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

Par : Monsieur Alonso Erwan

Titre du mémoire : Utilisation des modèles additifs généralisés pour la construction de table d'expérience en arrêt de travail

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

Membres présents du jury de la filière : Signature :	Entreprise : Nom : Guillaume BIESSY Signature : 
	Directeur de mémoire en entreprise Nom : Guillaume BIESSY Signature : 
Membres présents du jury de l'Institut des Actuaires : Signature :	Invité : Nom : Signature :
	Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels (après expiration de l'éventuel délai de confidentialité) Signature du responsable entreprise :  Signature du candidat : 

Remerciements

Je souhaite exprimer mes sincères remerciements à Michel Morcos El Douaihy, fondateur de LinkPact, pour m'avoir offert l'opportunité précieuse de réaliser mon mémoire au sein de son entreprise. Sa confiance et son soutien ont été des éléments déterminants dans cette expérience significative.

Un merci particulier à Guillaume Biessy, responsable R&D chez LinkPact, pour son encadrement attentif et la transmission généreuse de ses connaissances. Son dévouement a grandement enrichi la qualité de ce travail.

Je tiens à exprimer ma sincère gratitude envers notre partenaire pour la mise à disposition des données et pour son soutien précieux et sa collaboration étroite tout au long de cette étude. Leur disponibilité et leur engagement ont contribué à l'amélioration continue de la qualité des données et ont été essentiels pour répondre à mes questions.

Je tiens également à exprimer ma reconnaissance envers Olivier Lopez, ancien directeur de l'ISUP, pour ses réponses éclairées à mes questions, ainsi qu'à Maud Thomas, directrice de l'ISUP, pour ses conseils avisés qui ont été précieux dans l'élaboration de ce mémoire.

Un immense merci à ma famille et à mes proches pour leur soutien indéfectible tout au long de ce parcours. Leur encouragement et leur compréhension ont été les piliers de cette aventure académique.

Finalement, un soutien tout particulier à ma compagne Aurélie pour sa présence constante et son soutien inconditionnel. Son encouragement sans faille et sa patience m'ont été d'un réconfort inestimable et ont joué un rôle clé dans la réussite de ce projet.

Résumé

Le risque d'incapacité et d'invalidité est principalement évalué à travers les tables règlementaires du BCAC. Cependant, ces tables, établies en 1993 sur la base des portefeuilles d'assurance représentant la population française à cette époque, ne reflètent plus fidèlement les profils de risque actuels des assurés. En effet, les caractéristiques des assurés ont changé au fil du temps, ce qui fait que la population assurée aujourd'hui diffère considérablement de celle de 1993. Par conséquent, une mise à jour des méthodes d'évaluation du risque est nécessaire pour mieux prendre en compte les tendances actuelles en matière d'incapacité et d'invalidité. Ainsi, il est impératif de construire de nouvelles tables d'expérience refletant plus fidèlement le risque porté par les assureurs.

Ce mémoire commence par analyser les limitations des approches conventionnelles et propose une démarche novatrice reposant sur les modèles additifs généralisés. Cette approche vise à élaborer des tables d'expérience aussi bien pour l'incapacité que pour l'invalidité. L'originalité de cette méthode réside dans sa capacité à créer des solutions lissées, résultat optimale entre précision et robustesse, qui s'approchent davantage du phénomène étudié. Elle prend également en compte des facteurs tels que la censure et la troncature présentes dans les données. De plus, cette méthode offre la flexibilité de tenir compte de tables préexistantes et d'intégrer une ou plusieurs variables explicatives, ce qui permet d'aboutir à une modélisation plus fine du risque.

En conclusion, les résultats obtenus à partir des tables d'expérience sont directement appliqués dans le calcul des provisions liées au risque d'incapacité/invalidité. Cette approche se traduit par une réduction significative d'environ 20% des provisions, démontrant ainsi l'importance de la création de table d'expérience.

Mots clés : table d'expérience en incapacité, estimateur de Kaplan-Meier, estimateur de Turnbull, impact de la franchise, GAM, provisions prévoyance, modélisation du Risque Incapacité/Invalidité

Abstract

The assessment of short-term and long-term disability risks is primarily conducted using the regulatory tables of BCAC. However, these tables, designed in 1993 based on the French population, are now inadequate to faithfully reflect the risks borne by insurers. In this context, it becomes imperative to implement a more accurate and adaptable approach to better apprehend these risks.

This thesis begins by analyzing the limitations of conventional approaches and proposes an innovative approach based on generalized additive models. This approach aims to develop experience tables for both disability and invalidity. The uniqueness of this method lies in its ability to create smoothed solutions that come closer to the studied phenomenon. It also takes into account factors such as censorship and truncation present in the data. Furthermore, this method offers the flexibility to consider pre-existing tables and integrate one or more explanatory variables, resulting in a finer risk modeling.

Ultimately, the resulting tables are then utilized for calculating provisions related to disability/invalidity risk. For the calculation of these provisions, a dedicated section is devoted to formulating these formulas in modern and relevant notations.

Keywords: Disability Experience Table, Kaplan-Meier Estimator, Turnbull Estimator, Franchise Impact, GAM, Insurance Reserves, Disability/Invalidity Risk Modeling

Synthèse

Contexte réglementaire et théorique

Une nécessité fondamentale pour tout assureur consiste à évaluer le risque qu'il assume avec la plus grande précision possible. Dans le domaine de la prévoyance, cela implique d'estimer les fonctions de survie pour obtenir des tables de maintien en cas d'incapacité ou d'invalidité. À l'heure actuelle, les seules tables réglementaires disponibles ont été élaborées par le BCAC en 1993. Il est donc très probable que ces estimations ne reflètent plus fidèlement le risque tel qu'il se présente aujourd'hui. Dans ce contexte, il devient essentiel de créer des tables d'expérience afin de mettre à jour l'évaluation de ces risques.

La construction d'une table d'expérience nécessite de relever plusieurs défis, les premiers étant liés à la censure et à la troncature. En effet, ces phénomènes inhérents à l'étude des durées de survie introduisent des biais significatifs lorsque leur gestion n'est pas correcte. Ensuite, se pose la problématique de la franchise, un élément crucial à prendre en compte, car cette information affecte directement une variable importante dans l'estimation de la fonction de survie : le nombre de personnes à risque à un moment donné. Dans le but d'obtenir une estimation la plus précise possible, il est crucial d'intégrer l'apport d'informations provenant de variables explicatives additionnelles. Finalement, le dernier défi réside dans le phénomène de fluctuation d'échantillonnage. En effet nous savons que la probabilité de rester dans un état, que ce soit d'incapacité ou d'invalidité, évolue régulièrement selon l'âge ou l'ancienneté, ainsi il est souhaitable que l'évolution des taux de sortie du modèle suivent une évolution régulière, que ce soit pour l'âge ou pour l'ancienneté.

Pour répondre à ces défis, différents estimateurs existent, notamment les estimateurs non paramétriques de Kaplan-Meier et de Turnbull, ainsi que l'estimateur semi-paramétrique de Cox. Les deux premiers tiennent compte de la censure et de la troncature, et la franchise peut également être intégrée dans ces modèles. Cependant, ils présentent des limites en ce qui concerne l'intégration de variables explicatives et la régularité des résultats obtenus. Pour ces modèles, la prise en compte de variables explicatives ne peut se faire qu'en calibrant le modèle pour chaque modalité ou combinaison de modalités des variables, ce qui implique de diviser à chaque fois les données, augmentant ainsi l'incertitude sur les résultats. Concernant l'irrégularité il est possible de procéder à un lissage des taux obtenus.

Pour aborder la problématique des variables explicatives de manière plus robuste, l'estimateur semi-paramétrique de Cox apparaît comme une solution. Bien que ce modèle offre des solutions à tous les défis énoncés, il présente néanmoins une restriction : l'hypothèse d'indépendance de l'effet des variables dans le temps. Bien que théoriquement cette contrainte puisse être

contournée, en pratique, cela se traduit par une complexité algorithmique excessive qui rend l'exécution des calculs peu réaliste.

Face à ces enjeux, ce mémoire explore une approche novatrice dans l'estimation des tables d'expérience, à savoir les modèles additifs généralisés (GAM). Ces modèles semblent fournir une solution satisfaisante en offrant des tables régulières tout en intégrant des variables explicatives, le tout avec une mise en œuvre relativement simple. Un autre atout de ces modèles est leur capacité à intégrer des tables préexistantes pour calibrer un modèle relationnel entre la table et les données.

Traitement des données

Afin d'utiliser un modèle GAM Poisson, nous devons agréger nos données afin d'obtenir, pour chaque combinaison d'âge, d'ancienneté, et toute combinaison de modalités de variable qui nous intéresse :

- Le nombre de sorties observées
- L'exposition

La prise en compte de la censure, de la troncature, ainsi que de la franchise, se fait par le calcul de l'exposition. Nous nous sommes intéressés à la variable explicative correspondant au type de contrat. Cette variable est significative et correspond à 2 populations : les contrats collectifs et les travailleurs non salariés. Nous serons donc en mesure d'obtenir une table pour les contrats collectifs, ainsi qu'une table pour les travailleurs non salariés en calibrant un seul modèle.

La franchise n'étant pas présente explicitement dans nos données, un travail de recherche a été effectué pour retrouver cette information dans les données. Cela a abouti à la Figure 1. Les franchises retrouvées sont cohérentes et sont donc utilisées afin de déterminer avec précision la période d'observation de chaque individu pour laquelle une franchise a pu être clairement identifiée. De plus, afin de mesurer l'impact de la prise en compte de la franchise, nous avons constitué 2 bases de données supplémentaires à partir des sinistres dont nous avons désormais la connaissance de la franchise. La première base prend effectivement en compte la franchise, tandis que l'autre la considère nulle. Ainsi, pour chacune de ces bases, nous calibrons un modèle, dont les résultats seront comparés par la suite.

Résultats des modèles

Après avoir correctement formaté nos données, nous avons calibré le modèle GAM suivant à l'aide du package R `mgcv` :

$$\log(\mathbb{E}(\text{Sortie})) = \beta_0 + f_1(\text{Age}) + f_2(\text{Anc}) + f_3(\text{Age,Anc})$$

Avec :

- β_0 : représente l'effet constant, indépendamment des autres variables sur la variable Sortie.

Figure 1. : Histogramme des franchises pour les sinistres ayant donné lieu à une prestation en 2021 ou 2022

- f_1 : représente l'effet de l'âge sur la variable Sortie.
- f_2 : représente l'effet de l'ancienneté sur la variable Sortie.
- f_3 : représente l'effet de l'interaction entre l'âge et l'ancienneté sur la variable Sortie.

Ce modèle a été calibré en prenant en compte différents scénarios : en considérant la franchise, en la considérant nulle, et en incluant également le type de contrat en plus de la franchise. Il est important de souligner que l'incorporation du type de contrat s'effectue de manière extrêmement simple grâce au package que nous avons utilisé. Ce modèle peut être formulé comme suit :

$$\log(\mathbb{E}(\text{Sortie})) = \beta_0 + f_1(\text{Age}) + f_2(\text{Anc}) + f_3(\text{Age,Anc}) + \text{Type}$$

Dans ce modèle, nous utilisons une population de référence, à savoir les contrats collectifs, et ajustons un coefficient pour les contrats travailleurs non salariés (TNS).

De plus, nous avons illustré l'intégration d'une table préexistante dans notre modèle. Pour ce faire, nous avons utilisé la table BCAC de 2010 comme référence et l'avons intégrée à notre modèle comme offset, modifiant ainsi le modèle GAM précédent de la manière suivante :

$$\begin{aligned}\log(\text{Sortie}) &= log(\alpha * \mu_{\text{BCAC}} * e) = \log(\alpha) + \log(\mu_{\text{BCAC}}) + \log(e) \\ &= \beta_1 + f_4(\text{Age}) + f_5(\text{Anc}) + f_6(\text{Age,Anc})\end{aligned}$$

En considérant $\log(\mu_{\text{BCAC}}) + \log(e)$ comme un offset, la partie modélisée n'est plus le taux de sortie mais le coefficient de passage *alpha*, ce qui place la contrainte de lissage sur ces coefficients de passage. Pour obtenir la table de maintien de ce modèle, nous devons multiplier le coefficient de passage du modèle par le taux correspondant de la table BCAC. Cependant,

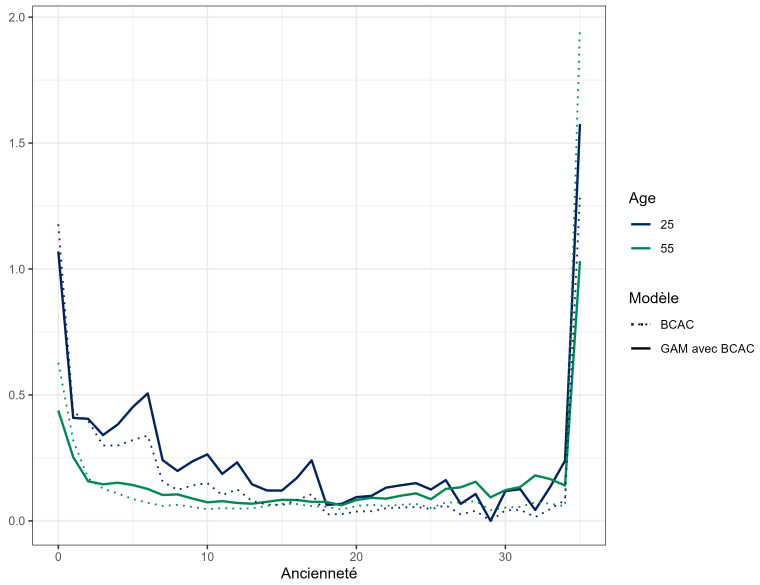


Figure 2. : Comparaison de la régularité des taux de sorties avec table du BCAC en offset

comme la table BCAC de 2010 n'est pas lisse, les taux de sortie ainsi obtenus ne le sont pas non plus. Cela est illustré à la Figure 2.

Cette observation nous a conduits à ne pas retenir ce modèle. Néanmoins, nous avons considéré qu'il s'agissait d'une possibilité intéressante et qu'il était pertinent de la mettre en avant. Notamment, avec une table d'entrée lisse, il s'agit d'une option tout à fait exploitable pour obtenir des taux de sortie lisses.

Après avoir calibré nos modèles GAM, nous les avons confrontés aux modèles suivants :

- Turnbull
- Kaplan-Meier
- Lissage de Whittaker-Henderson des données brutes
- Table réglementaire du BCAC

Les différentes comparaisons ont mis en évidence que le modèle GAM est le plus performant, en termes de régularité de la solution ainsi que de précision. Une mesure de performance que nous avons utilisée pour nos modèles GAM est l'AIC. Nous avons démontré que la prise en compte de la franchise et du type de contrat améliore significativement cette mesure, conduisant ainsi à identifier le modèle GAM utilisant le type de contrat et intégrant la franchise comme le meilleur.

Impact sur les provisions

Enfin, il restait à calculer les provisions pour l'incapacté en cours ainsi que l'invalidité en attente. Dans cette optique, nous avons en premier lieu établi les formules pour toutes les provisions liées aux risques d'incapacité, d'invalidité et de décès. Par la suite, nous avons présenté les diverses incidences de nos modèles sur ces provisions.

Nous avons constaté que l'utilisation de notre meilleur modèle entraîne une baisse significative de 20 % des provisions pour l'incapacité en cours et de 35 % pour l'invalidité en attente. De plus, nous avons relevé que l'impact de la prise en compte de la franchise est inférieur à 3 % pour ces deux types de provisions. Cette observation s'explique par le fait que l'incidence de la franchise intervient uniquement pour les périodes d'ancienneté inférieures aux durées de franchise. Étant donné que ces durées de franchise sont relativement courtes par rapport à la plage d'ancienneté possible, seuls les sinistres dont l'ancienneté est inférieure à 4 mois à la date de calcul sont touchés par cette incidence. Étant donné que ces sinistres ne représentent pas la majorité de notre portefeuille, l'effet de la prise en compte de la franchise sur les provisions est resté limité.

Conclusion

Nous pouvons donc conclure que l'utilisation des GAM dans la modélisation du risque incapacité/invalidité semble prometteuse. Elle répond de manière satisfaisante aux différents problèmes posés par la création de tables d'expérience, que ce soit en termes de cohérence des taux de sortie, de lissage et de précision. Le fait qu'il soit simple de prendre en compte des variables explicatives supplémentaires constitue un véritable atout pour ce modèle.

Cependant, il serait tout à fait possible de pousser la modélisation à un niveau plus élevé en utilisant des données sur l'invalidité, notamment les transitions vers l'invalidité. De telles données permettraient, en utilisant les GAM, de modéliser des taux de sortie par causes de sortie. Par exemple, pour l'incapacité, il serait possible de modéliser les sorties pour cause de reprise, les sorties liées au passage en invalidité et les sorties pour cause de décès. Cette modélisation approfondie, qui nécessite davantage de volume de données, permettrait d'obtenir directement différentes tables, telles que des tables de transitions vers l'invalidité et des tables de décès en incapacité. Un avantage significatif serait que ces tables seraient cohérentes entre elles. En d'autres termes, les taux de sorties globaux seraient égaux à la somme des taux de sorties pour chacune des causes de sortie possibles que nous avons citées pour l'incapacité.

Summary

Regulatory and Theoretical Context

One of the fundamental concepts for any insurer is to assess the risk it assumes as accurately as possible. In the field of insurance, this involves estimating survival functions to obtain incapacity maintenance table in cases of disability or invalidity maintenance table in cases of invalidity. Currently, the only available regulatory tables were developed by BCAC in 1993. It's highly likely that these estimates no longer accurately reflect the risk as it stands today. In this context, it becomes essential to create experience tables to update the evaluation of these risks.

Constructing an experience table involves overcoming several challenges, with the first ones being related to censorship and truncation. These inherent phenomena in survival duration studies introduce significant biases when not managed correctly. Next, the issue of franchise arises, a crucial element to consider, as this information directly affects a vital variable in survival function estimation: the number of individuals at risk at a given time. To achieve the most accurate estimation, it's crucial to incorporate information from additional explanatory variables. Lastly, the challenge lies in the phenomenon of sampling fluctuation. We know that the probability of staying in a state, whether it's disability or invalidity, evolves regularly based on age or seniority. Hence, it's desirable that the evolution of the model's exit rates follows a consistent pattern, be it with respect to age or seniority.

To address these challenges, various estimators exist, including non-parametric estimators like Kaplan-Meier and Turnbull, as well as the semi-parametric Cox estimator. The first two consider censorship and truncation, and franchise can also be integrated into these models. However, they have limitations regarding incorporating explanatory variables and the regularity of obtained results. For these models, considering explanatory variables can only be done by calibrating the model for each variable's modality or combination, which involves dividing the data each time, thus increasing result uncertainty. Regarding irregularity, it's possible to smooth the obtained rates.

To tackle the issue of explanatory variables more robustly, the semi-parametric Cox estimator emerges as a solution. Although this model provides solutions to all mentioned challenges, it comes with a restriction: the assumption of independence of variable effects over time. While theoretically, this constraint can be circumvented, in practice, it translates to excessive algorithmic complexity that renders calculation execution unrealistic.

Given these challenges, this dissertation explores an innovative approach in experience table estimation, namely Generalized Additive Models (GAM). These models seem to offer a sa-

tisfactory solution by providing regular tables while incorporating explanatory variables, all with a relatively simple implementation. Another advantage of these models is their ability to integrate pre-existing tables to calibrate a relational model between the table and data.

Data Processing

In order to use a Poisson GAM model, we need to aggregate our data to obtain, for each combination of age, seniority, and any combination of variable modalities of interest:

- The number of observed exits
- The exposure

Taking into account censoring, truncation, as well as franchise, is achieved through exposure calculation.

We have focused on the explanatory variable corresponding to the contract type. This variable is significant and corresponds to two populations: group contracts and self-employed workers. Thus, we will be able to obtain a table for group contracts as well as a table for self-employed workers by calibrating a single model.

Since franchise information is not explicitly present in our data, a research effort has been undertaken to retrieve this information from the data. This effort resulted in the Figure 3. The identified franchises are consistent and thus used to precisely determine the observation period for each individual for whom a franchise could be clearly identified. Moreover, to measure the impact of considering the franchise, we have created two additional databases from claims for which we now have franchise information. The first database indeed considers the franchise, while the other assumes it to be zero. Therefore, for each of these databases, we calibrate a model, the results of which will be compared subsequently.

Figure 3. : Histogram of franchises for claims resulting in a payment in 2021 or 2022

Results of the Models

After properly formatting our data, we calibrated the following GAM model using the R package `mgcv`:

$$\log(\mathbb{E}(\text{Exit})) = \beta_0 + f_1(\text{Age}) + f_2(\text{Seniority}) + f_3(\text{Age,Seniority})$$

With:

- β_0: represents the constant effect, independent of other variables on the Exit variable.
- f_1: represents the age effect on the Exit variable.
- f_2: represents the seniority effect on the Exit variable.
- f_3: represents the interaction effect between age and seniority on the Exit variable.

This model was calibrated considering various scenarios: considering the franchise, considering it as null, and also including the contract type in addition to the franchise. It is important to highlight that incorporating the contract type is extremely straightforward thanks to the package we used. This model can be formulated as follows:

$$\log(\mathbb{E}(\text{Exit})) = \beta_0 + f_1(\text{Age}) + f_2(\text{Seniority}) + f_3(\text{Age,Seniority}) + \text{Type}$$

In this model, we use a reference population, specifically the group contracts, and adjust a coefficient for the self-employed worker contracts.

Furthermore, we demonstrated the integration of a pre-existing table into our model. To achieve this, we utilized the 2010 BCAC table as a reference and integrated it into our model as an offset, thereby modifying the previous GAM model as follows:

$$\begin{aligned}\log(\text{Exit}) &= \log(\alpha * \mu_{\text{BCAC}} * e) = \log(\alpha) + \log(\mu_{\text{BCAC}}) + \log(e) \\ &= \beta_1 + f_4(\text{Age}) + f_5(\text{Seniority}) + f_6(\text{Age,Seniority})\end{aligned}$$

Considering $\log(\mu_{\text{BCAC}}) + \log(e)$ as an offset, the modeled part is no longer the exit rate but the passage coefficient *alpha*, which places the smoothing constraint on these passage coefficients. To obtain the maintenance table for this model, we need to multiply the model's passage coefficient by the corresponding rate from the BCAC table. However, since the 2010 BCAC table is not smooth, the resulting exit rates are also not smooth. This is illustrated in Figure 4.

This observation led us not to adopt this model. Nevertheless, we considered it an interesting possibility and deemed it relevant to highlight. Particularly, with a smooth input table, it presents a feasible option to obtain smooth exit rates.

After calibrating our GAM models, we compared them to the following models:

- Turnbull
- Kaplan-Meier
- Whittaker-Henderson smoothing of raw data
- Regulatory BCAC table

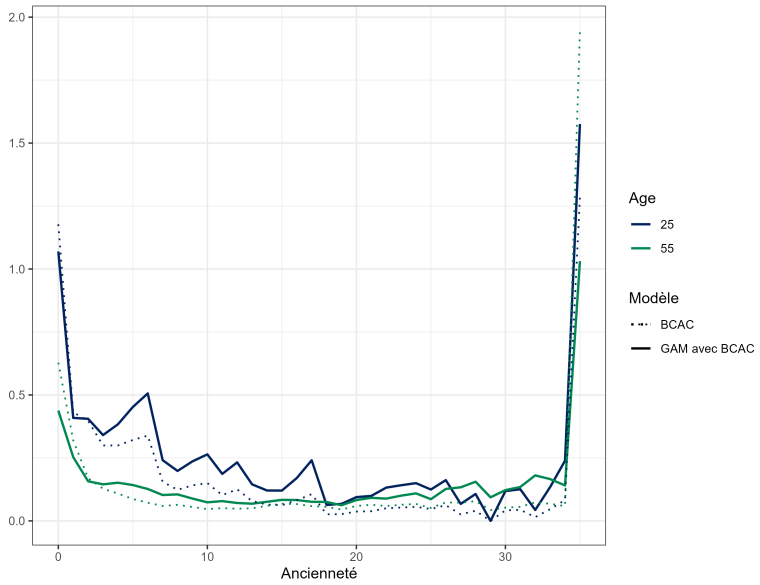


Figure 4. : Comparison of the smoothness of exit rates with the BCAC table offset

The different comparisons revealed that the GAM model performs the best in terms of solution regularity and accuracy. One performance measure we used for our GAM models is AIC. We demonstrated that incorporating the franchise and contract type significantly improves this measure, thereby identifying the GAM model using the contract type and incorporating the franchise as the best one.

Impact on reserves

Finally, we needed to calculate reserves for ongoing disability as well as pending disability. To achieve this, we first established the formulas for all reserves related to disability, invalidity, and death risks. Subsequently, we presented the various impacts of our models on these reserves.

We observed that using our best model leads to a significant decrease of 20% in reserves for ongoing disability and 35% for pending disability. Additionally, we noted that the impact of considering the franchise is less than 3% for both types of reserves. This observation is due to the fact that the impact of the franchise only affects periods of seniority shorter than the franchise durations. Given that these franchise durations are relatively short compared to the potential range of seniority, only claims with seniorities less than 4 months at the calculation date are affected by this impact. Since these claims do not represent the majority of our portfolio, the effect of considering the franchise on reserves remained limited.

Conclusion

In conclusion, the use of GAMs in modeling disability/invalidity risk appears promising. It adequately addresses the various challenges posed by the creation of experience tables, encom-

passing concerns related to exit rate consistency, smoothing, and accuracy. The simplicity of incorporating additional explanatory variables serves as a notable strength of this model.

However, it would be entirely feasible to elevate the modeling to a higher level by incorporating disability data, particularly transitions into invalidity. Such data, when applied in conjunction with GAMs, would enable the modeling of exit rates based on different exit causes. For instance, for disability, it would be possible to model exits due to resumption of work, exits linked to transition into invalidity, and exits due to death. This advanced modeling, necessitating more extensive data volume, would allow the direct generation of various tables, including transition-to-invalidity tables and death-in-disability tables. An important advantage would be the coherence among these tables. In other words, the overall exit rates would sum up to the total of exit rates for each of the potential exit causes we've identified for disability.

Table des matières

Liste des Figures

Liste des Tables

Introduction

Dans le secteur de l'assurance en France, les risques associés à l'arrêt de travail et à l'invalidité sont évalués en se basant sur des tables réglementaires élaborées par le Bureau Commun des Assurances Collectives (BCAC). Cependant, étant donné que ces tables datent de 1993, des interrogations se posent quant à leur adéquation avec le contexte actuel. En effet, depuis cette période, d'importants changements, à la fois technologiques et socio-économiques, ont eu lieu, suscitant ainsi des questionnements sur l'évolution temporelle du risque. De plus, les modifications législatives de 2010, telles que l'élévation de l'âge de la retraite de 2 ans, ont contraint le BCAC à extrapoler les tables de 1993 sur une période de 2 ans, sans modifier les valeurs non extrapolées. En résulte donc que les tables de 2010 sont en réalité des extrapolations des tables de 1993, étendues de 2 ans. Cette situation soulève des interrogations sur la précision et la pertinence actuelles de ces tables. Ainsi, pour une meilleure compréhension des caractéristiques du risque d'incapacité et d'invalidité dans le contexte contemporain, il devient essentiel de disposer de tables d'expérience plus récentes, spécifiquement adaptées à chaque assureur. Il est en effet judicieux d'ajuster ces tables en fonction des particularités des portefeuilles, tels que les types de populations comme les travailleurs non salariés et les employés d'entreprise. Cette nécessité de construire des tables d'expérience personnalisées justifie l'exploration de nouvelles méthodes pour une évaluation plus précise du risque, tout en prenant en considération les évolutions contextuelles. Cela représente un défi complexe, car les risques d'incapacité et d'invalidité sont influencés par une multitude de facteurs, ce qui rend leur modélisation et leur actualisation essentielles pour une gestion optimale de ces risques.

Dans le domaine de l'estimation des tables d'expérience, différentes méthodes sont utilisées. Parmi celles-ci, la méthode d'estimation de la fonction de survie proposée par Kaplan et Meier (1958) devenue une référence couramment utilisée, est illustrée dans les travaux de Bagui (2013). Toutefois, il est essentiel de noter que la recherche dans ce domaine ne se limite pas à une seule approche, et de nombreuses autres méthodes ont émergé au fil du temps. Une liste exhaustive de ces méthodes serait difficile à dresser, mais nous pouvons citer les principales méthodes en plus du modèle de Kaplan-Meier, que sont le lissage des taux bruts par la méthode de Whittaker-Henderson ou encore le modèle de Cox. Le mémoire de Petit (2017) a mis en avant pas moins de 5 modèles différents, illustrant la variété des approches possibles. Leurs distinctions résident souvent dans la manière dont ils intègrent des variables explicatives ou dans la mise en œuvre pratique des méthodes.

Notre objectif principal est de développer une méthode de construction de tables d'expérience qui tienne compte efficacement de la censure à droite, de la troncature à gauche et de plusieurs variables explicatives. Nous aspirons à proposer une approche simple et efficace permettant d'aboutir à des résultats pertinents et adaptés à tous types de portefeuilles.

Notre travail se démarque par son approche novatrice basée sur les modèles additifs généralisés, qui offre une flexibilité importante dans la modélisation des taux de sorties instantanés. Cette approche permet d'incorporer de manière naturelle les différentes variables explicatives et de mieux appréhender les effets de chaque facteur sur les risques d'incapacité et d'invalidité.

De plus, nous avons consciencieusement élaboré des formules dédiées aux provisions en cas d'incapacité, d'invalidité et de maintien de la garantie décès, en utilisant des notations modernes et appropriées. Cette étape est d'une importance significative, car ces formules constituent les bases des calculs de provisions en incapacité/invalidité, et il convient de noter qu'il existe actuellement peu de références en la matière.

Plan

Dans ce mémoire, nous débuterons par situer notre étude dans le contexte de la prévoyance en France. Nous définirons les termes clés et exposerons davantage la problématique que nous cherchons à résoudre, afin de bien cerner les enjeux de notre recherche.

Ensuite, nous explorerons le cadre théorique dans lequel s'inscrit notre travail. Nous mettrons en évidence deux approches particulièrement importantes et largement utilisées dans le domaine de la construction de tables d'expérience. Par la suite, notre attention se portera sur les Modèles Additifs Généralisés (GAM) (Hastie et Tibshirani 1987), qui constituera la base de notre approche.

Dans la troisième partie, nous présenterons en détail les données utilisées pour notre étude et les méthodologies employées pour leur traitement. Nous inclurons également des tests réalisés sur ces données et fournirons quelques statistiques descriptives pour mieux appréhender leurs caractéristiques. Enfin, nous expliquerons comment nous avons procédé pour mettre en forme nos données et les rendre compatibles avec notre modèle GAM.

Les résultats de notre recherche seront exposés dans la quatrième partie. Nous analyserons tout d'abord les résultats obtenus sur la base globale, c'est à dire la base comprenant tout nos sinistre et sans l'information de la franchise. Dans un second temps nous examinerons les résultats obtenus en incluant les franchises dans nos données pour finalement présenter ceux utilisant des covariables, afin d'affiner nos conclusions.

Enfin, nous évaluerons l'impact de notre travail sur les provisions en détaillant les formules que nous avons élaborées. Nous analyserons comment l'utilisation de nos tables d'expérience influence ces provisions.

Problématique

Dans le contexte que nous venons de décrire, les tables actuarielles jouent un rôle essentiel pour l'actuaire. Chaque risque est associé à un ensemble spécifique de tables. Par exemple, pour évaluer le risque de décès, nous utilisons la table de mortalité. Pour le risque d'invalidité, nous faisons référence à la table de maintien en invalidité et la table de mortalité en invalidité. Quant au risque d'incapacité, il est évalué à l'aide de la table d'incidence en incapacité, la table de maintien en incapacité, la table de passage en invalidité et la table de mortalité en incapacité. Cette situation peut être résumée dans la Figure 5.

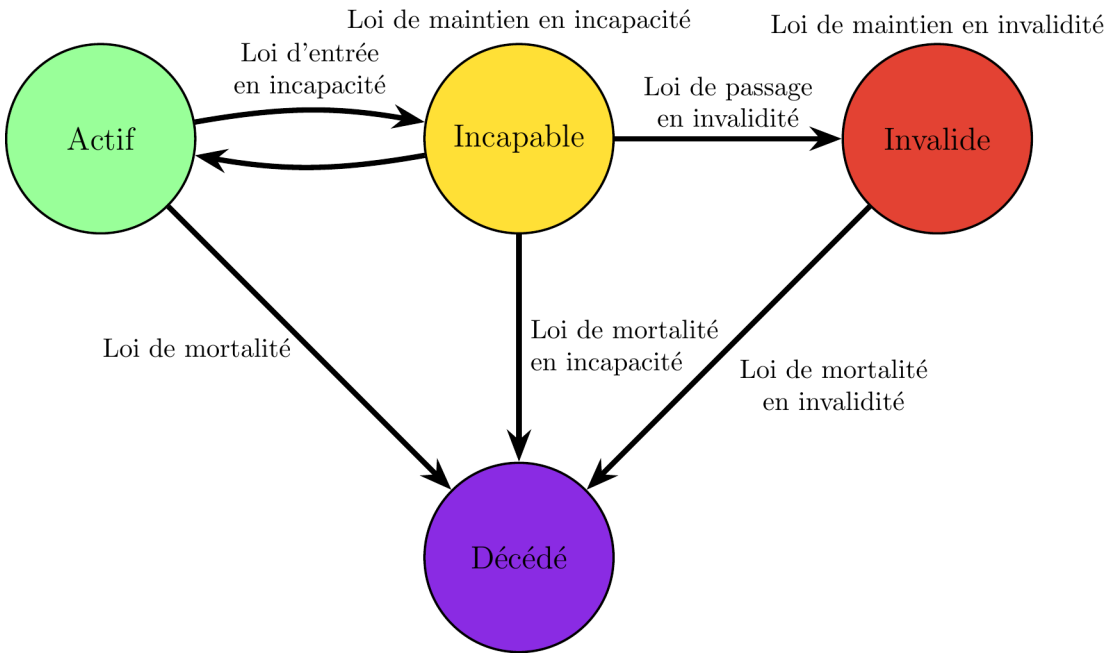


Figure 5. : Description du risque prévoyance

Il est essentiel de souligner que malgré qu'elle soit réglementaire, la table de maintien en incapacité du BCAC n'est pas exempte de défauts. En premier lieu, les tables réglementaires sont basées sur des données de population générale, ce qui peut ne pas correspondre exactement au profil de risque spécifique d'un assureur donné. De plus, nous pouvons observer que la table de maintien en incapacité présente des taux de sortie négatifs au-delà de 61 ans, ce qui est mathématiquement impossible. Face à ces limitations, mais surtout à la vétusté des tables réglementaires, il devient donc crucial pour les assureurs de créer leurs propres tables d'expérience en utilisant une méthode fiable et rigoureuse. Cela leur permettra de mieux appréhender le risque spécifique de leur portefeuille et d'améliorer la précision des évaluations actuarielles.

