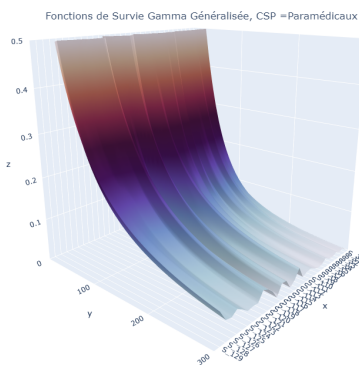


Figure 123: Gamma Généralisée Kinésithérapeutes

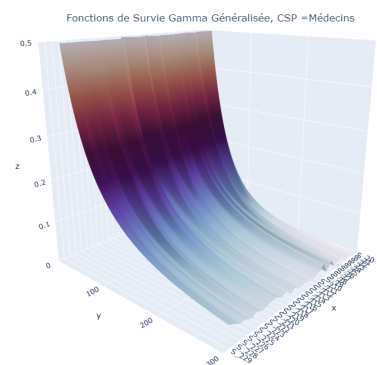


Figure 126: Gamma Généralisée Médecins

Lissage loi de Weibull et Log logistique

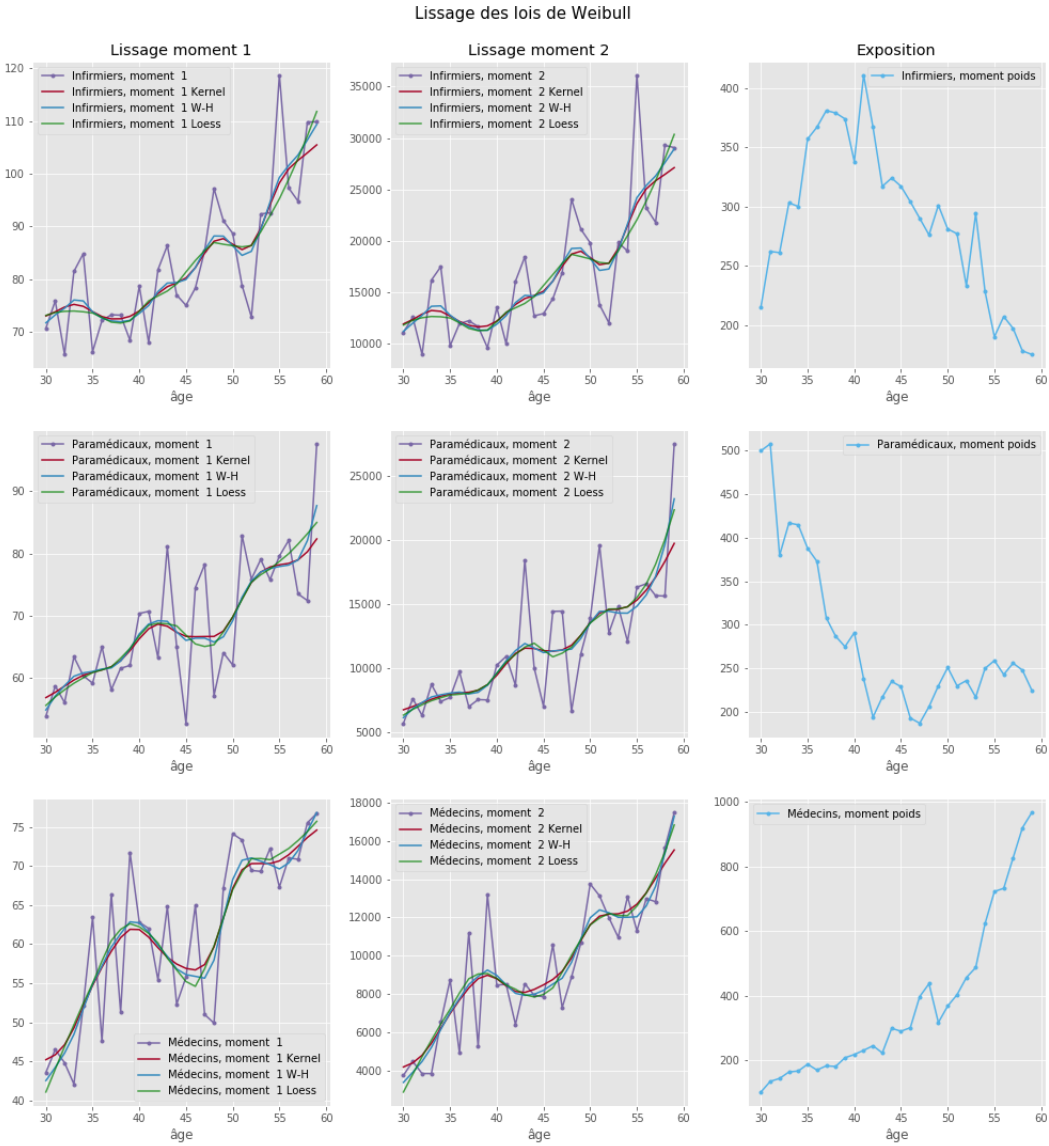


Figure 127: Lissage des moments de la loi de Weibull

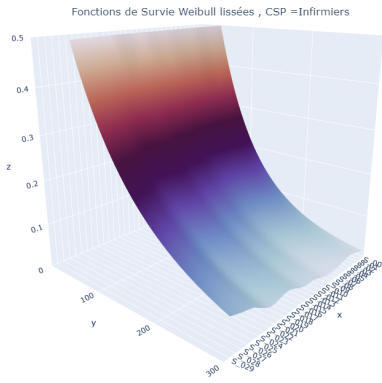


Figure 128: Weibull lissées Infirmiers

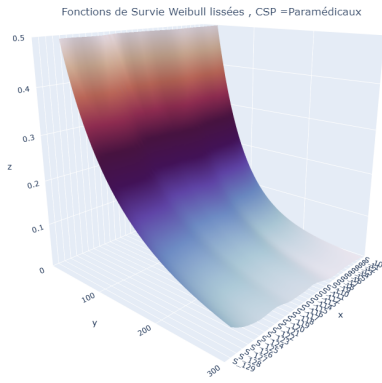


Figure 129: Weibull lissées Paramédicaux

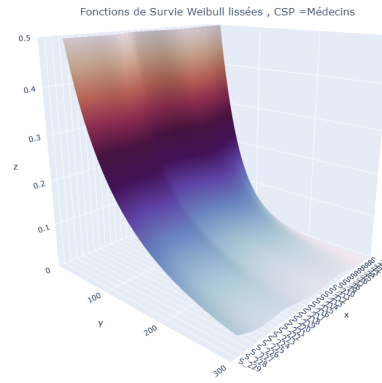


Figure 130: Weibull lissées Médecins

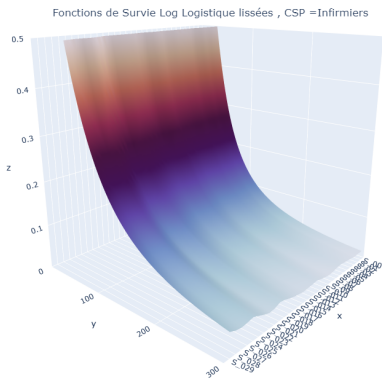


Figure 131: Log Logistique lissées Infirmiers

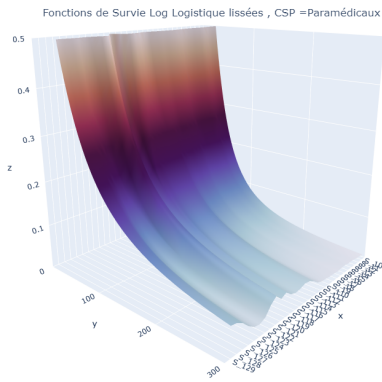


Figure 132: Log Logistique lissées Paramédicaux

