INSTITUT DES ACTUAIRES

Lyon 1

ISFA

Mémoire présenté le : 2 mai 2019

pour l'obtention du Diplôme Universitaire d'actuariat de l'ISFA et l'admission à l'Institut des Actuaires

Par : Léo Stefani

Titre : Méthodes de construction des barèmes tarifaires collectifs en arrêt de travail et apport des données de DSN

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

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

Membre présents du jury de l'Institut des Actuaires signature

S. Caraco

C. Fettig

I. Praud-Lion

Membres présents du jury de l'ISFA

C. Robert

Entreprise : ACTUARIS

Nom :

Signature :

Directeur de mémoire en entreprise :

Nom : C. Paradis

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

ACTUARIS S.A.S.
"Le Valvert" - 46 bis, Chemin du Vieux Moulin
69160 TASSIN LA DEMI LUNE
Tél. 04 72 18 58 58 - Fax 04 72 18 58 59
Siren 413 611 344 - APE 7022Z
SAS au capital de 100 000 €

Signature du candidat

Mots clefs

Arrêt de travail, incapacité, tarification, contrats collectifs, barème de tarification, modèles linéaires généralisés, taux d'entrée en incapacité, durée de maintien en incapacité, estimateur de Kaplan Meier, table d'expérience, Whittaker-Henderson, simulation, déclaration sociale nominative

Résumé

Les bases de données relatives à des contrats d'assurance en prévoyance individuelle et en prévoyance collective n'ont ni les mêmes caractéristiques, ni le même comportement. Là où les bases de données de contrats individuels possèdent généralement l'ensemble des informations relatives à la population assurée, tant sinistrée que non sinistrée, les bases de données des contrats collectifs ne disposent souvent que des informations détaillées relatives aux bénéficiaires sinistrés et sont complétées par des statistiques plus globales sur la population assurée. Il n'est donc traditionnellement pas possible d'appliquer les mêmes méthodes de modélisation pour construire des normes tarifaires personnalisées adaptées au risque arrêt de travail sur ces deux marchés.

Cependant, dans le cadre de ce mémoire d'actuariat réalisé en partenariat avec un organisme assureur, nous avons pu traiter une base de données relative à des contrats en prévoyance collective comportant à la fois des informations détaillées sur les bénéficiaires sinistrés mais également sur les bénéficiaires non sinistrés.

De ce fait, le premier objectif de ce mémoire est de vérifier si les méthodes de modélisation du risque arrêt de travail utilisées traditionnellement sur des portefeuilles individuels peuvent également être appliquées en prévoyance collective. Pour cela, nous retenons les 2 méthodes suivantes :

- une première méthode basée sur la modélisation du risque arrêt de travail à l'aide des modèles linéaires généralisés ;
- une seconde méthode basée sur la modélisation du risque arrêt de travail par simulation, à l'aide de tables d'expérience (passage et maintien).

Ces deux méthodes se heurteront à des problèmes d'hétérogénéités des données et de faibles volumétries de la base des prestations, problèmes fréquents sur des bases de données collectives du fait des multiples contrats présents, des durées de franchises plus ou moins longues, et des différentes populations assurées. De ce fait, les deux modélisations étudiées permettent d'obtenir des résultats « partiels » intéressants mais ne permettent pas d'obtenir un barème tarifaire complet.

Nous analyserons en seconde partie l'impact de l'exploitation des données de la déclaration sociale nominative sur ces méthodes (données que nous ne pouvions pas exploiter à ce jour puisque la DSN n'est partiellement fonctionnelle que depuis 2015). Nous observerons qu'une bonne utilisation de ces nouvelles données permettra de résoudre les difficultés rencontrées précédemment et d'améliorer les résultats.

Du fait des données disponibles et de leur profondeur, ce mémoire relatif à l'arrêt de travail n'aborde pas la tarification du risque invalidité.

Key words

*Sick leave, short term disability, pricing, group insurance policy, rate scale, generalized linear models, disability entrance rate, period of disability, Kaplan Meier's estimator, disability pricing tables, Whittaker-Henderson, simulation, "*déclaration sociale nominative*"*

Abstract

Databases of personal disability insurance contracts and databases of group disability insurance contracts are different. The first one contains all the detailed data about the insured people whereas the second one possesses detailed information only about people which have been seek at least once and global statistics for the rest of the population. Because of this difference, it is usually impossible to apply the same methods in both personal and group insurance in order to price disability contracts.

Nevertheless in this actuarial paper thanks to an insurance institution we own a database of group disability insurance contracts which contains all the insured people, affected and unaffected insured people.

Thanks to this database, the first goal of this paper is to study how personal disability insurance pricing methods can be applied to a database of group disability insurance contracts. We are going to study two pricing methods: the first one uses generalized linear models and the second method uses simulations and disability pricing tables.

Because of the heterogeneity and the lake of exposure of these group insurance contract databases, these methods haven't work perfectly and the expected results couldn't be reached. The heterogeneity of this kind of databases came from the multiples contracts, franchises, and population aggregated in one single data gathering.

However, the "déclaration sociale nominative" a French clustering of many social registrations may improve the quality of these pricing methods and solve the previous problem. The analysis of how these data can be used is the second goal of this paper.

Because of the short volume of data in the used databases, this paper deals only with short term disability and excludes the long term disability.

Remerciements

Je souhaite remercier tous ceux qui ont contribué de près ou de loi à l'élaboration de ce mémoire.

Plus particulièrement, je tiens tout d'abord à remercier le cabinet ACTUARIS et spécialement l'équipe Pricing & Data pour leur apprentissage et leur bonne humeur constante.

Je souhaite aussi sincèrement remercier Audrey PEYRILLER, Cécile PARADIS, Hélène JOURDAIN, Guerric BRAS, et Romain GAUCHON, qui se sont rendus disponible et m'ont aidé à mener à bien ce mémoire grâce à leurs conseils avisés.

Je tiens également à remercier Madame Anne EYRAUD-LOISEL, mon tuteur académique, pour ses conseils avisés.

Enfin, je remercie du fond du cœur ma famille et mes amis, particulièrement mes parents, Camille, Aymeric, et Quentin qui m'ont toujours soutenu durant cette période ce qui m'a permis de réaliser ce mémoire.

Sommaire

Introduction

Dans deux précédents mémoires reçus par l'Institut des Actuaires, PELLICIER J. (2010)[1] et LEBOSSE C. (2009)[2] ont développé deux méthodes de tarification du risque incapacité. Dans ces mémoires, PELLICIER J. a présenté une tarification du risque incapacité par Modèle Linéaire Généralisé *(MLG)* et LEBOSSE C. a introduit une tarification du risque incapacité par une méthode de simulation utilisant des tables d'expérience. Ces deux mémoires ont été réalisés sur une base de données mutualisant les informations de plusieurs organismes assureurs et principalement composée de travailleurs non-salariés. Il était donc question de données de contrats de prévoyance individuelle. L'important volume de données utilisées permet d'affirmer la véracité et la robustesse de ces deux méthodes de tarification.

Le premier objectif du présent mémoire est de transposer ces démarches à une base de données de contrats de prévoyance collective. Le comportement et la composition des bases de données relatives à des contrats de prévoyance individuelle et des contrats de prévoyance collective n'étant pas les mêmes, nous analyserons dans quelle mesure ces méthodes peuvent s'adapter. Nous reprendrons donc les deux méthodes de tarifications : la méthode par MLG et la méthode par simulation que nous appliquerons à des données collectives afin d'obtenir un barème de tarification pour un outil de souscription sur-mesure.

Puis, nous analyserons ce que la Déclaration Sociale Nominative *(DSN)* peut apporter aux deux modélisations présentées précédemment. Actuellement en arrêt de travail, à cause de la troncature gauche des données due à la présence de franchises, l'assureur n'a généralement pas connaissance des sinistres durant moins que la durée de franchise. Avec les déclarations évènementielles relatives à l'arrêt de travail contenues dans la DSN, l'assureur a la possibilité de connaitre tous les sinistres dès leur survenance. Nous étudierons comment ces déclarations vont pouvoir améliorer les processus de tarification précédents.

Afin de traiter les divers éléments exposés dans les paragraphes précédents, nous commencerons par présenter le contexte de ce mémoire avec une partie regroupant les généralités nécessaires à l'étude. Ensuite nous détaillerons la méthode de tarification par MLG suivie de celle par simulation. Et enfin dans une dernière partie nous développerons ce que peut apporter l'ajout de la DSN à la modélisation.

[1] PELLICIER J. (2010) : *Étude des facteurs discriminants en arrêt de travail pour les travailleurs non-salariés – Application à la tarification*, mémoire de l'Institut des Actuaires

[2] LEBOSSE C. (2009) : *Construction de barèmes de tarification d'arrêt de travail par une méthode de Simulation – Application à un portefeuille de Travailleurs Non-Salariés*, mémoire de l'Institut des Actuaires

Partie 1 – Généralités

L'objectif du présent mémoire est dans un premier temps l'obtention d'un barème de tarification du risque incapacité pour un outil de souscription sur-mesure, puis dans un second temps l'analyse de ce que la DSN peut apporter à la modélisation. Dans cette partie, nous présenterons donc le contexte du mémoire. Nous détaillerons tout d'abord le risque arrêt de travail et plus particulièrement le risque incapacité. Ensuite, nous introduirons les données qui seront utilisées pour l'étude, ceci en respectant leur confidentialité. Et enfin nous présenterons la DSN.

I. Le risque arrêt de travail

Lorsque l'activité professionnelle d'un assuré est interrompue à cause d'une maladie professionnelle, d'un accident du travail, ou encore d'un accident de la vie privée, nous parlons d'arrêt de travail. Cette interruption peut être partielle ou totale.

Nous différencions deux types d'arrêts de travail[3] :

- l'**incapacité temporaire**, généralement appelée « incapacité » ;
- l'**incapacité permanente**, généralement appelée « invalidité ».

Remarque : dans la suite du présent mémoire nous utiliserons les dénominations incapacité et invalidité uniquement.

A. Couverture du risque arrêt de travail

L'interruption partielle ou totale du travail de l'assuré engendre une perte partielle ou totale du salaire de ce dernier. Trois acteurs sont présents pour couvrir cette perte de revenus : le régime général de la Sécurité Sociale, l'employeur, et les organismes assureurs via les régimes complémentaires d'assurance. L'état d'arrêt de travail doit être énoncé par un médecin afin que la procédure d'indemnisation soit engagée.

La Sécurité Sociale assure un premier niveau de prestation dépendant du type d'arrêt de travail : arrêt lié à un accident du travail ou une maladie professionnelle, arrêt lié à un accident de la vie privée, ou arrêt pour maternité. En plus de cette distinction, des conditions telles que des délais de carence ou encore des plafonds de prestations vont s'appliquer.

Ensuite, afin de respecter la loi de mensualisation du 19 janvier 1978, l'employeur a l'obligation de maintenir pendant une durée déterminée une certaine proportion de la rémunération du salarié en arrêt de travail si ce dernier a plus de trois ans d'ancienneté. Cette loi a été révisée le 25 juin 2008 de telle sorte que l'ancienneté nécessaire pour bénéficier du maintien employeur est passée de trois ans à un an.

Ces deux interventions plafonnées vont être complétées par l'intervention des organismes assureurs. Trois types d'acteurs sont présents sur le marché de l'assurance :

- les sociétés d'assurance, régies par le Code des Assurances et à but lucratif ;
- les mutuelles, régies par le Code de la Mutualité et à but non lucratif ;

[3] Les caractéristiques d'un arrêt de travail sont définies par le régime de base de la Sécurité Sociale aux articles L321-1, R313-3, et R323-1 du Code de la Sécurité Sociale

- les institutions de prévoyance, régies par le Code de la Sécurité Sociale et à but non lucratif.

Les prestations qui seront versées par ces organismes devront respecter les minimums fixés par les Convention Collectives Nationales ou les accords de branche.

B. Le risque incapacité

Nous allons apporter quelques précisions sur le risque incapacité car ce dernier est au centre de ce mémoire.

Les prestations de la Sécurité Sociale relatives à l'incapacité diffèrent entre une incapacité ayant pour cause un accident du travail ou une maladie professionnelle *(AT/MP)*, une incapacité ayant pour cause un accident de la vie privée *(VP)*, et enfin une incapacité pour motif de maternité. Mis à part la maternité, le versement des indemnités journalières est limité à 1095 jours, soit trois ans. Au-delà de cette durée il y a consolidation de l'état d'incapacité et le salarié passe en invalidité. La sortie de l'état d'incapacité peut se réaliser par quatre évènements : le retour au travail, la retraite, le décès, et l'invalidité.

Bien que non négligeables, les prestations de la Sécurité Sociale restent faibles. D'où la nécessité – même en plus du maintien employeur – de souscrire un contrat de prévoyance complémentaire couvrant le risque arrêt de travail.

L'organisme qui regroupe officiellement les statistiques « marché » sur l'arrêt de travail en France est le Bureau Commun des Assurances Collectives *(BCAC)*. Il est chargé de réaliser les tables règlementaires pour le provisionnement du risque arrêt de travail. Il s'agit des tables de maintien en incapacité, de maintien en invalidité, et de passage de l'état d'incapacité à l'état d'invalidité. D'après les études du BCAC, l'incapacité est un risque court de fréquence où l'âge est un facteur clef de la modélisation. En effet, plus l'individu est vieux, moins il est probable qu'il quitte vite l'état d'incapacité et plus il est probable qu'il transite vers l'état d'invalidité.

C. La tarification du risque incapacité

Tel qu'énoncé dans la partie précédente, le BCAC est chargé de réaliser les tables règlementaires de provisionnement en arrêt de travail. L'article 143-12 du règlement ANC impose l'utilisation de ces tables (à défaut d'avoir une table d'expérience spécifique certifiée par un actuaire habilité par l'Institut des Actuaires). Cependant, en ce qui concerne la tarification, aucune norme n'est imposée. Les organismes assureurs peuvent réaliser leurs propres études de marché afin de réaliser leur barème de tarification.

Remarque : bien qu'aucune norme ne soit imposée, les tarificateurs se doivent de respecter la Gender Directive (2004/113/CE) : le processus de tarification peut être segmenté par sexe, cependant le tarif final doit être unisexe.

Le BCAC fournit quand même un barème de tarification, mais ce dernier n'est pas adapté pour deux raisons. Premièrement, le dernier barème a été publié en 2002 alors que le risque arrêt de travail est en constante évolution. Et deuxièmement, ce barème a été réalisé sur la population française dans son ensemble. Il ne capte donc pas les spécificités des diverses populations assurées. Il est plus sûr pour les organismes assureurs de créer leur propre barème basé sur leur population couverte.

L'étude qui est menée dans ce mémoire a vocation à créer un barème de tarification pour un outil de souscription sur mesure. Les méthodes de tarification qui seront développées seront présentées

dans les parties 2 et 3. Il s'agit dans un premier temps de l'application de MLG, puis dans un second temps de l'application d'une méthode de simulation.

D. Rappels sur les censures et troncatures

Un élément très important dans la tarification du risque arrêt de travail et particulièrement du risque incapacité est la gestion des censures et troncatures. Les données vont contenir une troncature gauche due à la présence de franchise et à la date de début d'observation des données, ainsi qu'une censure droite due à la date d'extraction des données.

La franchise d'un contrat d'arrêt de travail correspond à une durée exprimée en jour pendant laquelle l'assuré ne sera pas indemnisé. Durant cette période l'assureur n'a pas nécessairement connaissance de l'arrêt, d'où la présence de la troncature gauche. De plus, les données étant étudiées à partir d'une certaine date, les arrêts antérieurs à ce début d'observation ne peuvent être observés. Cette date correspond donc elle aussi à une troncature gauche des données.

Rappel : définition de la troncature
Soit X *une durée de survie. La variable* X *est dite tronquée à gauche (respectivement à droite) si elle n'est pas observable lorsqu'elle est inférieure à un seul* $c > 0$ *(respectivement supérieure à un seuil* $C > 0$*).*

La date d'extraction correspond à la date à laquelle l'organisme assureur partenaire nous a fourni ses données. De ce fait nous ne pouvons plus observer les arrêts après la date d'extraction, d'où la présence de censure droite.

Rappel : définition de la censure
Soient une durée de survie X *et* $C > 0$ *fixé. La variable* X *est dite censurée à droite si au lieu d'observer directement* X*, nous observons le couple* $(T; D)$ *tel que :*

$$T = C \wedge D \text{ et } D \begin{cases} 1 \text{ si } X \leq C \\ 0 \text{ si } X > C \end{cases}$$

Dans le cas de la censure gauche, nous n'observons X *que si elle a lieu après la date de censure* C*.*

Résumé du chapitre sur le risque arrêt de travail

Le risque arrêt de travail se divise entre le risque incapacité et le risque invalidité. Dans ce mémoire, seul le risque incapacité sera étudiée.

Dans le cadre des contrats de prévoyance collective, le risque incapacité est partagé entre trois acteurs : le régime général de la Sécurité Sociale, l'employeur, et les organismes assureurs. La sortie de l'état d'incapacité se réalise suite à la survenance d'un des quatre évènements suivants : le retour au travail, la retraite, le décès, et enfin l'invalidité. Le risque incapacité est un risque court de fréquence qui dépend fortement de l'âge.

Le BCAC a publié en 2013 un barème de tarification mais ce dernier n'est ni obligatoire ni fidèle au comportement des populations spécifiques des divers assureurs. Il est donc recommandé aux organismes assureurs de réaliser leur propre évaluation du risque pour créer leur barème de tarification. Deux méthodes pour y parvenir seront présentées dans ce mémoire.

II. Les données de l'étude

En partenariat avec un organisme assureur, nous avons obtenu une base de données de contrats de prévoyance collective. Les données reçues et utilisées ainsi que le nom de l'organisme assureur partenaire sont confidentiels. Ainsi dans cette partie de présentation des données et dans les parties pratiques suivantes, les explications et résultats resteront vagues afin de ne pas nuire à cette confidentialité.

A. Particularité

Commençons tout d'abord par rappeler la différence entre une base de données de contrats de prévoyance individuelle *(appelée dans la suite du mémoire base individuelle)* et une base de données de contrats de prévoyance collective *(appelée dans la suite du mémoire base collective)*.