Pour la création de tables, le BCAC utilise principalement l'estimateur de Turnbull, dont les détails techniques peuvent être consultés dans Turnbull (1976). En dehors de cela, les estimateurs de Kaplan-Meier et de Nelson-Aalen sont couramment utilisés pour estimer la fonction de survie. Bien que largement répandus, ces estimateurs ne représentent pas la meilleure méthode pour déterminer une fonction de survie en présence de censure et de troncature pour le risque d'incapacité/invalidité. Ils présentent des limites, notamment en ce qui concerne la censure à droite et la troncature à gauche, ainsi que l'incapacité à prendre en compte l'effet de variables supplémentaires.

Actuellement, lorsque nous souhaitons introduire une variable explicative telle que la catégorie socio-professionnelle (CSP), la méthode classique consiste à recourir au modèle de Cox ou à des modèles paramétriques dont les paramètres dépendent des variables explicatives. Le modèle de Cox présente certaines limites, notamment en ce qui concerne la modélisation de l'interaction entre l'âge et l'ancienneté. Cette limitation est directement liée à l'hypothèse de proportionnalité, qui suppose que l'effet des variables reste constant dans le temps. En conséquence, il n'est pas possible d'incorporer directement cette interaction dans le modèle de

Cox. Pour surmonter cette contrainte, il est possible d'utiliser des variables dépendantes du temps, mais cela entraîne une complexité algorithmique considérable. En effet, l'introduction de ces variables rend les calculs plus complexes et beaucoup plus exigeants en termes de temps et de ressources.

Face à ces considérations, il devient crucial de rechercher une méthode fiable, robuste et flexible pour estimer une fonction de survie qui permette d'obtenir des tables aussi fidèles que possible.

Objectif

Face aux constats que nous avons établis, notre objectif est de proposer un modèle qui réponde aux exigences suivantes :

- Modèle non paramétrique : Nous privilégions l'utilisation d'un modèle non paramétrique, car il permet de tirer pleinement parti de l'information contenue dans les données sans pour autant supposer de forme à priori.
- Prise en compte d'une table préexistante : Notre modèle sera conçu pour être capable d'intégrer une table d'expérience déjà construite, offrant ainsi la possibilité d'incorporer des informations a priori dans nos estimations.
- Modularité : Notre modèle est conçu pour s'adapter aisément à l'intégration de nouvelles variables explicatives. Cette flexibilité permettra un calibrage précis et une bonne interprétabilité des résultats du modèle.
- Robustesse : Malgré des quantités de données limitées, notre modèle continue de produire des résultats cohérents et fiables. Cette capacité à maintenir sa performance, même dans des conditions de données restreintes, renforce la crédibilité de nos analyses et la confiance dans les conclusions qui en découlent.
- Absence de biais lié à la censure et à la troncature : Nous veillerons à ce que notre modèle prenne en compte efficacement les phénomènes de censure à droite et de censure à gauche, qui sont caractéristiques du risque d'incapacité/invalidité.
- Cohérence : Notre modèle produit une fonction de survie décroissante et sans taux négatifs, garantissant ainsi des résultats cohérents et réalistes.

Un modèle qui répond à l'ensemble de ces critères est le modèle additif généralisé (GAM), lequel sera largement présenté dans la suite de ce mémoire.

Chapitre 1.

Contexte de la prévoyance en France

1.1. Définitions

La prévoyance

Dans la vie d'un travailleur, des événements imprévus peuvent survenir, comme une maladie ou un accident, qui le rendent incapable de travailler temporairement ou définitivement. Le travailleur est exposé à un risque de perte de revenus significative. L'incapacité à travailler peut non seulement compromettre les sources de revenus, mais également impacter la capacité à maintenir les obligations financières, comme des prêts par exemple. Dans ce contexte, le concept de prévoyance trouve sa place.

La prévoyance est un mécanisme visant à atténuer les risques financiers associés à la santé, l'incapacité, l'invalidité et le décès. Elle agit en complément des régimes de sécurité sociale existants. Bien que la Sécurité sociale offre des indemnités pour couvrir ces risques, elles peuvent souvent se révéler insuffisantes pour maintenir le niveau de vie. Pour l'incapacité, en moyenne ces indemnités correspondent à environ la moitié du salaire brut. Dans cette perspective, la souscription à un contrat de prévoyance peut s'avérer judicieuse pour assurer une couverture plus complète.

Les chiffres de la prévoyance en France

En 2022, le marché de l'assurance prévoyance en France a généré des cotisations atteignant 25,6 milliards d'euros, enregistrant une augmentation de 5,4% par rapport à l'année précédente. Ces chiffres ne tiennent pas compte des cotisations liées à l'assurance emprunteurs, qui se sont élevées à 10,2 milliards d'euros en 2022. Ces données soulignent l'importance de ce marché pour l'économie française.

La répartition de ces cotisations est disponible dans la Figure 1.1, tandis que l'évolution des cotisations, à l'exclusion de l'assurance emprunteur, est présentée dans la Figure 1.2.

Régimes de prévoyance

Pour se couvrir contre ces risques, différents régimes de prévoyance existent :

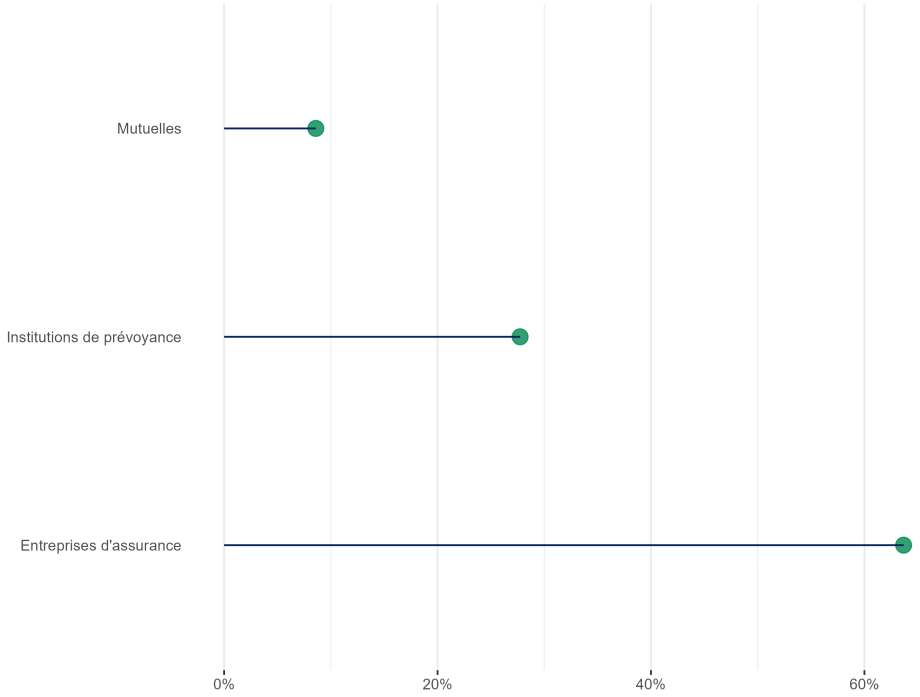


Figure 1.1. : Répartition des cotisations prévoyance entre acteurs de la prévoyance

- *Le régime de base de la Sécurité sociale* : Concerne tous les travailleurs salariés ainsi que les travailleurs indépendants depuis le 1er janvier 2018, ainsi que toute personne bénéficiant de droit au titre de la résidence (protection universelle maladie). Les indemnités journalières sont versées par l'Assurance Maladie après un délai de carence de 3 jours. L'objectif de ce régime est d'assurer à toute la population un montant minimal de revenu permettant de vivre dans des conditions décentes. Le montant des indemnités journalières est basé sur un salaire brut plafonné à 1,8 fois le Smic mensuel (soit 3 144,96 euros bruts au 1er mai 2023).
- *Le régime complémentaire obligatoire* : Créé en 1978 par la loi de "mensualisation" qui impose aux employeurs d'assurer, sous certaines conditions et avec un délai de carence, un certain niveau de salaire en cas d'arrêt de travail pour maladie ou accident aux salariés ayant au moins un an d'ancienneté. Son objectif est de garantir une meilleure protection sociale aux salariés en complément des prestations de la Sécurité sociale.
- *Le régime de prévoyance collective facultative* : Contrairement au régime complémentaire obligatoire, qui est imposé par la loi ou une convention collective, la prévoyance collective facultative est mise en place de manière volontaire par l'employeur. Elle vise à offrir une meilleure protection sociale complémentaire aux salariés. Les garanties et les prestations proposées peuvent varier en fonction du contrat souscrit, et les salariés ont généralement la possibilité de choisir les options qui correspondent le mieux à leurs besoins individuels.
- *La prévoyance individuelle* : Contrairement à la prévoyance collective qui est souscrite par l'employeur, la prévoyance individuelle est facultative et est souscrite directement par l'assuré pour couvrir ses propres besoins de protection. La prévoyance individuelle offre une plus grande flexibilité et personnalisation par rapport à la prévoyance collective,

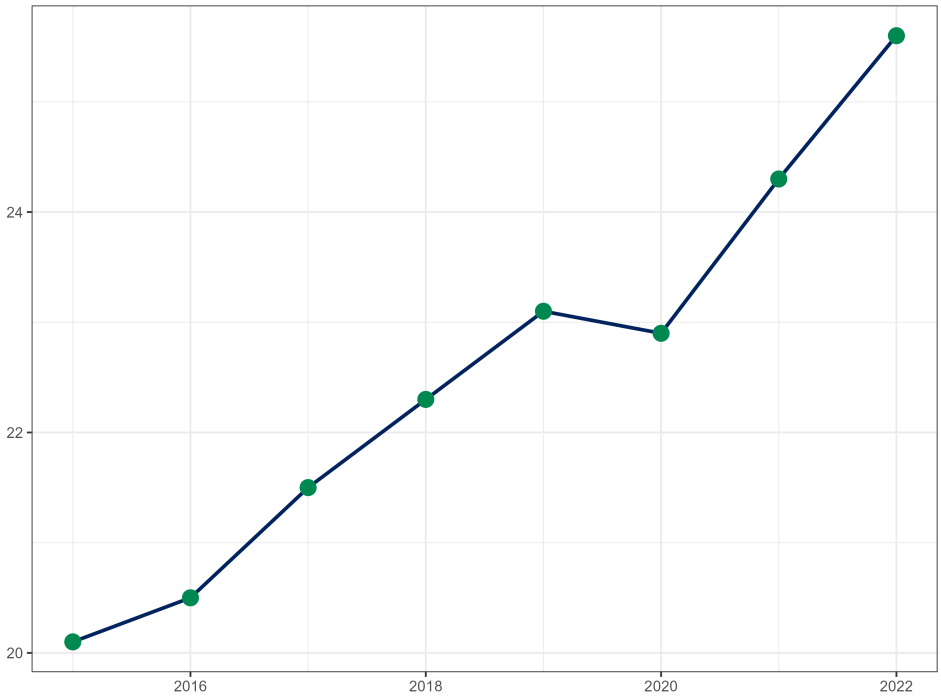


Figure 1.2. : Évolution des cotisations pour l'assurance prévoyance (en milliards d'euros) de 2015 à 2022

car l'assuré peut choisir les garanties et les niveaux de couverture qui correspondent à ses besoins spécifiques.

- *Le régime de prévoyance travailleurs non salariés* (TNS) : Conçu pour les travailleurs indépendants tels que les artisans, commerçants, professions libérales, micro-entrepreneurs et autres travailleurs non salariés. Étant donné que ces professionnels ne bénéficient pas des mêmes protections sociales que les salariés, la prévoyance TNS vise à compenser cette absence de couverture en leur offrant des garanties de prévoyance adaptées à leur situation. Il est important de souligner que la prévoyance TNS est facultative, ce qui signifie que les travailleurs non salariés doivent souscrire un contrat de prévoyance individuelle spécifique pour bénéficier de ces garanties.

Les principales garanties d'un contrat de prévoyance

Les principaux risques couverts par la prévoyance en France sont les suivants :

- *Incapacité temporaire de travail (ITT)* : Cette garantie prévoit le versement d'indemnités journalières à l'assuré en cas d'arrêt de travail temporaire causé par une maladie ou un accident. La durée de l'ITT est évaluée par le médecin traitant ou le médecin-conseil de l'assurance maladie et correspond à la période pendant laquelle l'assuré est incapable de poursuivre son activité professionnelle habituelle. L'objectif de cette garantie est de maintenir partiellement le revenu de l'assuré pendant son arrêt de travail.
- *Invalidité* : Cette garantie prévoit le versement d'une rente ou d'un capital en cas d'invalidité, qu'elle soit totale ou partielle, jusqu'à ce que l'assuré atteigne l'âge de

62 ans. Son objectif est de compenser la perte de revenu due à l'invalidité et de garantir un niveau de vie décent pour l'assuré jusqu'à sa retraite. L'invalidité est définie selon les critères de l'article L341-4 du code de la sécurité sociale et est évaluée par le médecin-conseil de l'Assurance Maladie. On distingue trois catégories d'invalidité :

  - Invalidité de 1ère catégorie : Elle concerne les assurés dont la capacité de travail est réduite d'au moins 2/3. Ces personnes peuvent exercer une activité rémunérée, mais avec des limitations importantes.
  - Invalidité de 2ème catégorie : Elle concerne les assurés dont la capacité de travail est réduite de manière absolue et définitive, les rendant totalement incapables d'exercer une profession.
  - Invalidité de 3ème catégorie : Elle concerne les assurés dont l'état de santé est tel qu'ils sont dans l'incapacité totale de se livrer à une quelconque activité professionnelle et nécessitent l'assistance constante d'une tierce personne.

- *Décès* : Cette garantie prévoit le versement d'un capital ou d'une rente aux bénéficiaires désignés en cas de décès de l'assuré. Son objectif est de protéger financièrement les proches et les bénéficiaires de l'assuré en leur fournissant un soutien financier en cas de disparition prématurée. Le capital ou la rente versés peuvent aider à couvrir les frais liés aux funérailles, à rembourser des dettes, à maintenir le niveau de vie de la famille ou à assurer l'avenir des enfants. La garantie décès peut être souscrite sous différentes formes, telles que le capital décès, la rente de conjoint ou la rente éducation. Le capital décès est un montant fixe qui est versé aux bénéficiaires désignés en cas de décès de l'assuré. Il peut être utilisé selon les besoins spécifiques de la famille, comme le règlement des frais funéraires ou le remboursement de dettes. La rente éducation, quant à elle, assure le versement d'une rente régulière pour soutenir les besoins financiers des enfants ou des personnes à charge afin de garantir leur éducation et leur avenir.

1.2. Evolution des tables publiées par le BCAC

La gestion du risque prévoyance représente un défi complexe, exigeant une élaboration minutieuse des tables pour appréhender avec précision le caractère du risque sous-jacent. Le Bureau Commun d'Assurances Collectives (BCAC) joue un rôle majeur dans cette démarche en élaborant des tables actuarielles réglementaires.

En 1993, le BCAC initie la conception des tables de maintien en incapacité, de passage en invalidité et de maintien en invalidité, lesquelles acquièrent leur statut réglementaire en 1996, marquant ainsi une étape significative dans l'évolution de la gestion du risque prévoyance.

En 2002, malgré l'introduction par le BCAC de tables de mortalité en incapacité et en invalidité, celles-ci n'obtiennent pas de statut réglementaire à cette époque.

En 2010, les tables de maintien en incapacité sont prolongées, une adaptation motivée par des changements législatifs liés aux retraites. Cette évolution souligne la réactivité du BCAC aux modifications du cadre légal et réglementaire.

En 2013, de nouvelles tables sont introduites, répondant à un besoin pressant de changer les modèles actuariels pour suivre l'évolution substantielle du risque depuis 1993. Bien que ces

tables ne sont à ce jour toujours pas réglementaire, leur introduction témoigne d'une réponse nécessaire aux transformations significatives dans le paysage des risques prévoyance.

Cette trajectoire chronologique met en lumière le rôle du BCAC dans l'évolution des tables actuarielles, illustrant à la fois sa réactivité aux changements législatifs en 2010 et la nécessité d'ajustements substantiels en 2013 pour suivre l'évolution réelle du risque depuis 1993. En conclusion, la rapidité des évolutions du risque prévoyance souligne l'importance d'une méthode efficace de création de tables, capable de s'adapter avec agilité aux fluctuations du risque mais aussi au transformation de notre société.

Chaque parution des tables actuarielles est accompagnée de limites d'âges spécifiques, répertoriées de manière dans la Table 1.1.

Table 1.1. : Résumé des limites d'âges des tables produites par le BCAC

Table	BCAC 1993	BCAC 2010[1]	BCAC 2013[2]
Maintien en incapacité			
Age de début	23 ans	23 ans	20 ans
Age de fin	65 ans	67 ans	66 ans
Maintien en invalidité			
Age de début	30 ans	30 ans	25 ans
Age de fin	60 ans	62 ans	65 ans
Passage incapacité vers invalidité			
Age de début	23 ans	23 ans	20 ans
Age de fin	60 ans	62 ans	63 ans

[1]Tables de 1993 prolongée de 2 ans; seules tables réglementaires à ce jour
[2]Tables construites à partir de nouvelles données

Chapitre 2.

Cadre théorique

2.1. Approche historique

Dans cette partie, nous présentons différentes approches historiquement utilisées dans le cadre de la construction de tables de survie, ou plus généralement, d'estimations du maintien dans un certain état. Nous commençons par exposer le modèle de Turnbull, utilisé par le BCAC lors de la création des différentes tables réglementaires de maintien, puis nous abordons les modèles largement utilisés que sont les estimateurs de Kaplan-Meier et de Nelson-Aalen.

2.1.1. Modèle de Turnbull

Comme évoqué précédemment, lors de la construction des tables en 1993 et 2013, le BCAC a opté pour l'utilisation de l'estimateur de Turnbull, qui est très bien documenté dans le mémoire de Wolfrum (1993). Nous allons prendre le temps ici de décrire cet estimateur, souvent méconnu dans le monde actuariel.

Il est important de noter avant d'utiliser l'estimateur de Turnbull, qu'il suppose que la **variable étudiée est discrète**. C'est une hypothèse fondamentale qui diffère considérablement de l'approche proposée par Kaplan-Meier. Cela contraste également avec la variable réellement étudiée, qui n'est pas véritablement discrète.

Il convient de souligner que cet estimateur prend en compte les censures et les troncatures, qu'elles soient à gauche ou à droite.

L'objectif est d'estimer la fonction de répartition F d'une variable aléatoire discrète X à partir d'un échantillon de taille N. Pour chaque individu, on dispose de la période d'observation B_i durant laquelle il était observable, ce qui permet de prendre en compte la troncature. Par exemple, pour un individu ayant une période de franchise de 3 mois, sa période observable associée sera $B_i = [3; 36]$.

Soit nos observations $X_i = x_i$ tirées de la loi tronquée $F(x, B_i) = P(X \leq x | X \in B_i)$. Ainsi, X_i est tronqué par B_i, ce qui signifie que l'individu n'aurait pas pu être observé si X_i n'appartenait pas à B_i.

Il est également possible que X_i ne soit pas observé de manière précise, et nous savons uniquement que X_i appartient à un certain ensemble A_i avec $A_i \subseteq B_i$. Par conséquent, notre

X_i est considéré comme censuré par A_i. De manière concrète, A_i correspond à l'espace des valeurs possibles prises par X_i. Dans le cas où nous connaissons la valeur exacte de X_i, alors $A_i = \{x_i\}$. En revanche, en cas de censure, nous n'observons pas la véritable valeur de X_i, mais nous savons simplement que X_i est supérieur à une certaine valeur. Par exemple, si une personne est censurée après 30 mois d'ancienneté, nous avons la variable A_i associée qui vaut : $A_i = [30; 36]$.

Finalement nos données observées consistent N paires $(A_1, B_1), (A_2, B_2), ..., (A_N, B_N)$.

Quelques précisons sur les B_i et A_i :

- Si $B_i = (-\infty, +\infty)$, cela signifie que l'individu i n'est pas tronqué, c'est-à-dire qu'il a été suivi pendant toute la période d'intérêt. (Remarque : Dans le cas où nous observons un évènement qui peut prendre un temps fini, nous pouvons considérer que $B_i = [0, \omega]$ correspond à un individu non tronqué, où ω représente la durée maximale de suivi, par exemple $\omega = 36$ pour l'incapacité.)
- Si A_i est un singleton $\{x_i\}$, cela signifie que l'individu i n'est pas censuré, c'est-à-dire que nous avons observé la valeur exacte de X_i.
- Si A_i est un intervalle de la forme $[L_i, R_i]$ alors X_i est dit censuré à droite si $R_i = +\infty$ ou $R_i = \omega$.

En supposant que chaque A_i s'écrit comme une union finie d'intervalles disjoints, c'est-à-dire $A_i = \bigcup_{j=1}^{k_i} [L_{ij}, R_{ij}]$, nous pouvons créer m intervalles dont les extrémités gauches et droites sont dans les ensembles $\{L_{ij}; 1 \leq j \leq k_i, 1 \leq i \leq N\}$ et $\{R_{ij}; 1 \leq j \leq k_i, 1 \leq i \leq N\}$ et qui ne contiennent aucun autre élément des ensembles $\{L_{ij}\}$ ou $\{R_{ij}\}$. On note ces intervalles $[q_1, p_1], [q_2, p_2], ..., [q_m, p_m]$. Concrètement, les q_i et p_i représentent tous les instants de sorties ou de censure des données. Pour notre risque, il est clair que nos intervalles $[q_j, p_j]$ correspondent simplement aux singletons de nos anciennetés allant de 0 à 36. Autrement dit : $\forall i \in [1, 37]$, $[q_j, p_j] = \{j - 1\}$.

Ainsi, nous nous intéressons à une variable aléatoire X discrète évoluant sur $C = \bigcup_{j=1}^{m} [q_j, p_j] = \bigcup_{j=1}^{m} \{j - 1\}$.

En définissant pour $1 \leq j \leq m$, $s_j = F(p_j^+) - F(q_j^-)$ la probabilité que X soit dans l'intervalle $[q_j, p_j]$ alors le vecteur $\mathbf{s} = (s_1, ...s_m)$ avec $\Sigma s_j = 1$ et $s_j \geq 0$ définit la loi de X. De cette manière, s_j correspond à la probabilité qu'un arrêt dure entre $j - 1$ et j mois.

Il est essentiel d'être extrêmement vigilant quant au sens donné aux notations. En effet, normalement, nous avons la définition suivante : $F(x) = P(X \leq x)$ et donc $S(x) = 1 - F(x) = P(X > x)$. Selon cette définition $S(36) = P(X > 36) = 0$, car un arrêt maladie ne peut dépasser 36 mois. Il peut atteindre 36 mois, mais en toute rigueur, il ne peut pas dépasser cette durée. Cependant, cela ne correspond pas à l'esprit et à l'utilisation des tables réglementaires de maintien en incapacité, car pour un âge donné, il existe un nombre strictement positif de survivants à 36 mois. Cela signifie que ces tables n'expriment pas $S(x)$ mais plutôt $\tilde{S}(x) = P(X \geq x)$. Cette fois-ci, en adoptant cette définition, nous avons effectivement $\tilde{S}(36) \neq 0$, ce qui correspond au comportement souhaité. Dans le cadre d'un problème continue, cela ne pose pas de problème, car $\forall x \quad P(X = x) = 0$ dans ce cas.

Dans le cadre de ce modèle, en utilisant différents théorèmes, on peut démontrer que maximiser la vraisemblance de notre modèle revient à maximiser la fonction suivante :

$$L(s_1, ..., s_m) = \prod_{i=1}^{N} \left(\frac{\sum_{j=1}^{m} \alpha_{ij} s_j}{\sum_{j=1}^{m} \beta_{ij} s_j} \right) \tag{2.1}$$

Avec :

- $\alpha_{ij} = 1$ si $[q_j, p_j] \subset A_i$, 0 sinon.
- $\beta_{ij} = 1$ si $[q_j, p_j] \subset B_i$, 0 sinon.

Nous cherchons donc **l'estimateur du maximum de vraisemblance**.

Turnbull propose alors un algorithme itératif convergeant vers cet estimateur.
Pour cela il utilise les quantités suivantes :

- $\mu_{ij}(\mathbf{s}) = \frac{\alpha_{ij} s_j}{\sum_{k=1}^{m} \alpha_{ik} s_k}$ qui représente la probabilité que l'individu i sorte à la période j.
- $\nu_{ij}(\mathbf{s}) = \frac{(1-\beta_{ij}) s_j}{\sum_{k=1}^{m} \beta_{ik} s_k}$
- $M(\mathbf{s}) = \sum_{i=1}^{N} \sum_{j=1}^{m} (\mu_{ij}(\mathbf{s}) + \nu_{ij}(\mathbf{s}))$
- $\pi_j(\mathbf{s}) = \frac{1}{M(\mathbf{s})} \sum_{i=1}^{N} \mu_{ij}(\mathbf{s}) + \nu_{ij}(\mathbf{s})$

Finalement, l'algorithme pour obtenir l'estimateur du maximum de vraisemblance est le suivant :

1. Initialiser les probabilités s_j^0. (par exemple en posant $s_j^0 = \frac{1}{m}$ pour $1 \leq j \leq m$)
2. Évaluer les quantités $\mu_{ij}(\mathbf{s^0})$ et $\nu_{ij}(\mathbf{s^0})$ pour $1 \leq i \leq N$ et $1 \leq j \leq m$, puis calculer $M(\mathbf{s^0})$ et $\pi_j(\mathbf{s^0})$.
3. Poser $s_j^1 = \pi_j(\mathbf{s^0})$ pour $1 \leq j \leq m$.
4. Retourner à l'étape 2 en remplaçant $\mathbf{s^0}$ par $\mathbf{s^1}$, etc.
5. S'arrêter lorsque l'évolution entre 2 étapes est suffisamment faible.

Il est possible, dans le cadre de notre problème, d'optimiser le temps de calcul en reformulant $\pi_j(\mathbf{s})$ de la manière suivante :

On remarque d'abord que pour une itération de l'algorithme h nous avons :

$$s_j^{h+1} = \pi_j(s^h) = \frac{\sum_{i=1}^{N} \left(\frac{\alpha_{ij}}{\sum_{k=1}^{m} \alpha_{ik} s_k^h} + \frac{(1-\beta_{ij})}{\sum_{k=1}^{m} \beta_{ik} s_k^h} \right)}{\sum_{i=1}^{N} \sum_{j=1}^{m} (\mu_{ij}(s^h) + \nu_{ij}(s^h))} s_j^h = \frac{A+B}{C} s_j^h$$

Nous allons décomposer le calcul de A, B et C.

Commençons par C :

$$\begin{aligned}
C &= \sum_{i=1}^{N}\left(\sum_{j=1}^{m}\frac{\alpha_{ij}s_j^h}{\sum_{k=1}^{m}\alpha_{ik}s_k^h}+\sum_{j=1}^{m}\frac{(1-\beta_{ij})s_j^h}{\sum_{k=1}^{m}\beta_{ik}s_k^h}\right)\\
&= \sum_{i=1}^{N}\left(\frac{\sum_{j=1}^{m}\alpha_{ij}s_j^h}{\sum_{k=1}^{m}\alpha_{ik}s_k^h}+\frac{\sum_{j=1}^{m}(1-\beta_{ij})s_j^h}{\sum_{k=1}^{m}\beta_{ik}s_k^h}\right)\\
&= \sum_{i=1}^{N}\left(1+\frac{\sum_{j=1}^{m}s_j^h-\sum_{j=1}^{m}\beta_{ij}}{\sum_{k=1}^{m}\beta_{ik}s_k^h}\right)\\
&= \sum_{i=1}^{N}\left(1+\frac{1-\sum_{j=1}^{m}\beta_{ij}}{\sum_{k=1}^{m}\beta_{ik}s_k^h}\right)\\
&= \sum_{i=1}^{N}\left(\frac{1}{\sum_{k=1}^{m}\beta_{ik}s_k^h}\right)\\
&= \sum_{i=1}^{N}\left(\frac{1}{\sum_{k=1}^{m}\beta_{ik}s_k^h}\right)\\
&= \sum_{\text{i non tronqué}}\left(\frac{1}{\sum_{k=1}^{m}\beta_{ik}s_k^h}\right)+\sum_{\text{i tronqué}}\left(\frac{1}{\sum_{k=1}^{m}\beta_{ik}s_k^h}\right)\\
&= \sum_{\text{i non tronqué}}1+\sum_{\text{i tronqué}}\left(\frac{1}{\sum_{k=1}^{m}\beta_{ik}s_k^h}\right) \text{ car } B_{ik}=1 \quad \text{pour tout k lorsque l'individu est non tronqué}\\
&= N_0+\sum_{\substack{\text{i tronqué}\\ \text{période 1}}}\left(\frac{1}{\sum_{k=2}^{m}s_k^h}\right)+\sum_{\substack{\text{i tronqué}\\ \text{période 2}}}\left(\frac{1}{\sum_{k=3}^{m}\beta_{ik}s_k^h}\right)+\cdots+\sum_{\substack{\text{i tronqué}\\ \text{période m-1}}}\left(\frac{1}{\sum_{k=m}^{m}\beta_{ik}s_k^h}\right)\\
&= N_0+\frac{P_1}{S(1)}+\frac{P_2}{S(2)}+\cdots+\frac{P_{m-1}}{S(m-1)}
\end{aligned}$$

Où N_0 désigne le nombre d'individus sans franchise, P_i désigne le nombre d'individus dont la franchise est i et $S(i)=P(X\geq i)=\sum_{j=i+1}^{m}s_j$.

Nous continuons maintenant avec le calcul de A :

$$\begin{aligned}
A &= \sum_{i=1}^{N}\frac{\alpha_{ij}}{\sum_{k=1}^{m}\alpha_{ik}s_k^h}\\
&= \sum_{\text{i non censuré}}\frac{\alpha_{ij}}{\sum_{k=1}^{m}\alpha_{ik}s_k^h}+\sum_{\substack{\text{i censuré}\\ \text{période 1}}}\frac{\alpha_{ij}}{\sum_{k=1}^{m}\alpha_{ik}s_k^h}+\cdots+\sum_{\substack{\text{i censuré}\\ \text{période j}}}\frac{\alpha_{ij}}{\sum_{k=j}^{m}\alpha_{ik}s_k^h}\\
&= \frac{n_j}{s_j^h}+\frac{\lambda_1}{S(0)}+\cdots+\frac{\lambda_j}{S(j-1)}
\end{aligned}$$

Où λ_i est le nombre de sortie entre l'instant $j-1$ et j pour les personnes censurées à la période i exactement.

Vient finalement le calcul de B :

$$\begin{aligned}
B &= \sum_{i=1}^{N} \frac{1-\beta_{ij}}{\sum_{k=1}^{m} \beta_{ik} s_k^h} \\
&= \sum_{\text{i non tronqué}} \frac{1-\beta_{ij}}{\sum_{k=1}^{m} \beta_{ik} s_k^h} + \sum_{\substack{\text{i tronqué} \\ \text{période j}}} \frac{1-\beta_{ij}}{\sum_{k=j+1}^{m} \beta_{ik} s_k^h} + \cdots + \sum_{\substack{\text{i tronqué} \\ \text{période m-1}}} \frac{1-\beta_{ij}}{\sum_{k=m}^{m} \beta_{ik} s_k^h} \\
&= 0 + \sum_{\substack{\text{i tronqué} \\ \text{période j}}} \frac{1}{S(j)} + \cdots + \sum_{\substack{\text{i tronqué} \\ \text{période m-1}}} \frac{1}{S(m-1)} \\
&= \frac{P_j}{S(j)} + \cdots + \frac{P_{m-1}}{S(m-1)}
\end{aligned}$$

Nous pouvons donc écrire $\pi_j(s^h)$ sous la forme :

$$\pi_j(s^h) = \frac{\frac{n_j}{s_j^h} + \frac{\lambda_1}{S(0)} + \cdots + \frac{\lambda_j}{S(j-1)} + \frac{P_j}{S(j)} + \cdots + \frac{P_{m-1}}{S(m-1)}}{N_0 + \frac{P_1}{S(1)} + \frac{P_2}{S(2)} + \cdots + \frac{P_{m-1}}{S(m-1)}} s_j^h$$

Après avoir examiné en détail l'estimateur de Turnbull, utilisé par le BCAC pour la création des tables réglementaires, nous constatons que son hypothèse de distribution discrète du risque peut rendre la prise en compte de la franchise imprécise. En effet, les franchises ne correspondent pas systématiquement à un nombre entier de mois elles doivent donc être arrondies à l'entier inférieur ce qui introduit inévitablement un biais dans les résultats obtenus. Nous nous tournons maintenant vers un autre estimateur fréquemment utilisé par les assureurs pour la construction de leurs tables d'expérience : l'estimateur de Kaplan-Meier.