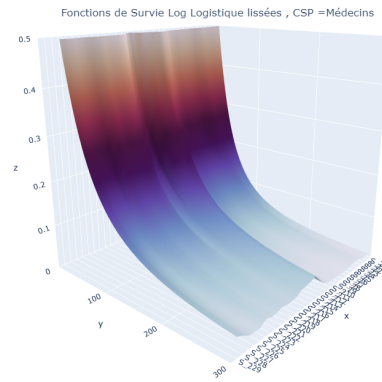


Figure 133: Log Logistique lissées Médecins

Critères d'informations AIC/BIC



Figure 134: AIC et BIC pour les infirmiers

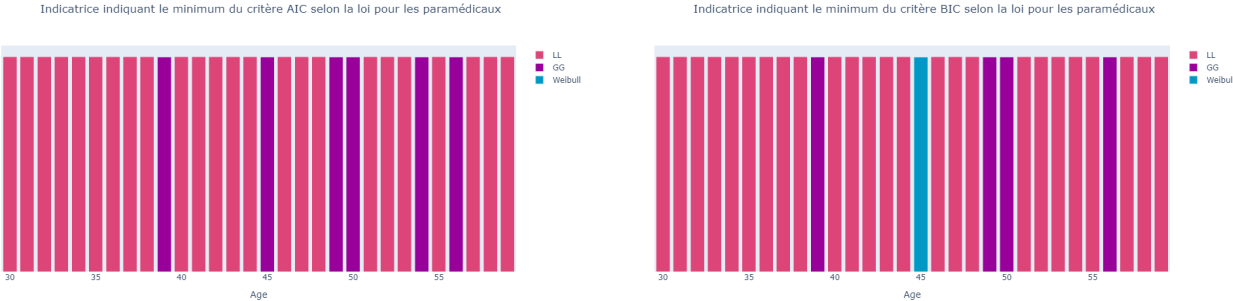


Figure 135: AIC et BIC pour les paramédicaux



Figure 136: AIC et BIC pour les médecins

Comparaison entre le modèle Cox et l'estimateur de Kaplan-Meier

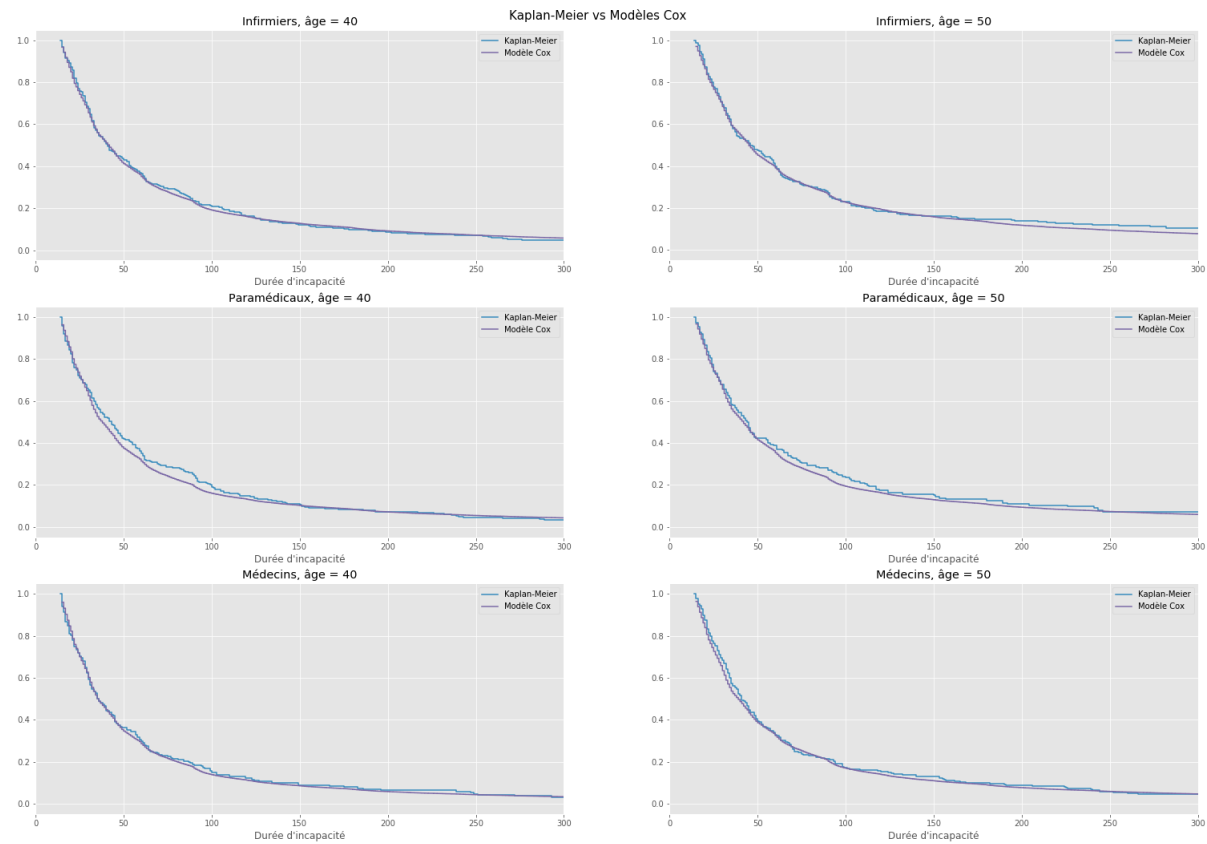


Figure 137: Modèle Cox vs estimateur Kaplan-Meier, BDD arrêts

Modélisation de la rechute des arrêts

Variables	Estimation	Std.Error	z	value
(Intercept)	-1.96680	0.12664	-15.531	<2e-16
factor(Grp_calibration)3	0.30390	0.05323	5.710	1.13e-08
factor(Grp_calibration)4	0.48248	0.05086	9.486	<2e-16
factor(age_maintien)31	0.48248	0.05086	9.486	<2e-16