Une base individuelle est telle qu'une ligne est associée à un individu. La clef identifiant chacune des lignes est le numéro associé à l'individu. A l'inverse, dans une base collective, une ligne est associée à un **contrat individualisé**, c'est-à-dire que nous ajoutons le numéro du contrat à la clef identifiant chacune des lignes. Plus précisément, dans une base individuelle chaque individu n'apparaîtra qu'une fois, alors que dans une base collective un individu peut avoir eu plusieurs contrats pendant sa période d'affiliation et donc apparaitre plusieurs fois.

Lors de la tarification du risque incapacité, la situation est différente entre la possession d'une base individuelle et la possession d'une base collective. Lorsque la tarification est réalisée à partir d'une base individuelle, les tarificateurs possèdent généralement toutes les informations détaillées sur les assurés, aussi bien sinistrés que non sinistrés. A l'inverse, lors de la tarification via une base collective les tarificateurs ne possèdent généralement que le détail des assurés sinistrés ainsi que des statistiques globales sur le reste de la population (la population non sinistrée). De plus, les bases collectives présentent une diversité de données bien plus importante que les bases de données individuelles. En effet, là où les bases individuelles ciblent souvent une population et quelques contrats – ce qui implique une bonne connaissance des assurés et de leurs niveaux de couverture – les bases collectives regroupent de multiples populations et de multiples contrats pouvant être très différents les uns des autres. Ceci ajoute une hétérogénéité non négligeable aux données.

Le point d'intérêt qui justifie les travaux réalisés dans ce mémoire est que nous possédons une base collective possédant l'ensemble des assurés. La base collective que nous utiliserons est composée des contrats individualisés sinistrés mais aussi des contrats individualisés non sinistrés. Ainsi, nous possédons la population complète pour chacun des contrats et personnes morales assurés. **Nous pouvons alors adapter les méthodes de tarification individuelle à une base collective et analyser les résultats obtenus**. Ceci nous permettra d'étudier l'adaptabilité de ces méthodes aux bases collectives et d'analyser si l'hétérogénéité de ces bases est un élément qui viendra perturber le processus de tarification ou non.

B. Présentation

Afin de pouvoir utiliser les données mises à disposition par l'organisme assureur partenaire, nous avons dû réaliser un grand nombre de retraitements. La base de données n'était pas du tout propice à l'emploi au moment de sa réception. Un rapide résumé des retraitements effectués est présenté en ANNEXE 1.

Une fois ces retraitements effectués nous avons à notre disposition deux bases de données.

- La première est la base des bénéficiaires. Composée d'environ 500 000 lignes, elle regroupe l'ensemble des informations relatives aux personnes physiques et aux contrats qui leurs sont associés.
- La seconde est la base des prestations et possède environ 30 000 lignes. Chacune de ces lignes représente une période d'incapacité pour l'un des contrats individualisés de la base des bénéficiaires.

Nous observons les données sur 8 années de 2009 à 2016. Cette période a été sélectionnée afin d'avoir une stabilité des données sur les années étudiées. En effet, les années antérieures à 2009 présentaient des problèmes d'exposition.

Sur la période retenue, nous observons environ 150 000 contrats individualisés par année avec une exposition moyenne de 0,85 ans. Ceci nous donne une exposition globale sur la période de 1 million d'années pour un âge moyen de 40 ans. Cette population est répartie sur l'ensemble du territoire (hors territoires d'outre-mer) et dans de nombreux secteurs d'activités différents. En termes de sinistralité, le taux d'indemnisation de cette population s'élève à environ 2,5%. La population sinistrée est de 3 ans plus vieille et plus féminine que la population globale.

Enfin, un point important de notre étude est que **la population étudiée est composée de deux sous populations distinctes**. La première est une population classique, la seconde est une population dépendant d'une Convention Collective Nationale *(CCN)* particulière. Cette population ajoute une hétérogénéité non négligeable à nos données car elle est en moyenne plus jeune, plus féminine, et moins sinistrée que la population principale.

C. Quelques remarques

Nous possédons une base collective à laquelle nous appliquerons des méthodes basées sur une base individuelle. Il est donc nécessaire de vérifier que les éléments spécifiques aux contrats collectifs ne vont pas venir influencer le comportement de nos données. Plus précisément, nous montrerons dans cette partie que notre base de données n'est influencée ni par le Maintien des Garanties Décès *(MGDC)* ni par la fin administrative des contrats.

Nous parlerons aussi du traitement des rechutes.

1. Maintien des garanties décès

Selon l'article 7-1 de la loi n° 89-1009 du 31 décembre 1989 (loi Evin), tel que modifié par la loi n° 2001-624 du 17 juillet 2001 applicable aux contrats en vigueur à compter du 1er janvier 2002, lorsqu'un salarié possède un contrat collectif couvrant l'arrêt de travail et le décès, et qu'il entre en arrêt de travail, ses garanties prévoyance doivent être maintenues, et ce quand bien même le contrat collectif serait résilié ou non-renouvelé.

Le MGDC aurait pu poser un problème si nous avions cherché à modéliser le coût moyen d'une Indemnité Journalière *(IJ)*. En effet, en fonction de comment sont saisies les données, la prestation due au titre du décès survenu pendant un arrêt de travail aurait pu être ajoutée à la prestation due au titre de l'arrêt de travail. Un tel comportement aurait perturbé la modélisation du coût moyen d'une IJ. Cependant, comme nous le verrons par la suite, nous ne rencontrerons pas de modèle de coût ce qui résout donc ce potentiel problème.

2. Maintien des prestations après résiliation du contrat

D'après l'article 7 de la loi Evin, la résiliation du contrat de prévoyance collective est sans effet sur le versement des prestations différées, acquises ou nées durant l'exécution du contrat. Ceci peut fortement impacter notre modélisation. En effet, il faut savoir si lorsqu'un contrat se termine administrativement alors que l'assuré est toujours en arrêt, la date de fin de période renseignée en base correspond à la date de fin de l'arrêt ou si elle correspond à la date de fin du contrat. Car si cette date correspond à la fin administrative du contrat alors il nous manquera des jours indemnisés.

Afin de vérifier cela, nous observons la répartition des mois des dates de fin d'incapacité. Si des dates administratives se sont glissées dans les dates de fin d'arrêt de travail, nous aurons une prépondérance du mois de décembre (mois le plus courant pour une fin de contrat). La répartition obtenue est présentée dans le graphique ci-dessous.

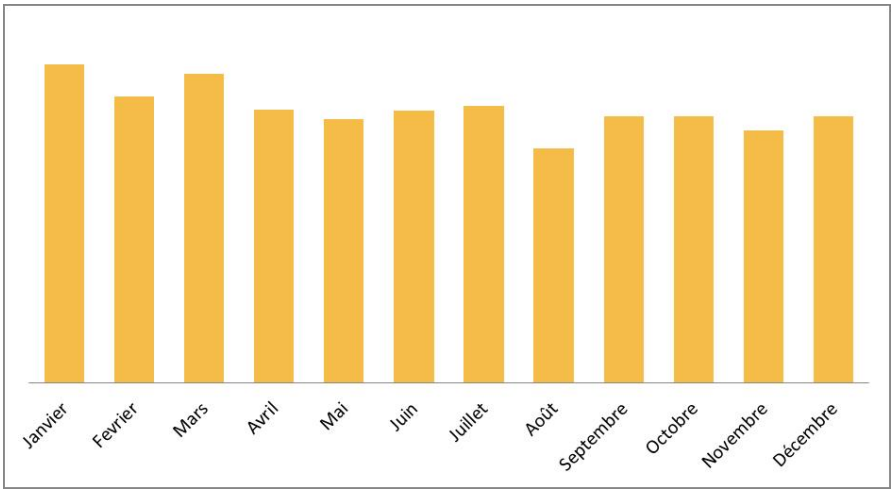


Figure 1 – Répartition des dates de fin d'incapacité par mois

Les dates de fin d'incapacité sont bien réparties sur les différents mois de l'année. Nous pouvons donc supposer que nous disposons de la date réelle de fin de l'incapacité et non de la date administrative de fin de contrat.

3. Gestion des rechutes

Nous précisons qu'à la réception des bases de données les rechutes avaient déjà été traitées. En base nous ne disposions (sauf erreurs corrigées par les retraitements) que d'une ligne par période d'incapacité pour un contrat individualisé.

Résumé du chapitre sur les données de l'étude

Dans cette partie nous avons présenté, dans le respect de leur confidentialité, les données mises à disposition par un organisme assureur partenaire afin d'être utilisées pour l'étude. Dans cette base de données de contrats de prévoyance collective, nous possédons à la fois les bénéficiaires sinistrés ainsi que les bénéficiaires non sinistrés. C'est ce dernier point qui justifie les travaux ici réalisés. En effet, généralement les bases de données collectives ne possèdent que les salariés sinistrés ainsi que des statistiques sur le reste de la population. De plus, les bases de données collectives possèdent une diversité de données bien plus importante que les bases individuelles. Nous pourrons donc étudier l'adaptabilité des méthodes de tarification individuelle à une base collective.

Notre base de données est composée de deux populations distinctes. La première est assimilable à une population classique alors que la seconde, plus spécifique, dépend d'une CCN particulière. Cette dernière est plus jeune, plus féminine, et moins sinistrée que la population classique. Elle introduit une hétérogénéité non négligeable dans nos données.

Enfin, nous avons montré que le maintien des garanties décès et le maintien des prestations d'incapacité après résiliation du contrat, obligations légales de la loi Evin, ne vont pas influencer notre étude. Aussi, les rechutes avaient déjà été traitées avant la réception des données.

III. La déclaration sociale nominative

Nous rappelons qu'après l'obtention d'un barème de tarification, le second objectif de ce mémoire est l'étude de ce que la DSN peut apporter aux méthodes de tarification du risque incapacité.

A. Présentation

La DSN est une déclaration dématérialisée des informations nécessaires à la gestion de la protection sociale des salariés. L'employeur doit la fournir mensuellement en se basant sur la fiche de paie des salariés. A cette déclaration mensuelle se rajoute des déclarations évènementielles faisant suite à la survenance d'évènements particuliers. Prise en charge par les logiciels comptables, la DSN permet de réduire les erreurs sur les déclarations de situation des salariés, d'automatiser ce processus, et de sécuriser ces données.

Cette DSN regroupe en un fichier unique de nombreuses déclarations qui précédemment devaient être envoyées les unes indépendamment des autres (telles que l'attestation employeur remise à Pôle Emploi par exemple). Les organismes et administrations concernés par cette DSN sont entre autres : la CPAM, l'URSSAF, les caisses de la MSA, les caisses des régimes spéciaux, l'AGIRC et l'ARRCO, les organismes complémentaires gestionnaires de contrats collectifs d'entreprises, Pôle Emploi, etc.

Les évènements particuliers nécessitant l'envoie d'une DSN évènementielle sont :

- la maternité ;
- la fin ou la modification du contrat de travail ;
- l'arrêt de travail et la reprise d'activité après un arrêt.

B. Périmètre

Comme indiqué précédemment, la DSN est une déclaration mensuelle avec ajout de déclarations ponctuelles lors de la survenance d'évènements particuliers.

La DSN mensuelle est à effectuer chaque mois avant le 5 du mois pour les entreprises d'au moins 50 salariés dont la paye est versée au cours du même mois que la période de travail, et avant le 15 du mois dans les autres cas. L'entreprise subira des pénalités financières en cas de retard. Les déclarations d'évènements tels que la survenance d'un arrêt de travail sont à effectuer dès que l'évènement est connu.

Une entreprise doit réaliser une DSN pour chacun de ses établissements. Ceci est possible grâce à la dénomination des entreprises par l'INSEE. Chaque entreprise possède un unique numéro SIREN et chaque établissement de cette entreprise possède un numéro SIRET. Ainsi, grâce à ces 14 chiffres (respectivement 9 et 5), chacun des établissements de chaque entreprise est indentifiable.

Pour le moment, quelques employeurs ne rentrent pas dans le périmètre de la DSN : les particuliers employeurs, les employeurs de la fonction publique, et certaines entreprises situées dans des zones géographiques non concernées par la DSN (telles que Monaco par exemple). Pour tous les autres, elle est obligatoire depuis le 1er janvier 2017. Cependant, la DSN monte progressivement en charge comme nous pouvons le voir sur le schéma suivant récupéré sur le site www.dsn-info.fr.

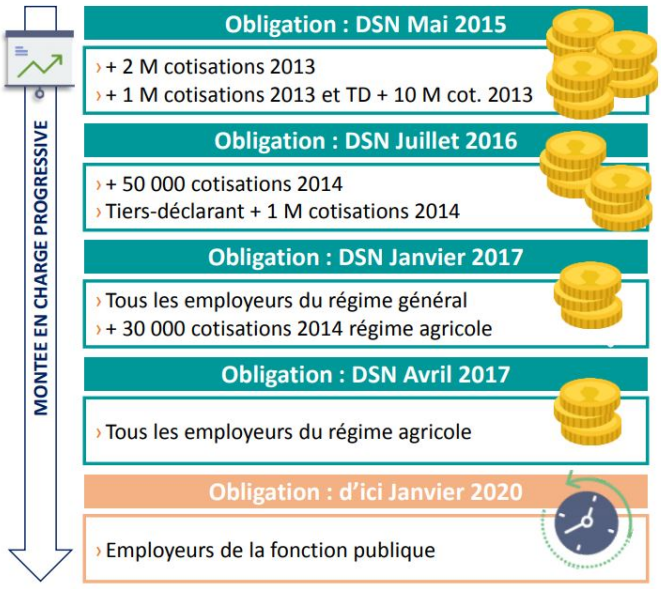


Figure 2 – Mise en place de la déclaration sociale nominative

Résumé du chapitre sur la déclaration sociale nominative

La DSN est une déclaration dématérialisée à la fois mensuelle et évènementielle permettant la gestion de la protection sociale des salariés. Les évènements qui doivent être déclarés sont la maternité, la modification du contrat de travail, et l'arrêt de travail ainsi que la reprise du travail après un arrêt. Depuis le 1er janvier 2017, tous les employeurs du régime général y sont sujets et son périmètre s'étend progressivement.

Partie 2 – Modélisation par modèle linéaire généralisé

Nous présenterons dans cette partie la tarification du risque incapacité par MLG. Tout d'abord, nous rappellerons brièvement quelques notions théoriques sur les MLG. Ensuite, nous présenterons les diverses méthodes de tarification qui peuvent être utilisées. Et enfin, nous analyserons la mise en œuvre de ces méthodes avec nos données.

I. Rappels sur les modèles linéaires généralisés

Avant de parler des MLG, rappelons pourquoi ils sont préférés aux modèles linéaires classiques.

A. Le modèle linéaire classique

Le modèle linéaire classique prend la forme suivante :

$$Y = X * \beta + \varepsilon$$

où :

- $Y = {}^t(Y_1; Y_2; \dots; Y_n)$ est le vecteur aléatoire à expliquer
- $X = \begin{pmatrix} 1 & X_{11} & \dots & X_{p1} \\ \vdots & X_{12} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & \cdots & X_{pn} \end{pmatrix}$ est la matrice des variables explicatives
- $\beta = {}^t(\beta_0; \beta_1; \dots; \beta_p)$ est le vecteur des paramètres de la régression
- $\varepsilon = {}^t(\varepsilon_1; \varepsilon_2; \dots; \varepsilon_n)$ est le vecteur des erreurs

Deux hypothèses sont malheureusement trop fortes pour pouvoir utiliser ces modèles en pratique.

1) Hypothèse de linéarité : Y et X doivent être linéairement corrélés. Cela veut dire que l'espérance de Y est une combinaison linéaire des variables explicatives X et des paramètres de régressions β.
2) Hypothèse sur l'erreur : pour chaque réalisation de X, les erreurs ε doivent être indépendantes et distribuées comme une variable aléatoire normale centrée $\mathcal{N}(0, \sigma^2)$, c'est-à-dire : $\mathbb{E}[\varepsilon|X = x] = 0$ et $\mathbb{V}[Y|X = x] = \mathbb{V}[\varepsilon|X = x] = \sigma^2$.

Ces hypothèses signifient que la variable aléatoire à expliquer doit être normale conditionnellement à la réalisation des variables explicatives, soit : $Y|X = x \sim \mathcal{N}(x * \beta, \sigma^2 * \mathbb{I})$ ($\mathbb{I}$ la matrice identité).

Cette hypothèse de normalité est trop forte pour la pratique de la tarification du risque incapacité sur des données réelles. Car par exemple, la majeure partie des périodes d'incapacité ne durent que quelques jours et quelques sinistres durent jusqu'à 1095 jours, donc le nombre de jours passés en incapacité ne peut pas être modélisé par une variable aléatoire symétrique.

Nous nous tournons donc vers les MLG. Bien que nécessitant aussi quelques hypothèses, ils sont bien plus souples que les modèles linéaires : Y peut appartenir à une autre loi de probabilité que la loi normale et la variance de Y ne doit plus nécessairement être constante.

De plus, avec un modèle linéaire classique il est impossible de construire un jeu de coefficients correcteurs alors que ceci est possible avec un MLG comme nous le verrons par la suite.

B. Les modèles linéaires généralisés

Avant de présenter la construction théorique des MLG, nous devons présenter la famille de lois de probabilité dans laquelle doit s'inscrire la variable à expliquer.

1. *La famille exponentielle*

Grace aux MLG nous pouvons modéliser des variables à expliquer suivant d'autres lois que la loi normale. Ces lois doivent appartenir à la famille exponentielle qui pour les paramètres θ et ϕ, a la fonction de densité :

$$f(y|\theta,\phi) = exp\left(\frac{y*\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right)$$

où :

- $a(\cdot)$, $b(\cdot)$ et $c(\cdot)$ sont des fonctions respectivement de $\mathbb{R} \to \mathbb{R}^*$, $\mathbb{R} \to \mathbb{R}$ et $\mathbb{R}^2 \to \mathbb{R}$, et $b \in \mathcal{C}^2$
- $\theta \in \mathbb{R}$ est appelé paramètre naturel
- $\phi \in \mathbb{R}$ est appelé paramètre de nuisance ou de dispersion
- y est une variable appartenant à $\mathbb{N}$ ou $\mathbb{R}$

De nombreuses lois connues entrent dans la famille exponentielle. Nous citerons par exemple : la loi normale, la loi de Bernoulli, la loi binomiale, la loi de Poisson, la loi Gamma, la loi Gaussienne inverse, la loi binomiale négative, etc.