2.1.2. Modèle de Kaplan-Meier

L'estimateur le plus utilisé dans l'analyse de survie est celui de Kaplan-Meier. Il s'agit d'un estimateur non paramétrique qui permet d'estimer la fonction de survie d'une variable donnée. Bien que sa construction soit largement connue, il est important de souligner que l'estimateur couramment appelé "l'estimateur de Kaplan-Meier" ne correspond pas strictement à l'estimateur original proposé par Kaplan et Meier en 1958, qui ne prend pas en compte la troncature. En réalité, il s'agit d'une amélioration ultérieure introduite par Wang, Jewell, et Tsai (1986), qui tient compte de la troncature.

L'estimateur est construit de manière progressive en utilisant une suite de produits. Pour une variable aléatoire X dont la fonction de survie S doit être estimée, l'estimateur se présente sous la forme suivante :

$$S_{KM}(t) = \prod_{t_j<t} \frac{n_j - d_j}{n_j}$$

Dans cette équation, les instants de sortie observés sont notés t_j, où n_j représente l'ensemble des individus à risque à l'instant t_j (c'est-à-dire toutes les personnes ayant au moins atteint leur franchise et n'ayant pas dépassé leur date de sortie ou de censure), et d_j correspond au nombre de sorties observées à l'instant t_j.

En d'autres termes, l'estimateur de Kaplan-Meier calcule la probabilité de survie à un instant donné en multipliant les probabilités de survie conditionnelle à chaque instant antérieur où une sortie a été observée. C'est à dire que les sauts se font à chaque sortie effectivement observée.

L'estimateur de Kaplan-Meier présente plusieurs propriétés importantes :

- **Consistent** : il converge vers la fonction de survie réelle lorsque le nombre d'individus tend vers l'infini.
- **Asymptotiquement gaussien** : il suit approximativement une distribution gaussienne lorsque le nombre d'observations devient très grand, ce qui facilite la construction d'intervalles de confiance et l'application de tests statistiques.

Concernant la variance, différentes méthodes existent pour estimer la variance de l'estimateur, comme la méthode de Greenwood et la méthode de Nelson-Aalen.

Nous rappelons ici très brièvement ces deux estimateurs de la variance de l'estimateur de Kaplan-Meier les plus populaires :

1. Estimateur de Greenwood:

$$\text{Var}[S_{KM}(t)] = \sum_{t_j<t} \frac{d_j}{n_j(n_j - d_j)} S_{KM}(t_j)^2$$

Où :

- t_j : instants de sortie observés
- n_j : nombre d'individus à risque à l'instant t
- d_j : nombre de sorties observées à l'instant t

2. Estimateur de Nelson-Aalen:

$$\text{Var}[S_{KM}(t)] = \sum_{t_j<t} \frac{d_j}{n_j} S_{KM}(t_j^-)^2$$

Où :

- t_j : instants de sortie observés
- n_j : nombre d'individus à risque à l'instant t
- d_j : nombre de sorties observées à l'instant t
- $S_{KM}(t_j^-)$: valeur de l'estimateur de Kaplan-Meier juste avant l'instant t

Le choix de l'estimateur de la variance dépend de plusieurs facteurs, tels que la taille de l'échantillon et la présence de censure, nous pouvons retenir les généralités suivantes :

- Estimateur de Greenwood : généralement plus précis que l'estimateur de Nelson-Aalen, mais peut être plus sensible à la censure.
- Estimateur de Nelson-Aalen : moins précis que l'estimateur de Greenwood, mais plus robuste à la censure.

2.1.3. Modèle de Nelson-Aalen

L'estimateur de Nelson-Aalen est un outil statistique largement utilisé en analyse de survie. Il permet d'estimer le cumul de la fonction de hasard (ou taux de risque cumulé) d'une variable aléatoire, à partir de données de survie brutes.

L'estimateur de Nelson-Aalen se distingue de l'estimateur de Kaplan-Meier par l'objet qu'il vise à estimer :

- **Estimateur de Nelson-Aalen** : Fonction de hasard cumulé ($H_{\mathrm{NA}}(t)$)
- **Estimateur de Kaplan-Meier** : Fonction de survie ($S_{\mathrm{KM}}(t)$)

La fonction de hasard cumulé $H(t)$ représente le risque cumulé d'un événement survenant avant l'instant t. Elle est définie comme l'intégrale de la fonction de risque h(t) sur l'intervalle $[0,\, t]$: $H(t) = \int_0^t h(u)\, du$

L'estimateur de Nelson-Aalen, noté $H_{\mathrm{NA}}(t)$, est une estimation non paramétrique de la fonction de hasard cumulé.

L'estimateur repose sur le principe de la sommation des quotients du nombre de sorties observées par le nombre d'individus à risque à chaque instant:

Formule de l'estimateur de Nelson-Aalen:

$$H_{\mathrm{NA}}(t) = \sum_{t_j < t} \frac{d_j}{n_j}$$

Où :

- t_j : Instants de sortie observés
- d_j : Nombre de sorties observées à l'instant t
- n_j : Nombre d'individus à risque à l'instant t, c'est-à-dire ceux n'ayant pas encore atteint l'événement d'intérêt, la sortie d'incapacité dans le cas de l'étude du maitien en incapacité

En d'autres termes, l'estimateur calcule le risque cumulé en additionnant le risque instantané à chaque instant où une sortie a été observée.

A partir de l'estimateur $H_{\mathrm{NA}}(t)$, la fonction de survie $S_{\mathrm{NA}}(t)$ peut être facilement déduite par la relation suivante :

$$S_{\mathrm{NA}}(t) = e^{-H_{\mathrm{NA}}(t)}$$

L'estimateur de Nelson-Aalen présente plusieurs avantages :

- Simplicité de calcul: Il est facile à calculer et à interpréter.
- Interprétabilité: Le cumul de la fonction de hasard peut être interprété comme le risque moyen d'un évènement survenant entre deux instants.

Dans la continuité de notre étude, nous avons pris la décision de ne pas intégrer le modèle de Nelson-Aalen pour les comparaisons ultérieures. Cette décision découle de notre constatation que le modèle de Nelson-Aalen, offre une approche relativement proche de l'estimateur de Kaplan-Meier. Ils se rapprochent tant en termes d'utilisation que de résultats finaux. En conséquence, l'inclusion du modèle de Nelson-Aalen n'apporterait pas de valeur ajoutée significative en tant qu'élément de comparaison supplémentaire par rapport à l'estimateur de Kaplan-Meier, qui a été sélectionné comme une référence pour notre analyse comparative. Ainsi, afin de maintenir la pertinence et la clarté de notre étude, nous avons choisi de nous concentrer exclusivement sur l'estimateur de Kaplan-Meier pour nos analyses ultérieures.

2.1.4. Modèle de Cox

Développé par le statisticien David Cox en 1972, le modèle de Cox (1972) permet d'examiner comment différentes variables explicatives influent sur le risque de survenance d'un évènement, tel que le décès ou la sortie de l'incapacité, tout en tenant compte de l'évolution dans le temps du risque sous-jacent. Contrairement à d'autres modèles de régression, le modèle de Cox est semi-paramétrique, il ne fait pas d'hypothèses restrictives sur la forme de la fonction de survie en fonction du temps mais l'impact des variables explicatives est lui paramétrique.

Le modèle de Cox est un modèle à risques proportionnels. Cela signifie que l'effet des variables explicatives sur le risque de survie est constant dans le temps, ou en d'autres termes, le rapport des risques entre deux individus est constant au fil du temps. Cette hypothèse de risques proportionnels est fondamentale dans le modèle de Cox et permet une interprétation directe des coefficients des variables explicatives en termes de rapport de risques. Ainsi, une augmentation d'une unité dans une variable explicative entraîne une multiplication du risque de survie par un facteur constant, quel que soit le temps écoulé depuis le début de l'observation.

Le modèle de Cox s'intéresse à la fonction de risque instantané $h(t)$, qui pour un individu i à un instant t est définie comme :

$$h_i(t) = h_0(t) \times \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip})$$

Où :

- $h_i(t)$ est la fonction de risque pour l'individu i à l'instant t.
- $h_0(t)$ est la fonction de risque de base, représentant le risque de base à l'instant t. Peut être le résultat de l'estimateur de Kaplan-Meier par exemple.
- $\beta_1, \beta_2, ..., \beta_p$ sont les coefficients de régression associés aux variables explicatives $x_{i1}, x_{i2}, ..., x_{ip}$.

Ce modèle utilise la maximisation de la fonction de vraisemblance partielle $\mathcal{L}(\beta)$ qui est définie comme :

$$\mathcal{L}(\beta) = \prod_{i \text{ non censuré}} \frac{\exp(\beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip})}{\sum_{j \in R(t_i)} \exp(\beta_1 x_{j1} + \beta_2 x_{j2} + ... + \beta_p x_{jp})}$$

Où :

- $R(t_i)$ est l'ensemble des individus à risque à l'instant t_i, c'est-à-dire ceux qui n'ont pas encore subi l'évènement ou ont été censurés à l'instant t_i.
- $\beta_1, \beta_2, \ldots, \beta_p$ sont les coefficients de régression à estimer.

L'estimateur des coefficients de régression $\hat{\beta}$ est obtenu en maximisant la fonction de vraisemblance partielle. Cela peut être réalisé à l'aide de méthodes d'optimisation telles que la méthode du maximum de vraisemblance ou les méthodes itératives généralisées. Une fois les coefficients estimés, ils peuvent être utilisés pour prédire le risque de survie pour de nouveaux individus en utilisant la fonction de risque instantané.

Pendant la poursuite de notre étude, nous avons décidé de ne pas retenir le modèle de Cox, principalement en raison de son incapacité à prendre en compte les interactions entre les variables. En particulier, nous nous intéressions grandement à la possibilité de modéliser une interaction entre l'âge et l'ancienneté. Aussi, bien qu'il soit théoriquement possible de rendre les effets des variables explicatives dépendants du temps, cela nécessite un temps de calcul considérablement long, car le nombre d'observations est alors proportionnel au produit du nombre d'individus et du nombre d'instants de sortie, ce qui est, dans notre cas, trop exigeant en termes de ressources. Ainsi, pour ces raisons, nous avons décidé de ne pas poursuivre l'utilisation du modèle de Cox dans notre étude ultérieure

2.1.5. Conclusion sur les modèles historiques

Tout ces estimateurs présentent une limitation supplémentaire que nous n'avons pas encore évoquée : ils fournissent une estimation à un âge fixé. En d'autres termes, pour créer une table d'expérience, l'idéal serait de calculer un estimateur pour chaque âge possible. Cette approche impliquerait une quantité astronomique de données, puisque chaque estimateur utiliserait uniquement les données des individus d'un âge spécifique. Cela rendrait l'établissement d'une table d'expérience extrêmement exigeant en termes de volume de données, nécessitant un nombre considérable de sorties pour chaque âge et chaque ancienneté.

Une alternative consisterait à regrouper les données en créant des tranches d'âge, mais cette approche est déjà insatisfaisante car elle nécessite une modélisation qui ne tient pas compte des variations de risque au sein de chaque tranche. Ainsi, ces modèles imposent au minimum une segmentation des données, ce qui ne permet pas une utilisation optimale de celles-ci. De plus, la segmentation peut entraîner des résultats très différents pour des personnes ayant des âges très proches mais appartenant à des tranches d'âge différentes. Cette incohérence est précisément ce que l'on cherche à éviter.

Dans le cadre de notre étude, la quantité limitée de données sur les incapacités nous a amenés à explorer une approche innovante dans la création de tables d'expérience, qui présente de nombreux avantages : le modèle additif généralisé (GAM).

Le modèle additif généralisé constitue une alternative prometteuse qui permet de surmonter les limitations précédemment mentionnées. Contrairement aux estimateurs précédents, le GAM offre une approche flexible, permettant de modéliser la relation entre les variables explicatives et le risque d'incapacité/invalidité de manière plus complète. Il permet également de prendre en compte simultanément plusieurs dimensions, telles que l'âge, l'ancienneté et d'autres caractéristiques individuelles, tout en exploitant pleinement les données disponibles.

Dans la prochaine section, nous décrirons en détail les principes fondamentaux du modèle GAM, son processus de modélisation et ses avantages spécifiques pour la création de tables d'expérience pour le risque d'incapacité/invalidité. Nous mettrons en évidence les différences clés par rapport aux méthodes traditionnelles, ainsi que les implications et les bénéfices potentiels de cette approche novatrice dans le domaine de la construction de tables.

2.2. Lissage de Whittaker-Henderson

Avant d'aborder un modèle plus sophistiqué tel que le modèle additif généralisé, examinons une alternative plus simple : le lissage de Whittaker-Henderson. Cette approche offre une manière plus rudimentaire d'obtenir des taux lisses tout en prenant en compte la troncature et la censure. Ici les taux bruts sont obtenus en divisant le nombre de sorties par l'exposition pour chaque combinaison d'âge et d'ancienneté. Le calcul de l'exposition sera expliqué en détail dans la Section 3.6.

Le lissage de Whittaker-Henderson constitue une méthode simple mais efficace pour adoucir les variations brutales des taux bruts. Elle tente de résoudre l'irrégularité des solutions inhérentes aux estimateurs de Turnbull et de Kaplan-Meier.

Explorons brièvement le fonctionnement de cette approche en dimension 1.

Supposons que y soit un vecteur d'observations, par exemple un vecteur de taux bruts pour un âge donné. L'approche du lissage vise à minimiser un critère qui intègre à la fois la fidélité aux données et la régularité. En notant $F(y, \theta, w)$ comme le critère de fidélité aux données et w comme un vecteur de poids, le critère de fidélité s'écrit :

$$F(y, \theta, w) = \sum_{i=1}^{n} w_i \times (y_i - \theta_i)^2$$

En notant $R_z(y)$ le critère de régularité, nous avons :

$$R_z(\theta) = \sum_{i=1}^{n-z} (\Delta^z \theta)_i^2$$

Ici, Δ^z représente l'opérateur de différence d'ordre z, et $(\Delta^z y)i = \sum k = 0^z \binom{z}{k} (-1)^{z-k} y_{i+k}$. Dans la pratique z est souvent pris égal à 2.

Le critère à minimiser $M_z(y, \theta, w, h)$ est alors une combinaison de ces deux critères, avec un paramètre multiplicateur h qui gère le niveau de lissage, exprimé de la manière suivante :

$$M_z(y, \theta, w, h) = F(y, \theta, w) + h \times R_z(\theta) \tag{2.2}$$

Ce critère peut être exprimé matriciellement en utilisant les notations suivantes :

- $W = \text{Diag}(w)$, la matrice diagonale des poids de dimensions $n \times n$.
- $D_{n,z}$, la matrice des différences d'ordre z de dimensions $(n - z) \times n$.

Le critère à minimiser Équation 2.2 peut alors être réécrit comme suit :

$$M_z(y, \theta, w, h) = (y - \theta)^\top W (y - \theta) + h \times \theta^\top D_{n,z}^\top D_{n,z}\, \theta$$

En notant alors $S_h = h \times D_{n,z}^\top D_{n,z}$ la matrice de pénalité, nos taux lisses peuvent être exprimés de cette manière :

$$\widehat{y} = \underset{\theta}{\operatorname{argmin}} \{(y - \theta)^\top W (y - \theta) + \theta^\top S_h\, \theta\}$$

Par un simple calcul de dérivé, nous pouvons obtenir une solution explicite suivante :

$$\widehat{y} = (W + S_h)^{-1} W\, y$$

Il conviendra de s'assurer que la matrice $(W + S_h)$ est bien inversible afin d'obtenir les taux lisses.

Il est important de noter que, bien que cette méthode produise des résultats lisses, sa simplicité peut limiter sa capacité à traiter des structures de données complexes et notamment elle ne permet pas d'intégrer des variables explicatives, tout comme les estimateurs de Turnbull et de Kaplan-Meier.

2.3. Modèle additif généralisé (GAM)

Finalement, nous détaillons la construction des modèles additifs généralisés, en fournissant les clés de compréhension essentielles pour leur utilisation. Ensuite, dans le contexte de l'incapacité, nous mettons en évidence le lien naturel qui existe entre les modèles additifs généralisés et la problématique d'estimation de la fonction de survie pour les personnes en incapacité. Ce lien, combiné aux diverses qualités des GAM, positionne ces derniers comme des modèles pertinents pour la construction de tables de maintien.

2.3.1. Introduction aux GAM

Un modèle additif généralisé est très similaire à un modèle linéaire généralisé, tant dans son expression que dans son utilisation. Son expression générale est la suivante :

$$g(\mu_i) = X_i \theta + f_1(x_{1i}) + \cdots + f_n(x_{ni})$$

Où $\mu_i = \mathbb{E}(Y_i)$, Y_i représente la variable réponse suivant une loi de la famille exponentielle, X_i est une ligne de la matrice du modèle paramétrique, θ est le vecteur de paramètres associé et f_i sont les fonctions régulières des covariables.

L'idée du modèle est que la relation entre $g(\mu_i)$ et les covariables n'est plus obligatoirement linéaire. Elle peut prendre une forme quelconque, qui sera représentée par les fonctions f_i.

Ce qui permet à ce modèle d'utiliser les méthodes classiques et bien connues des GLM est de représenter les fonctions régulières comme des combinaisons linéaires d'une base de fonctions connue et choisie. Nous avons donc :

$$f_j(x_{ji}) = \sum_{l=1}^{q_j} b_{lj}(x_{ji})\beta_l^j$$

De cette façon il est clair que nous obtenons un modèle linéaire que nous pouvons calibrer. Cependant il nous reste 2 problèmes : le choix des fonctions base ainsi que la valeur de q_j la dimension de la base.

On s'adresse d'abord au problème du choix de q_j. Pour cela, on utilise l'idée du lissage qui est de pénaliser l'irrégularité de la solution. Ainsi, là où dans un simple modèle linéaire nous minimisons par rapport à β :

$$||y - X\beta||^2$$

Nous souhaitons minimiser :

$$||y - X\beta||^2 + \sum_{i=1}^{n} \lambda_i \int_\Omega f_i''(x_{ji})dx_{ji}$$

Cette expression pénalise en effet les fonctions dont la dérivée seconde n'est pas nulle, c'est à dire que les fonctions affines ne sont pas pénalisées. On souligne que chaque fonction peut se voir attribuer un facteur de pénalisation λ_i différent.

Comme les fonctions f_j sont linéaires en β on peut toujours écrire cette intégrale sous la forme $\lambda\, \beta^\top S_j \beta$ où S_j est une matrice connues dépendant du choix de la base.

Ainsi le problème d'estimer le degré de lissage du modèle est maintenant le problème d'estimer les λ_i les paramètres de lissage. En effet en définissant lamatrice de pénalisation $S = \sum_{j=1}^{n} \lambda_j S_j$ notre problème se ré-écrit :

$$||y - X\beta||^2 + \beta^\top S\beta = \left\| \begin{bmatrix} y \\ 0 \end{bmatrix} - \begin{bmatrix} X \\ B \end{bmatrix} \beta \right\|^2$$

Où B est une matrice telle que $B^\top B = S$. On obtient ainsi de nouveau un modèle linéaire pénalisé. Etant donné les λ_i, le vecteur de paramètre β est estimé par un algorithme itératif (penalized iteratively re-weighted least squares) dont le détail est décrit dans Wood (2017).

On note que si $\lambda_i \to \infty$ alors f_i sera une droite et si $\lambda_i = 0$ alors on retrouve le problème de minimisation d'un modèle linéaire.

Les critères de sélection des λ_i se divisent en deux principales catégories. La première vise à minimiser l'erreur de prédiction du modèle en optimisant des critères tels que le critère d'information d'Akaike (AIC), la validation croisée ou la validation croisée généralisée (GCV). La deuxième considère les fonctions lisses comme des effets aléatoires, de sorte que les λ_i deviennent des paramètres de variance pouvant être estimés par maximum de vraisemblance (ML) ou par maximum de vraisemblance restreint (REML). Les différences entre ces deux approches ne sont pas l'objet de ce mémoire, et nous laissons au lecteur le soin d'en apprendre davantage dans Wood (2011). Cependant, selon les comparaisons effectuées par Reiss et Todd

Ogden (2009) entre la méthode REML et la méthode GCV, il semble que la méthode REML soit préférable. En effet, elle pénalise davantage l'overfitting, présente une variabilité plus faible du paramètre de lissage et une tendance réduite à produire plusieurs minima. C'est pourquoi, dans la suite de ce mémoire, nous utiliserons systématiquement la méthode REML.

Il reste désormais à aborder la question du choix de la base.

Tout d'abord, il est important de souligner que le choix de la base n'est heureusement pas d'une importance cruciale. Les différences de résultats entre deux choix de bases ne sont pas significatives, sauf dans des cas particuliers tels que les données cycliques.

Sans entrer dans les détails, nous pouvons mentionner différentes bases qui méritent d'être explorées :

Les splines de lissage cubiques

L'expression de la base des splines de lissage n'est pas nécessaire pour notre étude, mais elle est disponible dans Hastie et al. (2009) section 5.2. En revanche, son origine et sa propriété fondamentale nous intéressent. En effet, la base des splines de lissage cubique est intégrée dans la définition de notre objectif : minimiser $\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int g''(x)^2 dx$.

La solution g de ce problème est un spline de lissage, et ainsi la base des splines cubiques découle naturellement de la spécification de l'objectif de lissage. Elle n'est pas choisie à l'avance, mais émerge plutôt de la recherche de la fonction qui minimise notre objectif.

Dans la pratique, la base utilisée diffère légèrement de celle-ci, étant plutôt appelée "cubic regression splines". Leur construction est plus élaborée, ce qui permet d'éviter l'inconvénient des splines de lissage cubiques, où chaque observation ajoute un degré de liberté. Cela est réalisé en sélectionnant un nombre réduit de points de données, tout en veillant à ce que les valeurs des covariables dans cet ensemble plus restreint soient distribuées de manière cohérente pour refléter la distribution des covariables dans l'ensemble de données d'origine.

Les P-splines

Pour la construction des P-splines, on se place dans le cadre des splines de degré m (par exemple, $m = 2$ pour des splines cubiques). Soit $k + m + 1$ nœuds notés $x_1, ..., x_{k+m+1}$. Une base B-splines pour représenter une fonction avec k B-splines est alors donnée par :

$$\forall i \in \{1, ..., k\} \quad B_i^m(x) = \frac{x - x_i}{x_{i+m+1} - x_i} B_i^{m-1}(x) + \frac{x_{i+m+2} - x}{x_{i+m+2} - x_{i+1}} B_{i+1}^{m-1}(x)$$

et $B_i^{-1}(x) = \mathbb{I}_{x_i < x < x_{i+1}}$

Nous pouvons donc écrire $f(x) = \sum_{i=1}^{k} \beta_i B_i^m(x)$

Les P-splines sont définies lorsque les nœuds sont espacés régulièrement. Elles utilisent une pénalité sur les paramètres β_i pour contrôler l'irrégularité de la solution.

Pour illustrer, en utilisant une pénalité portant sur le carré des différences en β_i, nous obtenons une pénalité $P = \sum_{i=1}^{k-1}(\beta_{i+1} - \beta_i)^2$, qui s'écrit matriciellement comme suit :

$$P = \beta^\top \begin{bmatrix} 1 & -1 & 0 & . & . \\ -1 & 2 & -1 & . & . \\ 0 & -1 & 2 & . & . \\ . & . & . & . & . \\ . & . & . & . & . \end{bmatrix} \beta$$

Ce qui nous donne comme matrice de pénalité pour calibrer le modèle : $S = P^\top P$.

Nous devons aborder un dernier point avant de conclure notre brève présentation des modèles GAM, qui concerne l'estimation de fonctions lisses à plusieurs paramètres. En effet, il peut être nécessaire de modéliser une fonction régulière dépendant de plusieurs variables, comme un terme d'interaction par exemple.

Nous allons décrire brièvement le principe de construction des fonctions lisses à plusieurs paramètres, mais pour plus de détails, je vous invite à consulter la section 4.1.8 de l'ouvrage de Wood (2017).

Sans perte de généralité, prenons l'exemple de deux fonctions marginales, puis construisons la fonction f_{xy} représentant l'interaction. Ainsi, nous avons les fonctions f_x et f_y qui peuvent s'écrire de la manière suivante :

$$f_x(x) = \sum_{i=1}^{I} \alpha_i a_i(x) \qquad \text{et} \qquad f_y(y) = \sum_{k=1}^{K} \beta_k b_k(y)$$

Où les α_i et β_k sont les paramètres, et a_i et b_k sont les fonctions de bases connues.

La fonction f_x est déjà régulière selon x, donc notre objectif est de partir de cette expression et construire une fonction f_{xy} qui soit également régulière selon y. Pour cela, nous permettons au paramètre α_i de varier de manière régulière selon y, ce qui donne $\alpha_i(y)$. Étant donné que nous connaissons l'expression des fonctions régulières selon y, nous obtenons $\alpha_i(y) = \sum_{k=1}^{K} \beta_{ik} b_k(y)$. Ainsi, nous obtenons l'expression de f_{xy} suivante :

$$f_{xy}(x, y) = \sum_{i=1}^{I} \sum_{k=1}^{K} \beta_{ik} b_k(y) a_i(x)$$

Ainsi, les paramètres deviennent les β_{ik} et les fonctions de base pour l'interaction sont le produit des fonctions de base marginale. Nous disposons désormais de tous les éléments nécessaires pour calibrer un modèle GAM.

Dans la pratique, pour estimer les coefficients, le package R `mgcv` utilise un algorithme itératif appelé P-IRLS (Penalized Iteratively Reweighted Least Squares), dont les détails sont disponibles dans la section 3.4 de Wood (2017).

2.3.2. Lien avec le risque incapacité

Dans cette section, nous établissons le lien entre le risque d'incapacité que nous étudions et les modèles additifs généralisés. Ce lien découle d'une relation naturelle que nous allons expliciter ici. Cette étape est largement inspirée de Biessy (2022).

Commençons par introduire les notations utilisées :

- Anc_i représente la caractéristique donnant l'ancienneté de l'individu i. Cette caractéristique dépend du temps.
- C_i désigne l'ensemble des caractéristiques constantes dans le temps pour l'individu i, telles que le sexe, la catégorie socio-professionnelle (CSP) ou encore l'âge à l'entrée en incapacité..
- χ_i représente l'ensemble des caractéristiques de l'individu i, c'est-à-dire $\chi_i = C_i, \text{Anc}_i$. Notons que $\chi_i + t = C_i, \text{Anc}_i + t$ puisque C_i ne dépend pas du temps.
- T_χ désigne la durée de l'arrêt d'un individu ayant les caractéristiques χ.
- $S_\chi(t)$ est la fonction de survie associée à l'instant t.
- $f_{\chi_i}(t)$ représente la densité associée liée à la fonction de survie par la relation $f_\chi(t) = -\frac{d}{dt}S_\chi(t)$.
- μ_χ désigne la force de sortie instantanée définie par $\mu_\chi = \frac{1}{2}\lim_{h\to 0}\mathbb{P}(t \leq T_\chi < t + h)$.
- δ_i est l'indicateur de censure, valant 1 s'il n'y a pas de censure et 0 sinon.

Nous pouvons établir facilement les deux relations suivantes :

$$f_\chi(t) = S_\chi(t)\mu_{\chi+t} \qquad et \qquad S\chi(t) = \exp\left(-\int_{u=0}^{t}\mu_{\chi+u}du\right) \tag{2.3}$$

Finalement, il nous reste à écrire la log-vraisemblance :

$$\begin{aligned} l(\theta) &= \log\left\{\prod_{i=1}^{n} f_{\chi_i}(t_i,\theta)^{\delta_i} S_{\chi_i}(t_i,\theta)^{1-\delta_i}\right\} \\ &= \sum_{i=1}^{n}\left\{\delta_i \log\mu_{\chi_i+ti}(\theta) - \int_{u=0}^{t_i}\mu_{\chi_i+u}(\theta)du\right\} \end{aligned}$$

où l'on a utilisé la notation donnée par Équation 2.3 pour exprimer la vraisemblance uniquement en fonction de la force de sortie instantanée et avec la convention suivante :

- χ_i correspond aux caractéristiques de l'individu **au début de son observation**, c'est-à-dire passé le délai de franchise.
- t_i est la **durée d'observation** de l'individu i.

Il est essentiel de souligner qu'à ce stade, aucune hypothèse n'a été faite, à l'exception de l'hypothèse classique d'indépendance entre la variable étudiée et sa censure. Nous allons maintenant procéder à notre première et unique hypothèse concernant la force de sortie instantanée : nous supposerons qu'elle est constante par morceaux entre 2 âges entiers et 2 mois d'ancienneté consécutifs. Concrètement, cela signifie que $\mu_{\chi+\epsilon} = \mu_\chi; \forall\epsilon \in [0;1[$. Cette hypothèse permet, d'une part, de rester cohérent avec la construction de tables donnant le nombre de survivants par âge et par ancienneté, et d'autre part, de faciliter le calcul suivant :

$$
\begin{aligned}
l(\theta) &= \sum_{i=1}^{n} \left\{ \delta_i \log \mu_{\chi_i+t_i}(\theta) - \int_{u=0}^{t_i} \mu_{\chi_i+u}(\theta) du \right\} \\
&= \sum_{i=1}^{n} \left\{ \left[\sum_{\chi} \mathbb{1}_{\{\chi \leq \chi_i+t_i < \chi+1\}} \right] \delta_i \log \mu_{\chi_i+t_i}(\theta) - \int_{u=0}^{t_i} \left[\sum_{\chi} \mathbb{1}_{\{\chi \leq \chi_i+t_i < \chi+1\}} \right] \mu_{\chi_i+u}(\theta) du \right\} \\
&= \sum_{\chi} \left\{ \log \mu_{\chi}(\theta) \sum_{i=1}^{n} \mathbb{1}_{\{\chi \leq \chi_i+t_i < \chi+1\}} \delta_i - \mu_{\chi}(\theta) \sum_{i=1}^{n} \int_{u=0}^{t_i} \mathbb{1}_{\{\chi \leq \chi_i+u < \chi+1\}} du \right\} \\
&= \sum_{\chi} \{ d_{\chi} \log \mu_{\chi}(\theta) - \mu_{\chi}(\theta) e_{\chi} \}
\end{aligned}
$$

Avec :

- d_{χ} le nombre de sorties observées qui ne sont pas des censures
- e_{χ} l'exposition au risque, dont les détails du calcul sont expliqués dans la Section 3.6.

L'intérêt de ce calcul réside, d'une part, dans l'agrégation des données par combinaison des caractéristiques χ des individus, et d'autre part, dans la considérable réduction du temps de calcul pour la log-vraisemblance. En effet, précédemment, la vraisemblance se calculait par une somme sur tous les individus, tandis que désormais, il s'agit d'une somme sur les combinaisons possibles des caractéristiques, ce qui réduit significativement le nombre de termes à considérer.

C'est à ce stade que le lien avec les modèles additifs généralisés (GAM) peut être établi. En effet, si nous écrivons la log-vraisemblance d'un GAM Poisson, c'est-à-dire en faisant l'hypothèse que le nombre de sorties suit une loi de Poisson dont le paramètre est proportionnel d'une part à l'exposition et d'autre part au taux de sortie instantané, nous obtenons :

$$
\begin{aligned}
l(\theta) &= \log \left\{ \prod_{\chi} \mathbb{P}(D_{\chi} = d_{\chi}) \right\} \\
&= \log \left\{ \prod_{\chi} e^{\mu_{\chi}(\theta) e_{\chi}} \frac{(\mu_{\chi}(\theta) e_{\chi})^{d_{\chi}}}{d_{\chi}!} \right\} \\
&= \sum_{\chi} -\mu_{\chi}(\theta) e_{\chi} + d_{\chi} \log(\mu_{\chi}(\theta)) + d_{\chi} \log(e_{\chi}) - \log(d_{\chi}!)
\end{aligned}
$$

Nous remarquons que la somme porte sur les combinaisons car les réalisations de la loi de Poisson portent sur chaque combinaison. Ensuite, nous notons également que parmi les quatre termes de la somme, seuls deux nous intéressent dans le but de la maximiser par rapport à θ. En gardant uniquement ces termes, nous obtenons la même log-vraisemblance à maximiser que celle obtenue en faisant l'hypothèse des taux de sortie constants entre deux âges et deux anciennetés consécutifs.

L'ajout du GAM intervient dans la pénalisation de cette vraisemblance afin d'obtenir une solution lisse, plus proche de la réalité, car il est évident que notre taux de risque instantané réel ne saute pas d'âge en âge et d'ancienneté en ancienneté. En fin de compte, l'utilisation d'un GAM permet de maximiser la même vraisemblance que celle d'un modèle faisant uniquement l'hypothèse du taux de risque instantané constant entre deux changements d'âge ou

d'ancienneté, tout en ajoutant une pénalisation à celle-ci afin de rechercher une solution lisse, se rapprochant ainsi du véritable taux recherché.

Dans cette section consacrée au cadre théorique, nous avons exploré en détail les méthodes classiques et les approches alternatives pour modéliser les taux de sorties dans le contexte de la création de tables d'expérience pour le risque d'incapacité/invalidité. Cette exploration nous a permis de mettre en évidence les avantages et les limitations inhérents à chaque méthode, jetant ainsi les bases de notre choix du modèle additif généralisé (GAM).

Nous avons commencé par examiner l'estimateur de Turnbull, une méthode largement utilisée pour la construction de tables réglementaires. Cependant, sa nature discrète et son incapacité à tenir compte des variables explicatives en font une option moins adaptée pour des situations complexes, comme la prise en compte de la franchise.

Nous avons également évoqué l'estimateur de Kaplan-Meier, qui constitue un pilier de l'analyse de survie. Bien qu'efficace pour les données de survie, il n'est pas aussi flexible pour la prise en compte des variables explicatives dans le contexte de la construction de tables d'expérience pour le risque d'incapacité/invalidité.