factor(age_maintien)32	0.05490	0.16720	0.328	0.74265
factor(age_maintien)33	-0.05044	0.17380	-0.290	0.77165
factor(age_maintien)34	0.08654	0.16619	0.521	0.60255
factor(age_maintien)35	0.10003	0.16628	0.602	0.54747
factor(age_maintien)36	0.27846	0.16504	1.687	0.09156
factor(age_maintien)37	0.01451	0.17146	0.085	0.93258
factor(age_maintien)38	0.16715	0.16850	0.992	0.32119
factor(age_maintien)39	0.05143	0.17185	0.299	0.76473
factor(age_maintien)40	0.11077	0.16979	0.652	0.51415
factor(age_maintien)41	0.25577	0.16945	1.509	0.1312
factor(age_maintien)42	0.45499	0.16915	2.690	0.00715
factor(age_maintien)43	0.16287	0.17550	0.928	0.35338
factor(age_maintien)44	0.78429	0.16631	4.716	2.41e-06
factor(age_maintien)45	0.31380	0.17381	1.805	0.07102
factor(age_maintien)46	0.92867	0.16493	5.631	1.79e-08
factor(age_maintien)47	0.41731	0.17021	2.452	0.01422
factor(age_maintien)48	0.46092	0.16900	2.727	0.00639
factor(age_maintien)49	0.64764	0.16636	3.893	9.90e-05
factor(age_maintien)50	0.24828	0.17098	1.452	0.14648
factor(age_maintien)51	0.47598	0.16453	2.893	0.00382
factor(age_maintien)52	0.58246	0.16258	3.583	0.00034
factor(age_maintien)53	0.41434	0.16118	2.571	0.01015
factor(age_maintien)54	0.43149	0.15961	2.703	0.00686
factor(age_maintien)55	0.28832	0.16028	1.799	0.07203
factor(age_maintien)56	0.49493	0.15751	3.142	0.00168
factor(age_maintien)57	0.50043	0.15408	3.248	0.00116
factor(age_maintien)58	0.87977	0.15271	5.761	8.36e-09
factor(age_maintien)59	0.45706	0.15664	2.918	0.00352
factor(age_maintien)60	0.61892	0.15368	4.027	5.64e-05