L'espérance et la variance d'une telle loi sont :

$$\mathbb{E}[Y] = b'(\theta) = \mu$$
$$\mathbb{V}[Y] = b''(\theta) * \phi$$

La famille exponentielle offre la possibilité d'ajouter à chaque valeur de y une pondération ω, dans ce cas la densité devient :

$$f(y;\omega|\theta;\phi) = exp\left(\frac{y*\theta - b(\theta)}{\frac{a(\phi)}{\omega}} + c(y,\phi)\right)$$

2. *Le modèle linéaire généralisé*

Un modèle linéaire généralisé est composé des trois éléments suivants.

- La composante aléatoire : le vecteur aléatoire à expliquer : $Y = {}^t(Y_1; Y_2; \dots; Y_n)$
- La composante déterministe : la matrice X des variables explicatives. Les colonnes de cette matrice sont : $X_0 = {}^t(1; 1; \dots; 1)$, $X_1 = {}^t(X_{11}; X_{12}; \dots; X_{1n})$, …, $X_p = {}^t(X_{p1}; X_{p2}; \dots; X_{pn})$
- La fonction lien g_n déterministe, strictement monotone, définie sur $\mathbb{R}^n$ telle que :
$$g_n: \mathbb{R}^n \to \mathbb{R}^n \; ou \; g_n(y_1; y_2; \dots; y_n) = \big(g_n(y_1); g_n(y_2); \dots; g_n(y_n)\big)$$

Ces trois éléments sont liés les uns aux autres par la relation suivante :

$$g_n(\mathbb{E}[Y]) = X * \beta = \beta_0 + \beta_1 * X_1 + \dots + \beta_p * X_p$$

où :

- $\mathbb{E}[Y] = \mu$
- $\beta_0 + \beta_1 * X_1 + \dots + \beta_p * X_p = \eta$ est appelé le prédicteur linéaire ou le score du modèle avec ${}^t(\beta_0; \beta_1; \dots; \beta_p)$ les coefficients de la régression

Ainsi, nous pouvons réécrire le modèle :

$$g_n(\mu) = \eta$$

Les hypothèses pour appliquer un MLG sont énoncées ci-dessous.

1) Les Y_i où $i \in [\![1; n]\!]$ sont des variables aléatoires indépendantes appartenant à la famille exponentielle présentée dans la partie précédente.
2) Le score du modèle est tel que $\eta = X * \beta$.
3) Il existe une fonction g_n satisfaisant la relation précédente.

Il est possible d'ajouter un offset ξ au modèle. Dans ce cas, ce dernier se réécrit :

$$g_n(\mu) = \eta + \xi$$

Cet offset sert généralement à prendre en compte l'exposition car si la fonction lien est la fonction $ln(\cdot)$ et $\xi = ln(exposition)$ nous avons :

$$ln(\mu) = \eta + ln(exposition)$$
$$\Leftrightarrow ln\left(\frac{\mu}{exposition}\right) = \eta$$

L'estimation des paramètres du MLG est disponible en ANNEXE 2.

C. La tarification par modèle linéaire généralisé en pratique

Dans cette partie, nous ferons le lien entre la théorie énoncée précédemment et la tarification de contrats d'assurance.

1. Les modèles assurantiels

Afin d'estimer la variable à expliquer il est nécessaire de sélectionner plusieurs variables explicatives. Ces variables peuvent être qualitatives ou quantitatives. Pour les variables qualitatives, il faudra définir l'individu de référence qui aura son tarif contenu dans β_0, les autres β_i où $i \in [\![1; n]\!]$ permettront de modifier ce tarif de référence en fonction des autres modalités et valeurs des différentes variables.

En assurance nous retrouvons trois grandes classes de modèles.

- Les modèles de comptage vont servir à modéliser la fréquence ou le nombre de sinistres (variable a expliquer dans $\mathbb{N}$).
- Les modèles de propension vont servir à modéliser des taux (variable à expliquer catégorielle).
- Les modèles de coût vont servir à modéliser le coût moyen des sinistres (variable à expliquer continue).

L'objectif étant d'obtenir un barème de tarification, il est nécessaire de s'assurer de la justesse de ce barème. Il faut donc posséder de nombreux outils de vérifications des modèles. Ces outils vont être utilisés afin de vérifier la pertinence de la sélection des différentes variables explicatives, de vérifier la qualité des différentes modalités des variables catégorielles, et enfin de comparer les différents modèles entre eux.

La modélisation par MLG suit donc le processus suivant :

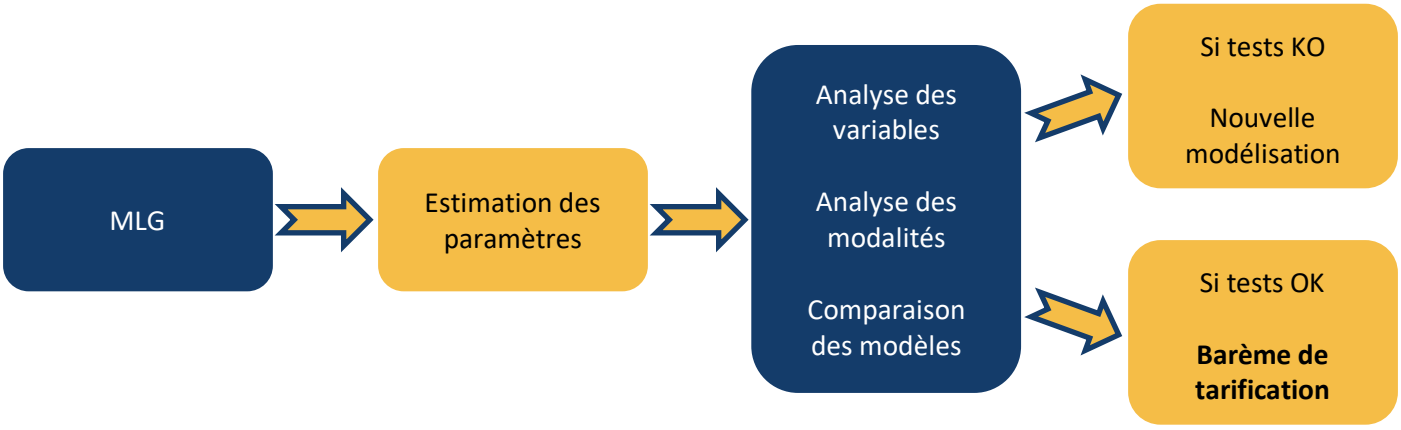


Figure 3 – Étapes de la modélisation par MLG

Nous rappelons que pour obtenir un MLG il est habituel de fractionner la base de données utilisée en une base d'apprentissage et une base de validation. La base d'apprentissage (généralement 70% de la base initiale) sert à obtenir le modèle. La base de validation sert à vérifier la justesse de la prédiction.

Dans les deux sous parties suivantes, nous présenterons tout d'abord ce qu'est un barème de tarification grâce à un exemple. Puis nous développerons une liste non exhaustive de quelques outils de validation des modèles.

2. *Le barème de tarification*

Prenons à titre d'exemple le modèle suivant (fictif). Nous cherchons à expliquer le coût moyen des consultations et visites médicales. Les variables explicatives que nous possédons sont :

- $Activité$ avec les modalités : $Primaire$, $Secondaire$ et $Tertiaire$
- $Collège$ avec les modalités : $Cadres$ et $Non\ cadres$
- $Sexe$ avec les modalités : $Homme$ et $Femme$

Notre individu de référence est un homme cadre travaillant dans le secteur tertiaire.

Ceci nous donne le modèle suivant :

$$\begin{aligned} ln(Coût\ moyen\ estimé) = \beta_0 \\ &+\beta_{Primaire} * \mathbb{I}_{Activité=Primaire} + \beta_{Secondaire} * \mathbb{I}_{Activité=Secondaire} \\ &+\beta_{\ Non\ cadres} * \mathbb{I}_{Collège\ =\ Non\ cadres} \\ &+\beta_{\ Femme} * \mathbb{I}_{Sexe\ =\ Femme} \end{aligned}$$

Posons :

$$\begin{aligned} \beta_{Activité} &= \beta_{Primaire} * \mathbb{I}_{Activité=Primaire} + \beta_{Secondaire} * \mathbb{I}_{Activité=Secondaire} \\ \beta_{Collège} &= \beta_{\ Non\ cadres} * \mathbb{I}_{Collège\ =\ Non\ cadres} \\ \beta_{Sexe} &= \beta_{\ Femme} * \mathbb{I}_{Sexe\ =\ Femme} \end{aligned}$$

Nous pouvons réécrire le modèle :

$$\begin{aligned} ln(Coût\ moyen\ estimé) &= \beta_0 + \beta_{Activité} + \beta_{Collège} + \beta_{Sexe} \\ \Leftrightarrow Coût\ moyen\ estimé &= e^{\beta_0} * e^{\beta_{Activité}} * e^{\beta_{Collège}} * e^{\beta_{Sexe}} \end{aligned}$$

Ici nous voyons apparaitre le concept du barème de tarification. Le coût de l'individu de référence est contenu dans e^{β_0} puis les correctifs $e^{\beta_{Activité}}$, $e^{\beta_{Collège}}$, et $e^{\beta_{Sexe}}$ vont venir moduler le tarif pour les autres individus en fonction de leurs caractéristiques.

Terminons cet exemple avec une application numérique du modèle précédent (toujours fictive). Nous supposons $\beta_0 = 3{,}5$ ce qui donne $e^{3,5} = 33€$ le coût de la consultation pour l'individu de référence.

$Coût\ moyen\ estimé = 33€ *$

$Activité$	$\beta_{Activité}$
$Primaire$	0,071
$Secondaire$	0,07
$Tertiaire$	0

$*$

$Collège$	$\beta_{Collège}$
$Non\ cadres$	0,156
$Cadres$	0

$*$

$Sexe$	β_{Sexe}
$Femme$	-0,028
$Homme$	0

Figure 4 – Exemple d'utilisation d'un barème tarifaire

Ainsi, pour une femme non cadre travaillant dans le secondaire le tarif sera :

$$33€ * e^{0.007} * e^{0,156} * e^{-0,028} = 38€$$

3. Outils de validation des MLG

Il existe de nombreux outils de validation des MLG. Dans cette partie nous n'en présenterons que quelques-uns. Nous commencerons par présenter le test entre modèles emboités, puis les critères d'information d'Akaike et de Bayes (ou critère AIC et BIC), et enfin l'analyse des résidus.

Test entre modèles emboités

Ce test est utilisé pour vérifier l'importance d'une variable dans le modèle. Généralement, il est utilisé après l'exclusion d'une variable pour justifier cette action.

Les hypothèses sont les suivantes :

- $H0 : \beta_{H0} = (\beta_0; \beta_1; \dots; \beta_q)$ où $q < p$
- $H1 : \beta_{H1} = (\beta_0; \beta_1; \dots; \beta_p)$

La statistique étudiée est : $\Delta = 2 * \left(ln\left(L_{\widehat{\beta_{H1}}}(y)\right) - ln\left(L_{\widehat{\beta_{H0}}}(y)\right)\right) \sim^{approx} \mathcal{X}^2_{p-q}$ et $H0$ est acceptée si $\Delta_{observé} < \mathcal{X}^2_{p-q;1-\alpha}$.

Dans le cas où $q = p - 1$, rejeter $H0$ signifie que la variable retirée était importante pour la modélisation, cette variable doit donc être conservée. A l'inverse, retenir $H0$ permet de justifier l'exclusion de la variable.

Les critères d'information d'Akaike et de Bayes

Les critères AIC et BIC sont utilisés pour comparer différents modèles entre eux. Ils permettent de quantifier un équilibre entre un bon ajustement du modèle (c'est-à-dire un petit biais) et une pénalisation due au nombre de paramètres. Nous chercherons à minimiser ces critères.

Les critères AIC et BIC sont les suivants :

- $AIC = -2 * ln(p) + 2 * p$
- $AICC = -2 * ln(p) + 2 * p * \frac{n}{n-p-1}$
- $BIC = -2 * ln(p) + p * ln(n)$

où p est le nombre de paramètres à estimer et n le nombre d'observations.

La différence entre ces critères concerne la pénalisation. Pour l'AIC, seul le nombre de paramètres estimés est pris en compte, alors que pour l'AICC (AIC Corrigé) et le BIC le nombre d'observations influence la pénalisation en plus du nombre de paramètres.

Analyse des résidus

Il existe différentes sortes de résidus :

- Les résidus ligne : $r_i = y_i - \mu_i$
- Les résidus de Pearson : $r_i^p = \frac{\sqrt{\omega_i} * (y_i - \mu_i)}{\sqrt{V(\mu_i)}}$ où V est la fonction variance
- Les résidus de déviance : $r_i^D = signe(y_i - \mu_i) * \sqrt{d_i}$ où d_i est la contribution de y_i à la déviance globale $D = \sum_{i=1}^{n} d_i$

Les résidus de déviance sont les plus utilisés car ils corrigent l'asymétrie de la distribution des deux précédents résidus. En effet : $\mathbb{V}(r_i) = \mathbb{V}(\mu_i) = \frac{\phi}{\omega_i} * V(\mu_i)$.

Les résidus du modèle doivent être centrés en 0 et ne pas présenter de forme particulière. Par exemple, dans les deux graphiques présentés ci-dessous, le modèle est correct dans le premier schéma et mal calibré dans le second.

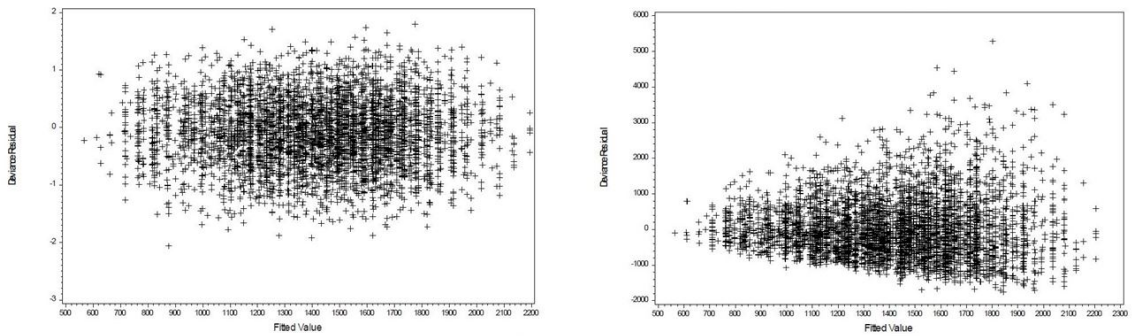


Figure 5 – Exemple de résidus de MLG

Afin de bien étudier les résidus, les diagrammes quantile-quantile sont des outils très utilisés.

Maintenant que nous avons traité la théorie sur les MLG ainsi que leur utilisation dans la tarification de contrats d'assurances, nous pouvons étudier la tarification spécifique du risque incapacité.

Résumé du chapitre de rappels sur les modèles linéaires généralisés

Dans cette partie, nous avons présenté quelques rappels sur la théorie des MLG. Tout d'abord nous avons vu que les MLG sont préférés aux modèles linéaires classiques à cause des trop fortes hypothèses de ces derniers. Ensuite, nous avons détaillé la construction des MLG avec la famille exponentielle et l'expression du modèle. Puis, nous avons présenté l'application générale des MLG à la tarification en présentant les trois modélisations couramment utilisées (modèle de taux, modèle de comptage, et modèle de coût). Ensuite nous avons exposé à l'aide d'un exemple ce qu'est un barème de tarification. Et finalement, quelques outils de validation des modèles ont été listés.

II. Méthode de tarification du risque incapacité par modèle linéaire généralisé

La tarification du risque incapacité par MLG se partage en deux éléments :

- la modélisation du nombre de jours indemnisés au titre des arrêts survenus dans l'année ;
- et la modélisation du coût moyen d'un jour indemnisé.

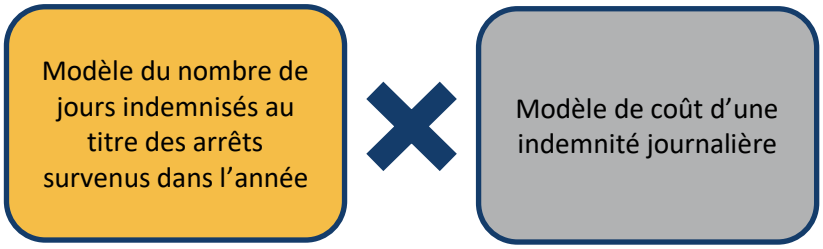


Figure 6 – Tarification du risque incapacité par MLG

Nous cherchons à réaliser un barème tarifaire pour 1€ de garantie journalière en incapacité. Ainsi, il n'est pas nécessaire de développer le modèle de coût. Grâce à la connaissance du salaire et de la garantie proposée pour ses bénéficiaires, l'assureur obtiendra le coût exact d'une IJ. Nous nous concentrons donc sur la modélisation du nombre de jours indemnisés au titre des arrêts survenus dans l'année.

Remarque : par nombre de jours indemnisés au titre des arrêts survenus dans l'année nous sous entendons que la durée de chaque arrêt est rattachée à son année de survenance. Si un arrêt commence en N et se termine en N+1, voire N+2, l'ensemble de sa durée sera rattachée à l'année N.

A. Modélisations retenues

Afin d'estimer le nombre de jours indemnisés au titre des arrêts survenus dans l'année nous détaillerons dans les paragraphes suivants quatre modélisations envisageables. Ces modélisations seront ensuite confrontées les unes aux autres dans la partie suivante.

Dans la suite de cette partie, l'indice a représentera les caractéristiques du bénéficiaire a.

1. 1ère modélisation

Nous cherchons ici à modéliser une unique variable aléatoire π_a.

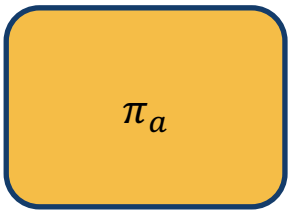


Figure 7 – 1ère modélisation du nombre de jours indemnisés

La variable à estimer π_a représente **le nombre de jours indemnisés au titre des arrêts survenus dans l'année**.

$\pi_a \in \mathbb{N}$ et modélise à la fois les bénéficiaires sinistrés et les bénéficiaires non sinistrés. Nous cherchons donc dans ce cas à modéliser une variable aléatoire prenant dans la majeure partie des cas 0, et dans quelques cas un entier positif. Un modèle de comptage classique ne sera pas adapté. Il faudra ici utiliser un modèle avec une surreprésentation des zéros : un *ZIM* pour *Zero-Inflated Model*.