Le lissage de Whittaker-Henderson, présenté ensuite, offre une approche simple pour lisser les taux bruts en prenant en compte la troncature et la censure. Bien qu'efficace pour réduire les variations brutales, il partage la limitation des estimateurs précédents en ne permettant pas d'intégrer de manière flexible des variables explicatives.

Enfin, nous avons abordé en profondeur le modèle additif généralisé (GAM), une solution novatrice pour aborder les problématiques complexes de censure, troncature et prise en compte des variables explicatives. En se basant sur la vraisemblance et en tenant compte de l'hypothèse de taux constants entre intervalles d'âge et d'ancienneté, les GAM offrent une approche souple et puissante pour modéliser les taux de sorties. Leur capacité à traiter des données avec lissage, tout en conservant une flexibilité pour inclure des variables explicatives, en fait une option privilégiée pour la construction de tables d'expérience.

Pour résumer ces distinctions, la Table 2.1 ci-dessous récapitule les limitations majeures de chaque modèle :

Table 2.1. : Résumé des avantages et inconvénients des différents modèles

Critère	Turnbull	Kaplan-Meier	Lissage WH	Cox	GAM
Discrétisation	oui	non	oui	non	oui
Données incomplètes	oui	oui	oui	oui	oui
Résultat lisse	non	non	oui	non	oui
Covariables	non	non	non	oui	oui

Les critères sont à comprendre de la manière suivante :

- Discrétisation : Le modèle utilise des données agrégées comme entrée.
- Données incomplètes : Le modèle gère correctement les données censurées et/ou tronquées.
- Résultat lisse : Le modèle produit un résultat lisse même avec une quantité de données limitée.
- Covariables : Le modèle peut prendre en compte des variables explicatives.

Chapitre 3.

Données et méthodologies

3.1. Présentation des données

Les données utilisées dans le cadre de cette étude pour la construction du modèle comprennent l'ensemble des prestations enregistrées sur une période allant du 01/01/2019 au 01/03/2023 d'une mutuelle française. La base de données initiale regroupe des prestations liées à l'incapacité, l'invalidité, le décès, ainsi que les exonérations de cotisations. Elles sont issues d'une extraction du système de gestion à l'échelle de périodes d'indemnisations, puis le résultat d'une agrégation des différentes périodes d'indemnisations liées au même sinistre.

Plus spécifiquement, notre intérêt se porte sur les prestations liées à l'incapacité, qui représentent un total de 10 129 lignes. Chaque ligne correspond à la prestation totale d'un arrêt en incapacité, caractérisée par 20 variables.

Nous donnons une description plus précise des variables que nous utiliserons par la suite :

- **CONTRAT** : Cette variable indique le numéro du contrat d'assurance auquel est rattachée la prestation en incapacité. Elle permet de lier la prestation aux caractéristiques spécifiques du contrat, mais aussi aux éventuelles prestations d'un autre type, par exemple un décès ou une invalidité. Elle sera utile pour déterminer la cause de sortie de l'incapacité.
- **TITRE** : Cette variable représente le titre de la personne bénéficiaire de la prestation, telle que M. (Monsieur) ou Mme (Madame). Elle permet d'identifier le genre de l'assuré et peut être utilisée pour étudier les différences de risque entre les hommes et les femmes.
- **DATE_NAISSANCE** : Cette variable indique la date de naissance de l'assuré concerné par la prestation en incapacité. Elle sera utilisée pour calculer l'âge de l'assuré au moment de la survenance de l'incapacité.
- **DATE_SURVENANCE** : Cette variable correspond à la date de survenance de l'incapacité, c'est-à-dire le début de la période d'arrêt de travail. Elle permet de déterminer le début de l'arrêt. et peut éventuellement détecter des erreurs avec la variable suivante.

- **DEBUT_INDEMNISATION** : Cette variable détermine également le début de l'arrêt de travail. Elle permettra éventuellement de détecter des erreurs avec la variable précédente.
- **FIN_INDEMNISATION** : Cette variable représente la date à laquelle l'indemnisation de l'assuré pour son arrêt en incapacité s'est terminée. Elle permet de calculer la durée totale de l'arrêt.
- **DEBUT_INVAL** : Cette variable indique la date à laquelle l'assuré a été reconnu en situation d'invalidité, le cas échéant. Elle est égale à **DATE_SURVENANCE** s'il n'y a pas d'invalidité.
- **MONTANT** : Cette variable représente le montant de l'indemnité versée à l'assuré pour son arrêt en incapacité. Elle sera utile surtout à des fins de compréhension des données. Elle permet notamment d'identifier des arrêts annulés (montant nul).
- **PRODUIT** : Cette variable fait référence au type de produit auquel le contrat est rattaché. Elle sera utilisée pour identifier différentes populations.

3.2. Traitement des données

À partir des variables mentionnées précédemment, nous avons créé les variables suivantes :

- **AGE** : Cette variable correspond à l'âge entier atteint à la survenance, calculé à partir des variables **DATE_NAISSANCE** et **DATE_SURVENANCE**.
- **SINISTRE_CLOS** : Cette variable indique si le sinistre est toujours en cours ou non au 01/03/2023. Pour déterminer si un sinistre est toujours en cours, nous avons utilisé la convention selon laquelle si une prestation a eu lieu moins de 3 mois avant la date d'extraction (01/03/2023), alors le sinistre est en cours.
- **DUREE** : Cette variable correspond à la durée, en fractions de mois, de l'arrêt. Elle est calculée à partir des variables **DATE_SURVENANCE** et **FIN_INDEMNISATION**.
- **PERIMETRE** : Cette variable segmente les contrats en contrats *collectifs* ou *TNS* (Travailleur Non Salarié). Elle permettra d'étudier les éventuelles différences de risque existant entre ces deux populations. Elle est construite à partir de l'information contenue dans la variable `PRODUIT`. Elle correspond à ce qu'on appellera dans la suite le **type de contrat**.

Les traitements suivants ont été appliqués :

- **Censure** : Toutes les dates de fin d'arrêt dépassant le 01/03/2023 ont été censurées au 01/03/2023.
- **Sinistre annulé** : Nous faisons l'hypothèse de travail qu'un sinistre dont le montant est nul est un sinistre annulé.
- **Contrat individuel** : Les sinistres attachés à des contrats individuels étant très peu nombreux, ils ont été assimilés à des contrats TNS.

3.3. Tests sur les données

Dans cette partie, nous répertorions tous les tests qui ont été effectués sur la base de données et qui ont conduit à la suppression d'une partie de celle-ci. Toutes les suppressions de données sont répertoriées dans la Figure 3.2.

Tout d'abord, nous nous sommes intéressés à l'âge des assurés. En effet, la table du BCAC de 2010 prend en compte les âges compris entre 20 ans et 67 ans, nous avons donc décidé de nous conformer à ce cadre. Nous présentons une représentation visuelle de la répartition des âges dans la base de données dans la Figure 3.1. Seuls les sinistres concernant des personnes dont l'âge est compris entre 20 et 67 ans ont été conservés.

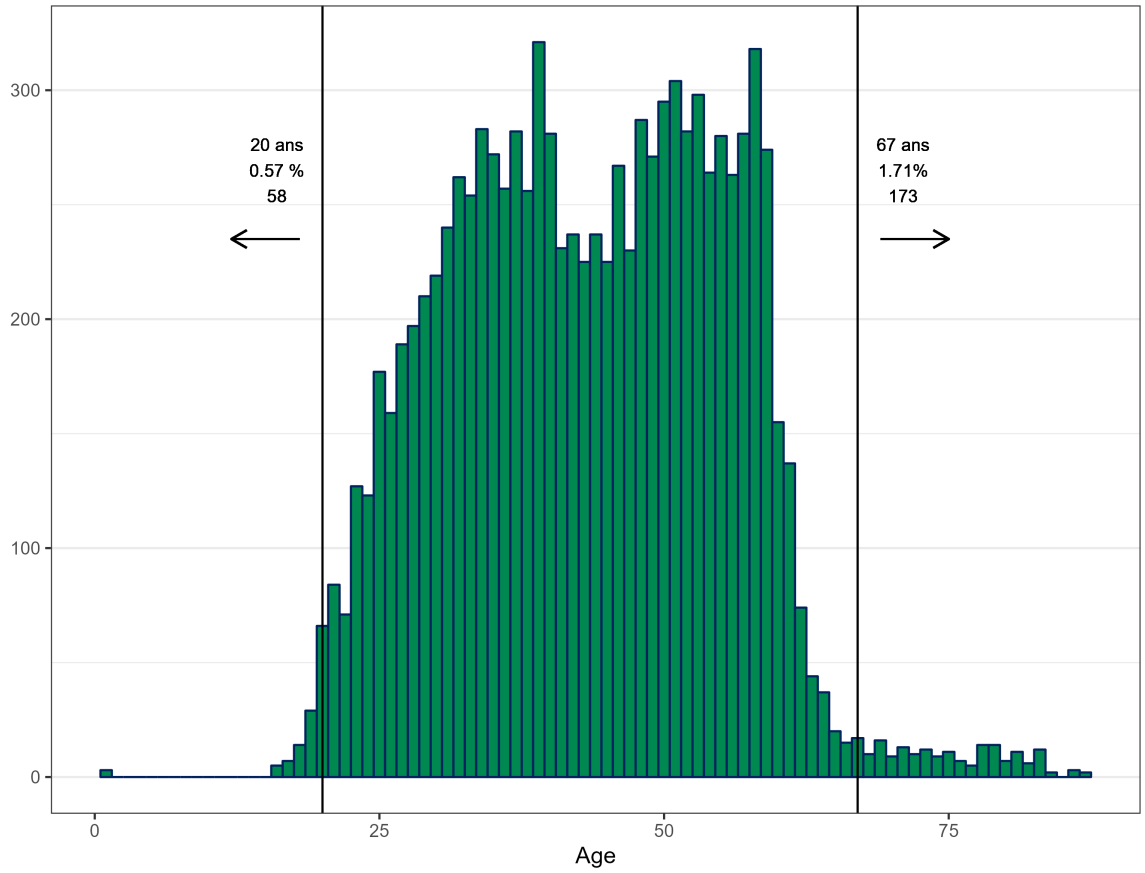


Figure 3.1. : Histogramme des âges des personnes en incapacité

Ensuite, en plus des sinistres dont le montant est nul, nous avons décidé de supprimer les arrêts dont le montant est négatif. Un montant négatif peut se produire dans le cas d'une annulation de prestations, qui a été identifiée comme un sinistre lors de l'agrégation. Cependant, ces sinistres sont en réalité des annulations d'une partie d'un vrai sinistre, et ne constituent pas de véritables sinistres.

Finalement, nous avons observé dans la base de données des arrêts dont la durée est supérieure à 36 mois (allant jusqu'à 9 ans). Il s'agit principalement de sinistres pour lesquels nous avons perdu les informations, et selon cette hypothèse, ils ont été retirés de la base de données. Ainsi, nous conservons uniquement les sinistres dont la durée est inférieure ou égale à 36 mois.

Finalement, les résultats des tests de cohérence effectués sur la base sont présentés dans la Figure 3.2. Ces tests ont entraîné une perte de 13,5 % de la base, ce qui nous laisse avec un total de 8 761 lignes.

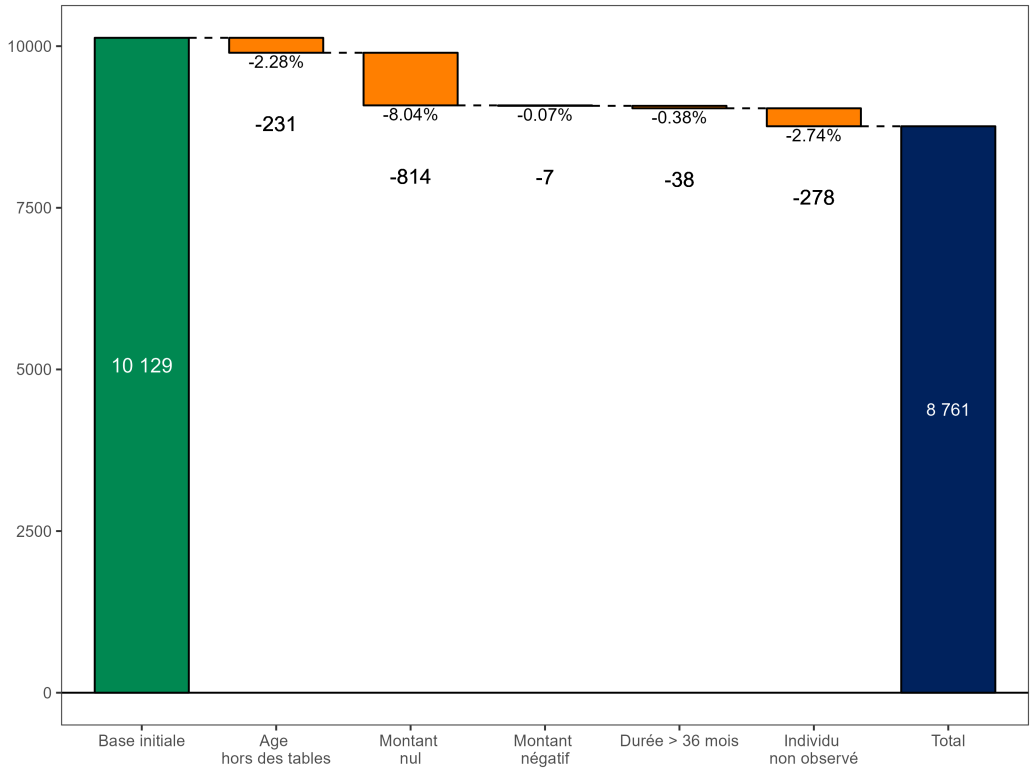


Figure 3.2. : Processus de perte des données

3.4. Statistiques decriptives

Parmi les variables de notre base de données, deux variables qualitatives retiennent notre attention : le sexe et le périmètre. La répartition de ces variables est illustrée dans la Figure 3.3. Nous remarquons une répartition relativement équilibrée pour la variable sexe, tandis que la variable périmètre présente une disparité marquée, avec une prédominance des contrats collectifs, représentant plus de 85% de la base de données. Ces variables seront utilisées par la suite pour analyser les éventuelles différences de risque entre ces populations.

Après avoir effectué une filtration de nos données, nous nous concentrons sur l'exploration de deux variables essentielles dans notre analyse : la distribution des âges (voir Figure 3.4) et la distribution des durées des arrêts (voir Figure 3.5).

La distribution des âges (Figure 3.4) nous permet d'obtenir une meilleure connaissance de la répartition des individus selon leur tranche d'âge. Nous observons que la majorité des individus se situent dans la tranche d'âge entre 25 et 60 ans, avec une diminution significative au-delà de 62 ans.

La distribution des durées des arrêts (Figure 3.5) met en évidence un sursaut des arrêts à 35 mois, ce qui fera l'objet d'une attention particulière dans la suite.

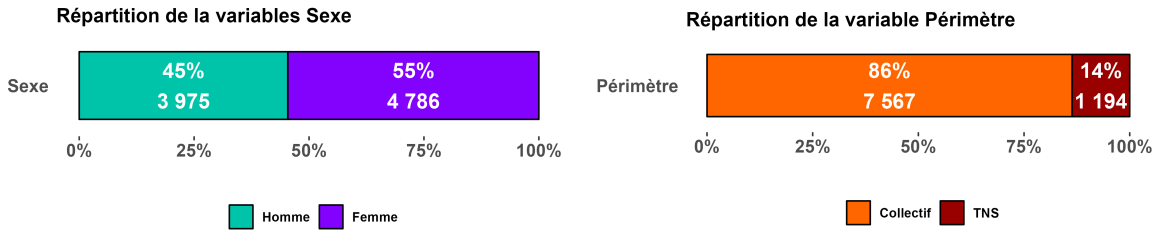


Figure 3.3. : Répartition des variables Sexe et Périmètre

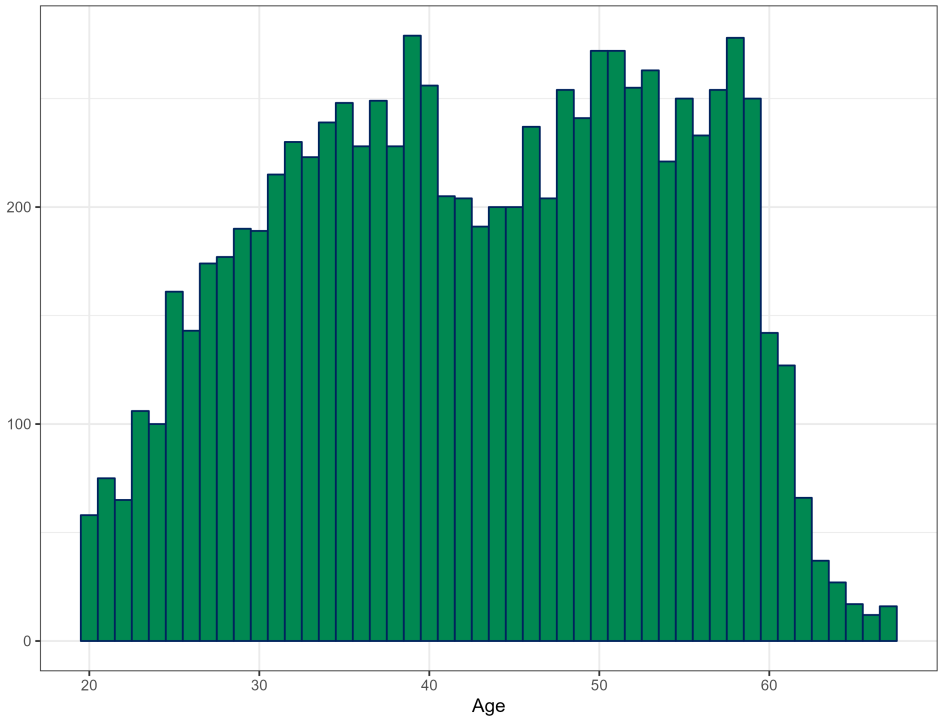


Figure 3.4. : Histogramme des âges après nettoyage des données

3.5. Imputation de la franchise : Méthodologie et résultats

Comme le montrera ce mémoire, la prise en compte de la franchise est un élément crucial dans la construction d'une table d'expérience fiable pour évaluer les risques d'incapacité et d'invalidité. Dans notre étude, nous avons effectué un travail conséquent pour retrouver cette information manquante mais essentielle.

La franchise n'étant pas une information que nous avions, nous avons du exploiter l'extraction brute du système de gestion afin de l'imputer.

Notre base de données est constituée de lignes pour chaque période d'indemnisation. Ainsi, une seule incapacité peut potentiellement être représentée par plusieurs lignes, décrivant des périodes successives. La base couvre les indemnisations ayant eu lieu au cours des années 2020, 2021 et une partie de l'année 2022.

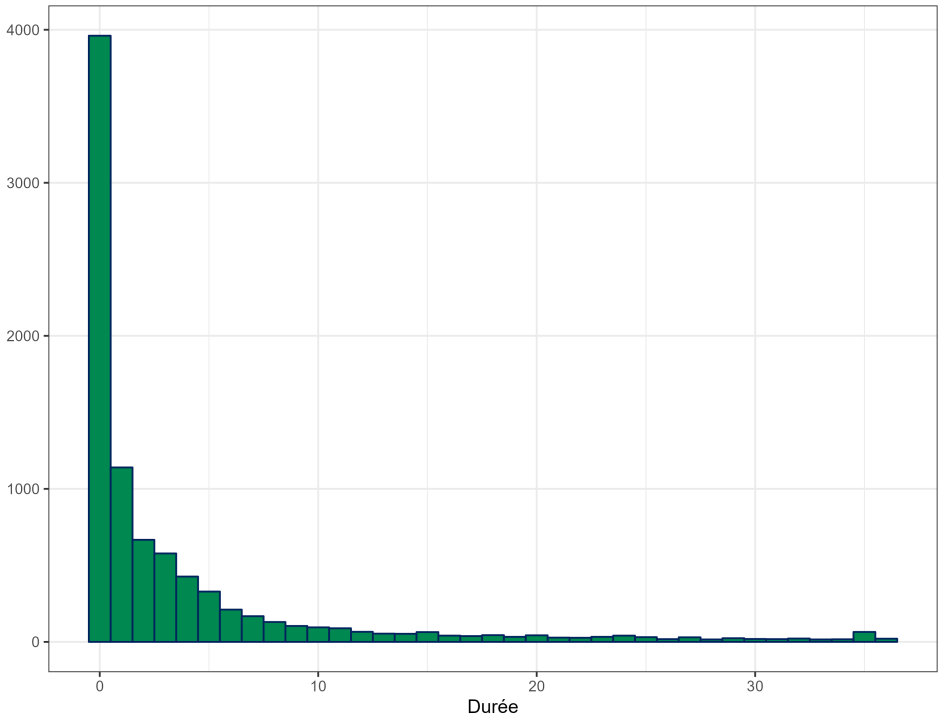


Figure 3.5. : Histogramme des durées des arrêts de travail après nettoyage des données

Pour la majorité des sinistres dans cette base, nous disposons d'une période d'indemnisation décrite par une date de début et de fin distinctes. Pour chaque période, nous possédons les informations suivantes :

- Un montant de prestations pour la période
- Une colonne `CONTEXTE` indiquant soit "Franchise", soit rien

Cependant, cette situation est spécifique aux années 2021 et 2022, car la colonne `CONTEXTE` n'est pas présente pour l'année 2020.

Ainsi, nous pouvons aisément retrouver la franchise pour les sinistres indemnisés en 2021 ou 2022.

Cependant un problème intervient : il arrive que, pour un même numéro de contrat, la méthode décrite donne différentes franchises entre un sinistre survenu en 2021 et un autre en 2022. Pour résoudre cette problématique, nous avons adopté la convention suivante :

- Un contrat est associé à une seule franchise.
- Nous avons retenu la plus grande franchise compatible avec les sinistres de ce contrat. Ainsi, si un sinistre pour un contrat a eu lieu sur une période de 1 mois, nous n'avons pas retenu de franchise dépassant 1 mois.

La répartition des franchises ainsi obtenue est illustrée dans la Figure 3.6. Ce graphique présente le nombre de jours non indemnisés, et pour obtenir la franchise du contrat, nous ajoutons ensuite un jour.

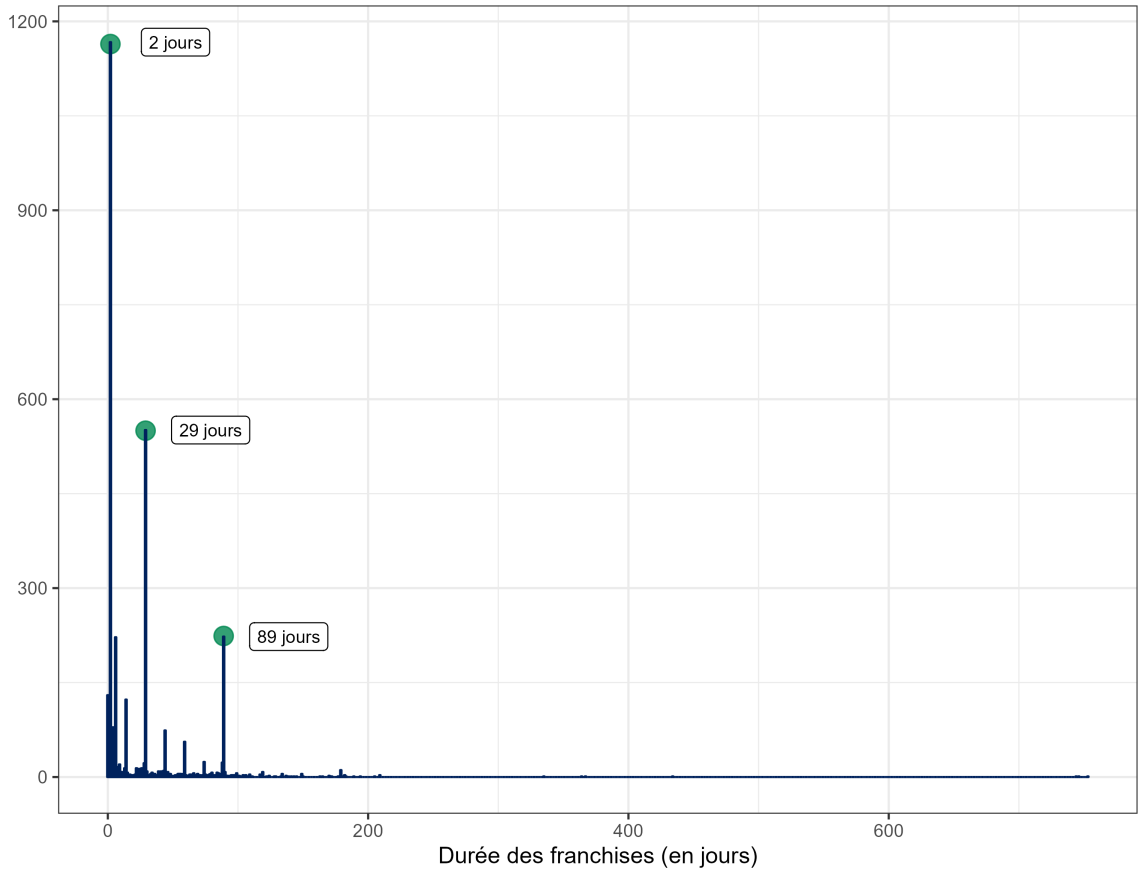


Figure 3.6. : Histogramme des franchises pour les sinistres ayant donné lieu à une prestation en 2021 ou 2022

Pour évaluer l'impact de la prise en compte de la franchise, aussi bien pour les tables que pour les provisions, nous avons créé deux bases supplémentaires en ne conservant que les sinistres pour lesquels nous disposons de l'information sur la franchise.

Ainsi, nous disposons de deux bases de données distinctes : l'une utilisant la franchise, et l'autre considérant une franchise nulle. Chacune de ces bases compte 4 647 lignes.

En comparant directement les résultats issus de ces deux bases, nous serons en mesure de mesurer précisément l'impact de la prise en compte de la franchise. Cette analyse comparative constituera une étape essentielle pour comprendre pleinement l'effet de la franchise sur nos modèles.

3.6. Construction des variables pour le GAM

Pour modéliser le risque d'incapacité par un GAM, nous avons vu que cela revient à considérer que le nombre de sorties est une variable aléatoire suivant une loi de Poisson, dont le paramètre, qui représente le nombre moyen de sorties pour l'âge et l'ancienneté considérés, est le produit

de la force de sortie (μ) et de l'exposition au risque (e), que l'on peut écrire :

$$\text{Sortie} \sim \text{Poisson}(\mu \times e)$$

L'exposition, tout comme les sorties et la force de sortie, est calculée pour chaque combinaison d'âge et d'ancienneté, et correspond au nombre de mois d'observation. Sa formulation est la suivante :

$$\begin{aligned}
e_{\text{Age,Anc}} &= \sum_i \int_{d_i}^{f_i} \mathbb{1}_{\{\text{age}_i=\text{Age}\ ,\ \text{Anc}\leq u\leq \text{Anc}+1\}} du \\
&= \sum_i \int_{d_i}^{f_i} \mathbb{1}_{\{\text{age}_i=\text{Age}\}} \mathbb{1}_{\{d_i\leq f_i\}} \mathbb{1}_{\{\text{Anc}\leq u\leq \text{Anc}+1\}} du \\
&= \sum_i \mathbb{1}_{\{\text{age}_i=\text{Age}\}} \int_{\max(d_i,\text{Age})}^{\min(f_i,\text{Age}+1)} 1 du \\
&= \sum_i \mathbb{1}_{\{\text{age}_i=\text{Age}\}} \left[\min(f_i, \text{Age}+1) - \max(d_i, \text{Age})\right]^+
\end{aligned}$$

Avec les notations suivantes :

- Age : l'âge d'entrée en incapacité considéré pour le calcul
- Anc : l'ancienneté considérée pour le calcul
- d_i : l'ancienneté exacte au début de l'observation
- f_i : l'ancienneté exacte à la fin de l'observation
- age_i : l'âge atteint à l'entrée en incapacité de l'individu i
- a^+ : Vaut a si $a > 0$ et 0 sinon

Nous souhaitons appliquer le modèle additif généralisé suivant à nos données :

$$\log(\mathbb{E}(\text{Sortie})) = \log(\mu_{\text{Age,Anc}} * e_{\text{Age,Anc}}) = f(\text{Age}, \text{Anc})$$

Dans ce modèle, nous cherchons à modéliser la relation entre la variable de sortie (Sortie) et les variables explicatives (Age et Ancienneté). Nous utilisons un modèle GAM qui nous permet de capturer les effets non linéaires et les interactions potentielles entre les variables.

Et la meilleur manière (meilleure au sens de l'AIC ainsi que du coefficient de R2 ajusté) de modéliser la fonction f dans un GAM est la suivante :

$$\log(\mathbb{E}(\text{Sortie})) = \beta_0 + f_1(\text{Age}) + f_2(\text{Anc}) + f_3(\text{Age,Anc})$$

Où :

- β_0 : représente l'effet constant, indépendamment des autres variables sur la variable Sortie.
- f_1 : représente l'effet de l'âge sur la variable Sortie.
- f_2 : représente l'effet de l'ancienneté sur la variable Sortie.

- f_3 : représente l'effet de l'interaction entre l'âge et l'ancienneté sur la variable Sortie.

La décomposition de la fonction f en ces trois composantes nous offre une plus grande flexibilité pour modéliser les effets spécifiques de l'âge, de l'ancienneté et de leur interaction sur la variable de sortie. De plus, cette approche permet une meilleure interprétation des résultats en identifiant les contributions individuelles de chaque composante au nombre de Sortie.

Nous avons évoqué le calcul de l'exposition sans parler de son utilisation. En reprenant l'expression de notre modèle, nous pouvons aller plus loin en utilisant la propriété fondamentale du logarithme :

$$\log(\mathbb{E}(\text{Sortie})) = log(\mu * e) = \log(\mu) + \log(e)$$

En utilisant $\log(e)$ comme un offset dans le modèle, nous sommes en mesure de calibrer directement les taux de sortie instantanés. Cette approche astucieuse offre également la possibilité de calibrer les taux de passage par rapport au BCAC, comme nous l'exposerons en détail dans la Section 4.1.2.

Finalement, une fois que nous serons en mesure de prédire les taux de sortie instantanés pour toutes les combinaisons d'âge et d'ancienneté, nous pourrons établir le lien avec une table de maintien en utilisant d'une part la relation suivante :

$$\mu_{\text{Age,Anc}} = -\frac{d}{dt} log(P_{\text{Age,Anc}})$$

et d'autre part, l'hypothèse que la force de sortie instantanée est constante entre deux ancienetés consécutives, ce qui donne :

$$P_{\text{Age,Anc}} = \exp(-\mu_{\text{Age,Anc}}) \tag{3.1}$$

Ainsi, pour pouvoir calibrer le modèle, nous devons agréger nos données au niveau de l'âge et de l'ancienneté et connaître, pour chaque combinaison existante dans la base, l'exposition ainsi que le nombre de sorties.

De cette manière, nous obtenons une base de données de 1 480 lignes et 4 colonnes pour l'ensemble des contrats, et de 1 263 lignes et 4 colonnes pour les contrats possédant une franchise connue.

3.7. Analyse de l'exposition

À ce stade de notre analyse, il est pertinent de visualiser l'exposition, car elle constitue une mesure directe de l'information disponible dans nos données. Une représentation graphique de l'exposition nous offre une meilleure compréhension de sa répartition selon les différentes combinaisons d'âge et d'ancienneté. En scrutant les variations d'exposition, nous serons en mesure d'identifier les zones où notre quantité d'informations est plus ou moins abondante, ce qui pourrait influencer la fiabilité de nos estimations et de nos conclusions. Cette analyse de l'exposition nous fournira des éléments clés pour interpréter de manière plus précise les résultats de notre modèle.

De plus, étant donné que nous utilisons différents ensembles de données pour évaluer notre modèle, notamment pour mesurer l'impact de la prise en compte de la franchise, il devient essentiel de visualiser l'exposition pour chacune de ces bases.

Rappelons brièvement les différentes bases à notre disposition :

- Base globale : Cette base inclut l'ensemble des données, sans considérer la franchise.
- Base avec franchise : Cette base se compose uniquement des sinistres pour lesquels nous disposons de l'information sur la franchise. À partir de ces données, nous obtenons deux sous-bases, l'une avec la prise en compte de la franchise et l'autre considérant la franchise nulle
- Base avec le type de contrat : Cette base est la plus fine, car elle inclut uniquement les sinistres pour lesquels la franchise est connue et prend également en compte le type de contrat. Ainsi elle tient compte de la franchise et du type de contrat.

Nous débutons en présentant l'exposition de la base globale, comme illustré dans la Figure 3.7. Cette représentation visuelle nous permet de constater la répartition typique de l'exposition, où la majorité de celle-ci se situe entre les âges de 30 et 60 ans et les anciennetés relativement faibles. Nous observons également que pour chaque âge, l'exposition diminue à mesure que l'ancienneté augmente, jusqu'à devenir très faible, voire nulle, pour les anciennetés élevées. On notera à titre de comparaison que la maximum d'exposition pour un âge et une ancienneté donnée est d'environ 200 sur la base globale.

Ensuite, nous procéderons à une comparaison de cette exposition avec la base où nous ne gardons que les sinistres pour lesquels nous connaissons la franchise. À ce stade, nous considérons la franchise nulle, afin d'analyser spécifiquement l'effet de la perte de sinistres. La Figure 3.8 illustre cette comparaison, où l'échelle reste commune et identique à celle de la base globale pour référence.