Table 35: Résultats totaux de la régression Binomiale Négative

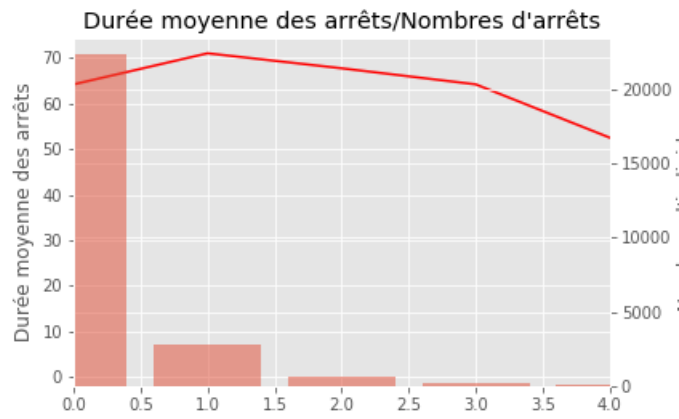


Figure 138: Durée des arrêts moyenne vs nombre d'arrêts

6.4 Lissage

Dans cette partie, nous allons présenter une démonstration plus théorique de l'estimateur Naraya-Watson.

L'objectif est donc de trouver un estimateur de l'espérance conditionnelle.

$$\mathbb{E}[Y|X = x] = m(x) = \int_{-\infty}^{+\infty} yf(y|x)dy \quad (95)$$

Nous allons développer ce résultat en terme de densités non-conditionnelles en utilisant le théorème de bayes à l'aide des densités jointes et marginales.

$$\mathbb{E}[Y|X = x] = \int_{-\infty}^{+\infty} \frac{yf(y, x)}{f(x)}dy = \frac{\int_{-\infty}^{+\infty} yf(y, x)dy}{\int_{-\infty}^{+\infty} f(y, x)dy} \quad (96)$$

A présent, pour estimer la densité jointe située au numérateur et dénominateur de la formule ci-dessus, nous allons utiliser l'estimateur de deux noyaux univariés.

$$\hat{f}(x, y) = \frac{1}{nh^2} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right) K\left(\frac{y - y_i}{h}\right) \quad (97)$$

En utilisant cet estimateur pour le dénominateur, nous obtenons,

$$\int_{-\infty}^{+\infty} \hat{f}(x, y)dy = \frac{1}{nh^2} \sum_{i=1}^n \left[K\left(\frac{x - x_i}{h}\right) \right] \underbrace{\left[\int_{-\infty}^{+\infty} K\left(\frac{y - y_i}{h}\right) dy \right]}_{=1} \quad (98)$$

Comme la fonction noyaux possède les mêmes propriétés qu'une densité de probabilité, à savoir, que l'intégrale est égale à 1, alors le terme à droite est égal à 1.

Pour le numérateur, nous trouvons le résultat suivant :

$$\int_{-\infty}^{+\infty} y\hat{f}(x, y)dy = \frac{1}{nh^2} \sum_{i=1}^n \left[K\left(\frac{x - x_i}{h}\right) \right] \underbrace{\left[\int_{-\infty}^{+\infty} yK\left(\frac{y - y_i}{h}\right) dy \right]}_{=y_i} \quad (99)$$

Le terme à droite est cette fois-ci égal à l'espérance de la fonction noyau. Or, nous savons que cette fonction est symétrique et centrée en y_i , donc l'espérance est égale à y_i .