Afin de modéliser π_a nous retiendrons les deux modèles suivants :

- le modèle *ZIP* pour *Zero Inflated Poisson* ;
- et le modèle *ZINB* pour *Zero Inflated Negative Binomial.*

2. 2nd modélisation

Cette seconde modélisation sépare le modèle précédent en deux sous parties. Nous modélisons ici deux variables aléatoires : p_a et n_a.

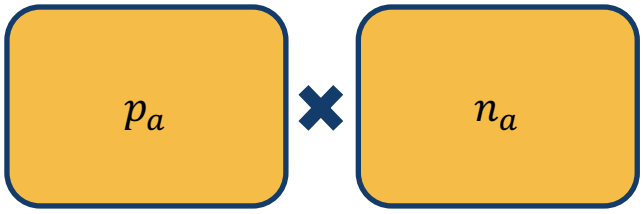


Figure 8 – 2nd modélisation du nombre de jours indemnisés

Dans ce modèle :

- p_a représente la **probabilité d'entrée en indemnisation durant l'année** ;
- et n_a représente **le nombre de jours indemnisés au titre des arrêts survenus dans l'année sachant que le bénéficiaire est sinistré**.

Ainsi, $p_a \in [0;1]$ et $n_a \in \mathbb{N}^*$. Dans ce cas les modèles de surreprésentation des zéros ne sont plus nécessaires. Pour modéliser p_a nous utiliserons un modèle de propension. Nous retiendrons la loi binomiale et les fonctions liens logit, probit, et cloglog. Pour n_a, nous modéliserons un modèle de comptage classique. La fonction lien sera la fonction $ln(\cdot)$ et les lois peuvent être la loi de Poisson, la loi de Poisson sur-dispersée, et la loi binomiale négative.

Profitons de cette introduction du taux d'entrée en indemnisation pour rappeler la différence entre le taux d'entrée en arrêt de travail et le taux d'entrée en indemnisation. Un sinistre est connu dès

lors qu'il est indemnisé, c'est-à-dire dès lors que sa durée dépasse la durée de franchise. Donc à cause de cette franchise, il existe des périodes où les bénéficiaires sont en arrêt sans être indemnisés. L'assureur n'a pas nécessairement connaissance de ces périodes. La présence de franchise crée donc une troncature gauche des données. La troncature n'étant pas gérée par les MLG, dans la suite de cette méthode nous parlerons de taux d'entrée en indemnisation et non de taux d'entrée en arrêt de travail, et de bénéficiaires indemnisés et non de bénéficiaires sinistrés. De ce fait, la durée de franchise devra nécessairement être retenue en tant que variable explicative.

3. 3ème modélisation

Ici, nous modélisons le même taux d'entrée en indemnisation que dans la partie précédente, mais nous fractionnons la variable n_a en deux variables aléatoires. Il y a donc ici trois variables à modéliser : p_a, N_a, et α^a.

$$p_a \times \sum_{t=1}^{N_a} \alpha^a$$

Figure 9 – 3ème modélisation du nombre de jours indemnisés

Nous avons donc :

- p_a la même variable aléatoire que dans la 2nd modélisation ;
- N_a représente **le nombre d'arrêts <u>survenus dans l'année</u> sachant que le bénéficiaire est sinistré** ;
- α^a représente **le nombre moyen de jours indemnisés <u>par arrêt survenu dans l'année</u> sachant que le bénéficiaire est sinistré**.

$N_a \in \mathbb{N}^*$ et $\alpha^a \in \mathbb{N}^*$. Ces deux variables seront modélisées par un modèle de comptage : fonction lien $ln(\cdot)$ et loi de Poisson, de Poisson sur-dispersée, et binomiale négative.

4. 4ème modélisation

Ce dernier niveau de modélisation nécessite de nombreux MLG. Tout d'abord il faut modéliser p_a et N_a comme lors des parties précédentes. Puis il faudra modéliser les nombreuses variables aléatoires α_t^a où $t \in \{1; 2; 3; \dots\}$.

$$p_a \times \sum_{t=1}^{N_a} \alpha_t^a$$

Figure 10 – 4ème modélisation du nombre de jours indemnisés

Dans cette modélisation p_a et N_a sont les mêmes variables que dans la partie précédente. Ensuite, les variables aléatoires suivantes devront être modélisées :

- α_1^a représente **le nombre moyen de jours indemnisés <u>au titre du premier arrêt de l'année</u> sachant que le bénéficiaire est sinistré** ;

- α_2^a représente **le nombre moyen de jours indemnisés au titre du deuxième arrêt de l'année sachant que le bénéficiaire est sinistré** ;
- Etc.

Les $\alpha_t^a \in \mathbb{N}^*$ et $\forall\, t \in \{1; 2; 3; \dots\}$ sont modélisées par les mêmes modèles de comptage que dans les parties précédentes.

B. Comparaison des modélisations

Les avantages et inconvénients évoluent dans des sens opposés au fur et à mesure qu'est complexifiée la modélisation. La 1ère modélisation est la plus directe. Elle présente l'avantage d'être moins sujette aux problèmes d'expositions. Cependant, comme nous ne modélisons qu'une seule et unique variable aléatoire, nous ne pouvons pas identifier l'origine d'un fort ou faible nombre de jours indemnisés. A l'opposé, la dernière modélisation est la plus précise mais est très complexe à mettre en place. Il est nécessaire de posséder suffisamment de données pour chacun des « numéros » d'arrêts. Or dans la pratique, nous observons que la majeure partie des bénéficiaires ont un arrêt dans l'année, peu des bénéficiaires ont deux arrêts dans l'année, et très peu des bénéficiaires ont trois arrêts ou plus dans l'année. Donc cette méthode se heurtera à un important problème d'exposition. Elle a cependant l'avantage de décomposer le risque en de nombreuses composantes ce qui permet une meilleure analyse du résultat obtenu. Les 2ème et 3ème modélisations sont un entre-deux entre les problèmes et avantages précédents. Le schéma ci-dessous présente la comparaison des modélisations.

Remarque : par meilleure analyse du résultat, nous entendons analyse de l'origine du risque. Il est intéressant de savoir si un nombre important de jours indemnisés provient de nombreux arrêts courts ou bien de quelques longs arrêts. Une telle analyse est permise par la dernière modélisation mais ne l'est pas par la première.

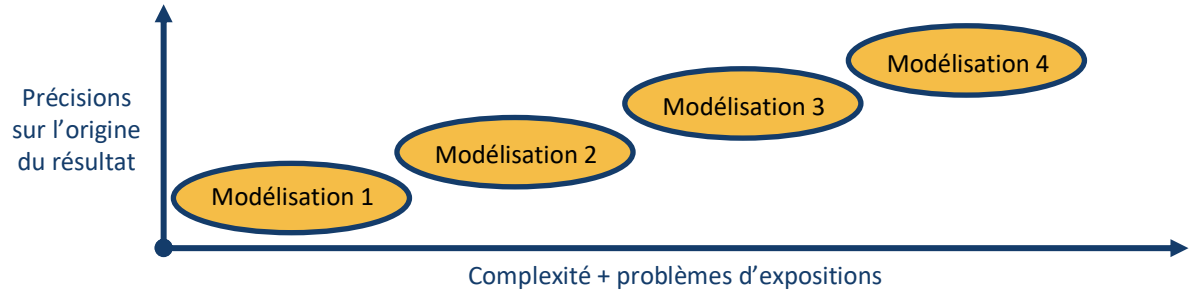


Figure 11 – Avantages et inconvénients des divers modélisations par MLG

Nous précisons aussi que dans la 4ème modélisation, un biais est introduit car en réalité les α_t^a où $t \in \{1; 2; 3; \dots\}$ ne sont pas indépendants les uns des autres, alors que nous avons supposé cette indépendance pour pouvoir décomposer le α^a en α_t^a.

C. Consolidation du modèle

La consolidation est la dernière étape de la modélisation par MLG. Nous voyons avec les méthodes précédentes que nous pouvons avoir besoin de plus d'un MLG afin d'obtenir notre nombre moyen de jours indemnisés. Or un barème de tarification se réalise à partir d'un unique MLG. Ainsi, les résultats précédents vont être regroupés dans une unique variable finale grâce à une formule de consolidation. Puis, cette variable finale sera modélisée par un nouveau MLG. C'est depuis ce dernier

modèle que nous pourrons obtenir le barème de tarification. Cette formule de consolidation a aussi l'avantage de permettre l'ajout de paramètres extérieurs à la modélisation tels que la prise en compte des coûts de gestion par exemple et de permettre le choix des variables à retenir pour le barème de tarification.

Le processus de consolidation est résumé dans le schéma suivant.

Figure 12 – Consolidation des MLG

Résumé du chapitre sur les méthodes de tarification du risque incapacité

Dans cette partie nous avons étudié les différentes méthodes qui peuvent être utilisées afin de tarifer le risque incapacité. Nous avons vu que ces méthodes se décomposent en une modélisation de coût moyen d'une IJ et une modélisation du nombre de jours indemnisés. L'assureur connait les taux d'indemnisation et le salaire de ses bénéficiaires, il connait donc exactement le coût d'une IJ. Ainsi, nous nous concentrons sur le modèle de comptage.

Nous avons présenté quatre modélisations du nombre de jours indemnisés au titre des arrêts survenus dans l'année, de la moins détaillée présentant des problèmes de surreprésentation de zéros mais étant la plus simple à mettre en place, à la plus détaillée présentant des problèmes d'exposition mais permettant d'expliquer précisément l'origine du risque. Après avoir comparé ces différents modèles, nous avons étudié le processus de consolidation permettant d'obtenir le barème tarifaire final.

III. Tarification par modèle linéaire généralisé en pratique

Dans cette partie, nous présenterons l'application des méthodes précédentes à nos données collectives.

Pour chacun des modèles qui seront présentés ci-dessous, les étapes de l'initialisation sont les suivantes :

- obtention de la variable à expliquer (agrégation de lignes, regroupements des bases, gestion des censures, …) ;
- regroupement à priori de valeurs pour les variables catégorielles ;
- test de corrélation et exclusion des variables corrélées ;
- exclusion des variables sous exposées ;
- création des bases d'apprentissage et de validation ;
- définition de l'individu de référence.

L'une des étapes importantes du premier point présenté ci-dessus est la gestion des censures. Nos données présentaient une censure à droite due à leur extraction le 14/03/2017. Afin de rendre cette censure non informative, nous avons considéré comme date de censure la date d'extraction moins

cinq mois, soit le 14/10/2016. Ainsi, tous les sinistres non clos à cette date sont considérés comme en cours. Afin de compenser l'effet de cette censure droite nous avons extrapolé les durées restantes d'indemnisation à partir d'une table d'espérance résiduelle en incapacité obtenue à partir du logiciel ***addactis® PM Expert®*** depuis la table règlementaire de maintien en incapacité du BCAC avec un taux d'actualisation nul (table disponible en ANNEXE 3). Cette extrapolation était nécessaire car les MLG ne gèrent pas les censures.

Remarque : nous précisons que lors des étapes précédentes, une attention particulière a été portée à l'étude des corrélations et des expositions afin d'éviter tout problème de modélisation à ce niveau.

Nous commencerons par détailler les variables utilisées. Puis nous présenterons les modélisations développées pour expliquer nos données. Et enfin, nous exposerons les premiers résultats obtenus ainsi que les difficultés rencontrés. L'ensemble de ces travaux ont été réalisés sous R[4].

A. Les variables retenues pour l'étude

Nous avons à notre disposition 46 variables dans la base des bénéficiaires et 65 variables dans la base des prestations. Cependant, après sélection des variables utilisables et étude de la corrélation et de l'exposition, il nous reste uniquement les variables présentées ci-dessous.

- Au niveau des informations sur le contrat, nous possédons :
  - les taux d'indemnisation ;
  - la valeur de la franchise VP ;
  - l'indicatrice de présence d'une franchise spécifique :
    - pour cause d'hospitalisation ;
    - et pour cause d'accident.
- Au niveau des informations sur la personne physique, nous possédons :
  - la région ;
  - le code NAF ;
  - le secteur d'activité ;
  - le sexe ;
  - et l'âge durant l'année étudiée ou à la souscription du contrat.

Remarque : il s'agit des variables utilisables, aucun MLG n'est initialisé avec l'ensemble des variables précédentes.

[4] Les packages utilisés sont :
- SPINU V. et al. : *lubridate* *(version 1.7.4)*
- WICKHAM H. : *stringr* *(version 1.3.1)*
- DOWLE M. et al. : *data.table* *(version 1.11.4)*
- JACKMAN et al. : *pscl* *(version 1.5.2)*
- HOTHORN T. et al. : *lmtest* *(version 0.9-36)*
- STASINOPOULOS M. et al. : *gamlss* *(version 5.0-8)*
- MEYER D. et al. : *vcd* *(version 1.4-4)*
- SING T. et al. : *ROCR* *(version 1.0-7)*
- KOHL M. : *MKmisc* *(version 1.0)*
- FOX J. et al. : *car* *(version 3.0-0)*
- WICKHAM H. et al. : *dplyr* *(version 0.7.5)*
- PLAFF B. et al. : *evir* *(version 1.7-4)*

Afin de donner plus d'informations à leur sujet, nous les présenterons tout d'abord dans les sous-parties suivantes, puis nous terminerons par donner les caractéristiques de notre individu de référence.

1. Les taux d'indemnisation

Initialement nous avions trois variables concernant le taux d'indemnisation : l'indemnisation sur la tranche A, sur la tranche B, et sur la tranche C. L'étude des corrélations nous a contraints à exclure la seconde. Ainsi nous conservons les taux d'indemnisation sur la tranche A et sur la tranche C. Ces variables sont quantitatives et prennent leurs valeurs dans l'intervalle $[\![0,100]\!]$.

Rappel : la tranche A comprend les salaires de 0 à 1 PASS, la tranche B les salaires de 1 à 4 PASS, la tranche C les salaires de 4 à 8 PASS, et la tranche D les salaires au-delà de 8 PASS.

2. Valeur de la franchise VP

Généralement les franchises diffèrent entre les contrats AT/MP et VP. Il peut aussi exister des franchises réduites si la cause de l'arrêt est un évènement particulier tel qu'une hospitalisation ou un accident. A titre d'exemple, une situation classique en assurance incapacité VP serait :

- franchise principale : 90 jours ;
- franchise hospitalisation : 30 jours ;
- franchise accident : 3 jours.

Dans la base de données se trouvaient et la valeur de franchise VP et la valeur de franchise AT/MP. Ces deux variables étant fortement corrélées entre elles, nous n'avons conservé que la franchise VP car elle présentait le plus grand nombre de niveaux suffisamment exposés.

Nous avons réalisé les regroupements à priori suivants : 0, $]0;5]$, $]5;10]$, $]10;20]$, $]20;35]$, $]35;50]$, $]50;75]$, $]75;90]$, et $]90;365]$. Ces regroupement étaient nécessaire car notre base de données comportait de nombreuses franchises atypiques. Là où généralement nous étudions des franchises telles que 3, 30, 45, 60, 90, nous avions par exemple la valeur 7 ou encore la valeur 65.

De plus, nous avions à la fois des franchises continues et discontinues. Afin d'harmoniser nos données, nous avons converti les franchises discontinues en franchise continues grâce à la table de conversion des franchises présente dans le logiciel ***addactis® Prévoyance Office®***.

3. Variable sur les franchises spécifiques

Initialement nous possédions trois catégories de franchises pour cause spéciale d'arrêt de travail :

- les franchises si la cause de l'incapacité est une hospitalisation ;
- les franchises si la cause de l'incapacité est un accident ;
- les franchises si la cause de l'incapacité est une hospitalisation à la suite d'un accident.

La dernière catégorie a été exclue d'office car elle n'était pas suffisamment exposée. Puis, pour les franchises hospitalisation et accidents, nous possédions à la fois une indicatrice de présence de la franchise et la valeur de cette franchise. Pour des raisons de bonne exposition et d'absence de corrélation nous avons conservé uniquement les indicatrices.

Ainsi, les variables de présence de franchises spécifiques prennent les valeurs « O » ou « N » pour respectivement la présence ou l'absence d'une telle franchise dans le contrat observé.

4. *Les régions*

La variable de localisation présente initialement dans notre base de données était le code postal des personnes morales. Afin d'obtenir un nombre raisonnable de modalités, nous avons remplacé ces codes postaux par le numéro de département auquel ils correspondent. Puis nous avons regroupé ces départements dans la nouvelle nomenclature des régions présentée ci-dessous (carte obtenue sur le site www.interieur.gouv.fr).

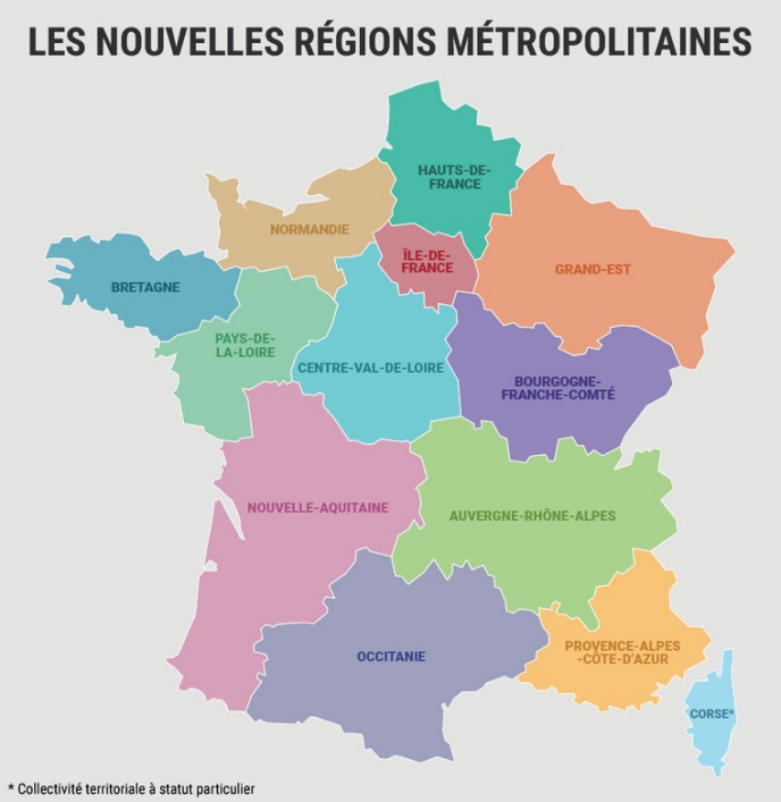


Figure 13 – Carte des nouvelles régions de France

Nous rappelons que nous n'avons pas de contrat individualisé sur l'un des territoires d'outre-mer dans notre base de données.