Sur la Figure 3.8, nous remarquons une baisse logique de l'exposition de manière générale. L'exposition sur les premiers mois d'ancienneté ne semble pas avoir baissé drastiquement. En revanche, pour les grandes anciennetés, nous n'avons plus d'exposition à certaines combinaisons d'âge et d'ancienneté avec la base réduite, là où elle était faible mais non nulle sur la base globale. Ceci s'explique par le fait que les arrêts longs sont rares, et donc seul un petit nombre de sinistres permettait d'obtenir de l'exposition. La perte de certains sinistres comprenait certains de ces arrêts prolongés, dont l'exposition était très faible.

Passons à présent à l'analyse de l'impact de la prise en compte de la franchise sur l'exposition. La Figure 3.9 compare l'exposition de la base réduite selon que la franchise soit prise en compte ou non. L'échelle prend cette fois si comme référence la base réduite sans prendre en compte la franchise. On note que l'exposition maximale dans ce cas est d'environ 100.

Nous remarquons que l'impact de la prise en compte de la franchise se concentre, de manière logique, sur les premiers mois d'ancienneté. En effet, la franchise a pour effet qu'une personne ne participe pas à l'exposition tant qu'elle est sous franchise. Par conséquent, la franchise n'a aucun impact sur l'exposition pour les anciennetés supérieures à la durée de la franchise. Ainsi, la franchise n'entraîne pas de perte d'information pour les arrêts de longue durée. Cependant, nous observons une réduction d'un facteur d'environ deux de l'exposition pour les anciennetés les plus faibles. Cette diminution de l'exposition ne présente pas un problème majeur, car elle affecte principalement les combinaisons d'âge et d'ancienneté où l'exposition est déjà élevée, et les données restantes restent bien plus abondantes que pour les anciennetés plus élevées.

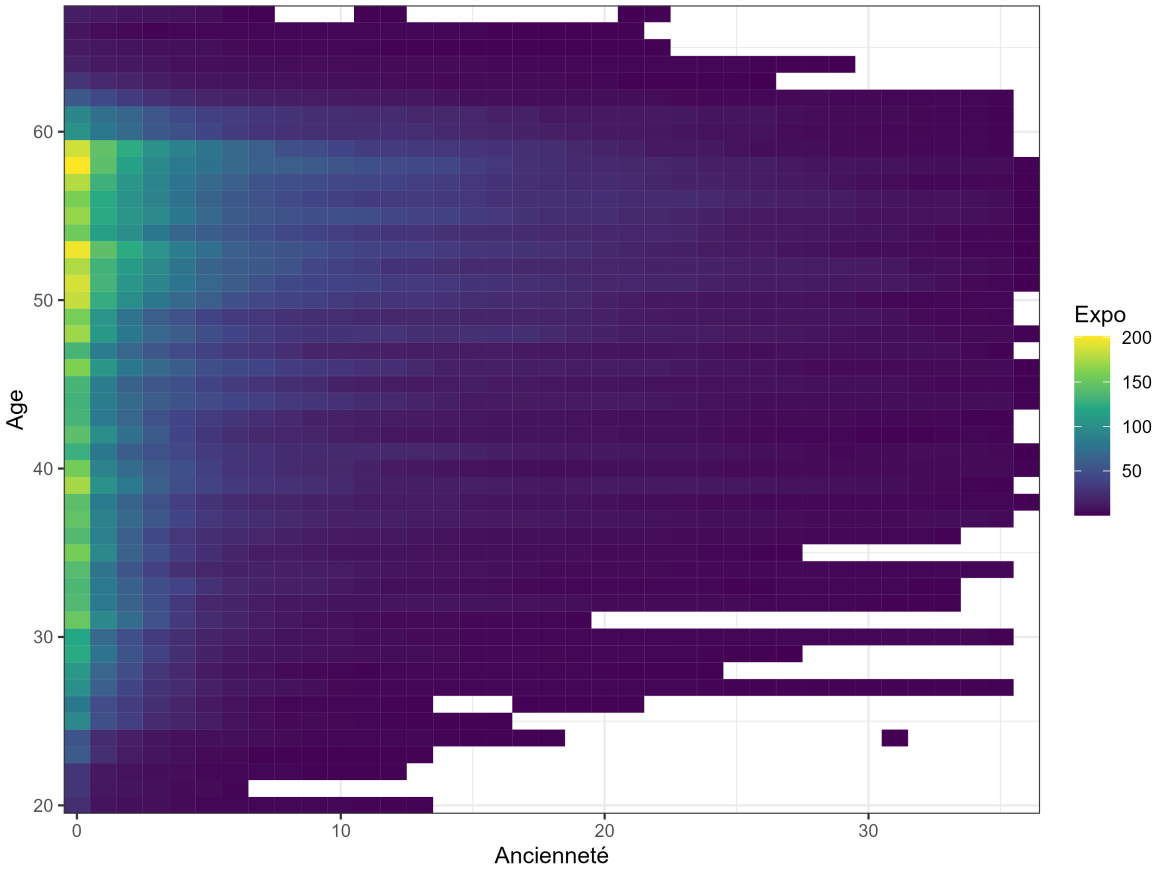


Figure 3.7. : Représentation 2D de l'exposition pour l'incapacité sur la base globale

Enfin, nous examinons la répartition de l'exposition de la base réduite en fonction de la variable `PERIMETRE`, qui représente le type de contrat. La Figure 3.10 illustre cette répartition, en présentant l'exposition pour chaque type de contrat, avec les contrats collectifs servant de référence pour l'échelle.

Nous remarquons tout d'abord que les contrats collectifs concentrent la majeure partie de l'exposition, représentant environ 90 % du total. Cette observation souligne la pertinence de choisir les contrats collectifs comme population de référence pour notre modèle, afin d'étalonner notamment le coefficient correspondant à la population des travailleurs non salariés (TNS). En outre, il est intéressant de noter que l'écart d'exposition entre les anciennetés faibles et élevées se révèle nettement moins marqué chez les TNS que chez les contrats collectifs. Cette tendance suggère que les travailleurs non salariés ont tendance à présenter des arrêts de plus longue durée.

Ainsi nous avons observé l'exposition pour toutes les bases que nous allons utiliser. Ceci nous permettra d'avoir une vision plus éclairée sur nos résultats.

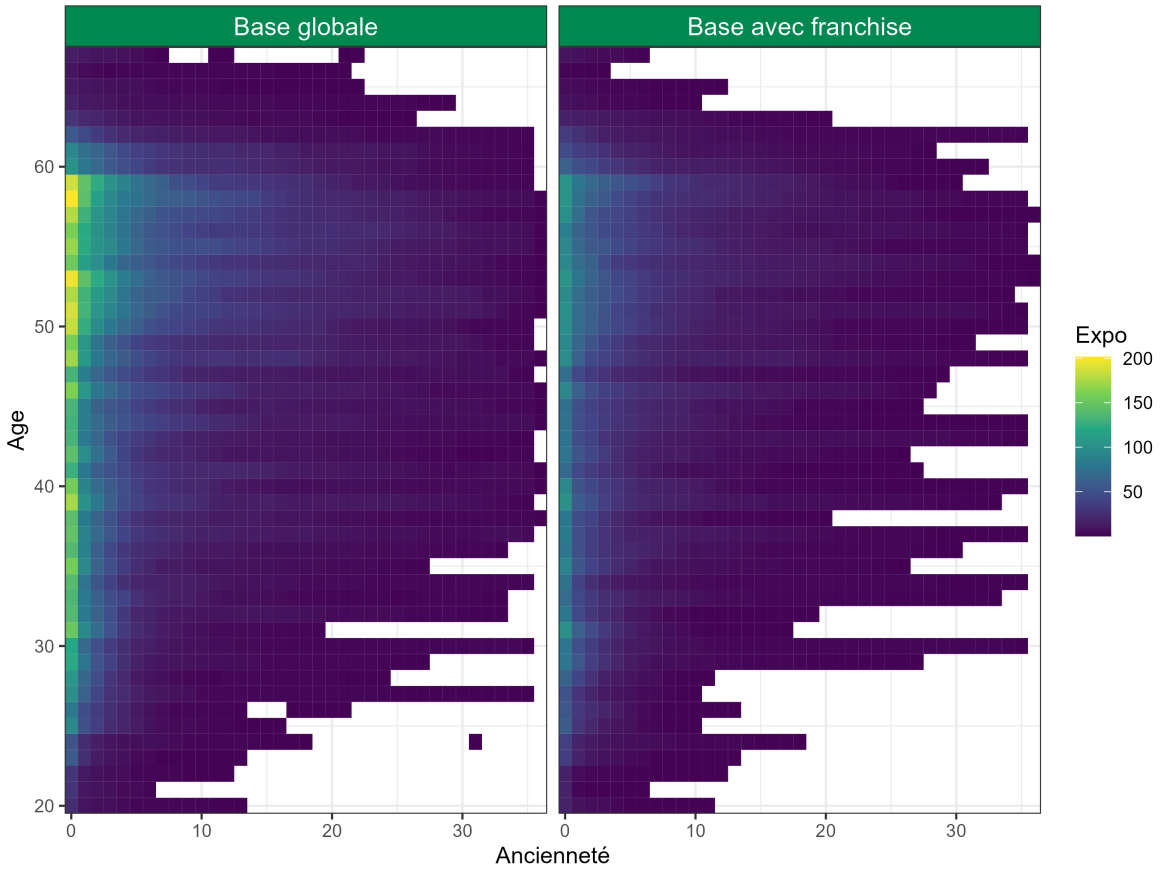


Figure 3.8. : Comparaison 2D de l'exposition entre la base globale et celle avec franchise, où la franchise est considérée nulle

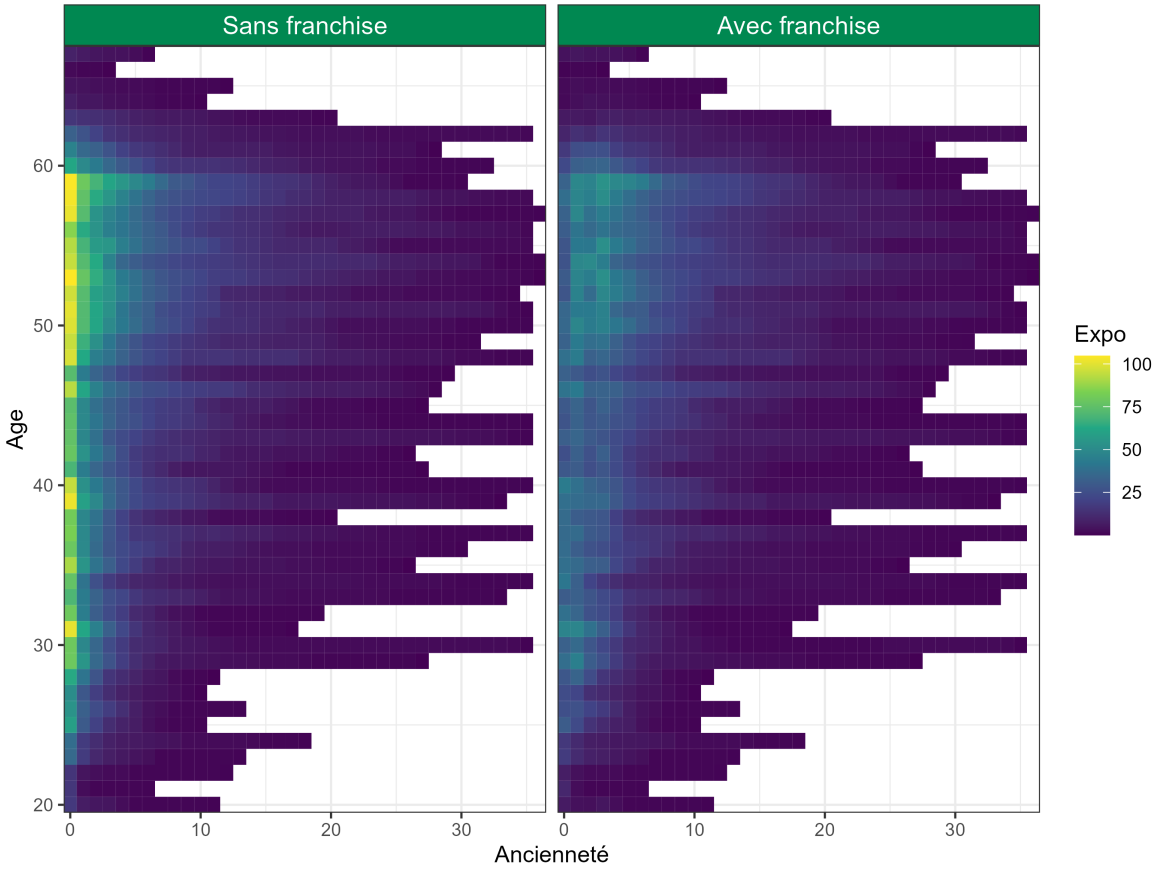


Figure 3.9. : Impact de la prise en compte de la franchise sur l'exposition

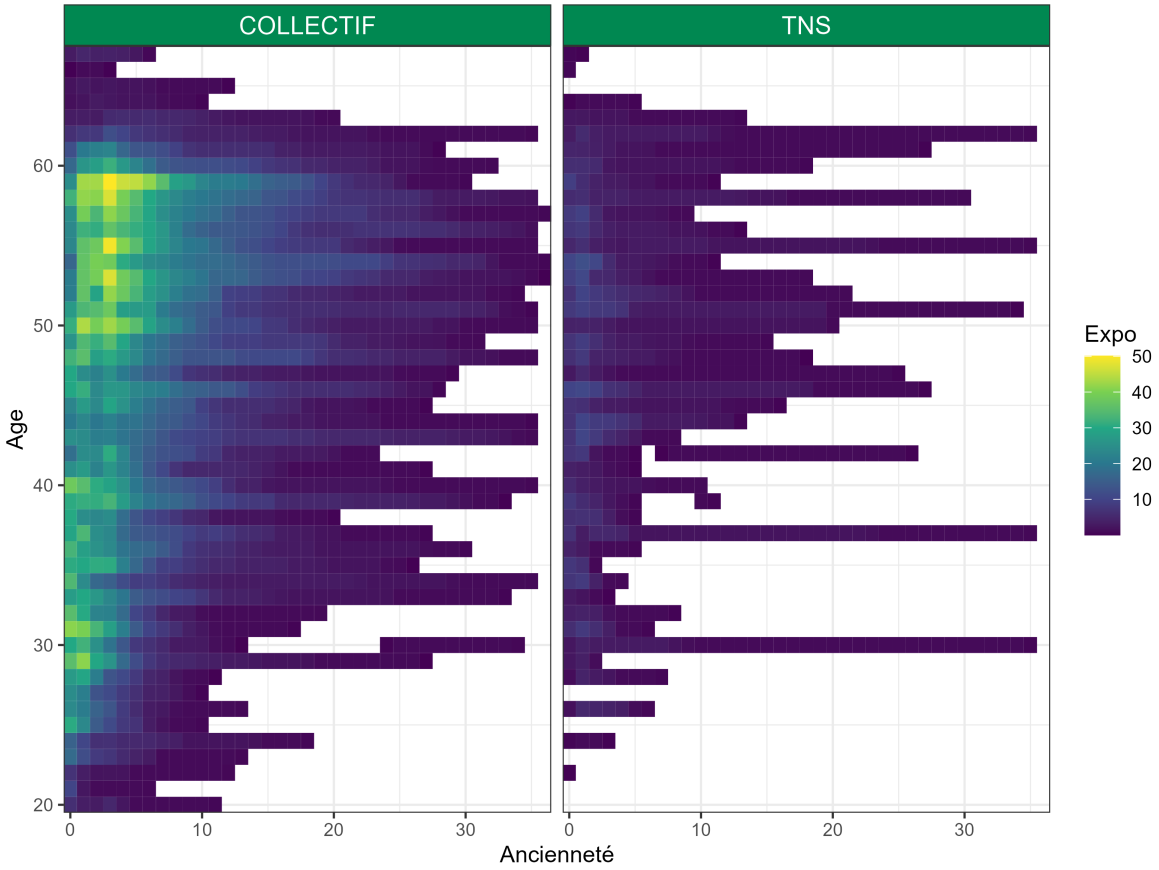


Figure 3.10. : Répartition de l'exposition selon le type de contrat

Chapitre 4.

Résultats

4.1. Base globale

Dans cette section, nous présentons l'ensemble des résultats obtenus à partir des modèles suivants :

- GAM utilisant uniquement nos données
- GAM utilisant la table réglementaire en tant qu'offset
- Lissage de Whittaker-Henderson appliqué à nos données

Cette approche nous permettra d'observer le décalage entre la loi suivie par notre portefeuille et la table réglementaire, ainsi que d'analyser les conséquences de l'utilisation de la table réglementaire en tant qu'offset sur notre modèle. Enfin, nous confrontons notre modèle au lissage de Whittaker-Henderson, une méthode largement utilisée dans la construction de tables d'expérience. Nous effectuons une comparaison qualitative à l'aide de graphiques, ainsi qu'une évaluation quantitative à l'aide de mesures pertinentes telles que l'AIC, la log-vraisemblance, ou encore le nombre de degrés de liberté utilisés.

4.1.1. Modèle GAM

Pour calibrer notre modèle GAM, nous utilisons le package `mgcv` dans le logiciel R en suivant les étapes suivantes :

Tout d'abord, nous divisons notre base de données en deux parties distinctes :

- Base 1 : Cette base comprend toutes les combinaisons d'âge pour les anciennetés inférieures ou égales à 34.
- Base 2 : Cette base comprend toutes les combinaisons d'âge pour l'ancienneté 35.

La raison de cette division est due à la présence d'un pic de sortie à l'ancienneté 35, qui n'est pas directement lié au risque étudié en fonction de l'âge et de l'ancienneté. Ce pic est plutôt d'ordre administratif. Il est préférable de le séparer du reste des données, car notre modèle réalise un lissage. Un pic dans un lissage entraîne une augmentation du nombre de degrés de liberté, ce qui ne correspond pas à la régularité réelle du phénomène étudié. En séparant la base de données comme décrit, nous permettons à notre modèle d'être suffisamment lisse pour

les 35 premières anciennetés, sans compromettre la représentation précise du pic à l'ancienneté 35.

Ensuite, nous procédons à la calibration du premier modèle en utilisant la fonction `gam` avec une base de cardinalité 13 pour l'âge et l'ancienneté. Cela signifie que nous aurons une combinaison linéaire de 12 fonctions de base pour nos prédicteurs, en raison d'une contrainte d'identifiabilité. De plus, nous appliquons la même cardinalité pour la base de l'interaction, ce qui entraîne l'utilisation d'un produit tensoriel de deux bases de cardinalité 12 (13-1 en raison de la contrainte d'identifiabilité).

Puis, nous calibrons le deuxième modèle uniquement à partir de la dernière ancienneté pour chauqe âge. Cette fois-ci nous utilisons une base de cardinalité 4 pour l'âge. Il n'y a pas de modélisation selon l'ancienneté puisque nous sommes ici à ancenneté fixée à 35 mois.

Les paramètres de lissage sont estimés en maximisant la vraisemblance restreinte (REML). Nous utilisons le logarithme de l'exposition comme offset et calibrons nos modèles sur chacune des deux bases. Ensuite, nous utilisons la fonction `predict` pour prédire le taux de sortie pour chaque âge et chaque ancienneté.

Les prédicteurs des taux lisses sont présentés dans la Figure 4.1, tandis que la représentation des taux lisses pour deux âges différents est montrée dans la Figure 4.2. On peut y discerner, d'une part, les taux lisses obtenus à partir de notre modèle GAM sur la base globale, tracés en trait continu, et d'autre part, les taux bruts de la table de maintien en incapacité du BCAC, indiqués en pointillés. Cette illustration met en évidence la différence entre les taux non lisses de la table du BCAC et les taux lisses que notre modèle génère. Finalement on observe une certaine proximité entre les taux obtenus et ceux du BCAC.

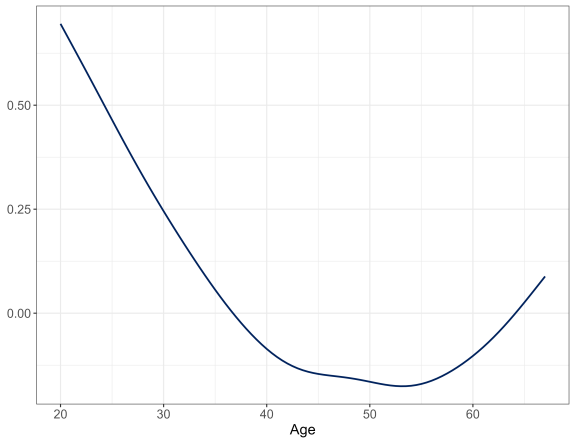
Ensuite, nous obtenons les probabilités de maintien en utilisant Équation 3.1 : $P = \exp(-\mu)$.

Finalement, pour obtenir la table de maintien correspondante, il suffit de calculer, pour chaque âge, le vecteur des produits cumulés sur une base de 10 000 individus. Une illustration de cette table est disponible pour divers tranches d'âge dans la Figure 4.3. À nouveau, la référence du BCAC est tracée en pointillés. Il est notable que notre modèle, qui intègre l'ensemble des données, se positionne au-dessus de la table du BCAC pour les périodes d'arrêt inférieures à un an environ, et qu'au-delà de ce seuil, il se situe légèrement en dessous.

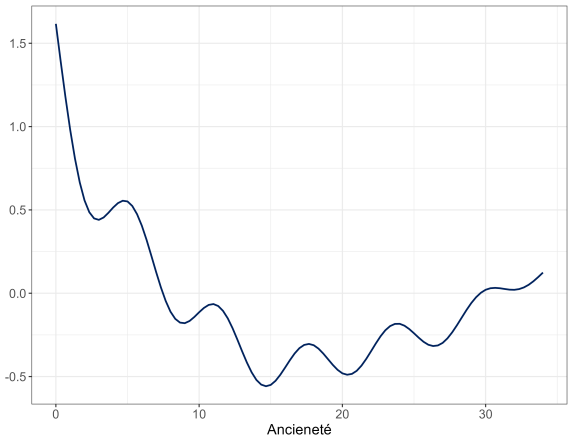
Comme évoqué précédemment, notre modèle a la capacité d'intégrer les données d'une table préexistante. Avant de présenter les résultats issus de notre modèle GAM en utilisant les données de la table de maintien en incapacité du BCAC de 2010, nous souhaitons présenter une visualisation des durées moyennes des arrêts pour chaque âge, ainsi que la durée moyenne globale pondérée par le nombre d'arrêts par âge, obtenue sans l'apport des données du BCAC.

Cette présentation nous permettra ensuite de faire une comparaison avec l'utilisation des données du BCAC.

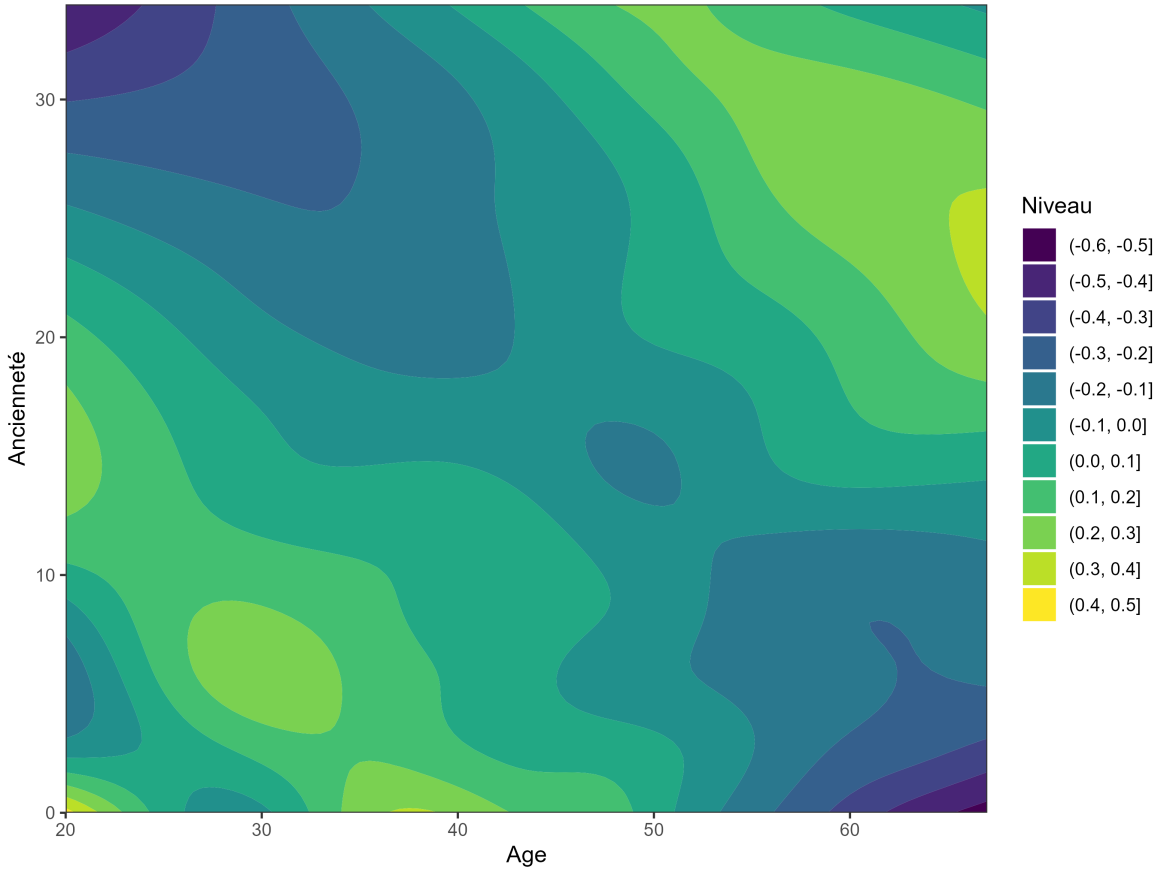
Cette représentation est disponible Figure 4.4. On y constate que la table d'expérience obtenue est très proche de la table du BCAC. Donc, pour ce portefeuille, ne pas prendre en compte la franchise dans l'estimation de la loi de maintien en incapacité donne une table proche de celle du BCAC.



(a) Effet de l'âge



(b) Effet de l'ancienneté



(c) Effet de l'interaction

Figure 4.1. : Prédicteurs des taux lisses

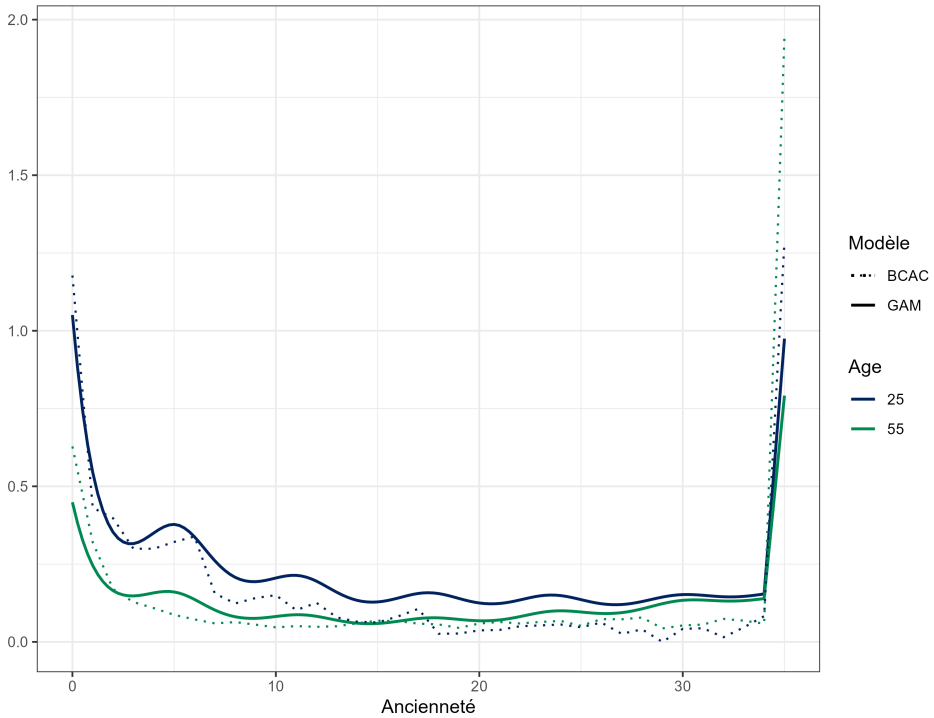


Figure 4.2. : Représentation des taux lisses obtenus pour notre modèle prenant en compte la franchise aux âges 25 et 55 ans

4.1.2. À partir des taux du BCAC

Une des particularités que nous visons pour notre modèle est qu'il puisse aisément intégrer une table de référence préexistante. Cette particularité peut être réalisée au moyen de modèles GAM en utilisant astucieusement l'offset. Ce modèle relationnel revêt un intérêt majeur, car il ne se focalise plus sur le calibrage des taux de sortie, mais directement sur l'ajustement des coefficients de passage, permettant ainsi de se positionner par rapport à la table de référence.

La formulation du modèle relationnel se présente comme suit :

$$\log(\mathbb{E}(\text{Sortie})) = log(\alpha * \mu_{\text{BCAC}} * e) = \log(\alpha) + \log(\mu_{\text{BCAC}}) + \log(e)$$

En considérant cette fois $\log(\mu_{\text{BCAC}}) + \log(e)$ comme un offset du modèle, nous pouvons directement calibrer les coefficients α. Ainsi, la contrainte de lissage porte sur les coefficients α plutôt que sur les taux de sortie eux-mêmes. Cependant, puisque la table du BCAC n'est pas lisse, la multiplication de ses taux par des coefficients lisses engendre des taux non lisses, comme illustré dans la Figure 4.5.

En outre, nous présentons les résultats des durées moyennes d'arrêts par âge obtenus grâce à ce modèle dans la Figure 4.6. Lorsque nous examinons cette figure à la lumière de la Figure 4.4, nous constatons une convergence remarquable des résultats obtenus. Bien que la moyenne globale puisse différer, cette divergence s'explique par l'utilisation de la table du BCAC en entrée, qui impose des contraintes sur les âges possibles. En effet nous rappelons que la table de maintien en incapacité du BCAC de 2010 couvre uniquement les âges de 23 à 66 ans.

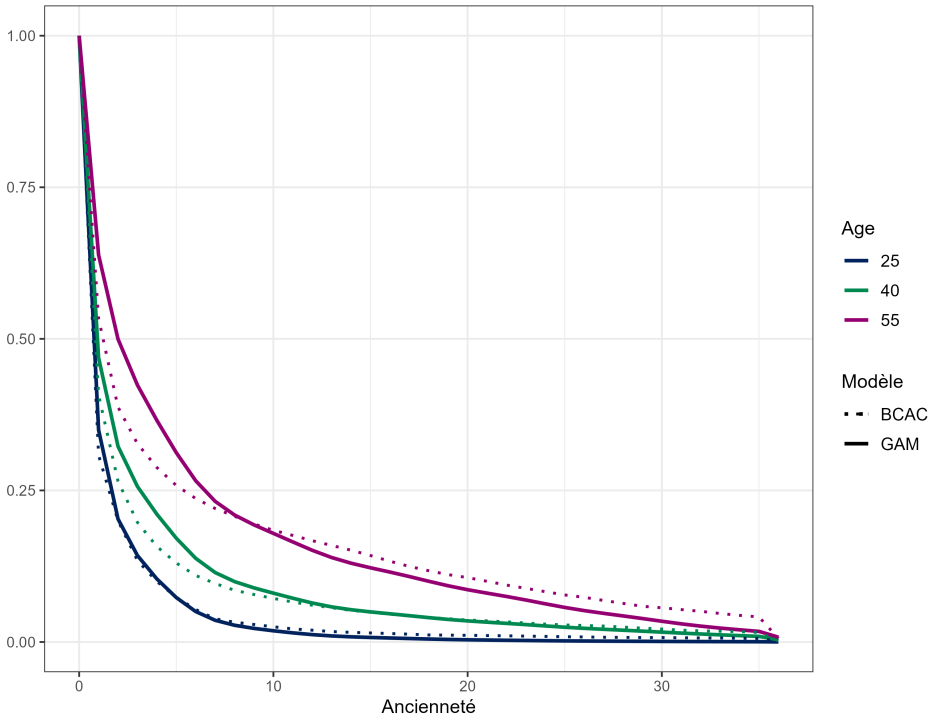


Figure 4.3. : Représentation des fonctions de survie pour les âges 25, 40 et 55 ans

Avec la concertation de l'entreprise il a été décidé de ne pas utiliser ce modèle par la suite pour les raisons suivante :

- Les résultats obtenus sont très proches de ceux obtenus sans l'utilisation de la table du BCAC, ce qui renforce la confiance des résultats obtenus sans son inclusion.
- La contrainte de lissage est plus pertinente lorsqu'appliquée aux taux de sortie plutôt qu'aux coefficients de passage.
- La souplesse d'éviter les restrictions liées aux tranches d'âge s'avère importante, puisque la table de 2010 ne couvre pas la plage d'âge sur lequel porte le risque.

Il est toutefois important de noter que cette possibilité reste intéressante et pourrait être utilisée, par exemple, avec une table d'expérience construite par le passé. Cette alternative constitue une solution tout aussi satisfaisante au calibrage des taux de sortie.

4.1.3. Approche par lissage

Maintenant nous allons confronter notre modèle à une approche bien plus classique, à savoir l'application du lissage de Whittaker-Henderson (Henderson 1924). On rappelle que les taux bruts sont obtenus en divisant le nombre de sorties par l'exposition pour chaque combinaison d'âge et d'ancienneté.

Nous allons appliqué le lissage de Whittaker-Henderson selon 2 approches différentes. La première consitste simplement à lisser la matrice des taux bruts. La seconde s'applique à la matrice des sorties observées ainsi que la matrice d'exposition et utilise la vraisemblance du modèle de Poisson.