Nous obtenons donc l'estimateur suivant :

$$\hat{m}(x) = \frac{\int_{-\infty}^{+\infty} yf(y, x)dy}{\int_{-\infty}^{+\infty} f(y, x)dy} = \frac{\sum_{i=1}^n \left[K\left(\frac{x - x_i}{h}\right) \right] y_i}{\sum_{i=1}^n \left[K\left(\frac{x - x_i}{h}\right) \right]} \quad (100)$$

Cet estimateur correspond exactement à l'estimateur de Nadaraya-Watson.

Remarque : Nous pouvons remarquer qu'au niveau du dénominateur, nous avons un estimateur de $f(x)$. L'estimation sera donc imprécise lorsque x sera proche de 0. Cet estimateur sera donc utilisable seulement pour des valeurs de x avec une densités suffisamment élevée. En conclusion, cet estimateur n'est pas utilisable pour faire de la prévision.

Lissage LOESS/LOWESS

La régression Loess permet de mieux tenir compte des valeurs extrêmes par rapport à l'approche Nadaraya-Watson.

L'idée reste similaire : nous cherchons une fonction de lien, notée $m(x)$, qui établit la relation entre les variables y et x .

$$\begin{aligned}y &= m(x) + \epsilon \\ \mathbb{E}[y|x] &= m(x)\end{aligned}$$

La différence entre la méthode à kernel et cette approche est que la fonction $m(x)$ est cette fois-ci le résultat d'une droite de régression et non d'une moyenne empirique des données. Par exemple, pour une valeur au point x_0 , nous allons chercher une fonction $m(x_0)$ qui sera évaluée par une fonction paramétrique évaluée localement au voisinage de $N(x_0)$.

On estimera donc la fonction suivante :

$$\hat{m}(x_0) = \hat{a}(x_0) + \hat{b}(x_0)x_0 \quad (101)$$

où les estimateurs \hat{a} et \hat{b} sont les résultats du programme de minimisation suivant :

$$\text{Min}_{(a(x_0), b(x_0))} \sum_{x_i \in N(x_0)} (y_i - a(x_0) - b(x_0)x_i)^2 \quad (102)$$

Il existe des variantes où il est possible d'attribuer des poids aux données. La version sans poids correspond à la régression Loess (LOcal rEGRESSion), tandis que celle avec des poids correspond à la régression Lowess (LOcally WEighted Scatterplot Smoothing).

En ce qui concerne la régression Lowess, le programme de minimisation est le suivant :

$$\text{Min}_{(a(x_0), b(x_0))} \sum_{x_i \in N(x_0)} (y_i - a(x_0) - b(x_0)x_i)^2 K\left(\frac{x_i - x_0}{\lambda}\right) \quad (103)$$

Avec λ un paramètre de lissage.

6.5 Réassurance

Dans cette section de l'annexe, nous détaillons littérairement les étapes de l'algorithme de simulation.

1. Générer NS variables aléatoires selon une loi de Bernoulli, de paramètre q_x :
2. Pour chaque variable aléatoire égale à 1, générer une variable aléatoire selon la loi de maintien :
 - Loi de durée totale : Générer une variable aléatoire supérieure à 14
 - Loi de durée des arrêts :
 - (a) Générer un nombre n de rechutes
 - (b) Générer n réalisations de la loi de durée des arrêts. Si la somme des réalisations à 14, alors simuler à nouveau des réalisations.
3. Pour chaque variable aléatoire de durée supérieure à 0, générer une variable aléatoire selon une loi de Bernoulli en utilisant le taux d'incidence du passage en invalidité associé à l'individu.
4. Pour chaque simulation, appliquer tous les contrats (Incapacité et Invalidité) et les garanties associées.
5. Calculer la somme totale des coûts pour chaque simulation.
6. Appliquer les traités de réassurance individuelle pour chaque simulation.
7. Calculer la moyenne des coûts par simulation et par tranche de réassurance.
8. La moyenne de chaque tranche correspond à une tarification spécifique.

L'exemple ci-dessous représente un cas d'application avec la loi de maintien totale.