5. *Le secteur d'activité*

Cette variable se fractionne en trois niveaux : industrie, tertiaire, et ni industrie ni tertiaire. Au cours des différentes modélisations, afin de la rendre significative, elle sera souvent transformée en : industrie et autres.

6. *Sexe*

La variable sexe prend la valeur 1 pour les hommes et 2 pour les femmes.

7. Le code NAF

Nous possédions le code NAF de chacune des personnes morales présentes dans la base de données. Cependant, comme pour les codes postaux, le code NAF représente trop de modalités. Il n'est donc pas possible de garder cette variable telle quelle. Ainsi, nous n'avons retenu que le premier niveau de la nomenclature NAF, la lettre. Notre variable NAF prend donc les valeurs suivantes :

Lettre du code NAF	Correspondance
A	Agriculture, sylviculture et pêche
B	Industries extractives
C	Industrie manufacturière
D	Production et distribution d'électricité, de gaz, de vapeur et d'air conditionné
E	Production et distribution d'eau ; assainissement, gestion des déchets et dépollution
F	Construction
G	Commerce ; réparation d'automobiles et de motocycles
H	Transports et entreposage
I	Hébergement et restauration
J	Information et communication
K	Activités financières et d'assurance
L	Activités immobilières
M	Activités spécialisées, scientifiques et techniques
N	Activités de services administratifs et de soutien
O	Administration publique
P	Enseignement
Q	Santé humaine et action sociale
R	Arts, spectacles et activités récréatives
S	Autres activités de services
T	Activités des ménages en tant qu'employeurs ; activités indifférenciées des ménages en tant que producteurs de biens et services pour usage propre
U	Activités extraterritoriales

Figure 14 – Premier niveau des codes NAF

Remarque : la variable NAF ne sera que peu utilisée par la suite pour cause de problème d'exposition.

8. Âge ou âge à l'entrée

Les variables d'âge ont été regroupées dans les modalités suivantes : $[16; 25[$, $[25; 35[$, $[35; 40[$, $[40; 45[$, $[45; 50[$, $[50; 55[$, $[55; 60[$, et ≥ 60.

9. *Individu de référence*

Maintenant que nous avons présenté les différentes modalités de nos variables catégorielles, nous pouvons présenter notre individu de référence pour les modélisations à venir. Voici ses caractéristiques :

Variable	Valeur de référence
Franchise VP	]20; 35]
Présence d'une franchise spécifique	N
Région	Île-de-France
NAF	
Secteur d'activité	Tertiaire
Sexe	1 ou 2 en fonction du modèle
Âge	[40,45[

Figure 15 – Individu de référence pour la modélisation par MLG

Remarque : le code NAF le plus représenté n'a pas été divulgué afin de ne pas corrompre la confidentialité des données.

B. Modélisation en pratique

Nous avons implémenté les trois premières modélisations sur les quatre présentées précédemment. La dernière modélisation a été exclue d'office à cause d'un manque de données.

Afin de modéliser p_a la base de données a été fractionnée par année et seules les années entièrement exposées ont été conservées. Ainsi, si un bénéficiaire est présent pendant trois années et demi, il correspondra à trois enregistrements en base. Ceci introduit un biais. En effet, l'une des hypothèses des MLG est que les variables à expliquer soient indépendantes les unes des autres. Or ici, entre deux exercices, l'indépendance n'est pas rigoureusement observée. Le modèle finalement retenu pour p_a est un MLG avec une loi binomiale et une fonction lien probit. L'AUC de notre modélisation est 73 %, ce qui d'après l'usage n'est pas mauvais mais n'est pas excellent non plus.

Remarque : nous précisons que ce sont les caractéristiques des modèles de propension qui nous obligent à ne conserver que les années entièrement exposées. En effet, la variable à expliquer de notre modèle de propension est une indicatrice prenant les valeurs 1 ou 0 (1 dans le cas d'un bénéficiaire entré en indemnisation dans l'année, 0 sinon), et nous cherchons à modéliser une probabilité d'entrée en indemnisation annuelle. Donc pour les années non entièrement exposées il faudrait ajouter comme poids au modèle l'exposition de ces années (exposition différente de 1) afin de prendre en compte ce manque d'information. Or une telle exposition remplacerait les 1 de la variable à expliquer par des valeurs non entières ce qui empêche le modèle de fonctionner.

Pour la modélisation de N_a nous retenons un MLG avec une loi de Poisson sur-dispersée et la fonction lien $ln(\cdot)$. Et enfin, pour la modélisation du α^a, du n_a et du π_a nous retenons des MLG avec la loi binomiale négative et la fonction lien $ln(\cdot)$.

Bien que les modèles de comptage obtenus présentent une significativité correcte, ils ne sont pas convainquant en termes de résultats. En effet, les graphiques quantile-quantile des divers résidus

présentent des divergences au niveau des queues de distribution. A titre d'exemple, voici le graphique quantile-quantile des résidus de déviance du modèle retenu pour n_a :

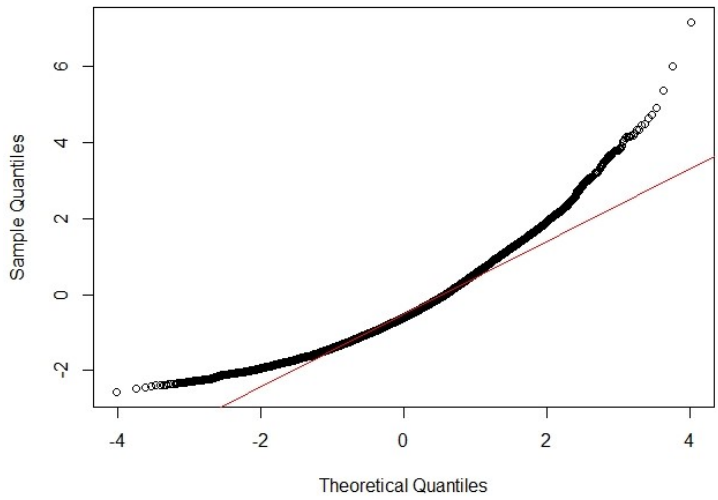


Figure 16 – Exemple de graphique quantile-quantile sur le MLG retenu pour modéliser n_a

Comme nous pouvons le visualiser sur le graphique précédent, et comme ceci s'observe sur les valeurs des prédictions, nous modélisons très bien la moyenne mais nous avons de véritables difficultés de prédiction au niveau des queues de distribution.

Toujours en étudiant le cas de la modélisation de n_a, la loi discrète qui correspond le mieux à la répartition de nos données est une loi binomiale négative. Cependant, si nous comparons la loi binomiale négative optimale pour modéliser nos données et notre répartition empirique nous obtenons les schémas suivants. A gauche l'histogramme de nos données et à droite l'histogramme de la binomiale négative optimale.

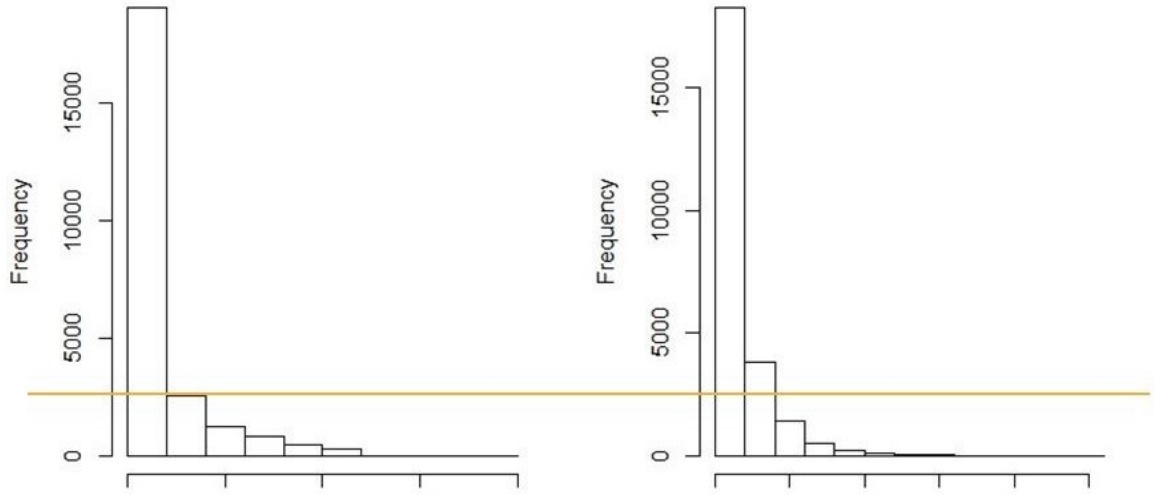


Figure 17 – Comparaison entre distribution empirique et optimale pour modélisation de n_a

Comme nous pouvons le voir, la queue de distribution de nos données est beaucoup plus épaisse que celle de la loi binomiale négative optimale pour les représenter. Ceci est très probablement l'un des éléments expliquant nos difficultés de modélisation sur les modèles de comptage.

C. Pistes de résolution des difficultés rencontrées

Suite à ces complications rencontrées sur la modélisation du nombre de jours indemnisés, nous avons essayé d'harmoniser nos données afin d'obtenir des résultats de meilleure qualité. Notre harmonisation s'est portée sur le nombre de jours indemnisés en fonction de la valeur de la franchise VP et est présenté dans le schéma ci-dessous.

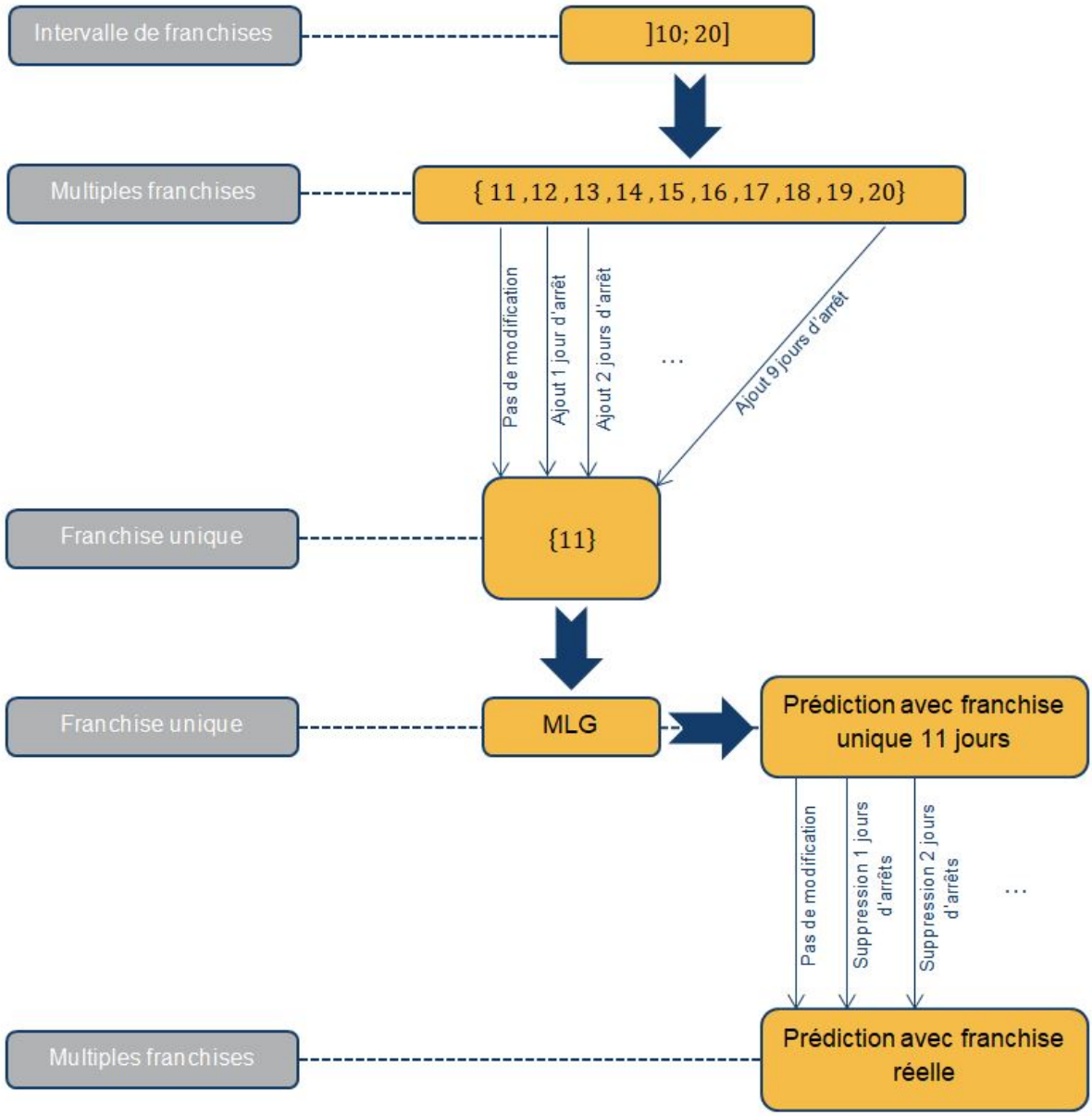


Figure 18 – Méthode d'harmonisation des franchises

Afin d'explicité ce processus d'harmonisation, prenons l'exemple de la classe de franchises $]10;20]$. Au sein de cette classe, nous considérons des nombres de jours indemnisés associés à une franchise 11 jours au même titre que des nombres de jours indemnisés associés à une franchise 20 jours. Ceci donne beaucoup d'hétérogénéité dans une même classe de franchise. L'idée de cette harmonisation est que si un sinistre avec une franchise de 20 jours avait été indemnisé, il aurait été indemnisé 9 jours de plus que s'il avait eu 11 jours de franchise. C'est de cette manière que nous augmentons nos nombres de jours indemnisés afin de se ramener pour la modélisation par MLG à une franchise unique par classe.

Cependant, même avec cette harmonisation du nombre de jours indemnisés au sein d'une classe de franchises, les résidus ne sont pas améliorés. A ce stade, plusieurs évolutions pouvaient être envisagées afin d'obtenir de meilleurs résultats :

- séparer la loi en deux partie et utiliser les lois déjà obtenues pour la majeure partie des données puis une loi spécifique pour les queues de distribution ;
- utiliser la discrétisation de lois continues plutôt que des lois discrètes ;
- modéliser le nombre de jours indemnisés par semaine.

Nous avons cependant préféré passer à la méthode suivante sur laquelle nous portions plus d'espoir plutôt que continuer le développement de la tarification par MLG.

Résumé du chapitre sur la tarification par modèle linéaire généralisé en pratique

Nous avons tout d'abord présenté l'initialisation de notre modélisation avec en particulier la gestion des censures. En effet, les MLG ne permettant pas leur intégration automatique, nous avons retraité l'ensemble des sinistres encore en cours à la date de censure en extrapolant pour chacun le nombre de jours d'incapacité restants à couvrir.

Puis, après avoir développé les différentes variables utilisables ainsi que notre individu de référence, nous avons présenté les différents modèles retenus pour l'étude.

Finalement, nous avons étudié les premiers résultats obtenus puis les difficultés rencontrées. En effet, l'une des distributions que nous cherchons à modéliser possède une queue de distribution trop épaisse pour les lois classiques des MLG.

Partie 3 – Modélisation par simulation

Nous détaillerons dans cette partie la tarification du risque incapacité par simulation. Afin de faciliter la compréhension de cette méthode, nous commençons par un aperçu global du process de tarification, puis nous le détaillerons point par point.

I. Présentation générale de la tarification par simulation

La méthode de tarification du risque incapacité par simulation permet d'obtenir la sinistralité d'un bénéficiaire grâce aux divers états par lesquels il transite pendant l'année. Il s'agit de l'état au travail, de l'état en incapacité, et de la sortie du portefeuille. Ce dernier état regroupe le décès, la retraite, la résiliation du contrat, et le passage en invalidité du bénéficiaire suite à une période d'incapacité. L'engagement de l'assureur sera alors obtenu en fonction de l'évolution, jour par jour, de l'état du bénéficiaire. Le processus étudié est présenté dans le schéma ci-dessous.