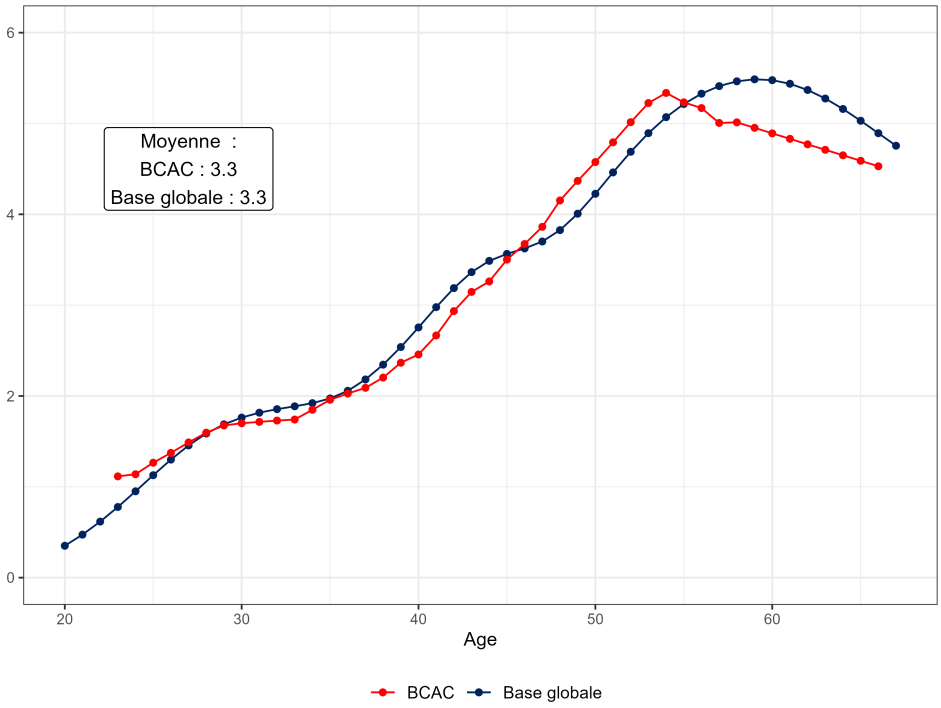


Figure 4.4. : Espérances des durées d'arrêts en mois sur la base globale. Les moyennes globales sont pondérées par le nombre d'arrêts

Pour effectuer ce lissage, nous avons utilisé le package R WH. Pour des détails plus approfondis sur le lissage de Whittaker-Henderson et notamment l'approche par la vraisemblance, veuillez consulter Biessy (2023).

Dans la pratique, l'utilisation du package WH permet de déterminer le paramètre de lissage h optimal en employant un critère. Nous avons choisi d'utiliser le critère d'optimisation REML, comme pour nos modèles GAM.

4.1.3.1. Comparaison des 2 approches du lissage de Whittaker-Henderson

Nous cherchons d'abord à comparer les deux approches de lissage de Whittaker-Henderson que nous avons évoquées précédemment. Il convient de rappeler que la comparaison des taux se fait sur les 35 premières anciennetés.

D'une part, nous divisons la matrice des sorties d'incapacité par la matrice d'exposition afin d'obtenir une matrice des taux bruts. C'est sur cette matrice que nous appliquons le lissage de Whittaker-Henderson. D'autre part, nous utilisons l'approche par vraisemblance du modèle de Poisson à l'aide du package `WH`.

Une première comparaison graphique des taux lissés est alors possible. Nous les visualisons sur la figure Figure 4.7, où le lissage des taux bruts est appelé "Lissage taux bruts" et l'approche utilisant la vraisemblance "Lissage WH".

Nous constatons alors que le lissage des taux bruts est peu satisfaisant, car les taux ainsi obtenus sont très erratiques, surtout pour les anciennetés élevées, où les taux bruts sont plus

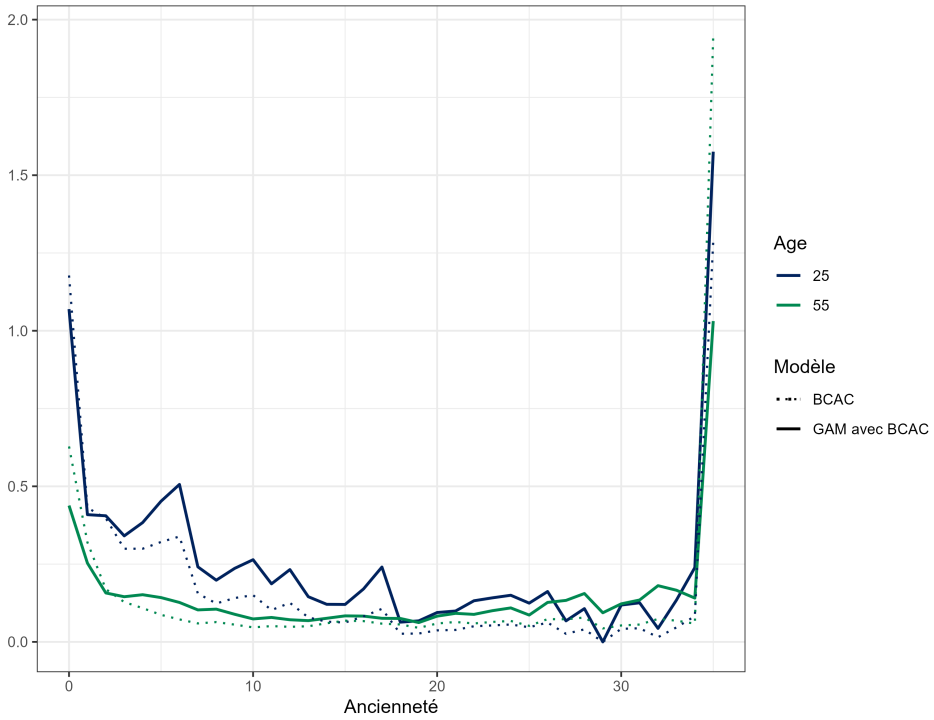


Figure 4.5. : Taux du modèle GAM avec table du BCAC en offset

variables. Ceci traduit un sous-lissage, qu'il serait nécessaire de corriger manuellement en augmentant arbitrairement les paramètres de lissage. Nous rappelons qu'ici, ils sont choisis automatiquement selon le même critère que celui utilisé pour le GAM. De plus, augmenter arbitrairement les paramètres de lissage peut entraîner une perte d'informations importantes dans les données. En revanche, les taux lissés par l'autre approche se révèlent bien plus réguliers.

Un autre point notable est la surestimation des taux par le lissage des taux bruts. Cette tendance à surestimer les taux de sorties n'est pas avantageuse, car cela conduit à construire une table moins prudente.

Finalement, il est intéressant de comparer ces deux approches quantitativement. Cette comparaison est présentée dans la Table 4.1.

Table 4.1. : Résumé des avantages et inconvénients des différents modèles

Modèle	Log-vraisemblance	Degrés de liberté	AIC
Lissage des taux bruts	-4802	$57,84$	9719
Lissage WH	-4522	$64,16$	9171

La différence est alors claire : l'approche utilisant la vraisemblance maximise davantage la Log-vraisemblance que le lissage des taux bruts et conduit ainsi à un AIC bien plus faible. L'augmentation du nombre de degrés de liberté reste très raisonnable par rapport à l'amélioration de la précision apportée.

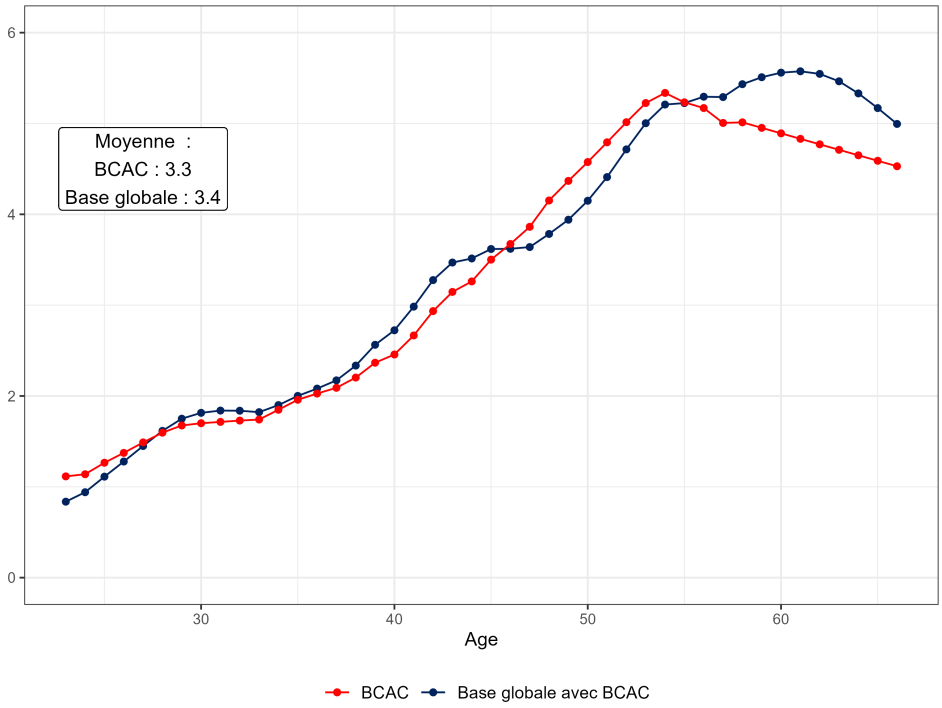


Figure 4.6. : Espérances des durées d'arrêts en mois sur la base globale avec la table du BCAC en offset. Les moyennes globales sont pondérées par le nombre d'arrêts

Nous observons alors que le lissage de Whittaker-Henderson peut produire des résultats tout à fait différents selon l'approche utilisée. Il s'avère que l'approche la plus simple de ce lissage n'exploite pas tout le potentiel du lissage. Il est donc pertinent de se tourner vers des approches plus élaborées lorsque les données le permettent.

4.1.3.2. Comparaison entre le lissage de Whittaker-Henderson et le GAM

Nous comparons maintenant notre GAM avec l'approche la plus performante du lissage de Whittaker-Henderson. Une amélioration des performances par rapport à cette approche permettra d'établir que notre GAM apporte de la valeur dans la modélisation du risque d'arrêt de travail par rapport au lissage de Whittaker-Henderson.

Pour commencer cette deuxième comparaison, un graphique est réalisée afin de comparer les taux lissés issus de notre modèle GAM et les taux lissés obtenus par la méthode de Whittaker-Henderson. Cette comparaison est effectuée pour les âges de 25 ans et 55 ans, comme illustré dans la Figure 4.8.

Nous observons alors une convergence des solutions obtenues. En effet, particulièrement pour l'âge de 55 ans où l'exposition est élevée, les taux lissés sont similaires dans les deux approches. En revanche, nous constatons un écart plus important à 25 ans, où l'exposition est plus faible.

Ensuite, nous pouvons comparer quantitativement la performance de ces deux modèles à l'aide d'une mesure commune : l'AIC. Pour ce faire, nous construisons une base de comparaison

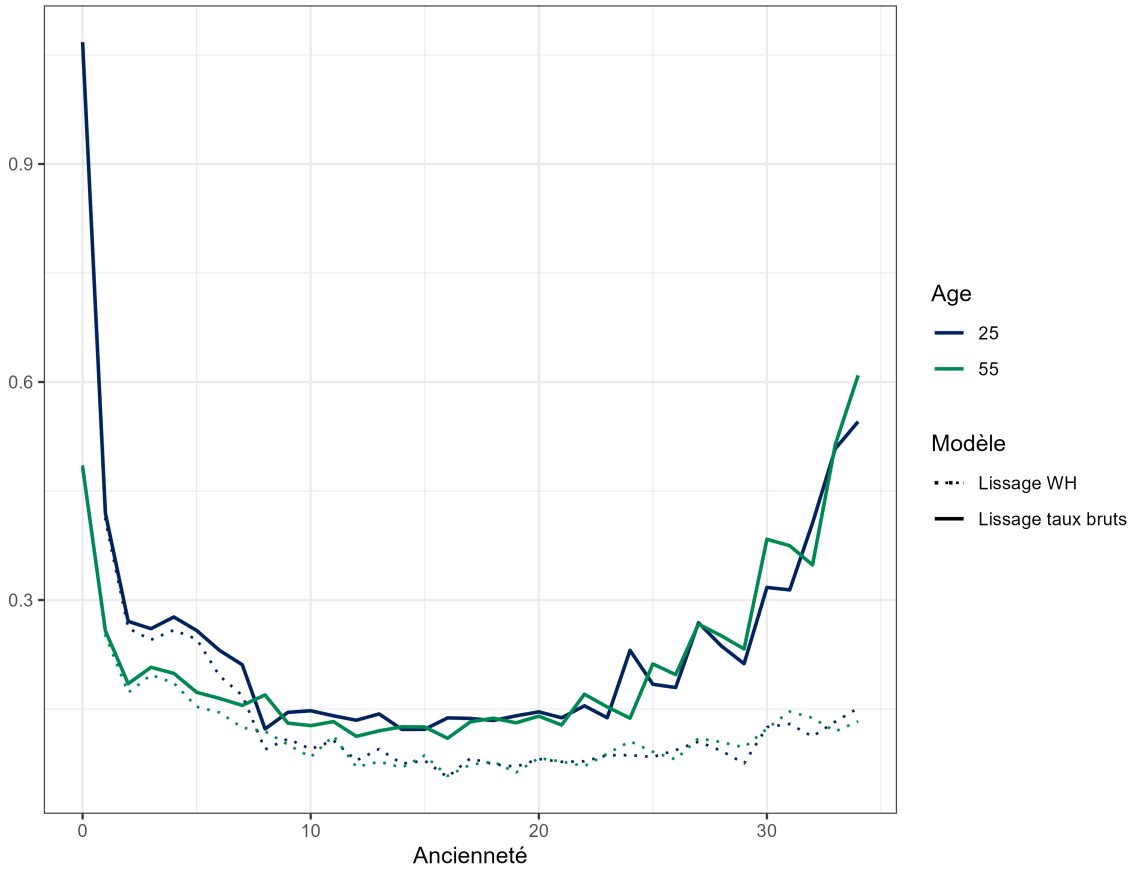


Figure 4.7. : Comparaison des taux lissés selon deux approches de lissage de Whittaker-Henderson

correspondant aux données agrégées au niveau de l'âge, de l'ancienneté, du sexe et du type de contrat. Pour chaque combinaison de ces éléments, nous disposons des taux de sorties issus de nos deux modèles, et en tenant compte de l'exposition, nous obtenons ainsi l'équivalent du nombre de sorties prédites par le modèle. Nous pouvons ensuite calculer la log-vraisemblance de Poisson :

$$l(y, \hat{y}) = \sum_{\chi} y_i \times \log(\hat{y}_i) - \sum_{\chi} \hat{y}_i - \sum_{\chi} \log(y_i)$$

Où χ représente l'ensemble des combinaisons d'âge, d'ancienneté, de sexe et de type de contrat, y représente le vecteur du nombre de sorties observé pour toutes les combinaisons de χ et $\hat{y}$ représente le vecteur du produit entre l'exposition et le taux de sortie prédit par le modèle pour chaque combinaison de χ. Finalement, connaissant le nombre de degrés de liberté (ddl)de chaque modèle, nous obtenons l'AIC du modèle m :

$$\text{AIC}(m) = 2 \times \text{ddl}_m - 2 \times l_m(y, \hat{y}_m)$$

Les résultats sont résumés dans la Table 4.2. Nous constatons tout d'abord que les valeurs de l'AIC sont très proches, ce qui n'est pas surprenant étant donné que les deux modèles

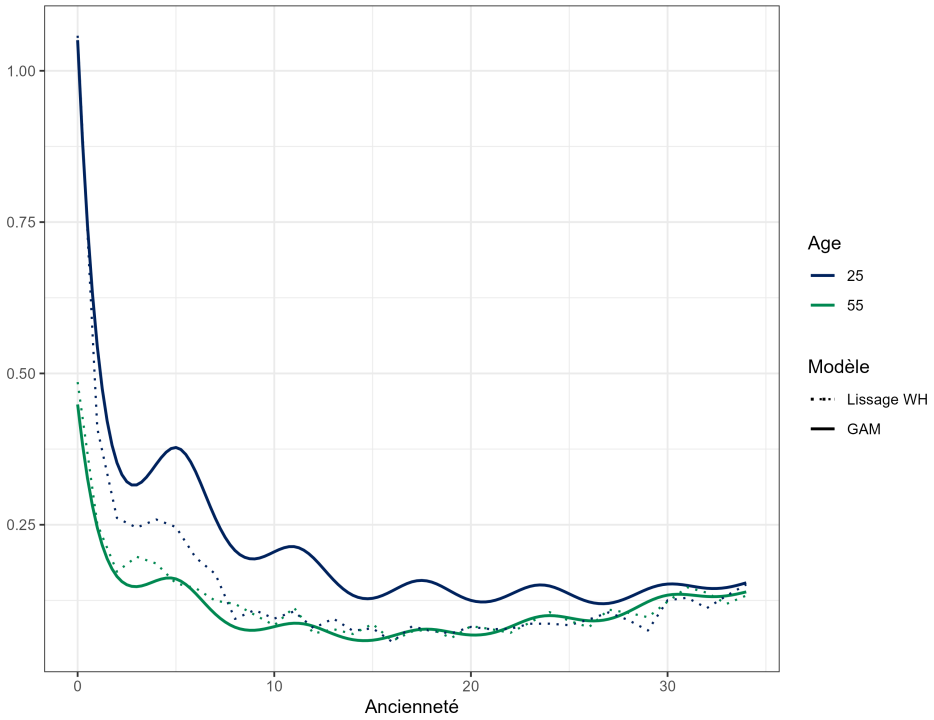


Figure 4.8. : Comparaison des taux lisses obtenus pour notre GAM et par lissage de Whittaker-Henderson sans prise en compte de la franchise aux âges 25 et 55 ans

sont calibrés à partir des mêmes données, sans prendre en compte ni la franchise ni d'autres variables que l'âge et l'ancienneté. De plus, ils utilisent tous les deux la méthode REML pour déterminer leur paramètre de lissage. Cependant, le lissage WH, bien qu'il présente une vraisemblance plus élevée, utilise plus du double des degrés de liberté de notre GAM, ce qui conduit finalement, au regard de l'AIC, à privilégier le modèle GAM, sachant qu'à ce stade, aucune variable supplémentaire n'est prise en compte. Ainsi, l'idée de construire une table simplement en appliquant le lissage de Whittaker-Henderson aux données brutes ne permet pas d'obtenir des résultats aussi précis et cohérents que ceux offerts par un GAM, surtout si l'on souhaite faire intervenir des variables explicatives. En revanche, nous avons montrer qu'en l'absence de variable explicatives, le lissage de Whittaker-Henderson utilisé avec une bonne approche peut s'avérer satisfaisant.

Table 4.2. : Comparaison quantitative entre le lissage de Whittaker-Henderson et le modèle GAM

Modèle	Log-vraisemblance	Degrés de liberté	AIC
GAM	−4538	35, 86	9148
Lissage WH	−4522	64, 16	9171
Lissage taux bruts	−4802	57, 84	9719

Après avoir examiné la comparaison des taux de sorties, il est intéressant de se pencher sur la comparaison en termes de durées moyennes des arrêts. La Figure 4.9 illustre de légères irrégularités entre les âges de 55 et 60 ans, là où l'exposition est limitée, une zone où le GAM reste stable. Cela souligne davantage la capacité des GAM à générer des résultats lisses.

En conclusion, il est possible d'affirmer que le simple lissage des données bruts via la méthode de Whittaker-Henderson ne constitue pas une solution optimale en comparaison avec les modèles GAM, bien qu'elle peut s'avérer satisfaisante suivant l'approche utilisée. Le processus de lissage produit une solution présentant des irrégularités lorsque l'exposition est faible, un phénomène non rencontré par les GAM utilisant la même méthode de sélection de paramètre

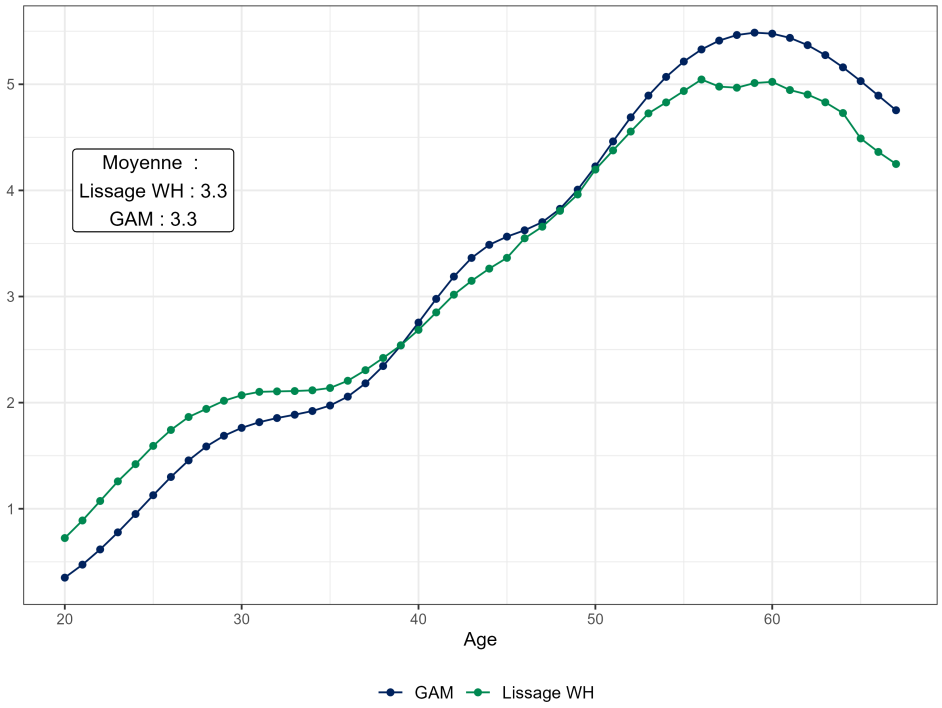


Figure 4.9. : Comparaison des espérances des durées d'arrêts en mois entre le modèle GAM et le lissage de Whittaker-Henderson. Les moyennes globales sont pondérées par le nombre d'arrêts

De plus, la méthode de lissage de Whittaker-Henderson ne permet pas d'intégrer des variables explicatives autrement que par la segmentation des données en fonction des modalités de la variable. Une autre limitation cruciale du lissage réside dans sa dépendance à une base de données complète, sans lacunes. En d'autres termes, il est impératif de disposer d'une base de données suffisamment étendue pour couvrir l'intégralité de la plage de données d'intérêt. Cette contrainte est particulièrement forte pour les ensembles de données de petite taille, et davantage encore si l'inclusion d'une ou de plusieurs variables explicatives est souhaitée. Par conséquent, l'approche utilisant les GAM est justifiée au vu de ces observations.

4.2. Base avec franchise

Nous avons précédemment mentionné que nous disposons de données concernant la franchise pour une partie de nos contrats. Dans cette partie, nous allons étudier l'impact de la prise en compte de la franchise.

La prise en compte de la franchise se reflète dans le calcul de l'exposition, en utilisant comme caractéristiques initiales les caractéristiques de l'individu une fois qu'il a dépassé la franchise.

Afin de mesurer l'impact de la prise en compte de la franchise, nous allons calibrer notre modèle sur la base des 4 647 lignes dont nous connaissons la franchise. Nous allons le faire en prenant en compte la franchise d'une part, et en la considérant nulle d'autre part.

Nous représentons les deux fonctions de survie dans la Figure 4.10 pour un âge donné, et nous fournissons également une représentation de l'espérance de la durée d'un sinistre dans ces deux cas Figure 4.11. Cette métrique est pertinente car les provisions pour un sinistre sont justement proportionnelles à la durée de vie résiduelle de ce sinistre.

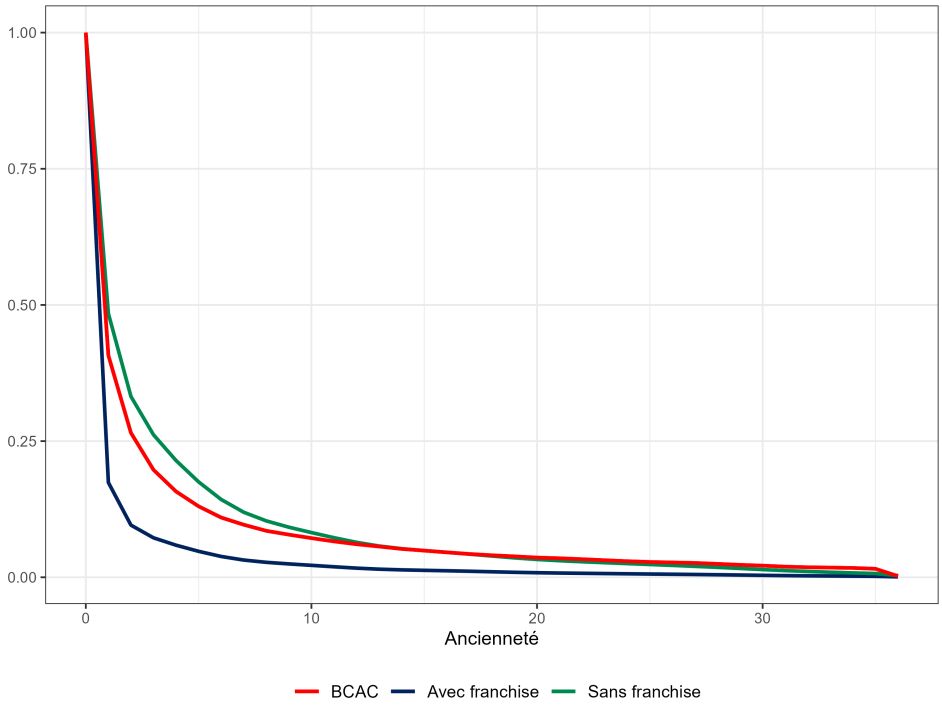


Figure 4.10. : Comparaison des fonctions de survie pour l'incapacité à 40 ans : Effet de la franchise sur la survie

L'espérance est calculée de la manière suivante :

$$\mathbb{E}(\mathbf{X}) = \sum_{k=1}^{36} \prod_{i=1}^{k} p_i$$

Où p_i représente la probabilité de rester en incapacité à l'ancienneté i.

La différence significative dans l'espérance des arrêts entre la prise en compte et l'absence de prise en compte de la franchise est clairement illustrée dans la figure Figure 4.11. Cette différence s'explique principalement par son impact important sur les premières anciennetés, qui sont susceptibles d'être masquées par la présence de franchises. Lorsque nous prenons en compte la franchise, les individus sont exclus de l'exposition pendant leur période de franchise, ce qui entraîne une augmentation du taux de sortie (défini comme le nombre de sorties divisé par l'exposition). Or, les premières anciennetés jouent un rôle crucial dans la détermination de l'espérance de la durée des arrêts, ce qui explique l'écart significatif observé. Cette constatation justifie pleinement la nécessité de prendre en compte la franchise dans la suite de ce mémoire. Par conséquent, nous utiliserons la base de données composée de 4 647 lignes.

De plus, il convient de souligner que les résultats obtenus concernant les durées moyennes des arrêts à partir de la base réduite, dans laquelle les effets des franchises sont exclus, présentent

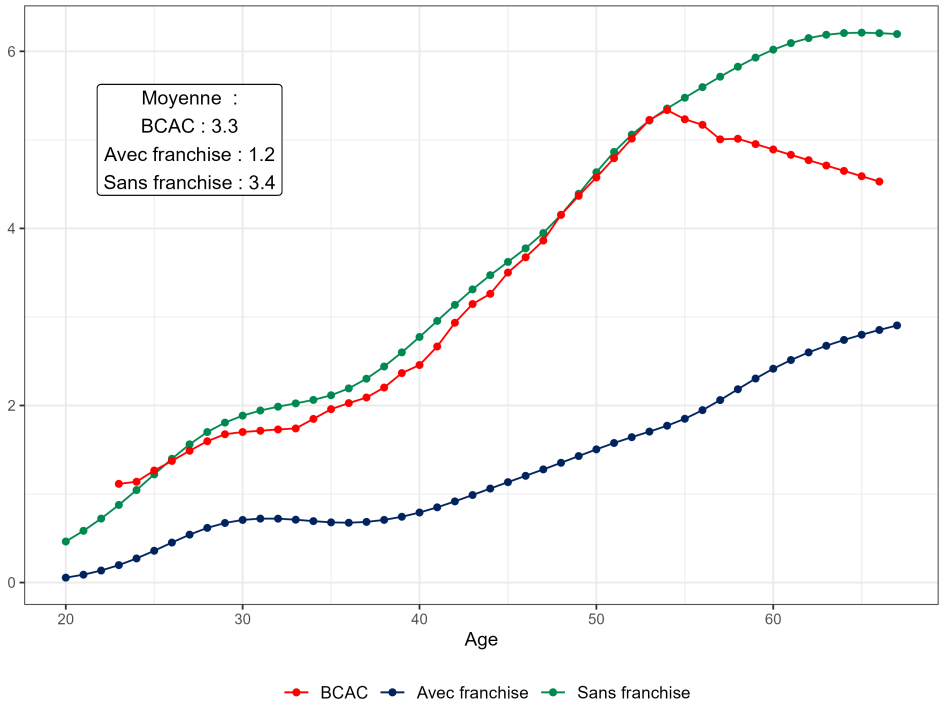


Figure 4.11. : Comparaison des espérances des durées d'arrêts en mois. Les moyennes globales sont pondérées par le nombre d'arrêts

une correspondance étroite avec ceux obtenus sur la base globale. Cette concordance renforce la confiance dans la représentativité de la base réduite, suggérant que malgré l'exclusion de certaines données, les conclusions tirées de l'analyse sur la base réduite seront robustes et dignes de confiance.

4.3. Prise en compte de covariable

Les tables construites servant à calculer les provisions il est intéressant est pertinent de prendre en compte une variable utilisée pour le calcul des provisions : le type de contrats.

Nous focaliserons notre analyse uniquement sur le type de contrat pour deux principales raisons :

- Premièrement, les provisions sont calculées en fonction du type de contrat, et notre expérience montre que le niveau de risque varie significativement en fonction de ce critère.
- Deuxièmement, les autres variables se sont avérées moins significatives pour améliorer la précision du modèle. Par exemple, le sexe des assurés n'a pas augmenté de manière significative les performances du modèle (en termes d'AIC ou de log-vraisemblance), mais a ajouté des degrés de liberté au modèle.

Le portefeuille est composé de contrats Collectif et de contrats TNS, et les provisions sont calculées séparément pour ces deux populations.

Ainsi, nous calculons l'exposition et le nombre de sorties en fonction de l'âge, de l'ancienneté et du type de contrat. Nous souhaitons examiner spécifiquement l'impact du type de contrat sur les sorties. Pour cela, nous ajustons notre modèle en utilisant la méthode et les paramètres décrits précédemment (voir Section 4.1).

Le modèle final prend la forme suivante :

$$\log(\mathbb{E}(\text{Sortie})) = \beta_0 + f_1(\text{Age}) + f_2(\text{Anc}) + f_3(\text{Age,Anc}) + \text{Type}$$

Ce modèle repose sur l'utilisation d'une population de référence (en l'occurrence, les contrats collectifs car il s'agit de la population majoritaire représentant 90,02% des données) et ajuste un seul coefficient pour les contrats TNS, permettant ainsi d'évaluer quantitativement l'écart entre ces deux populations en termes de taux de sorties.

Un modèle plus ambitieux est possible en utilisant les GAM :

$$\log(\mathbb{E}(\text{Sortie})) = \beta_0 + f_1(\text{Age}|\text{Type}) + f_2(\text{Anc}|\text{Type}) + f_3(\text{Age,Anc}|\text{Type}) + \text{Type}$$

Cependant, cette approche requiert un volume de données beaucoup plus important pour pouvoir calibrer le comportement spécifique de chaque population. Dans notre cas, nous avons opté pour le modèle précédemment mentionné, où un seul coefficient est calibré pour refléter l'impact global de la variable du type de contrat. Cette approche plus simple nous permet d'obtenir des résultats significatifs tout en utilisant la quantité de données disponible.

On peut observer dans la Figure 4.12 un écart significatif au niveau de l'espérance des durées des arrêts entre les contrats TNS et les contrats collectifs. En effet, les arrêts TNS sont en moyenne trois fois plus longs que ceux des contrats collectifs.

Une autre mesure d'intérêt concerne le coefficient attribué à la population TNS dans notre modèle. Sur l'échelle du prédicteur linéaire, pour les 35 premières anciennetés (rappelons que nous utilisons un modèle séparé pour modéliser la dernière ancienneté), le coefficient est de -0.383. Ce qui signifie à l'échelle du taux lisses un coefficient à $\exp(-0.383) \simeq 0,68$. Cette valeur peut s'interpréter de la manière suivante : En moyenne, pour chaque ancienneté inférieure ou égale à 34, les taux de sorties instantanés des contrats TNS sont inférieurs de 32% par rapport à ceux des contrats collectifs.

En utilisant notre modèle final, qui inclut la variable du type de contrat, nous avons pu obtenir des tables d'expérience distinctes pour chaque type de contrat. Ce résultat a été réalisé simplement en ajoutant la variable pertinente à la fonction `gam` du package `mgcv`. Cette approche nous permet de quantifier l'écart entre les contrats collectifs et les contrats TNS en termes de taux de sorties instantanés. Nous constatons une différence significative, avec des taux de sorties instantanés inférieurs de 32% en moyenne pour les contrats TNS par rapport aux contrats collectifs, pour chaque ancienneté inférieure ou égale à 34. Cette analyse approfondie des différences entre les deux types de contrats nous donne une vision plus précise de leur comportement en matière de sorties.