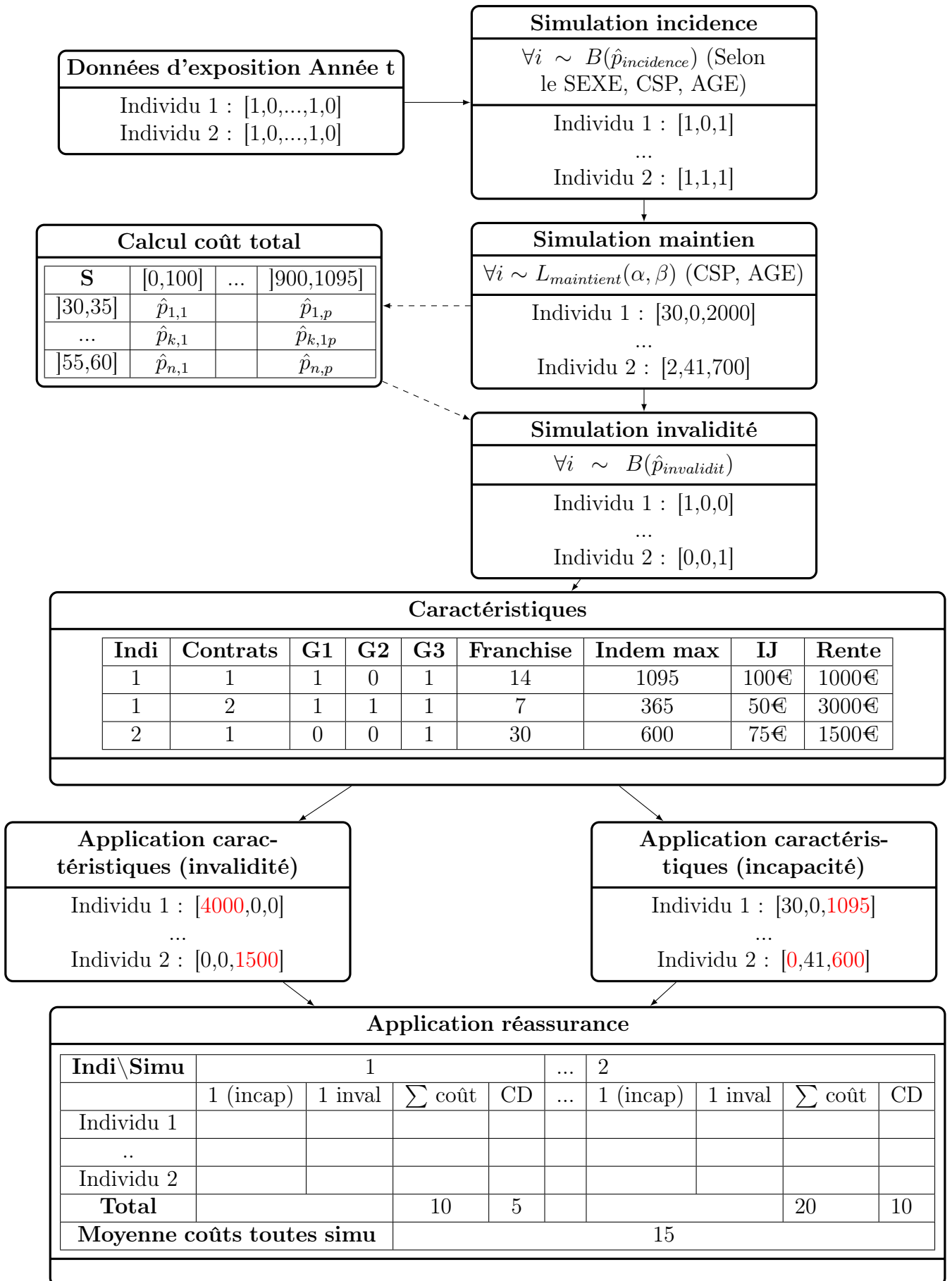


Figure 139: Exemple illustratif de la simulation du portefeuille