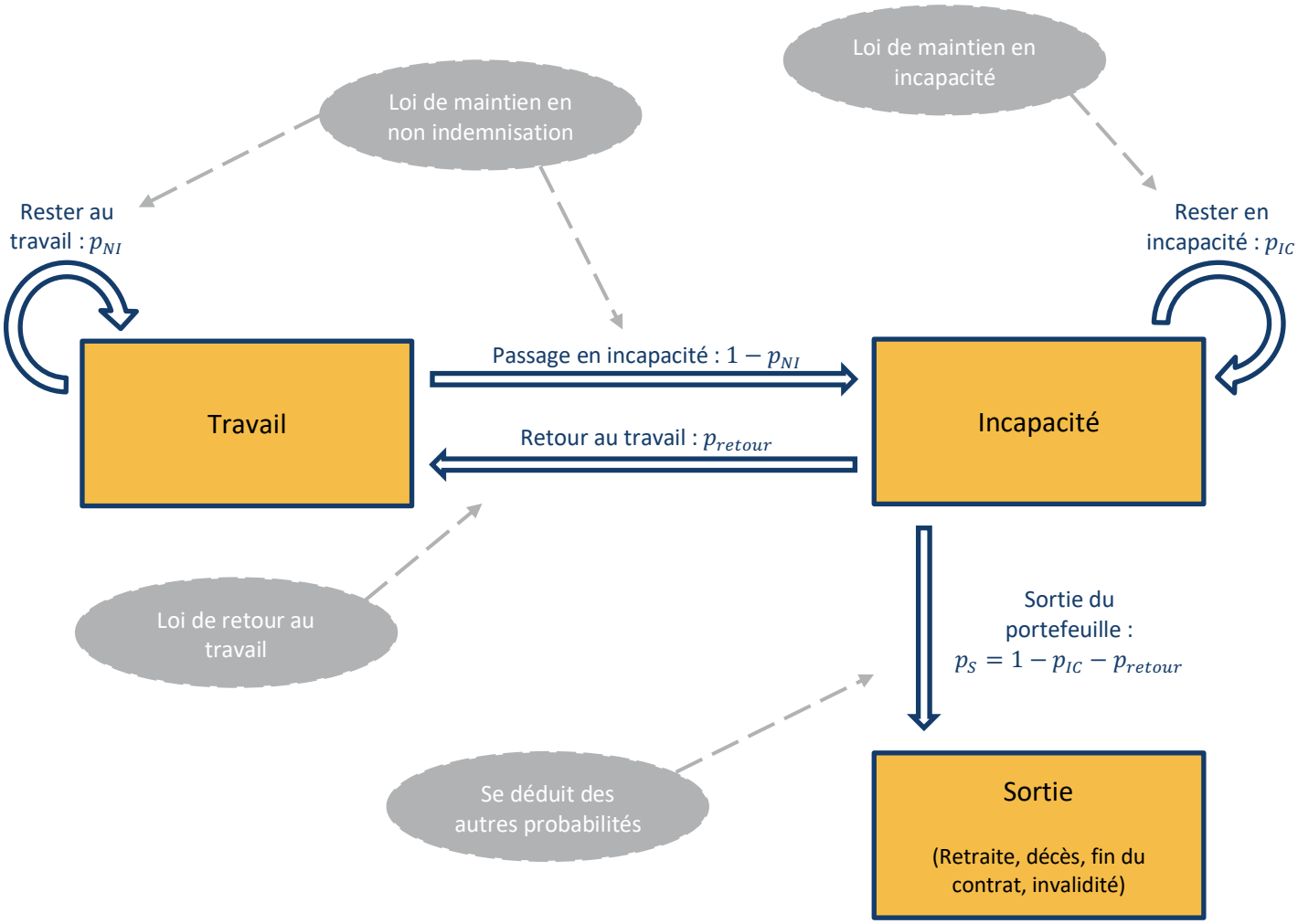


Figure 19 – États de la simulation

Afin de déterminer en tout temps la position du bénéficiaire, nous devons connaître les probabilités de maintien dans les états et de passage entre états. Comme indiqué sur le schéma ci-dessus, ces

probabilités vont être obtenues à partir de lois de maintien et de passage. Ainsi, nous devrons construire trois lois :

- une loi de maintien en incapacité ;
- une loi de retour au travail ;
- et enfin une loi de maintien en non indemnisation.

Ces lois seront construites grâce à l'estimateur de Kaplan Meier, présenté en début de partie suivante.

Remarque : une fois entré dans l'état de sortie du portefeuille, aucun retour n'est envisagée. Ceci ne pose pas de problème pour le décès, la retraite, ou la résiliation du contrat. Cependant, un biais est introduit pour le cas de l'invalidité. En effet, il existe quelques cas de retour au travail suite à un état d'invalidité. Cependant, ils sont rares et ne représentent pas au sein de notre étude une exposition suffisante pour être modélisés. Ces cas seront donc négligés dans la suite de notre étude.

En simulant un grand nombre de fois les déplacements du bénéficiaire entre les états au sein d'une année, et plus particulièrement ses passages par l'état incapacité, nous pourrons déterminer la Valeur Actuelle Probable *(VAP)* des engagements de l'assureur pour une garantie de 1€ versée par jour dans l'état en incapacité.

Dans les paragraphes suivants nous présenterons les méthodes d'obtention des différentes lois ainsi que le détail de la méthode de tarification par simulation, et enfin nous appliquerons cette méthode à nos données afin d'obtenir un barème de tarification du risque incapacité.

Nous précisons que la méthode de tarification par simulation est retenue par rapport à la recherche d'une formule fermée à partir des lois de maintien et de passage car aucune formule fermée relativement simple ne peut être obtenue. Sous l'hypothèse que le bénéficiaire ne peut avoir qu'un sinistre durant l'année, une formule relativement facile d'utilisation est atteignable. Cependant, dès que nous considérons qu'un bénéficiaire peut avoir plus d'un sinistre dans l'année, les formules deviennent difficiles d'utilisations. Pour une justification plus précise de ce point, nous dirigeons le lecteur vers LEBOSSE C. (2009)[5].

Remarque : nous clarifions enfin que ce mémoire est un mémoire de tarification. De ce fait nous chercherons uniquement à modéliser l'incapacité en cours. La notion d'invalidité en attente nécessaire au provisionnement ne sera pas étudiée ici. Ce détail nous permet d'agréger l'état de sortie (décès, retraite, résiliation) et l'état invalidité.

Afin de réaliser cette simulation, nous nous placerons dans le cas d'un modèle markovien. Dans un tel modèle, l'état de l'individu est modélisé par une chaine de Markov, c'est-à-dire que la prédiction de l'état futur du bénéficiaire ne dépend que de l'état actuel de ce dernier.

Rappel : définition d'une chaine de Markov
Une chaine de Markov est un processus stochastique $\{X_t; t \in \{1; 2; \dots\}\}$ *à temps discret, définit sur un espace d'état* S *fini ou dénombrable et vérifiant la propriété de Markov suivante :*

$$P[X_t = s|X_0, \dots, X_{t-1}] = P[X_t = s|X_{t-1}] \; \forall s \in S \text{ et } \forall t \geq 1$$

[5] LEBOSSE C. (2009) : *Construction de barèmes de tarification d'arrêt de travail par une méthode de Simulation – Application à un portefeuille de Travailleurs Non-Salariés*, mémoire de l'Institut des Actuaires

Résumé du chapitre de présentation générale de la tarification par simulation

La tarification du risque incapacité par simulation repose sur la construction de trois lois distinctes :

- une loi de maintien en incapacité ;
- une loi de retour au travail ;
- et enfin une loi de maintien en non indemnisation.

Ces lois vont nous permettre en tout temps de déterminer si le bénéficiaire est au travail, s'il est en incapacité, ou s'il est sorti du portefeuille suite à une période d'incapacité (décès, résiliation, retraite, invalidité). En simulant de nombreuses fois l'évolution du bénéficiaire nous pourrons déterminer la valeur actuelle probable des engagements de l'assureur pour une garantie de 1€ versée par jour dans l'état incapacité.

II. Construction des lois

Dans les sous parties suivantes nous introduirons tout d'abord l'estimateur de Kaplan Meier. Ensuite nous détaillerons son application afin d'obtenir chacune des trois lois à estimer. Puis, ces lois étant généralement erratiques, nous présenterons les méthodes de lissage, de validation des lissages, et enfin de structuration des lois. Ces méthodes nous permettront d'obtenir les lois utilisables pour la simulation.

A. L'estimateur de Kaplan Meier

Afin de construire les diverses lois utiles pour la simulation, nous utiliserons l'estimateur statistique non paramétrique de Kaplan Meier. Cet estimateur est souvent choisi lors de la construction de table d'expérience grâce à trois propriétés avantageuses. Tout d'abord, c'est un estimateur non paramétrique, ce qui permet une modélisation sans à priori sur les données. Ensuite, cet estimateur prend en compte la présence de censures et de troncatures synonymes de la modélisation du risque arrêt de travail. Enfin, c'est un estimateur discret et les arrêts de travail se modélisent de façon discrète. De plus, si nous rajoutons l'hypothèse que nos censures et nos troncatures n'influencent pas les durées modélisées, l'estimateur de Kaplan Meier devient l'estimateur du maximum de vraisemblance, donc un estimateur non biaisé.

Remarque : le fait que l'estimateur soit non paramétrique est un avantage mais aussi un inconvénient. En effet, si le nombre de données utilisées n'est pas suffisant, la loi obtenue par l'estimateur de Kaplan Meier sera très chaotique. De ce fait, il faudra porter un intérêt particulier à la volumétrie de la base de données utilisée ainsi qu'aux méthodes de lissage utilisées.

L'estimateur de Kaplan Meier nous permet d'obtenir une estimation de la fonction de survie discrétisée. Son expression pour un âge à l'entrée dans l'état étudié x et une durée passée dans l'état étudié m est la suivante :

$$\hat{S}(x;m) = \prod_{t=duree_{min}}^{m} \left(1 - \frac{n(x;t)}{e(x;t)}\right) \; ; \; \forall m \in \{duree_{min}; duree_{max}\}$$

où :

- $duree_{min}$ est la durée minimale envisagée pour le risque étudié
- $duree_{max}$ est la durée maximale envisagée pour le risque étudié

- $e(x;m)$ est l'effectif du portefeuille exposé au risque pour l'âge x et la durée passée dans le risque m
- $n(x;m)$ est le nombre de sorties du risque non censurées à l'âge x et la durée passée dans le risque m

Remarque : nous précisons que l'estimateur présenté précédemment n'est pas rigoureusement l'estimateur de Kaplan Meier. En effet, l'une des hypothèses de ce dernier est qu'il ne peut y avoir qu'une sortie par pas de temps. Hors, dans la modélisation que nous mettrons en place par la suite il peut y avoir plusieurs sorties par pas de temps.

La variance de l'estimateur précédent est approchée par l'estimateur de Greenwood :

$$\mathbb{V}\left[\hat{S}(x;m)\right] = \hat{S}(x;m) * \sum_{t=1}^{m} \frac{n(x;t)}{e(x;t) * \left(e(x;t) - n(x;t)\right)}$$

Cet estimateur nous permettra dans un premier temps d'exclure les âges extrêmes non suffisamment exposés (âges pour lesquels l'estimateur prendra des valeurs trop élevées). Puis dans un second temps il nous permettra d'obtenir un intervalle de confiance de niveau α de l'estimateur de Kaplan Meier. Pour un âge à l'entrée x et une ancienneté m, nous ne pouvons pas utiliser la relation :

$$IC_{\alpha}(x\,;m) = \left[\hat{S}(x\,;m) \pm z_{1-\frac{\alpha}{2}} * \sqrt{\mathbb{V}\left[\hat{S}(x\,;m)\right]}\right]$$

car $\hat{S}(x;m)$ est proche de 0 ou de 1 et les bornes pourraient donc dépasser l'une de ces valeurs. De ce fait, nous retenons l'intervalle de confiance de Rothman qui permet de contourner ce problème :

$$IC_{\alpha}^{R}(x;m) = \frac{K(x;m)}{K(x;m) + \left(z_{1-\frac{\alpha}{2}}\right)^2} * \left(\hat{S}(x;m) + \frac{\left(z_{1-\frac{\alpha}{2}}\right)^2}{2 * K} \pm z_{1-\frac{\alpha}{2}} * \sqrt{\mathbb{V}\left[\hat{S}(x;m)\right] + \frac{\left(z_{1-\frac{\alpha}{2}}\right)^2}{4 * K^2}} \right)$$

où :

- $z_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $\left(1-\frac{\alpha}{2}\right)$ de la loi normale centrée réduite
- et $K(x;m) = \frac{\hat{S}(x;m) * \left(1-\hat{S}(x;m)\right)}{\mathbb{V}[\hat{S}(x;m)]}$

Remarque : la durée passée dans le risque est aussi appelée ancienneté dans le risque. Ce terme sera noté m *ou* anc *dans la suite du mémoire.*

Dans les parties de construction des lois qui vont suivre, nous reprendrons les notations de LEBOSSE C. (2009)[6].

B. Loi de maintien en incapacité

Afin de construire cette loi de maintien en incapacité, il faudra nécessairement connaitre pour chaque sinistre les éléments suivants.

- A propos de l'individu :
  - la date de naissance ddn ;

[6] LEBOSSE C. (2009) : *Construction de barèmes de tarification d'arrêt de travail par une méthode de Simulation – Application à un portefeuille de Travailleurs Non-Salariés*, mémoire de l'Institut des Actuaires

- à propos du contrat :
  - les dates de début et de fin de garantie $t1$ et $t2$;
  - la durée maximale d'indemnisation $dmax$;
  - la franchise applicable au sinistre f ;
- à propos de chaque sinistre :
  - la date de survenance de l'arrêt de travail AT ;
  - les dates de début et de fin d'indemnisation pj et dj ;
  - la date de passage en invalidité inv ;
- à propos de la gestion des données :
  - la date de censure non informative cs ;
  - la date de troncature tc.

Les notations précédentes seront spécifiées de la sorte :
- un indice fera référence à un contrat individualisé ;
- un exposant fera référence à un sinistre.

Maintenant que nous avons défini les variables utilisées pour l'étude, passons à la construction de la loi de maintien en incapacité.

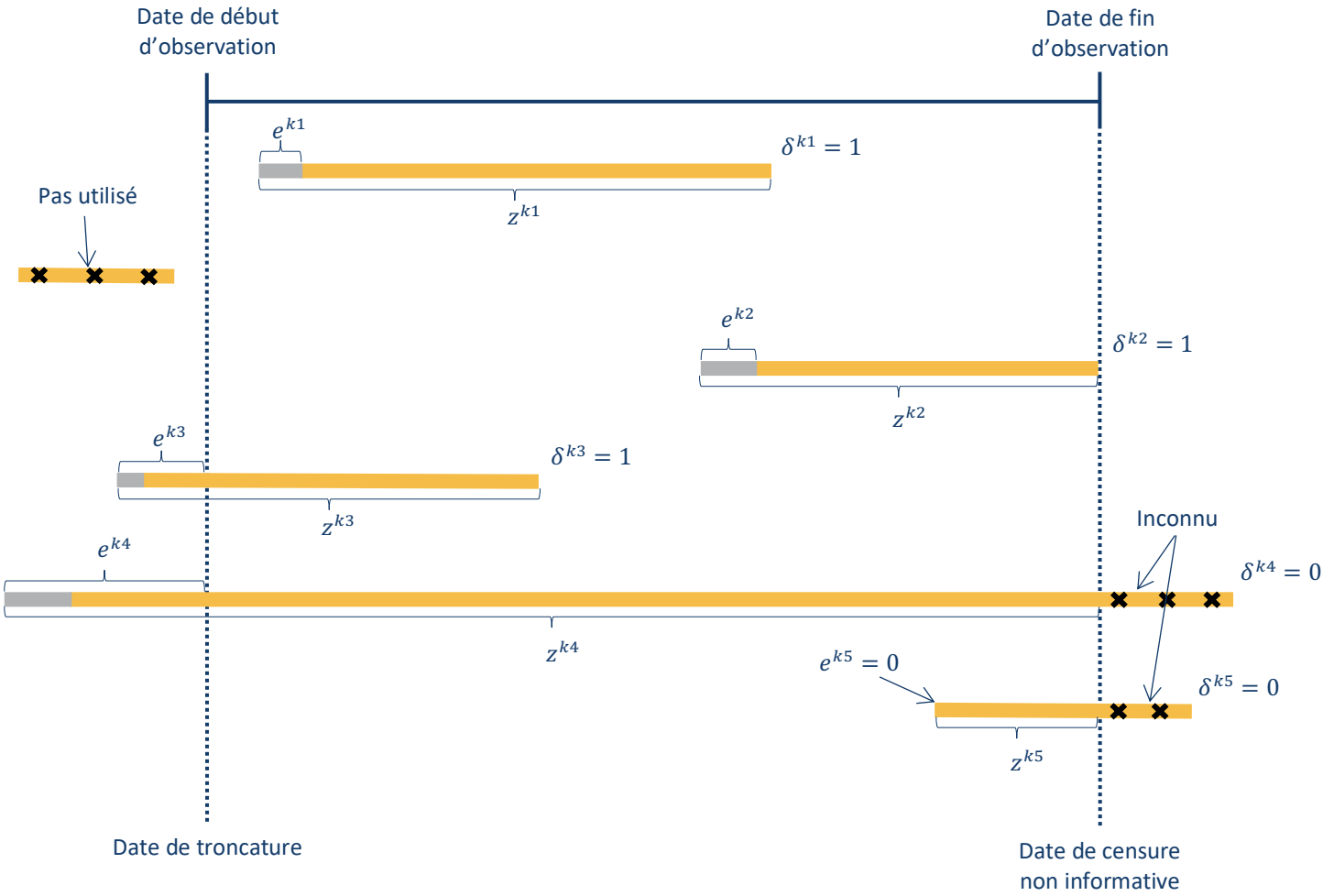
Soit V_x le portefeuille regroupant l'ensemble des sinistres qui vont être utilisés pour la construction de la loi de maintien en incapacité pour l'âge x en année. Ce vecteur est de dimension $n \times 1$ où n représente le nombre d'incapacités survenues en base pour un bénéficiaire d'âge x.

$$V_x = \begin{pmatrix} \{e_1; z_1; \delta_1\} \\ \vdots \\ \{e_n; z_n; \delta_n\} \end{pmatrix}$$

Les différents V_x sont supposés indépendants mais ne sont pas identiquement distribués.

Pour chaque sinistre $k \in [\![1; n]\!]$, les variables contenues dans V_x sont :
- e^k la durée en nombre de jours dans l'état d'incapacité au début des observations :
$$e^k = max(pj^k; tc) - AT^k$$
- z^k la durée en nombre de jours dans l'état d'incapacité à la fin des observations :
$$z^k = min(dj^k; cs; inv^k - 1) - AT^k$$
- δ^k l'indicateur des sorties non censurées, c'est-à-dire les sorties pour motif de passage en invalidité, de décès, de résiliation du contrat, de retraite, ou de retour au travail :
$$\delta^k = \begin{cases} 1 \; si \; fin \; de \; l'incapacité \\ 0 \; sinon \end{cases}$$

L'objectif de ces variables est d'isoler pour chaque sinistre l'information utilisable pour la construction de notre fonction de survie de Kaplan Meier. Plus précisément, voyons ceci avec les quelques exemples présentés dans le schéma ci-dessous.



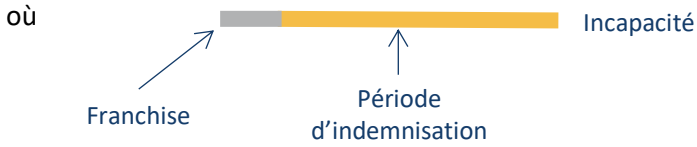


Figure 20 – Exemple variables utilisées pour Kaplan Meier

Sur le schéma ci-dessus nous voyons que les sinistres $k1$ et $k2$ pourront être utilisés dans leur globalité. A l'inverse, les sinistres $k3$ et $k4$ ne seront utilisés qu'à partir de la durée d'indemnisation e^{k3} et e^{k4}. Enfin, malgré la censure nous pourrons utiliser une partie des sinistres $k4$ et $k5$. Nous avons ici observé comment l'estimateur de Kaplan Meier permet de prendre en considération les données censurées et tronquées.