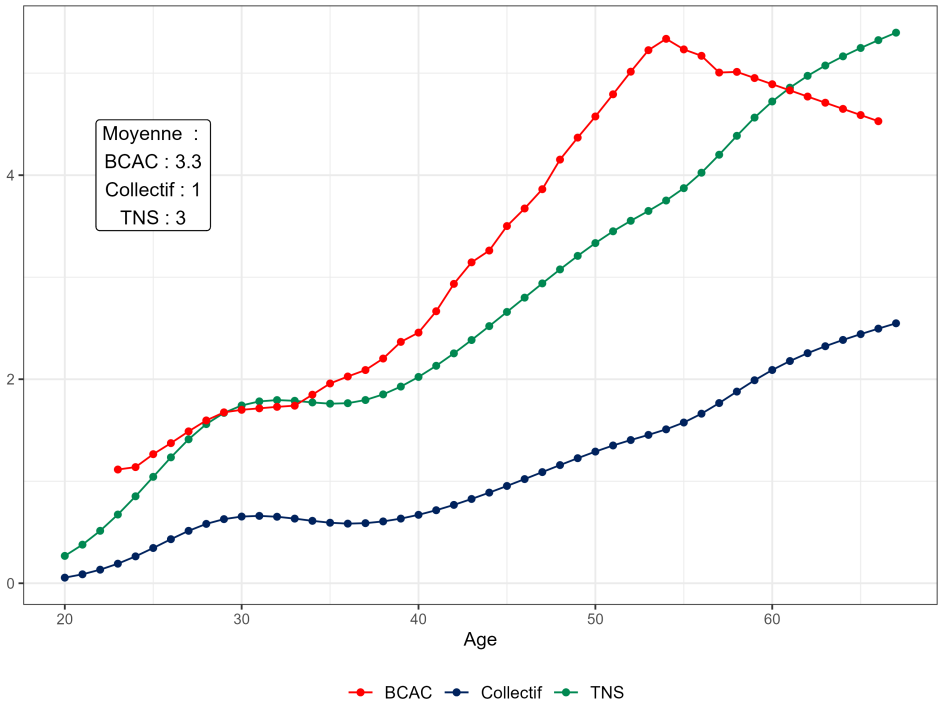


Figure 4.12. : Comparaison des espérances des durées d'arrêts en mois par type de contrat. Les moyennes globales sont pondérées par le nombre d'arrêts

4.4. Analyse des résultats

Dans cette section, nous présentons les résultats obtenus à partir des différents modèles que nous avons utilisés, à savoir :

- Le modèle de Turnbull
- Le modèle de Kaplan-Meier
- Le modèle GAM sans franchise
- Le modèle GAM avec franchise
- Le modèle GAM avec le type de contrat (avec franchise)

Il convient de noter que comparer des modèles de durées aussi hétérogènes n'est pas une tâche aisée. En effet, ces modèles estiment des caractéristiques différentes : certains estiment une fonction de survie, tandis que d'autres estiment la probabilité de sortie à un mois. En ce qui concerne les modèles GAM, ils estiment les taux de sortie instantanés.

Malgré cette hétérogénéité, nous pouvons tout de même rapprocher ces modèles sur une caractéristique commune, à savoir la fonction de survie. Pour faciliter la comparaison, nous présentons une représentation graphique des fonctions de survie obtenues par ces différents modèles (voir Figure 4.13).

Pour les modèles de Turnbull et de Kaplan-Meier, comme mentionné précédemment, ils doivent être calibrés âge par âge. Cependant, en raison du nombre limité de données dont nous disposons, nous avons regroupé les âges par tranche d'âge (40-50 ans sur la Figure 4.13)

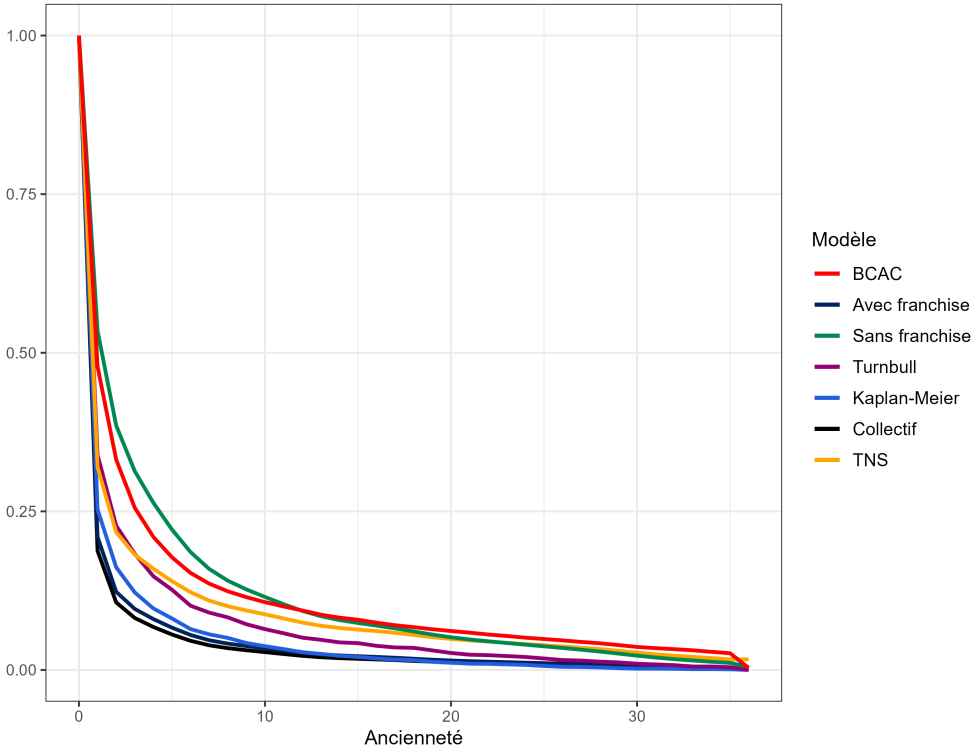


Figure 4.13. : Comparaison des fonctions de survie à 45 ans entre les différents modèles

pour ces modèles et nous représentons nos modèles GAM à l'âge moyen de la tranche (45 ans) afin de pouvoir nous rapprocher au mieux d'une comparaison avec ces modèles.

Sur la Figure 4.13, nous observons que le modèle excluant la franchise se situe logiquement au-dessus des autres modèles et, conformément à nos observations précédentes, au-dessus des prédictions du BCAC pour la première année. Cette mise en évidence souligne clairement l'importance cruciale de la prise en compte de la franchise dans le processus de construction d'une table d'expérience.

En revanche, en considérant le modèle de Turnbull, qui tient compte de la franchise, nous constatons une fonction de survie significativement supérieure à celle obtenue par la méthode de Kaplan-Meier. En effet il est important de souligner que la prise en compte de la franchise dans le modèle de Turnbull se révèle peu satisfaisante. Étant donné que le modèle de Turnbull est discret, il exige une franchise exprimée en mois entiers. Or, comme en témoigne la Figure 3.6, nos données ne suivent pas cette contrainte, ce qui pourrait expliquer cette différence de résultats.

Finalement, la courbe générée par Kaplan-Meier se rapproche de celle obtenue par notre modèle incluant la franchise. Cette observation renforce notre confiance dans les résultats obtenus par notre approche.

Ensuite, nous avons comparé nos modèles GAM à l'aide de mesures statistiques telles que l'AIC et le R^2 ajusté.

Nous rappelons rapidement le calcul du R^2 ajusté :

$$R^2_{\text{ajusté}} = 1 - \frac{\ell_{\text{sat}} - \ell_{\text{modèle}}}{\ell_{\text{sat}} - \ell_{\text{nul}}} \times \frac{n-1}{n-d_l}$$

Avec :

- $\ell_{\text{modèle}}$ la log-vraisemblance de notre modèle.
- ℓ_{sat} la log-vraisemblance du modèle saturé, utilisant autant de paramètres que d'individus.
- ℓ_{nul} la log-vraisemblance du modèle nul, c'est-à-dire celui n'utilisant qu'une seule constante.
- n le nombre d'individus.
- d_l le nombre de degrés de liberté.

Ces mesures ont été calculées à partir d'une base de données commune, agrégée sur la maille Age, Ancienneté et Type de contrat. Ainsi, pour chaque combinaison de ces caractéristiques, nous disposons d'une exposition et d'un nombre de sorties observées, ce qui nous permet de calculer ces mesures. Pour mieux comprendre les variations d'AIC entre les modèles, nous avons choisi de décomposer cette mesure en termes de log-vraisemblance et de nombre de degrés de liberté.

Table 4.3. : Comparaison des différents modèles GAM

Modèle	AIC	Log-vraisemblance	Degrés de liberté	R^2 (ajusté)
Sans franchise	7631	−3790	25,51	0,568
Avec franchise	6252	−3100	26,24	0,874
Avec périmètre	6192	−3068	27,55	0,889

Nous constatons ainsi que le modèle prenant en compte à la fois la franchise et le type de contrat est le meilleur en termes d'AIC et de R^2 ajusté. La prise en compte du type de contrat améliore significativement la vraisemblance du modèle avec seulement un degré de liberté supplémentaire, ce qui se traduit par un meilleur AIC. Dans la section suivante (Chapitre 5), nous examinerons l'impact de ces différents modèles sur le montant des provisions.

Un autre résultat important concerne le modèle prenant en compte le type de contrat. La prise en compte de cette variable s'effectue à l'aide d'un coefficient additif sur l'échelle des prédicteurs pour la population TNS. Sur l'échelle des taux de sorties instantanés, cela se traduit par un coefficient multiplicatif qui peut être interprété. Le coefficient de la variable TNS est estimé à 0,682, avec un intervalle de confiance à 95% compris entre 0,616 et 0,754. Cette estimation signifie que la population des travailleurs non-salariés présente un taux de risque instantané inférieur de 32% par rapport à la population des contrats collectifs. Un taux de risque instantané plus faible indique moins de sorties et, par conséquent, des arrêts plus longs. C'est ce que l'on peut observer sur la Figure 4.13 où la courbe des TNS est bien au dessus de la courbe des collectifs.

En conclusion, nos résultats montrent que le modèle GAM avec franchise et prise en compte du type de contrat est le plus performant selon les mesures AIC et R^2 ajusté. Ce modèle prend en compte de manière flexible les caractéristiques importantes du risque d'incapacité, ce qui lui permet d'obtenir une meilleure adéquation aux données. De plus, nous avons mis en évidence l'impact significatif du type de contrat sur les taux de sortie instantanés, avec une estimation

de coefficient indiquant que la population des TNS présente un taux de risque instantané inférieur de 32% par rapport à la population des contrats collectifs. Cette information nous permettra, dans la section suivante, de calculer les provisions en fonction du type de contrat, en utilisant une table d'expérience spécifique à chaque type.

Finalement, dans cette partie, nous avons observé que le modèle GAM parvient à produire des résultats lisses, une caractéristique clé pour une modélisation précise des taux d'incapacité. L'utilisation d'une table d'expérience en tant qu'offset a été explorée, montrant les limites de cette approche lorsque la table en offset présente des irrégularités. De même, l'approche par lissage de Whittaker-Henderson a été évaluée, révélant ses limitations en termes de régularité et d'intégration de variables explicatives.

En tenant compte de la franchise, nous avons identifié son impact significatif sur les durées d'arrêt, qui s'explique par son effet sur l'exposition aux premières anciennetés. L'ajout de covariables, en particulier le type de contrat, a permis de segmenter les résultats et de mettre en évidence des disparités importantes entre les différents types de contrats. Enfin, une analyse comparative approfondie a confirmé la performance du modèle GAM par rapport à d'autres méthodes, renforçant son statut en tant qu'outil de choix pour la modélisation des taux d'incapacité.

Maintenant que nous avons identifié le modèle le plus performant pour estimer les taux de sortie en incapacité, il est temps de passer à une étape cruciale : évaluer l'impact concret de ce modèle sur le calcul des provisions. En utilisant le modèle GAM qui intègre à la fois la franchise et le type de contrat, nous allons pouvoir déterminer comment ces facteurs influencent les provisions. De plus, nous allons également explorer comment la prise en compte de la franchise peut affecter les résultats des provisions. Cette analyse nous fournira une vision approfondie de l'efficacité de notre approche dans le domaine du provisionnement des risques liés à l'incapacité et à l'invalidité.

Chapitre 5.

Impact sur les provisions

5.1. Formules pour le calcul des provisions

Nous débutons cette section en répertoriant toutes les provisions liées aux risques d'incapacité, d'invalidité et de décès :

- Pour l'incapacité :
  - La provision pour l'incapacité en cours, reflétant le coût probable de l'incapacité restante.
  - La provision pour l'invalidité en attente, représentant le coût d'une éventuelle invalidité si la personne passe en invalidité.
  - La provision pour le maintien de la garantie décès en cas d'incapacité.
  - La provision pour le maintien de la garantie décès en cas d'invalidité en attente.
- Pour l'invalidité :
  - La provision pour invalidité en cours, reflétant le coût probable de l'invalidité restante.
  - La provision pour le maintien de la garantie décès en cas d'invalidité.

Les notations utilisées dans la suite sont toutes répertoriées dans l'Annexe A.

Incapacité en cours

On s'intéresse tout d'abord à la provision pour l'incapacité en cours pour une prestation **mensuelle** de 1 euro. Il s'agit de l'espérence de la durée probable de l'arrêt, actualisée au taux annuel i. Pour une personne d'âge age et en incapacité depuis anc mois à la date de calcul, ceci s'écrit de cette façon :

$$\mathrm{PM}_{\mathrm{inc}}(\mathrm{age},\mathrm{anc}) = \int_{u=\mathrm{anc}}^{36} {}_{u-\mathrm{anc}}p^{\mathrm{inc}}_{\mathrm{age},\mathrm{anc}} \times \nu^{\frac{u-\mathrm{anc}}{12}}\, du \tag{5.1}$$

En utilisant les notations suivantes :

- $\nu = \frac{1}{1+i}$ avec i le taux d'actualisation annuel.
- ${}_{u-\text{anc}}p^{\text{inc}}_{\text{age,anc}}$ la probabilité conditionnelle de maintien à l'ancienneté u, pour une personne entrée en incapacité à l'âge age et d'ancienneté anc mois à la date de calcul.

Cependant, nous sommes limités à l'utilisation de probabilités discrètes à partir des tables, ce qui nous oblige à approximer l'intégrale en utilisant la méthode des rectangles milieu. Cette approximation suppose essentiellement que les sorties de l'incapacité se produisent de manière uniforme chaque mois.

En utilisant cette méthode, nous obtenons la formule suivante :

$$\text{PM}_{\text{inc}}(\text{age}, \text{anc}) \simeq \frac{1}{2} \sum_{\text{k=anc}}^{35} {}_{\text{k}-\text{anc}}p^{\text{inc}}_{\text{age,anc}} \times \nu^{\frac{\text{k}-\text{anc}}{12}} + {}_{\text{k+1}-\text{anc}}p^{\text{inc}}_{\text{age,anc}} \times \nu^{\frac{\text{k+1}-\text{anc}}{12}}$$

Avec :

- age désignant l'âge atteint par l'assuré au moment de la survenance de l'incapacité
- anc représentant l'ancienneté atteinte à la date de calcul
- ${}_{\text{k}-\text{anc}}p^{\text{inc}}_{\text{age,anc}}$ la probabilité conditionnelle de maintien en incapacité à 1 mois à l'ancienneté k, pour une personne entrée en incapacité à l'âge age et d'ancienneté anc mois à la date de calcul.

Pour écrire cette dernière formule avec les coefficients des tables nous utilisons :

- ${}_{\text{k}-\text{anc}}p^{\text{inc}}_{\text{age,anc}} = \frac{L^{\text{inc}}_{\text{age,k}}}{L^{\text{inc}}_{\text{age,anc}}}$

Et nous obtenons notre formule (Équation 5.2) pour calculer la provision pour l'incapacité en cours en utilisant les tables :

$$\text{PM}_{\text{inc}}(\text{age}, \text{anc}) \simeq \frac{1}{2 \times L^{\text{inc}}_{\text{age,anc}}} \sum_{\text{k=anc}}^{35} L^{\text{inc}}_{\text{age,k}} \times \nu^{\frac{\text{k}-\text{anc}}{12}} + L^{\text{inc}}_{\text{age,k+1}} \times \nu^{\frac{\text{k+1}-\text{anc}}{12}} \tag{5.2}$$

Où nous avons utiliser la notation :

- $L^{\text{inc}}_{\text{age},k}$ le nombre de survivants, dans la table de maintien en incapacité, après une ancienneté de k mois depuis l'entrée en incapacité à l'âge age.

Invalidité en attente

Maintenant nous détaillons le calcul pour la provision pour invalidité en attente.

Nous avons donc le calcul suivant :

$$\begin{aligned}
\mathrm{PM}_{\mathrm{pass}}(\mathrm{age},\mathrm{anc}) &= \int_{u=\mathrm{anc}}^{36} {}_{u-\mathrm{anc}}p^{\mathrm{inc}}_{\mathrm{age},\mathrm{anc}} \times \mu^{\mathrm{inc}\to\mathrm{inv}}_{\mathrm{age},u} \times \nu^{\frac{u-\mathrm{anc}}{12}} \times \mathrm{PM}_{\mathrm{inv}}(\mathrm{age}+\frac{u}{12},0)\,du \\
&= \sum_{\mathrm{k}=\mathrm{anc}}^{35} \int_{u=\mathrm{k}}^{\mathrm{k}+1} {}_{u-\mathrm{anc}}p^{\mathrm{inc}}_{\mathrm{age},\mathrm{anc}} \times \mu^{\mathrm{inc}\to\mathrm{inv}}_{\mathrm{age},u} \times \nu^{\frac{u-\mathrm{anc}}{12}} \times \mathrm{PM}_{\mathrm{inv}}(\mathrm{age}+\frac{u}{12},0)\,du \\
&\simeq \sum_{\mathrm{k}=\mathrm{anc}}^{35} \nu^{\frac{\mathrm{k}+0.5-\mathrm{anc}}{12}} \times \mathrm{PM}_{\mathrm{inv}}(\mathrm{age}+\frac{\mathrm{k}+0.5}{12},0) \times \int_{u=\mathrm{k}}^{\mathrm{k}+1} {}_{u-\mathrm{anc}}p^{\mathrm{inc}}_{\mathrm{age},\mathrm{anc}} \times \mu^{\mathrm{inc}\to\mathrm{inv}}_{\mathrm{age},u}\,du
\end{aligned}$$

Avec :

- $\mu^{\mathrm{inc}\to\mathrm{inv}}_{\mathrm{age},u}$ la force de passage instantanée à l'ancienneté u pour une personne entrée en incapacité à l'âge age.

Ici nous avons pris la valeur de l'actualisation et de la PM invalidité en milieu de mois sous l'hypothèse que les transitions vers l'invalidité se produisent uniformément au cours de chaque mois. Cette hypothèse est crédible dans le sens où la valeur de l'actualisation et de la provision varie très peu entre deux mois consécutifs.

Finalement il suffit de remarquer que

$$\begin{aligned}
\int_{u=\mathrm{k}}^{\mathrm{k}+1} {}_{u-\mathrm{anc}}p^{\mathrm{inc}}_{\mathrm{age},\mathrm{anc}} \times \mu^{\mathrm{inc}\to\mathrm{inv}}_{\mathrm{age},u}\,du &= \int_{u=\mathrm{k}}^{\mathrm{k}+1} {}_{\mathrm{k}-\mathrm{anc}}p^{\mathrm{inc}}_{\mathrm{age},\mathrm{anc}} \times {}_{u-\mathrm{k}}p^{\mathrm{inc}}_{\mathrm{age},\mathrm{k}} \times \mu^{\mathrm{inc}\to\mathrm{inv}}_{\mathrm{age},u}\,du \\
&= {}_{\mathrm{k}-\mathrm{anc}}p^{\mathrm{inc}}_{\mathrm{age},\mathrm{anc}} \times \int_{u=\mathrm{k}}^{\mathrm{k}+1} {}_{u-\mathrm{k}}p^{\mathrm{inc}}_{\mathrm{age},\mathrm{k}} \times \mu^{\mathrm{inc}\to\mathrm{inv}}_{\mathrm{age},u}\,du \\
&= {}_{\mathrm{k}-\mathrm{anc}}p^{\mathrm{inc}}_{\mathrm{age},\mathrm{anc}} \times p^{\mathrm{inc}\to\mathrm{inv}}_{\mathrm{age},\mathrm{k}}
\end{aligned}$$

Avec :

$$ {}_{\mathrm{k}-\mathrm{anc}}p^{\mathrm{inc}}_{\mathrm{age},\mathrm{anc}} = \frac{L^{\mathrm{inc}}_{\mathrm{age},\mathrm{k}}}{L^{\mathrm{inc}}_{\mathrm{age},\mathrm{anc}}} \qquad \text{et} \qquad p^{\mathrm{inc}\to\mathrm{inv}}_{\mathrm{age},\mathrm{k}} = \frac{N^{\mathrm{inc}\to\mathrm{inv}}_{\mathrm{age},\mathrm{k}}}{L^{\mathrm{inc}}_{\mathrm{age},\mathrm{k}}}$$

En utilisant les notations suivantes :

- $p^{\mathrm{inc}\to\mathrm{inv}}_{\mathrm{age},\mathrm{k}}$ la probabilité de passage en invalidité entre le mois k et k+1 pour une personne entrée en incapacité à l'âge age.
- $N^{\mathrm{inc}\to\mathrm{inv}}_{\mathrm{age},\mathrm{k}}$ le nombre de passages dans la table de passage, après une ancienneté de k mois depuis l'entrée en incapacité à l'âge age.

Remarque : Si une table de maintien en incapacité a été réalisée sans la création d'une table de passage en invalidité, ce qui est notre cas ici, le calcul de $p^{\mathrm{inc}\to\mathrm{inv}}_{\mathrm{age},\mathrm{k}}$ doit se faire **uniquement** avec les tables du BCAC (Maintien en incapacité ainsi que passage en invalidité).

Et nous obtenons la formule de la provision pour l'invalidité en attente (Équation 5.3) :

$$\mathrm{PM}_{\mathrm{pass}}(\mathrm{age},\mathrm{anc}) \simeq \sum_{\mathrm{k}=\mathrm{anc}}^{35} {}_{\mathrm{k}-\mathrm{anc}}p^{\mathrm{inc}}_{\mathrm{age},\mathrm{anc}} \times p^{\mathrm{inc}\to\mathrm{inv}}_{\mathrm{age},\mathrm{k}} \times \nu^{\frac{\mathrm{k}+0.5-\mathrm{anc}}{12}} \times \mathrm{PM}_{\mathrm{inv}}(age+\frac{\mathrm{k}+0.5}{12},0) \quad (5.3)$$

Où $\mathrm{PM}_{\mathrm{inv}}(age + \frac{\mathrm{k}+0.5}{12}, 0)$ s'obtient par interpolation linéaire entre $\mathrm{PM}_{\mathrm{inv}}(\lfloor age + \frac{\mathrm{k}+0.5}{12} \rfloor, 0)$ et $\mathrm{PM}_{\mathrm{inv}}(\lfloor age + \frac{\mathrm{k}+0.5}{12} \rfloor + 1, 0)$ (où $\lfloor\ \rfloor$ désigne la partie entière inférieure).

Pour le calcul de la provision en invalidité voir Invalidité en cours.

Maintien de la garantie décès en incapacité

Il faut calculer la provision pour le maintien de la garantie décès en incapacité, qui est proportionnelle au **capital décès garanti**.

Pour cela, il est nécessaire d'utiliser la table de mortalité en incapacité construite par le BCAC en 2002.

Nous détaillons le calcul :

$$\begin{aligned}
\mathrm{PM}^{\mathrm{DC}}_{\mathrm{inc}}(\mathrm{age}, \mathrm{anc}) &= \int_{u=\mathrm{anc}}^{36} {}_{u-\mathrm{anc}}p^{\mathrm{inc}}_{\mathrm{age,anc}} \times \mu^{\mathrm{inc}\to\mathrm{DC}}_{\mathrm{age,k}} \times \nu^{\frac{\mathrm{u\text{-}anc}}{12}}\, du \\
&= \sum_{\mathrm{k}=\mathrm{anc}}^{35} \int_{u=\mathrm{k}}^{\mathrm{k}+1} {}_{u-\mathrm{anc}}p^{\mathrm{inc}}_{\mathrm{age,anc}} \times \mu^{\mathrm{inc}\to\mathrm{DC}}_{\mathrm{age,k}} \times \nu^{\frac{\mathrm{u\text{-}anc}}{12}}\, du \\
&\simeq \sum_{\mathrm{k}=\mathrm{anc}}^{35} \nu^{\frac{\mathrm{k}+0.5-\mathrm{anc}}{12}} \times \int_{u=\mathrm{k}}^{\mathrm{k}+1} {}_{u-\mathrm{anc}}p^{\mathrm{inc}}_{\mathrm{age,anc}} \times \mu^{\mathrm{inc}\to\mathrm{DC}}_{\mathrm{age,k}}\, du
\end{aligned}$$

L'actualisation se justifie par l'hypothèse que les décès se produisent de manière uniforme au cours du mois. Autrement dit, en moyenne, les décès ont lieu au milieu de chaque mois.

Cette fois-ci on remarque que :

$$\begin{aligned}
\int_{u=\mathrm{k}}^{\mathrm{k}+1} {}_{u-\mathrm{anc}}p^{\mathrm{inc}}_{\mathrm{age,anc}} \times \mu^{\mathrm{inc}\to\mathrm{DC}}_{\mathrm{age,k}}\, du &= \int_{u=\mathrm{k}}^{\mathrm{k}+1} {}_{\mathrm{k}-\mathrm{anc}}p^{\mathrm{inc}}_{\mathrm{age,anc}} \times {}_{\mathrm{u}-\mathrm{k}}p^{\mathrm{inc}}_{\mathrm{age,k}} \times \mu^{\mathrm{inc}\to\mathrm{DC}}_{\mathrm{age,k}}\, du \\
&= {}_{\mathrm{k}-\mathrm{anc}}p^{\mathrm{inc}}_{\mathrm{age,anc}} \times \int_{u=\mathrm{k}}^{\mathrm{k}+1} {}_{\mathrm{u}-\mathrm{k}}p^{\mathrm{inc}}_{\mathrm{age,k}} \times \mu^{\mathrm{inc}\to\mathrm{DC}}_{\mathrm{age,k}}\, du \\
&= {}_{\mathrm{k}-\mathrm{anc}}p^{\mathrm{inc}}_{\mathrm{age,anc}} \times q^{\mathrm{inc}}_{\mathrm{age,k}}
\end{aligned}$$

Avec :

$$ {}_{\mathrm{k}-\mathrm{anc}}p^{\mathrm{inc}}_{\mathrm{age,anc}} = \frac{L^{\mathrm{inc}}_{\mathrm{age,k}}}{L^{\mathrm{inc}}_{\mathrm{age,anc}}} \qquad \text{et} \qquad q^{\mathrm{inc}}_{\mathrm{age,k}} = 1 - \frac{L^{\mathrm{inc,DC}}_{\mathrm{age,k}+1}}{L^{\mathrm{inc,DC}}_{\mathrm{age,k}}}$$

En utilisant les notations suivantes :

- $\mu^{\mathrm{inv}\to\mathrm{DC}}_{\mathrm{age},u}$ la force de mortalité instantanée en invalidité à l'ancienneté u pour une personne entrée en incapacité à l'âge age.
- $q^{\mathrm{inc}}_{\mathrm{age,k}}$ la probabilité de décès à 1 mois à l'ancienneté k pour une personne entrée en incapacité à l'âge age.
- $L^{\mathrm{inc,DC}}_{\mathrm{age,k}}$ le nombre de survivants dans la table de mortalité en incapacité, après une ancienneté de k mois depuis l'entrée en incapacité à l'âge age.

Finalement nous obtenons la formule de la provision pour le maintien de la garantie décès en incapacité (Équation 5.4) :

$$\mathrm{PM}^{\mathrm{DC}}_{\mathrm{inc}}(\mathrm{age},\mathrm{anc}) \simeq \sum_{\mathrm{k=anc}}^{35} {}_{\mathrm{k-anc}}p^{\mathrm{inc}}_{\mathrm{age,anc}} \times q^{\mathrm{inc}}_{\mathrm{age,k}} \times \nu^{\frac{\mathrm{k}+0.5-\mathrm{anc}}{12}} \tag{5.4}$$

Maintien de la garantie décès pour l'invalidité en attente

Pour la provision en cas de décès pendant l'invalidité en attente, il nous suffit de reprendre la formule pour la provision d'invalidité en attente et de remplacer dans la formule la provision pour invalidité par la provision pour décès en invalidité.

On obtient alors la formule de la provision pour le maintien de la garantie décès pour l'invalidité en attente (Équation 5.5) incapacité:

$$\mathrm{PM}^{\mathrm{DC}}_{\mathrm{pass}}(\mathrm{age},\mathrm{anc}) \simeq \sum_{\mathrm{k=anc}}^{35} {}_{\mathrm{k-anc}}p^{\mathrm{inc}}_{\mathrm{age,anc}} \times p^{\mathrm{inc}\to\mathrm{inv}}_{\mathrm{age,k}} \times \nu^{\frac{\mathrm{k}+0.5-\mathrm{anc}}{12}} \times \mathrm{PM}^{\mathrm{DC}}_{\mathrm{inv}}(\mathrm{age}+\frac{\mathrm{k}+0.5}{12},0) \tag{5.5}$$

Avec :

$$ {}_{k-\mathrm{anc}}p^{\mathrm{inc}}_{\mathrm{age,anc}} = \frac{L^{\mathrm{inc}}_{\mathrm{age,k}}}{L^{\mathrm{inc}}_{\mathrm{age,anc}}} \qquad \text{et} \qquad p^{\mathrm{inc}\to\mathrm{inv}}_{\mathrm{age,k}} = \frac{N^{\mathrm{inc}\to\mathrm{inv}}_{\mathrm{age,k}}}{L^{\mathrm{inc}}_{\mathrm{age},k}}$$

Et $\mathrm{PM}^{\mathrm{DC}}_{\mathrm{inv}}(\mathrm{age}+\frac{k+0.5}{12},0)$ est obtenu par interpolation linéaire entre $\mathrm{PM}^{\mathrm{DC}}_{\mathrm{inv}}(\lfloor \mathrm{age}+\frac{k+0.5}{12}\rfloor,0)$ et $\mathrm{PM}^{\mathrm{DC}}_{\mathrm{inv}}(\lfloor \mathrm{age}+\frac{k+0.5}{12}\rfloor+1,0)$

Pour le calcul de la provision pour le maintien de la garantie décès en invalidité voir Maintien de la garantie décès en invalidité.

Nous avons ainsi toutes les formules nécessaires pour calculer la provision pour l'incapacité globale. En combinant les provisions pour l'incapacité en cours, pour l'invalidité en attente, ainsi que les provisions pour le maintien de la garantie décès en incapacité et en invalidité, nous pouvons estimer de manière adéquate les montants nécessaires pour couvrir les risques liés à l'incapacité.

Invalidité en cours

Nous passons à l'invalidité en commençant par l'invalidité en cours. La provision pour l'invalidité en cours pour une prestation **annuelle** de 1 euro peut être exprimée comme suit :

$$\mathrm{PM}_{\mathrm{inv}}(\mathrm{age},\mathrm{anc}) = \int_{u=\mathrm{anc}}^{62-\mathrm{age}} {}_{u-\mathrm{anc}}p^{\mathrm{inv}}_{\mathrm{age,anc}} \times \nu^{u-\mathrm{anc}}\, du$$

En utilisant méthode des rectangles milieu nous obtenons la formule de la provision pour l'invalidité en cours (Équation 5.6) :

$$\mathrm{PM}_{\mathrm{inv}}(\mathrm{age},\mathrm{anc}) \simeq \frac{1}{2} \sum_{\mathrm{k=anc}}^{62-\mathrm{age}-1} {}_{\mathrm{k-anc}}p^{\mathrm{inv}}_{\mathrm{age,anc}} \times \nu^{\mathrm{k-anc}} + {}_{\mathrm{k+1-anc}}p^{\mathrm{inv}}_{\mathrm{age,anc}} \times \nu^{\mathrm{k+1-anc}} \tag{5.6}$$

Avec :

$${}_{\mathrm{k-anc}}p^{\mathrm{inv}}_{\mathrm{age,anc}} = \frac{L^{\mathrm{inv}}_{\mathrm{age,k}}}{L^{\mathrm{inv}}_{\mathrm{age,anc}}} \qquad \text{et} \qquad {}_{\mathrm{k+1-anc}}p^{\mathrm{inv}}_{\mathrm{age,anc}} = \frac{L^{\mathrm{inv}}_{\mathrm{age,k+1}}}{L^{\mathrm{inv}}_{\mathrm{age,anc}}}$$

En utilisant les notations suivantes :

- ${}_{u-\mathrm{anc}}p^{\mathrm{inv}}_{\mathrm{age,anc}}$ la probabilité conditionnelle de maintien en invalidité à l'ancienneté u, pour une personne entrée en invalidité à l'âge age et d'ancienneté anc années à la date de calcul.
- ${}_{\mathrm{k-anc}}p^{\mathrm{inv}}_{\mathrm{age,anc}}$ la probabilité conditionnelle de maintien en invalidité à 1 an à l'ancienneté k, pour une personne entrée en invalidité à l'âge age et d'ancienneté anc années à la date de calcul.
- $L^{\mathrm{inv}}_{\mathrm{age,k}}$ le nombre de survivants dans la table de maintien en invalidité, après une ancienneté de k années depuis l'entrée en invalidité à l'âge age.

Maintien de la garantie décès en invalidité

Il reste à calculer la provision en cas de décès pour l'invalidité. Cette provision est proportionnelle au capital décès garanti et nécessite l'utilisation d'une table de mortalité en invalidité. (Construite par le BCAC en 2002)

Pour un capital décès garanti de 1 euro, la provision en cas de décès à constituer au titre de l'invalidité est donnée par la formule suivante :

$$\mathrm{PM}^{\mathrm{DC}}_{\mathrm{inv}}(\mathrm{age},\mathrm{anc}) = \int_{u=\mathrm{anc}}^{62-\mathrm{age}} {}_{u-\mathrm{anc}}p^{\mathrm{inv}}_{\mathrm{age,anc}} \times \mu^{\mathrm{inv}\to\mathrm{DC}}_{\mathrm{age,anc}} \times \nu^{u-\mathrm{anc}}\, du$$

Les calculs sont identiques que pour Maintien de la garantie décès en incapacité.

Nous obtenons donc finalement la formule de la provision pour le maintien de la garantie décès en invalidité (Équation 5.7):

$$\mathrm{PM}^{\mathrm{DC}}_{\mathrm{inv}}(\mathrm{age},\mathrm{anc}) \simeq \sum_{\mathrm{k=anc}}^{62-\mathrm{age}-1} {}_{\mathrm{k-anc}}p^{\mathrm{inv}}_{\mathrm{age,anc}} \times q^{\mathrm{inv}}_{\mathrm{age,k}} \times \nu^{\mathrm{k+0.5-anc}} \tag{5.7}$$

Avec :

$${}_{\mathrm{k-anc}}p^{\mathrm{inv}}_{\mathrm{age,anc}} = \frac{L^{\mathrm{inv}}_{\mathrm{age,k}}}{L^{\mathrm{inv}}_{\mathrm{age,anc}}} \qquad \text{et} \qquad q^{\mathrm{inv}}_{\mathrm{age,k}} = 1 - \frac{L^{\mathrm{inv,DC}}_{\mathrm{age,k+1}}}{L^{\mathrm{inv,DC}}_{\mathrm{age,k}}}$$

En utilisant les notations suivantes :

- $\mu^{\mathrm{inv}\to\mathrm{DC}}_{\mathrm{age},u}$ la force de mortalité instantanée en invalidité à l'ancienneté u pour une personne entrée en incapacité à l'âge age.

- $q^{\text{inv}}_{\text{age,k}}$ la probabilité de décès à 1 an à l'ancienneté k, pour une personne entrée en invalidité à l'âge age.
- $L^{\text{inv}}_{\text{age,k}}$ le nombre de survivants dans la table de maintien en invalidité, après une ancienneté de k années depuis l'entrée en invalidité à l'âge age.
- $L^{\text{inv,DC}}_{\text{age,k}}$ le nombre de survivants dans la table de mortalité en invalidité, après une ancienneté de k années depuis l'entrée en incapacité à l'âge age.

Nous avons ainsi toutes les formules nécessaires pour calculer la provision pour l'invalidité globale. En combinant les provisions pour l'invalidité en cours ainsi que la provision pour le maintien de la garantie décès en invalidité, nous pouvons estimer de manière adéquate les montants nécessaires pour couvrir les risques liés à l'invalidité

5.2. Modalité de calcul des provisions

Les équations précédentes sont conçues pour le calcul des provisions à des âges et des anciennetés entiers. Cependant, dans le portefeuille, les individus n'ont généralement pas des âges et des anciennetés entiers au moment du calcul. Pour obtenir des estimations précises des provisions, deux approches sont envisageables.

La première méthode consiste à arrondir les âges et les anciennetés au nombre entier le plus proche. Cette méthode est la plus simple à mettre en place est donne un résultat en moyenne juste.

L'alternative consiste en une interpolation bilinéaire, qui vise à mieux approcher la valeur réelle de la provision. Bien qu'un peu plus complexe à implémenter, cette méthode offre des estimations plus précises au niveau individuel. Nous avons opté pour cette méthode dans le calcul de nos provisions, étant donné qu'elle représente une approche plus fidèle à la réalité. Nous détaillons plus en profondeur la mise en œuvre de cette méthode.

En prenant en compte les paramètres x et y, correspondant respectivement à l'âge exact en années et à l'ancienneté exacte en mois, nous définissons les valeurs suivantes :

- $x_1 = \lfloor x \rfloor$, l'âge entier immédiatement inférieur à l'âge à l'entrée en incapacité ou en invalidité.
- $y_1 = \lfloor y \rfloor$, l'ancienneté entière immédiatement inférieure à l'ancienneté à la date de calcul.
- $x_2 = x_1 + 1$, l'âge entier immédiatement supérieur à l'âge à l'entrée en incapacité ou en invalidité.
- $y_2 = y_1 + 1$, l'ancienneté entière immédiatement supérieure à l'ancienneté à la date de calcul.

Notons f la fonction correspondant à l'une des formules précédemment, le coefficient peut être obtenu selon la méthode suivante :

$$\begin{aligned} f(x,y) = {} & f(x_1,y_1) + (x-x_1) \times [f(x_2,y_1) - f(x_1,y_1)] \\ & + (y-y_1) \times [f(x_1,y_2) - f(x_1,y_1)] \\ & + (x-x_1) \times (y-y_1) \times [f(x_1,y_1) + f(x_2,y_2) - f(x_2,y_1) - f(x_1,y_2)] \end{aligned}$$

Cette méthodologie se révèle être la meilleure approche pour réaliser le calcul des provisions de manière optimale et précise.

5.3. Impact de la table d'expérience sur les provisions

Nous avons donc calculé d'une part les provisions pour les contrats TNS et collectifs à partir de la table du BCAC, puis avec nos modèles n'utilisant pas la franchise, celui utilisant la franchise et celui utilisant le type de contrats. Nous présentons les résultats pour l'incapacité en cours et pour l'invalidité en attente.

Nous pouvons résumer nos résultats dans la Table 5.1 :

Table 5.1. : Mesure de l'impact de l'utilisation des tables d'expérience sur les provisions par rapport à la table du BCAC 2010

	Incapacité en cours	Invalidité en attente
Contrats TNS		
Table du BCAC	100%	100%
Table d'expérience	74,94%	62,8%
Table avec type de contrat	97,35%	89,39%
Contrats Collectifs		
Table du BCAC	100%	100%
Table d'expérience	76,19%	61,2%
Table avec type de contrat	74,23%	58,9%
Tous les contrats		
Table du BCAC	100%	100%
Table d'expérience	75,91%	61,56%
Table avec type de contrat	79,45%	65,62%

Nous observons que l'utilisation de notre table d'expérience a un impact d'environ 25% à la baisse sur l'ensemble des types de contrats pour les incapacités en cours. Cependant, lorsque nous prenons en compte le type de contrat, nous constatons une augmentation significative des provisions pour les contrats TNS, ce qui réduit l'impact de la table d'expérience à 20% au niveau global.

En ce qui concerne les invalidités en attente, l'impact est encore plus marqué, avec un écart de 40% par rapport à la première table d'expérience, et de 35% en prenant en compte le type de contrat.

Il est important de souligner que la prise en compte du type de contrat rend les provisions globales plus prudentes. Cette approche plus fine de l'estimation du risque permet de distinguer les risques propres aux contrats collectifs des risques liés aux travailleurs non-salariés (TNS).

Par la suite, nous nous penchons sur l'impact de la prise en compte de la franchise sur les provisions. Cette fois-ci, nous utiliserons comme référence la table d'expérience qui ne tient pas compte du type de contrat. Les résultats sont représenter dans la Table 5.2

Table 5.2. : Mesure de l'impact de la prise en compte de la franchise sur les provisions

	Incapacité en cours	Invalidité en attente
Contrats TNS		
Table d'expérience avec franchise	100%	100%
Table d'expérience sans franchise	103, 44%	102, 69%
Contrats Collectifs		
Table d'expérience avec franchise	100%	100%
Table d'expérience sans franchise	102, 5%	102, 79%
Tous les contrats		
Table d'expérience avec franchise	100%	100%
Table d'expérience sans franchise	102, 71%	102, 77%

Nous constatons un résultat apparemment surprenant, avec un écart très faible, inférieur à 3%, que l'on pourrait considérer comme non significatif. Cette observation s'explique par le fait que les provisions sont proportionnelles à l'espérance de la durée résiduelle actualisée de l'arrêt. Cependant, l'importance de la franchise se révèle principalement pour les anciennetés faibles (inférieures à 4 mois). Pour évaluer cette influence, nous avons effectué le calcul des provisions uniquement sur les arrêts dont l'ancienneté est inférieure ou égale à 4 mois à la date de calcul, et nous avons alors observé un écart de 10% sur les provisions.

Par conséquent, même si l'impact peut sembler négligeable dans ce cas précis, il ne doit pas être sous-estimé. En réalité, il dépend de la structure des contrats sélectionnés, et une modification de cette structure peut avoir des répercussions beaucoup plus significatives sur les provisions. Il est essentiel de prendre en compte la franchise dans la construction des tables malgré que l'impact sur les provisions peut s'avérer être faible.

Finalement, notre étude démontre l'importance cruciale de prendre en compte à la fois le type de contrats et la franchise dans la création de table d'expérience. En incluant ces deux éléments, nous parvenons à des provisions plus justes et à une vision "best-estimate" améliorée. La significativité du coefficient pour les travailleurs non-salariés valide l'impact de cette variable sur le risque en incapacité.

De plus, la prise en compte de la franchise permet de réduire considérablement la probabilité de maintien pour les arrêts de faible ancienneté, offrant ainsi une meilleure appréhension du risque.

En conclusion, en adoptant une approche plus sophistiquée qui intègre le type de contrats et la franchise dans notre modèle, nous améliorons la fiabilité de nos estimations de provisions.

Chapitre 6.

Invalidité

6.1. Présentation des données d'invalidité

En ce qui concerne l'invalidité, un travail considérable et itératif d'amélioration et de fiabilisation de la base des personnes invalides a été entrepris. Cette démarche a permis d'établir une base de données comprenant 436 cas d'invalidité, pour lesquels les dates de début, de fin, ainsi que les causes de cessation ont été soigneusement validées.

Sur la Figure 6.1, tout comme pour l'incapacité, sont exposés les divers filtres appliqués à la base de données pour obtenir notre ensemble de travail. Il est essentiel de souligner que la base ainsi obtenue répond à des critères de cohérence stricts, garantissant la qualité des données analysées.

À partir de cette base de données, nous avons initié une première analyse en examinant les tranches d'âge présentes. Pour ce faire, nous avons élaboré l'histogramme illustré sur la Figure 6.2. Nos observations révèlent que l'individu le plus jeune à avoir fait face à une invalidité avait 28 ans au moment de son occurrence, tandis que le plus âgé avait 63 ans.

La distribution des âges, telle que présentée dans l'histogramme, s'avère cohérente avec les attentes associées à ce type de risque. En effet, l'invalidité tend à se manifester à des âges plus avancés, la majorité des cas survenant après l'âge de 50 ans. Cette constatation confirme la tendance générale de l'invalidité à être plus fréquente chez les individus plus âgés.

Finalement, une analyse approfondie a été réalisée sur la durée des invalidités ainsi que sur leurs causes de fin, et les résultats sont présentés à travers la Figure 6.3. Cette analyse met en évidence des tendances significatives : les reprises ainsi que les décès au sein de notre portefeuille surviennent rapidement après le début de l'invalidité, se produisant tous dans les 5 premières années (à l'exception d'un décès). Par ailleurs, la répartition des causes de cessation offre des informations pertinentes. Avec 72% des sinistres censurés, il est notable que très peu de sorties d'invalidité non censurées sont observées. Cette constatation souligne la prédominance des sinistres censurés dans notre ensemble de données, ce qui peut avoir des implications importantes pour l'application de notre modèle.

En effet, le modèle que nous avons appliqué pour l'incapacité nécessite non seulement une quantité de données adéquate, mais également un nombre suffisant d'événements, c'est à dire de fin d'invalidité qui ne soit pas une censure. Les résultats de cette analyse laissent entrevoir

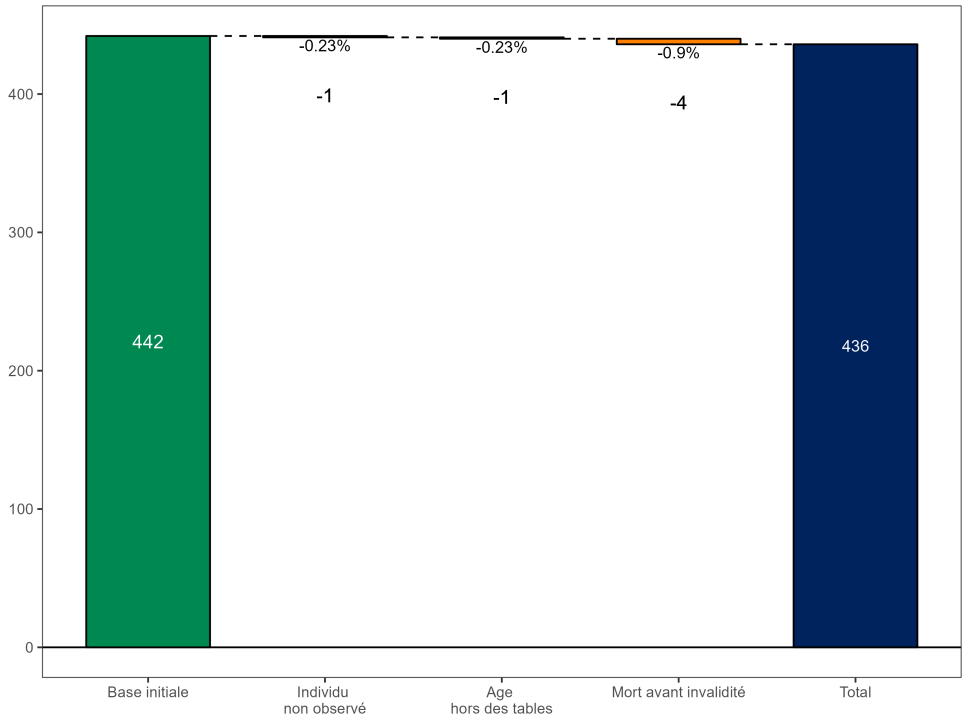


Figure 6.1. : Filtrage des données d'invalidité

une insuffisance du nombre d'événements pour appliquer le même modèle de manière effective. Ainsi, nous sommes confrontés à la nécessité d'étudier l'exposition, afin d'évaluer la faisabilité de l'application de ce modèle. L'analyse de l'exposition permettra d'éclairer la décision quant à la viabilité de l'application du modèle choisi dans le contexte de notre étude sur l'incapacité.

La Figure 6.4 illustre l'exposition au sein de notre base d'invalidité. De manière frappante, elle met en évidence un déficit notable de données, en particulier pour les tranches d'âge inférieures à 40 ans. Globalement, l'exposition demeure significativement limitée avant l'âge de 50 ans, ne permettant pas une couverture adéquate pour une partie substantielle d'une table de maintien en invalidité.

6.2. Comparaison à la table règlementaire

Les différentes études que nous avons menées nous conduisent à la conclusion qu'il est impossible d'établir un modèle similaire à ceux appliqués pour l'incapacité. Cependant, une alternative viable se présente : la calibration du modèle suivant :

$$\log(\mathbb{E}(\text{Sortie})) = \log(\alpha) + \log(\mu_{\text{BCAC}}) + \log(\text{Expo})$$

Cette fois-ci, le coefficient α estimé est constant et ne dépend ni de l'âge ni de l'ancienneté.

L'intérêt de ce modèle réside dans sa capacité à situer notre portefeuille par rapport à la loi du BCAC. En effet, nous calibrons un seul coefficient multiplicatif sur les taux de sorties par rapport au BCAC.

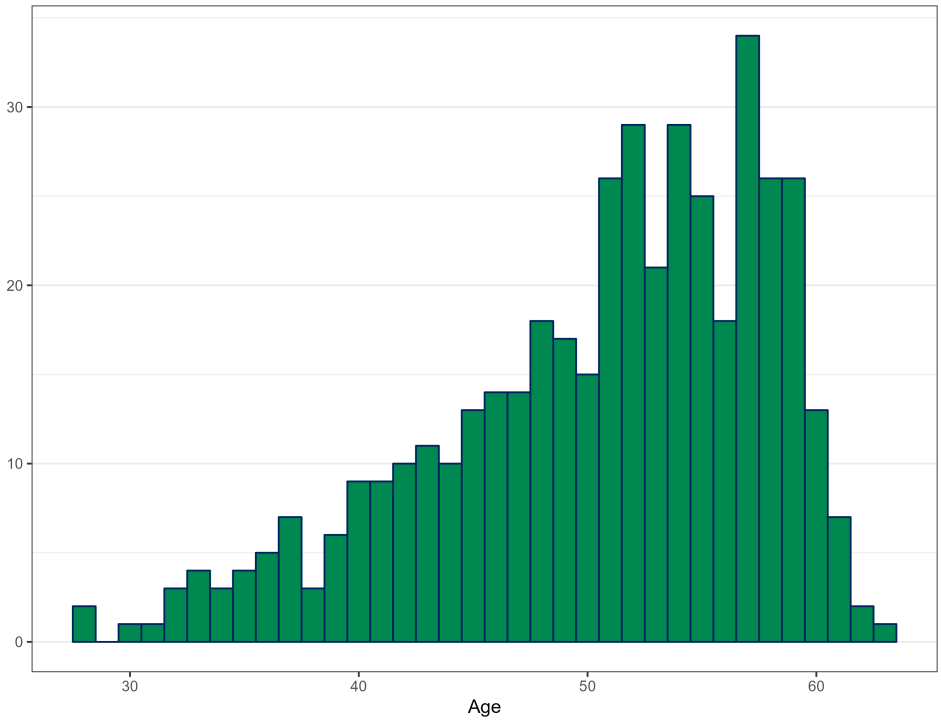


Figure 6.2. : Histogramme des âges pour les invalides

Ce qui retient notre attention dans ce modèle, c'est l'intervalle de confiance fourni pour notre coefficient. Nous pourrons l'interpréter de la manière suivante : Si l'intervalle contient 1, alors ce test suggère qu'avec les données disponibles, il n'est pas possible d'affirmer qu'il existe une différence statistique entre la loi de maintien en invalidité de notre portefeuille et celle du BCAC. En revanche, si l'intervalle n'inclut pas 1, les données indiquent qu'il est probable qu'il existe une différence significative entre la loi suivie par notre portefeuille et celle du BCAC.

6.3. Interprétation des résultats du modèle

Après avoir effectué la calibration de notre modèle, les résultats obtenus sont les suivants :

- Le coefficient est estimé à 2,69.
- L'intervalle de confiance à 95% est [2,15; 3,38].

Étant donné que la valeur 1 ne se trouve pas dans l'intervalle de confiance, nous pouvons en déduire que, a priori, le portefeuille ne suit pas la loi du BCAC. Cette constatation incite à la collecte de données plus étendues et à la réalisation d'une étude approfondie une fois un historique de données plus conséquent disponible. De plus, une estimation du coefficient supérieur à 1 suggère que, a priori, les assurés ont une probabilité de sortie instantanée plus élevée, ce qui se traduirait par des arrêts plus courts que ceux prévus par le BCAC. Cela pourrait entraîner une baisse des provisions avec une nouvelle loi basée sur les données spécifiques du portefeuille.

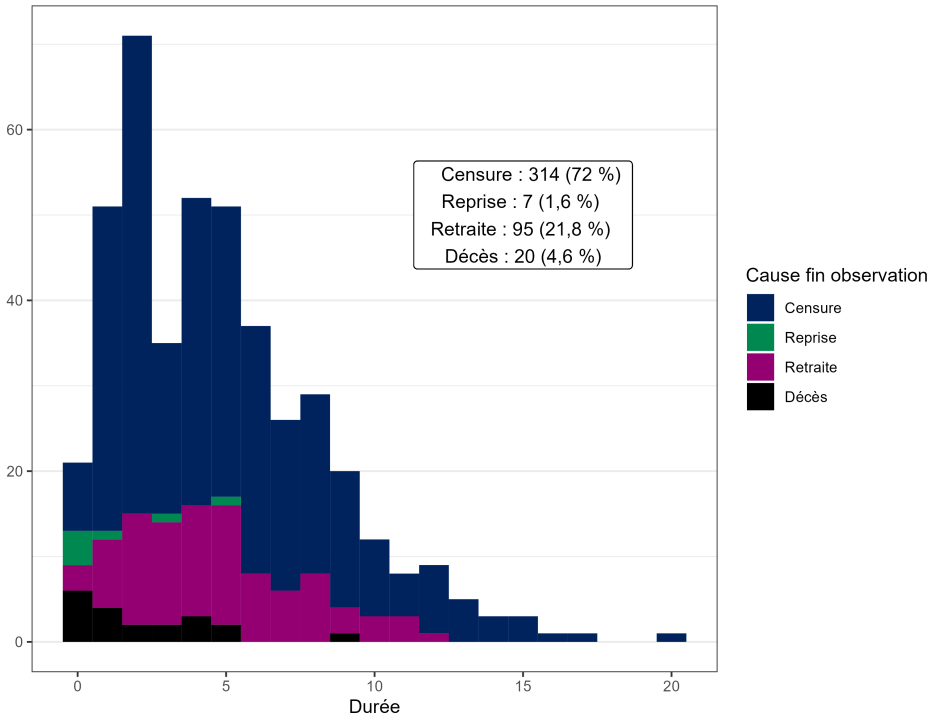


Figure 6.3. : Histogramme des durées des invalidités

Aussi, compte tenu du manque de données que l'on a évoqué, ainsi que de la largeur de l'intervalle de confiance, il n'est pas prudent de créer une table d'expérience et de l'utiliser pour les calculs de provisions. Les tables du BCAC concernant le maintien en invalidité sont jugées prudentes d'après cette étude, ce qui justifie leur utilisation dans nos calculs de provisions précédents.

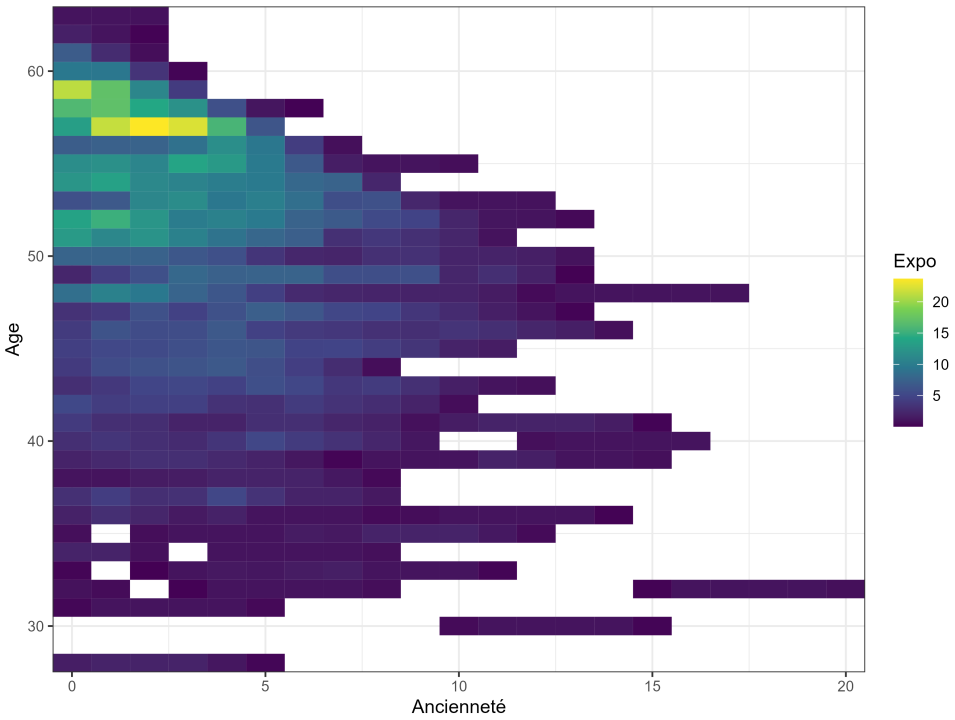


Figure 6.4. : Représentation 2D de l'exposition pour les invalides

Conclusion

Synthèse des résultats

Dans ce mémoire, nous avons présenté une méthode de construction de table d'expérience à la fois simple et polyvalente. Cette approche permet une prise en compte efficace de la censure à droite, de la troncature à gauche, ainsi que de plusieurs variables explicatives. Pour arriver à la table d'expérience, il est nécessaire dans un premier temps de mener un calcul d'exposition sur nos données, puis de déterminer les hyperparamètres des modèles GAM. Les résultats se présentent sous la forme de taux de sortie instantanés, que nous avons ensuite convertis en probabilités de maintien grâce à une simple relation.

Nous avons également démontré l'importance de prendre en compte la franchise dans nos données. En effet, ne pas prendre en compte cette information conduit à surestimer de 200% le durée des arrêts de travail. Cet impact est néanmoins plus faible sur les provisions, car il se concentre principalement sur les premières anciennetés.

De plus, nous avons amélioré notre modèle, en termes d'AIC et de coefficient R^2 ajusté, en prenant en compte la variable explicative du périmètre (Collectif/TNS). La prise en compte de cette variable explicative s'est faite à travers l'estimation d'un coefficient dédié aux TNS.

Nous avons ensuite mesuré l'impact de ces modèles sur les provisions pour l'incapacité en cours et l'invalidité en attente, et il en ressort une diminution des provisions respectives de 20% et 35%. Ces écarts significatifs démontrent l'inadéquation des tables du BCAC et l'importance de construire des tables d'expérience.

De plus, nous avons réalisé un travail d'éclaircissement des formules de provisions pour le risque incapacité/invalidité. À ce jour, il n'existe pas de document statuant officiellement et de manière claire sur les formules à utiliser. Cependant, il est essentiel de prendre en compte certaines subtilités, notamment lorsque les tables de maintien en incapacité ne sont pas accompagnées de tables de passage en invalidité. Pour combler cette lacune, nous avons établi ces formules en partant de la forme générale sous forme d'intégrale des provisions que nous cherchons à estimer. Ce travail contribue ainsi à une meilleure compréhension et à une application plus précise des formules de provisions pour le risque d'incapacité/invalidité.

Une étude approfondie de la base des invalides a été menée, révélant une limite significative dans notre capacité de modélisation : le manque d'événements observés. Cette contrainte nous a empêchés d'utiliser le modèle précédemment adopté. Par conséquent, une comparaison avec la loi réglementaire de maintien en invalidité du BCAC a été entreprise. Cette analyse a révélé des différences substantielles entre notre portefeuille et la loi du BCAC. En outre, il est apparu que les tables fèglementaire étaient plus pessimistes que la loi de notre portefeuille.

En conclusion, il est établi que la table du BCAC adopte une approche prudente en matière de constitution des provisions relatives à l'invalidité.

Limites et perspectives

Cependant, malgré les avantages de notre approche basée sur les modèles additifs généralisés, il est important de reconnaître certaines limites et d'envisager des perspectives d'amélioration. Les Modèles Additifs Généralisés (GAM), tels qu'intégrés dans notre modèle, offrent des possibilités d'amélioration significatives avec une exploitation plus approfondie des données.

En effet, actuellement, la variable explicative est prise en compte par un coefficient pour chaque modalité de la variable, valable pour tous les âges et toutes les anciennetés. En disposant de volumes de données plus importants, une approche plus avancée consiste à ajuster non plus seulement un coefficient par modalité de variable explicative, mais plutôt à calibrer directement un prédicteur lisse pour chaque modalité. Cela signifie que l'effet de la variable dépendrait à la fois de l'âge de l'individu et de son ancienneté.

Cette démarche, bien que potentiellement plus puissante, nécessite une mise en garde importante. En effet, calibrer un prédicteur lisse pour chaque modalité de variable explicative augmente considérablement le nombre de paramètres à estimer. Cela implique une plus grande complexité dans le processus d'estimation. Par conséquent, une quantité de données importante devient cruciale pour garantir une estimation robuste et éviter des problèmes de sur-apprentissage. En résumé, bien que cette approche puisse offrir une précision accrue, elle est tributaire de l'abondance de données pour traiter efficacement la complexité ajoutée lors de l'estimation des paramètres.

Une perspective intéressante, qui nécessite également une quantité de données conséquente, consiste à modéliser les taux de sorties instantanés par type de sortie. Avec notre approche actuelle, il serait tout à fait envisageable d'obtenir des taux de sortie vers l'invalidité, le décès, ou la reprise de travail (en prenant l'incapacité comme exemple). Ce faisant, nous pourrions directement obtenir des lois de maintien et de passage pour les différentes sorties, tout en assurant une cohérence entre elles. Cette cohérence consisterait à ce que la somme des probabilités de sorties pour chaque type de sortie soit égale à la probabilité de sortie globale. En effet, il a été démontré par Hardy (2019) que cette propriété n'est pas vérifiée avec les tables actuelles du BCAC.

Ces perspectives d'amélioration nous invitent à poursuivre nos travaux dans le domaine de la modélisation des risques d'incapacité et d'invalidité en exploitant pleinement les potentialités offertes par les modèles additifs généralisés . En prenant en compte ces limitations et en explorant ces pistes d'amélioration, notre modèle pourrait évoluer vers une approche encore plus précise et cohérente pour l'estimation des provisions et l'évaluation du risque lié à l'incapacité et l'invalidité.

Bibliographie

Bagui, H. 2013. « Refonte des lois de maintien en incapacité temporaire de travail ». Mémoire d'actuariat, ISFA.

Biessy, G. 2022. « Etude R&D LinkPact sur la mortalité ». 2022. https://linkpact.fr/document/mortalite/notions.html.

———. 2023. « Revisiting Whittaker-Henderson Smoothing ». *arXiv preprint arXiv:2306.06932*.

Cox, David R. 1972. « Regression models and life-tables ». *Journal of the Royal Statistical Society: Series B (Methodological)* 34 (2): 187-202.

Hardy, E. 2019. « Mesure du risque de réserve des provisions mathématiques en arrêt de travail ». CEA, Institut du Risk Management.

Hastie, T, et R Tibshirani. 1987. « Generalized additive models: some applications ». *Journal of the American Statistical Association* 82 (398): 371-86.

Hastie, T, R Tibshirani, Jerome H Friedman, et Jerome H Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer.

Henderson, R. 1924. « A new method of graduation ». *Transactions of the Actuarial Society of America* 25: 29-40.

Kaplan, L, et P Meier. 1958. « Nonparametric estimation from incomplete observations ». *Journal of the American statistical association* 53 (282): 457-81.

Petit, C. 2017. « Construction d'une table de maintien en incapacité ». Mémoire d'actuariat, ISFA.

Reiss, Philip T, et R Todd Ogden. 2009. « Smoothing parameter selection for a class of semiparametric linear models ». *Journal of the Royal Statistical Society Series B: Statistical Methodology* 71 (2): 505-23.

Turnbull, Bruce W. 1976. « The empirical distribution function with arbitrarily grouped, censored and truncated data ». *Journal of the Royal Statistical Society: Series B (Methodological)* 38 (3): 290-95.

Wang, Mei-Cheng, Nicholas P Jewell, et Wei-Yann Tsai. 1986. « Asymptotic properties of the product limit estimate under random truncation ». *the Annals of Statistics*, 1597-1605.

Wolfrum, R. 1993. « Une alternative non paramétrique au calcul des provisions techniques en assurances invalidité ». Mémoire d'actuariat, ISUP.

Wood, Simon N. 2011. « Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models ». *Journal of the Royal Statistical Society Series B: Statistical Methodology* 73 (1): 3-36.

———. 2017. *Generalized additive models: an introduction with R*. CRC press.

Annexe A.

Notations

On répertorie ici toutes les notations utilisées pour les formules de provisions :

- $\nu = \frac{1}{1+i}$ avec i le taux d'actualisation annuel.
- ${}_{\text{k}-\text{anc}}p^{\text{inc}}_{\text{age,anc}}$ la probabilité conditionnelle de maintien en incapacité à 1 mois à l'ancienneté k, pour une personne entrée en incapacité à l'âge age et d'ancienneté anc mois à la date de calcul.
- ${}_{\text{k}-\text{anc}}p^{\text{inv}}_{\text{age,anc}}$ la probabilité conditionnelle de maintien en invalidité à 1 an à l'ancienneté k, pour une personne entrée en invalidité à l'âge age et d'ancienneté anc années à la date de calcul.
- $p^{\text{inc}\to\text{inv}}_{\text{age,k}}$ la probabilité de passage en invalidité entre le mois k et k$+1$ pour une personne entrée en incapacité à l'âge age.
- $q^{\text{inc}}_{\text{age,k}}$ la probabilité de décès à 1 mois à l'ancienneté k pour une personne entrée en incapacité à l'âge age.
- $q^{\text{inv}}_{\text{age,k}}$ la probabilité de décès à 1 an à l'ancienneté k, pour une personne entrée en invalidité à l'âge age.
- $L^{\text{inc}}_{\text{age,k}}$ le nombre de survivants dans la table de maintien en incapacité, après une ancienneté de k mois depuis l'entrée en incapacité à l'âge age.
- $L^{\text{inv}}_{\text{age,k}}$ le nombre de survivants dans la table de maintien en invalidité, après une ancienneté de k années depuis l'entrée en invalidité à l'âge age.
- $N^{\text{inc}\to\text{inv}}_{\text{age,k}}$ le nombre de passages dans la table de passage, après une ancienneté de k mois depuis l'entrée en incapacité à l'âge age.
- $L^{\text{inc,DC}}_{\text{age,k}}$ le nombre de survivants dans la table de mortalité en incapacité, après une ancienneté de k mois depuis l'entrée en incapacité à l'âge age.
- $L^{\text{inv,DC}}_{\text{age,k}}$ le nombre de survivants dans la table de mortalité en invalidité, après une ancienneté de k années depuis l'entrée en incapacité à l'âge age.
- $\mu^{\text{inc}\to\text{inv}}_{\text{age,k}}$ la force de passage instantanée à l'ancienneté k pour une personne entrée en incapacité à l'âge age.
- $\mu^{\text{inc}\to\text{DC}}_{\text{age},u}$ la force de mortalité instantanée en incapacité à l'ancienneté u pour une personne entrée en incapacité à l'âge age.
- $\mu^{\text{inv}\to\text{DC}}_{\text{age},u}$ la force de mortalité instantanée en invalidité à l'ancienneté u pour une personne entrée en incapacité à l'âge age.