Grâce à la connaissance des V_x regroupant l'information utilisable sur chacun de nos sinistres, nous pouvons construire la fonction de survie de Kaplan Meier discrétisée en version actuarielle. Soit m la durée en jour dans l'état d'incapacité allant de 0 à 1095 et k parcourant les différents sinistres retenus pour la construction de la loi pour l'âge x, nous avons :

- l'effectif du portefeuille exposé au risque pour l'âge x et l'ancienneté m :

$$e(x;m) = \sum_k \mathbb{I}_{e^k \leq m \leq z^k}$$

- le nombre de vraies sorties pour les bénéficiaires en incapacité à l'âge x et l'ancienneté m :

$$n(x;m) = \sum_k \mathbb{I}_{z^k=m} * \mathbb{I}_{\delta^k=1}$$

Ainsi, nous pouvons calculer la fonction de survie de Kaplan Meier discrétisée en version actuarielle :

$$\begin{cases} \hat{S}(x;m) = 1 & pour\ m = 0 \\ \hat{S}(x;m+1) = \hat{S}(x;m) * \left(1 - \dfrac{n(x;m)}{e(x;m)}\right) & pour\ m > 0 \end{cases}$$

Remarque : cette fonction de survie ne peut être calculée que pour $e(x;m) \neq 0$.

Nous obtenons ainsi la loi de maintien en incapacité empirique pour chaque âge x en année et ancienneté m en jour :

$$\begin{cases} L_{IC}(x;m) = 10\,000 & pour\ m = 0 \\ L_{IC}(x;m) = \hat{S}(x;m) * 10\,000 & pour\ m > 0 \end{cases}$$

La table de maintien en incapacité se présente de la sorte :

Age (en année) \ Durée de l'arrêt (en jour)	0	1	⋯	1095
$âge\ minimum$	10 000	⋯	⋯	⋯
$âge\ minimum + 1$	10 000	⋯	⋯	⋯
⋮	⋮	⋮	⋮	⋮

Figure 21 – Exemple table de maintien en incapacité

Elle indique pour chaque âge sur 10 000 bénéficiaires en incapacité à $t = 0$, combien de bénéficiaires seront toujours en incapacité au bout du premier jour, du second jour, …, du 1095ème jour.

C. Loi de retour au travail

Les données à connaitre pour la construction de la loi de retour au travail sont les mêmes que celles nécessaires à la construction de la loi de maintien en incapacité car ces deux lois sont fortement liées.

Tout d'abord nous devons obtenir le taux de retour empirique grâce à l'estimateur de Kaplan Meier. Les définitions des e^k et z^k sont les mêmes que dans la partie précédente, seul δ^k est modifié. Dans le cas présent, δ^k est l'indicateur des retours au travail non censurés :

$$\delta^k = \begin{cases} 1\ si\ retour\ au\ travail \\ 0\ sinon \end{cases}$$

Ensuite, en calculant $e(x;m)$ et $n(x;m)$ comme précédemment, nous pouvons obtenir les taux empiriques de retour au travail pour chaque âge x en année et ancienneté m en jour :

$$\widehat{Taux_{retour}}(x;m) = \frac{n(x;m)}{e(x;m)}$$

Finalement nous obtenons la loi empirique de retour au travail pour chaque âge x en année et ancienneté m en jour :

$$L_{retour}(x;m) = \widehat{Taux_{retour}}(x;m) * L_{IC}(x;m)$$

La table de retour au travail se présente de la même manière que la table de maintien en incapacité. Elle indique le nombre de sortie de l'état d'incapacité pour chaque combinaison d'âge et d'ancienneté dans l'arrêt.

D. Loi de maintien en non indemnisation

La dernière loi qu'il nous reste à construire est la loi de maintien en non indemnisation. Cette dernière étant plus spécifique, nous commencerons par définir ce qu'est une période de non indemnisation. Nous verrons entre autres qu'à la différence des deux tables précédentes, cette table sera construite par franchise. Puis nous présenterons la méthode de construction pour une franchise f particulière.

1. Définition de la non indemnisation

La période de non indemnisation est le complémentaire de la période d'indemnisation. Dans un premier temps, nous aurions pu penser que modéliser une loi de maintien au travail est préférable à la modélisation d'une loi de maintien en non indemnisation. Cependant, comme l'assureur n'a pas connaissance des sinistres durant moins que la durée de franchise, la totalité des données nécessaires à la construction d'une telle loi n'est pas connue. De ce fait, il est préférable de se tourner vers la seconde option. En effet, comme les périodes d'indemnisation sont bien définies, les périodes de non indemnisation le sont aussi.

Par exemple, dans le cas d'un bénéficiaire ayant eu un sinistre au cours de sa période d'affiliation (cette dernière bien comprise entre la date de troncature et la date de censure) nous observons la situation suivante :

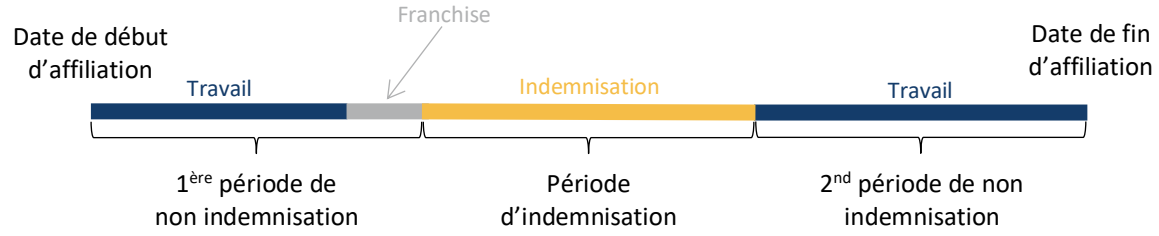


Figure 22 – Cas simple de séparation des périodes d'indemnisation et de non indemnisation

Le schéma précédent illustre pourquoi il est nécessaire de construire une loi de maintien en non indemnisation par franchise f et non au global. En effet, comme la franchise est comprise dans la période de non indemnisation, la loi de maintien en non indemnisation ne peut pas être la même entre un contrat ayant une franchise de 3 jours et un contrat ayant une franchise de 90 jours. Il y aura plus de maintien en non indemnisation pour la seconde franchise que pour la première.

Les dernières étapes avant de présenter la construction de cette loi par l'estimateur de Kaplan Meier sont la détection des périodes de non indemnisation ainsi que l'analyse de leurs troncatures et censures. Nous devons différencier plusieurs cas en fonction du nombre de sinistres survenus pour le bénéficiaire observé.

- Si le bénéficiaire n'est pas sinistré, la période de non indemnisation commence à la date de début de garantie (possiblement tronquée à la date de début d'observation) et se termine par une censure à la date de fin de garantie (possiblement elle aussi censurée à la date de fin

d'observation), censure car après cette date nous n'avons plus connaissance de l'état du bénéficiaire.

- Si le bénéficiaire a un sinistre, nous fractionnons l'unique période de non indemnisation précédente en deux périodes non indemnisées. La première débute comme précédemment mais prend fin la veille de la date de début d'indemnisation. La seconde débute le lendemain de la date de fin d'indemnisation et se termine comme l'unique période précédente.
- Si le bénéficiaire a plusieurs sinistres, nous suivons le même processus que précédemment pour les périodes de non indemnisation extrêmes (la première et la dernière). Puis pour les périodes de non indemnisation entre deux sinistres, elles débuteront le lendemain de la fin d'indemnisation du sinistre antérieur et se termineront la veille du jour de début d'indemnisation du sinistre suivant.
- Enfin, si le dernier sinistre du bénéficiaire est censuré, nous procèderons comme précédemment sauf que la dernière période de non indemnisation sera celle précédant le sinistre censuré.

2. Construction des lois de maintien en non indemnisation

A la différence des deux lois précédentes qui étaient construites uniquement sur la base des prestations, les lois de maintien en non indemnisation seront construite grâce à la base des prestations et à la base des bénéficiaires.

Il faudra nécessairement connaitre les éléments suivants pour chaque bénéficiaire i :

- à propos de l'individu :
  - la date de naissance ddn_i ;
- à propos du contrat :
  - les dates de début et de fin de garantie $t1_i$ et $t2_i$;
  - la durée maximale d'indemnisation $dmax_i$;
  - la franchise applicable au sinistre f_i ;
- à propos de chaque sinistre (k est ici l'indice des différents sinistres du bénéficiaire i) :
  - la date de survenance de l'arrêt de travail AT_i^k ;
  - les dates de début et de fin d'indemnisation pj_i^k et dj_i^k ;
  - la date de passage en invalidité inv_i^k ;
- à propos de la gestion des données :
  - la date de censure non informative cs ;
  - la date de troncature tc.

Nous redéfinissons tout d'abord la période d'observation de chaque bénéficiaire i. Cette étape consiste à borner les dates de début et de fin de garantie des bénéficiaires avec les dates de troncature et de censure non informative.

$$[\tau 1_i; \tau 2_i] = \begin{cases} [max(tc; t1_i); min(t2_i; cs)] \; si \; [tc; cs] \cap [t1_i; t2_i] \neq \emptyset \\ \emptyset \; sinon \end{cases}$$

Afin de calculer les périodes de non indemnisation, il faudra identifier chacune des périodes d'indemnisation de chaque bénéficiaire. Le calcul s'effectuera alors bénéficiaire par bénéficiaire. Comme présenté dans la partie précédente, le traitement des données changera en fonction du nombre de sinistres du bénéficiaire i.

Par franchise, cette table est uniquement définie en fonction de l'âge du bénéficiaire. La plage d'âges retenue sera notée $[\alpha; \beta]$. Ce choix dépendra fortement de l'exposition sur chaque âge. Il est fortement probable que les âges extrêmes de la base de données ne soient pas suffisamment exposés pour être conservés.

Maintenant que nous avons posé toutes les notations requises, nous pouvons présenter la méthode de construction d'une loi de maintien en non indemnisation. Notons V_x le portefeuille contenant l'ensemble des données qui seront utilisées pour la construction de la loi pour l'âge x en jour :

$$V_x = \begin{pmatrix} \{x_1^{min}; x_1^{max}; f_1; \delta_1\} \\ \vdots \\ \{x_n^{min}; x_n^{max}; f_n; \delta_n\} \end{pmatrix}$$

où :

- x_i^{min} et x_i^{max} sont l'âge de l'individu respectivement au début et à la fin de la période de non indemnisation
- f_i la franchise associée au contrat
- δ_i l'indicatrice des sorties non censurées, c'est-à-dire des sorties de l'état de non indemnisation engendrées par un début d'indemnisation :

$$\delta_i = \begin{cases} 0 \; si \; sortie \; censurée \\ 1 \; sinon \end{cases}$$

Les vecteurs V_x sont supposés indépendants mais non identiquement distribués.

Comme précédemment, nous utilisons l'estimateur de Kaplan Meier afin de construire la loi de maintien en non indemnisation. Pour x variant de α à β exprimé en jour et i parcourant l'ensemble des périodes de non indemnisation :

- l'effectif du portefeuille exposé au risque pour l'âge x est :

$$e(x) = \sum_i \mathbb{I}_{x_i^{min} + f_i \leq x < x_i^{max}} \left(1 - \mathbb{I}_{x_i^{max} = x+1}(1 - \delta_i)\right)$$

- le nombre de vraies sorties pour les bénéficiaires en incapacité pour l'âge x :

$$n(x; m) = \sum_k \mathbb{I}_{x_i^{max} = x+1} * \mathbb{I}_{\delta_i = 1}$$

Remarque : un bénéficiaire d'âge x *est exposé au risque de sortie de l'état de non indemnisation s'il travaille depuis suffisamment longtemps pour être sujet à cette sortie et s'il n'est pas censuré le lendemain. La première condition implique qu'il est non indemnisé depuis au moins la durée* f_i*, ce qui se retranscrit par la condition :* $x_i^{min} + f_i \leq x < x_i^{max}$*. La seconde condition est si* $x_i^{max} = x + 1$ *alors il faut que* $\delta_i \neq 1$*, elle peut être réécrite* $1 - \mathbb{I}_{x_i^{max} = x+1}(1 - \delta_i)$*.*

Ainsi, nous pouvons calculer la fonction de survie de Kaplan Meier discrétisée en version actuarielle :

$$\begin{cases} \hat{S}(x) = 1 & pour \; x = \alpha \\ \hat{S}(x) = \hat{S}(x-1) * \left(1 - \dfrac{n(x-1)}{e(x-1)}\right) & pour \; \alpha < x \leq \beta \end{cases}$$

Remarque : cette fonction de survie ne peut être calculée que pour $e(x) \neq 0$*.*

Nous pouvons donc maintenant obtenir la loi de maintien en non indemnisation empirique pour la franchise f notée $L_{NI,f}(x)$ pour chaque âge x :

$$\begin{cases} L_{NI,f}(x) = 10\,000 & pour \; x = \alpha \\ L_{NI,f}(x) = \hat{S}(x) * 10\,000 & pour \; \alpha < x \leq \beta \end{cases}$$

La table de maintien en non indemnisation se présente de la sorte :

Age (en jour)	$Franchise\ 1$	$Franchise\ 2$	$Franchise\ 3$	$\cdots$
α	10 000	10 000	10 000	$\cdots$
$\alpha + 1$	$\vdots$	$\vdots$	$\vdots$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$

Figure 23 – Exemple table maintien en non indemnisation

Elle indique pour chaque franchise retenue, sur 10 000 bénéficiaires au travail à l'âge α, combien seront encore au travail à l'âge $\alpha + 1$, à l'âge $\alpha + 2$, etc.

E. Réduction des tables

A moins d'avoir un nombre très conséquent de données, les tables en jour sont irrégulières du fait d'un manque d'exposition pour certains âges et durées dans l'arrêt. Ainsi, nous réduisons l'unité de la durée dans l'état d'incapacité en passant du jour au mois pour les tables de maintien en incapacité et de retour au travail, et en passant du jour à l'année pour les âges retenus dans la table de maintien en non indemnisation.

Pour réaliser cette réduction, nous considérons que les mois durent alternativement 30 et 31 jours, et les années 365 jours, 365 jours, 365 jours, puis 366 jours, et ainsi de suite. Nous réalisons donc les sélections suivantes.

- Pour la table de maintien en incapacité nous retiendrons pour chaque âge x :
  - $L_{IC}(x; 0)$;
  - $L_{IC}(x; 30)$;
  - $L_{IC}(x; 61)$;
  - … ;
  - $L_{IC}(x; 1095)$.
- Pour la table de retour au travail, s'agissant d'une table de passage il faut sommer les valeurs mensuellement. Nous retenons donc pour chaque âge x :
  - $L_{retour}(x; 30) = \sum_{j=1}^{30} L_{retour}(x; j)$;
  - $L_{retour}(x; 61) = \sum_{j=31}^{61} L_{retour}(x; j)$;
  - … ;
  - $L_{retour}(x; 1095) = \sum_{j=1068}^{1095} L_{retour}(x; j)$.
- Enfin, pour la table de maintien en non indemnisation pour chaque franchise f :
  - $L_{NI,f}(\alpha)$;
  - $L_{NI,f}(\alpha + 365)$;
  - $L_{NI,f}(\alpha + 730)$;
  - $L_{NI,f}(\alpha + 1095)$;
  - $L_{NI,f}(\alpha + 1461)$;
  - … ;
  - $L_{NI,f}(\beta)$.

  où α et β sont les âges minimums et maximums retenus pour la construction de la table.

Nous rappelons qu'il est nécessaire de porter une attention particulière à l'exposition des âges extrêmes. Si les données présentent trop peu d'exposition à leurs niveaux, il sera nécessaire de réduire la plage d'âges étudiée.

F. Lissage des lois

Les lois précédemment obtenues ont généralement un comportement erratique. Il faudrait un nombre très important de données pour obtenir dès l'utilisation de l'estimateur de Kaplan Meier des lois lisses. C'est pour cela que le lissage est indispensable à l'utilisation de nos lois.

1. Lissage de Whittaker-Henderson

Le lissage de Whittaker-Henderson est basé sur la minimisation d'une combinaison linéaire entre deux critères :

- un critère de fidélité ;
- et un critère de régularité.

L'objectif étant de trouver le meilleur compromis entre eux en faisant varier divers paramètres.

Le lissage de Whittaker-Henderson est initialement prévu pour lisser des tables unidimensionnelles. Mais une extension permet de l'appliquer à la dimension deux afin de lisser des tables bidimensionnelles comme une table de maintien en incapacité.

Dans le contexte de ce mémoire le lissage de Whittaker-Henderson en dimension deux sera appliqué aux taux bruts de sortie d'incapacité et aux taux bruts de retour au travail. Les tables lissées de maintien en incapacité et de retour au travail seront obtenues depuis ces taux lissés.

a. Whittaker-Henderson en dimension un

Comme expliqué précédemment, deux critères vont être minimisés : la fidélité F :

$$F = \sum_{i=1}^{I} \omega_i * (q_i^* - \widehat{q_i})^2$$

et la régularité R :

$$R = \sum_{i=1}^{I-z} (\Delta^z \, q_i^*)^2$$

où :

- I est le nombre de données de la table à lisser
- ω_i est le poids associé à chaque donnée à lisser
- q_i^* est la donnée lissée
- $\widehat{q_i}$ est la donnée non lissée
- Δ^z est l'opérateur différence avant tel que :
  - $\Delta \, q_i^* = q_{i+1}^* - q_i^*$
  - $\Delta^2 \, q_i^* = \Delta \, (\Delta \, q_i^*) = q_{i+2}^* - 2 * q_{i+1}^* + q_i^*$
  - …
  - $\Delta^n \, q_i^* = \sum_{j=0}^{n} \binom{n}{j} * (-1)^{n-j} * q_{i+j}^*$

  Cet opérateur mesure une distance entre les données lissées de q_i^* à q_{i+z}^*
- z est le paramètre d'ordre du modèle

Grâce à ces deux expressions, nous pouvons noter la moyenne de Whittaker-Henderson comme combinaison linéaire du critère de fidélité et du critère de régularité :

$$M = F + h * R$$

où h est le paramètre de poids de la régularité par rapport à la fidélité. Par rapport à 1, plus il est élevé, plus les données seront lissées (car plus il est élevé, plus le critère de régularité sera prédominant par rapport à celui de fidélité), et inversement.

Afin d'obtenir les q_i^*, nous chercherons à minimiser cette expression, c'est-à-dire à résoudre les équations $\forall\, i \in [\![1; I]\!]$:

$$\frac{\partial M}{\partial q_i} = 0$$

Pour cela, nous utilisons les notations matricielles suivantes :

- le vecteur des données non lissées est $\hat{q} = {}^t(\widehat{q_1}; \widehat{q_2}; \dots; \widehat{q_I})$
- le vecteur des données lissées est $q^* = {}^t(q_1^*; q_2^*; \dots; q_I^*)$
- la matrice des poids est $W = \begin{pmatrix} \omega_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \omega_I \end{pmatrix}$
- $\Delta^z\, q^* = {}^t(\Delta^z\, q_1^*; \Delta^z\, q_2^*; \dots; \Delta^z\, q_{I-z}^*)$
- K_z matrice de taille $I - z \times I$ dont les termes sont les coefficients binomiaux d'ordre z dont le signe alterne et commence positivement pour z pair.

Ainsi nous pouvons réécrire le critère de fidélité et le critère de régularité :

- $F = {}^t(q^* - \hat{q}) * W * (q^* - \hat{q})$
- $S = {}^t(\Delta^z\, q^*) * (\Delta^z\, q^*) = {}^t(K_z * q^*) * (K_z * q^*)$ car $\Delta^z\, q^* = K_z * q^*$

Ainsi :

$$M = F + h * R = {}^t(q^* - \hat{q}) * W * (q^* - \hat{q}) + h * {}^t(K_z * q^*) * (K_z * q^*)$$

En remplaçant l'expression à minimiser précédente par sa forme matricielle nous obtenons :

$$\frac{\partial M}{\partial q} = 0 \Leftrightarrow 2 * W * q^* - 2 * W * \hat{q} + 2 * h * {}^t K_z * K_z * q^*$$

La résolution de cette équation donne finalement :

$$q^* = (W + h * {}^t K_z * K_z)^{-1} * W * \hat{q}$$

Pour plus de précisions sur ce développement et quelques exemples, cf. PLANCHET F., THEROND P. (2011)[7].

b. Whittaker-Henderson en dimension deux

L'extension du lissage de Whittaker-Henderson à la dimension deux suit la même logique que le lissage en dimension un. L'expression du critère de fidélité est :

$$F = \sum_{i=1}^{I} \sum_{j=1}^{J} \omega_{ij} * \left(q_{ij}^* - \widehat{q_{ij}}\right)^2$$

et le critère de régularité est fractionné entre un critère de régularité verticale R_V et un critère de régularité horizontale R_H :

[7] PLANCHET F., THEROND P. (2011) : *Modélisation statistique des phénomènes de durée – Application actuarielles*, Economica

$$R_V = \sum_{j=1}^{J} \sum_{i=1}^{I-OV} \left(\Delta_V^{OV}\, q_{ij}^*\right)^2$$

$$R_H = \sum_{i=1}^{I} \sum_{j=1}^{J-OH} \left(\Delta_H^{OH}\, q_{ij}^*\right)^2$$

où :

- I est le nombre de lignes de la table à lisser
- J est le nombre de colonnes de la table à lisser
- ω_{ij} est le poids associé à chaque donnée à lisser
- q_{ij}^* est la donnée lissée
- $\widehat{q_{ij}}$ est la donnée non lissée
- Δ_V^{OV} est l'opérateur différence avant vertical (à j fixé dans la somme précédente)
- Δ_H^{OH} est l'opérateur différence avant horizontal (à i fixé dans la somme précédente)
- OV et OH sont les paramètres d'ordre vertical et d'ordre horizontal

Avec les notations suivantes nous pouvons réécrire la moyenne de Whittaker-Henderson :

$$M = F + \alpha * R_V + \beta * R_H$$

où α et β sont respectivement les paramètres de poids de la régularité verticale et de la régularité horizontale.

L'astuce pour résoudre l'équation $\frac{\partial M}{\partial q} = 0$ en dimension deux est de réarranger les éléments afin de se retrouver dans le cas de la dimension un. Nous posons ainsi de nouvelles notations matricielles :

- Nouveau vecteur des données lissées q^*
- Nouveau vecteur des données non lissées $u_{q*(i-1)+j} = \widehat{q_{ij}}$
- Nouveau vecteur des poids $W^*_{q*(i-1)+j,q*(i-1)+j} = W_{ij}$
- Et de même pour K^*_{OH} et K^*_{OV}

Avec ces notations et par le même procédé que celui présenté en dimension un, nous obtenons :

$$q^* = (W^* + \alpha * {}^tK^*_{OV} * K^*_{OV} + \beta * {}^tK^*_{OH} * K^*_{OH})^{-1} * W^* * u$$

Comme précédemment, pour plus de précisions sur l'extension en dimension deux du lissage de Whittaker-Henderson nous orientons le lecteur vers PLANCHET F., THEROND P. (2011)[8].

c. Choix des paramètres

La théorie présentée précédemment n'est pas suffisante pour appliquer le lissage de Whittaker-Henderson. En effet, il est nécessaire de savoir quels paramètres choisir. Dans le cas de la dimension deux, nous devons choisir cinq paramètres :

- le poids donné à la régularité verticale α ;
- le poids donné à la régularité horizontale β ;
- l'ordre vertical OV ;
- l'ordre horizontal OH ;
- et la matrice de poids de chacune des observations W.

[8] PLANCHET F., THEROND P. (2011) : *Modélisation statistique des phénomènes de durée – Application actuarielles*, Economica

Ces paramètres vont fortement influencer le lissage final. En fonction de leurs valeurs, le lissage sera plus ou moins fort et nous éloignera plus ou moins de la distribution empirique. Afin de valider leur sélection, nous nous baserons sur la littérature, des observations graphiques, des vérifications de cohérences, et enfin les tests du Chi 2 et du signe.

Plusieurs mémoires reçus par l'Institut des Actuaires tel que LEFRANC C. (2013)[9] suggèrent que les paramètres d'ordre varient entre 1 et 5. Ces paramètres influent le nombre de valeurs retenues dans le calcul de distance de l'opérateur différence avant. De ce fait ils ne peuvent pas prendre de valeur trop grande, d'où la limite à 5.

Les paramètres α et β sont eux des poids donnés aux critères de régularité horizontale et verticale par rapport au critère de fidélité. Ils permettent de donner plus de poids à l'un de ces critères s'ils sont supérieurs à un, ou moins de poids s'ils sont inférieurs à un. De ce fait, ils évoluent sur une plage de valeur beaucoup plus grande que les paramètres d'ordre : $\mathbb{R}^{+*}$ (il ne faut cependant pas prendre de trop grandes valeurs sinon le critère de fidélité sera totalement négligé).

Les vérifications de cohérence concernent les signes des données lissées obtenues. En effet, le lissage de Whittaker-Henderson n'exclue pas l'obtention de données négatives. Ainsi, lors de son application à des tables de maintien ou de passage, un jeu de paramètre renvoyant des données lissées négatives sera à proscrire (en arrêt de travail nous ne pouvons pas avoir de taux d'entrée ou de sortie négatifs).

Les tables que nous lissons par la méthode de Whittaker-Henderson en dimension deux possèdent l'âge en indice de ligne et la durée de l'arrêt en indice de colonne. Nous retiendrons donc les deux jeux de poids suivants.

- Soit un poids unique et unitaire pour chacune des observations.
- Soit des poids dépendant de l'exposition. Dans ce cas, nous ne regarderons que l'exposition par âge, chaque ancienneté aura donc le même poids pour un âge donné.

Le test du signe permettra de justifier la cohérence du lissage. Soit $d = \hat{q} - q^*$ la différence entre les valeurs initiales et les valeurs lissées. Dans les conditions d'application de l'approximation normale, cette différence est positive avec probabilité $\frac{1}{2}$ et négative avec probabilité $\frac{1}{2}$. De ce fait, le nombre de changement de signe des n valeurs observées suit la loi binomiale $\mathcal{B}\left(n-1;\frac{1}{2}\right)$ avec un nombre de changement de signe moyen fixé à $\frac{n-1}{2}$. Afin de vérifier cette hypothèse, nous observons la statistique :

$$signe = \frac{2 * k - (n-1)}{\sqrt{n-1}}$$

où k est le nombre de changement de singe observé. Nous comparerons la valeur empirique de cette statistique au quantile à 95% d'une loi normale centrée réduire. Si $signe$ est plus petite que cette valeur, alors notre lissage est cohérent au sens du test du signe.

Enfin, le test du Chi 2 sera utilisé pour vérifier que notre jeu de données lissées n'est pas trop éloigné en termes de distribution de notre loi de départ. La statistique que nous utiliserons est :

$$S = \sum_{x=1}^{I} \sum_{anc=1}^{J} \frac{\left(\widehat{N_{xanc}} - N^*_{xanc}\right)^2}{N^*_{xanc}}$$

où :

[9] LEFRANC C. (2013) : *Provisionnement des garanties incapacité et invalidité et problématiques associées*, mémoire de l'Institut des Actuaires

- $\widehat{N_{xt}}$ représente la donnée empirique à l'âge x et l'ancienneté anc ;
- N_{xt}^{*} représente la donnée lissée à l'âge x et l'ancienneté anc.

Les hypothèses de ce test sont :

- $H0$: la population empirique et la population lissée suivent la même loi ;
- $H1$: la population empirique et la population lissée sont différentes.

Nous accepterons $H0$ si $S < \chi^2_{ddl;1-\alpha}$ où $\alpha = 5\%$ le risque d'erreur et $ddl = I * J - 1$ le nombre de degrés de liberté de notre modèle.

2. *Lissage polynomial*

Le lissage polynomial est un lissage paramétrique unidimensionnel que nous appliquons aux taux bruts de sortie de non indemnisation.

La première étape de ce lissage consiste à obtenir les divers taux de sortie de l'état de non indemnisation empirique :

$$q_{x,f} = 1 - \frac{L_{NI,f}(x+1)}{L_{NI,f}(x)}$$

où f est la franchise observée et x l'âge de l'individu en année.

Pour chaque franchise f nous allons étudier trois modèles linéaires classiques :

- un modèle constant : $logit(q_{x,f}) = a$;
- un modèle de degré un : $logit(q_{x,f}) = a + b * x$;
- un modèle de degré deux : $logit(q_{x,f}) = a + b * x + c * x^2$.

Rappel : la fonction logit est telle que $\mathrm{logit(x)} = \ln\left(\frac{\mathrm{x}}{1-\mathrm{x}}\right)$ *où* $\mathrm{x} \in]0;1[$.

Une fois ces trois modèles obtenus, nous déterminons le meilleur modèle grâce à :

- l'étude de la significativité des modèles ;
- la minimisation des critères AIC et BIC ;
- le test entre modèles emboités ;
- et enfin une analyse graphique.

Remarque : le test entre modèles emboités et les critères AIC et BIC ont été présentés dans la partie sur les MLG.

Après sélection du meilleur modèle par les critères présentés ci-dessus, les valeurs prédites seront les logit lissés : $logit(q_{x,f}^{*})$.

Nous obtenons finalement les taux bruts de sortie de non indemnisation par la transformation :

$$q_{x,f}^{*} = \frac{e^{logit(q_{x,f}^{*})}}{1 + e^{logit(q_{x,f}^{*})}}$$

Après l'application de l'ensemble de ces méthodes de lissage nous posséderons :

- la table de maintien en incapacité lissée L_{IC}^{*} ;
- la table de retour au travail lissée L_{retour}^{*} ;
- les tables de maintien en non indemnisation pour les divers franchises f retenues $L_{NI,f}^{*}$.

G. Structuration de la loi de maintien en non indemnisation pour chaque franchise

Cette dernière partie ne concerne que la loi de maintien en non indemnisation. Comme nous l'avons vu précédemment, cette loi est construite par franchise. Or il est peu probable que l'ensemble des franchises soient suffisamment exposées afin de construire les lois de maintien en non indemnisation pour l'ensemble d'entre elles. Ainsi, seules les franchises les plus exposées seront retenues pour la construction. Mais il sera nécessaire de connaitre cette loi pour l'ensemble des franchises. De ce fait, depuis les lois construites sur les franchises les plus exposées, nous cherchons à déduire les lois des autres franchises. Cette étape est appelée structuration de la loi de maintien en non indemnisation pour chaque franchise.

Nous informons le lecteur qu'à cause d'un manque de données, cette structuration ne pourra pas être mise en place dans la partie pratique de ce mémoire. De ce fait, la théorie ne sera pas non plus présentée. Cependant, afin de ne pas laisser le processus incomplet, nous invitons un lecteur implémentant la méthode dans sa globalité à consulter LEBOSSE C. (2009)[10].

Résumé du chapitre de construction des lois

Afin de réaliser une tarification du risque incapacité par simulation, plusieurs lois sont nécessaires. Tout d'abord la loi de maintien en incapacité, puis la loi de retour au travail, et enfin les lois de maintien en non indemnisation (une pour chaque franchise suffisamment exposée au sein de la base de données). Ces lois sont toutes construites grâce à l'estimateur non paramétrique de Kaplan Meier.

Après leur construction empirique, ces lois auront probablement un comportement erratique. Il faudrait un nombre très important de données afin d'obtenir des lois empiriques lisses. De ce fait, nous devront tout d'abord réduire les unités des tables en passant du jour au mois pour les tables de maintien en incapacité et de retour au travail, et en passant du jour à l'année pour les âges retenus dans la table de maintien en non indemnisation. Puis nous appliquerons des méthodes de lissage sur ces lois empiriques. Les lois de maintien en incapacité et de retour au travail seront lissées par la méthode de Whittaker-Henderson en dimension deux et les lois de maintien en non indemnisation seront lissées par lissage polynomial.

Une fois que l'ensemble des lois seront lissées, il faudrait structurer les lois de maintien en non indemnisation par franchise. Mais à cause d'un manque d'exposition ceci ne pourra être réalisé.

III. Implémentation de la tarification

Comme introduit précédemment, l'objectif est l'obtention de la VAP des engagements de l'assureur pour une indemnité de 1€ par jour en incapacité. Pour réaliser cela, nous simulons à de nombreuses

[10] LEBOSSE C. (2009) : *Construction de barèmes de tarification d'arrêt de travail par une méthode de Simulation – Application à un portefeuille de Travailleurs Non-Salariés*, mémoire de l'Institut des Actuaires

reprises l'évolution entre les différents états d'un bénéficiaire sur une année. Ainsi, pour chacune des années simulées, nous connaitrons précisément les jours où le bénéficiaire sera en incapacité.

Nous commencerons par détailler l'algorithme qui nous permettra d'obtenir les multiples simulations des états du bénéficiaire au cours d'une année. Ensuite, nous développerons le processus de tarification du risque incapacité (c'est-à-dire l'obtention de la VAP des engagements de l'assureur) qui nous permettra d'obtenir une approximation de la prime pure. Et enfin nous parlerons de la déclinaison de la tarification par facteurs discriminants.

A. Algorithme de simulation

Nous allons simuler de nombreuses fois l'évolution du bénéficiaire entre les différents états au cours d'une année. Cet algorithme est initialisé avec les trois paramètres suivants qui vont définir la simulation et les caractéristiques du bénéficiaire concerné :

- nb_simu le nombre de simulations effectuées, chacune représentant une année ;
- age l'âge du bénéficiaire au début de la simulation en jour ;
- f la franchise du contrat du bénéficiaire.

Nous cherchons à obtenir nb_simu fois la sinistralité du bénéficiaire relative à l'année N. Donc pour chaque simulation nous étudierons les états de l'individu du $01/01/N$ au $31/12/N$. Mais à cause de la présence de franchise, les indemnisations relatives aux sinistres de cette période seront réparties du $01/01/N + f$ au $31/12/N + f$. Ainsi, nous effectuons nos simulations sur une durée $duree\_maximum = 365 + f - 1$.

Remarque : le -1 provient de la structure de l'algorithme. Comme nous le verrons par la suite, nous simulons l'état de l'assuré au jour $\mathrm{j} + 1$ en fonction de son état au jour j. Donc simuler sur $365 + \mathrm{f} - 1$ jours nous permet de connaitre notre bénéficiaire sur $365 + \mathrm{f}$ jours.

Nous allons maintenant détailler les étapes de l'algorithme. Nous commencerons par introduire les variables qui seront utilisées, puis nous présenterons l'algorithme en lui-même, et enfin nous exposerons la forme des résultats.

1. *Les variables*

Les variables qui seront utilisées au sein de l'algorithme sont :

- k allant de 1 à nb_simu le pas de la simulation
- t représentant les jours de l'années au sein d'une simulation $t \in [\![1; duree\_maximum]\!]$
- $p_{IC}(x; anc)$ la probabilité qu'un bénéficiaire entré en incapacité à l'âge x en jour et avec une ancienneté en incapacité anc en jour reste en incapacité le jour suivant. Son expression est :
$$p_{IC}(x; anc) = \frac{L_{IC}(x; anc + 1)}{L_{IC}(x; anc)}$$
- $p_{retour}(x; anc)$ la probabilité qu'un bénéficiaire entré en incapacité à l'âge x en jour et avec une ancienneté en incapacité anc en jour retourne au travail le jour suivant. Son expression est :
$$p_{retour}(x; anc) = \frac{L_{retour}(x; anc)}{L_{IC}(x; anc)}$$
- $p_{NI,f}(x)$ la probabilité qu'un bénéficiaire d'âge x en jour avec une franchise f reste en non indemnisation le jour suivant. Son expression est :

$$p_{NI,f}(x) = \frac{L_{NI,f}(x+1)}{L_{NI,f}(x)}$$

- r, u et v sont des réalisations des variables aléatoires U, V, et R identiquement distribuées selon une loi uniforme $\mathcal{U}([0;1])$
- $debut\_travail$ est l'instant t auquel le bénéficiaire débute une période dans l'état au travail
- age_{incap} est l'âge du bénéficiaire au début d'une période d'incapacité
- anc est l'ancienneté en incapacité d'un bénéficiaire dans une période d'incapacité
- enfin une matrice de dimension $duree\_maximum \times nb\_simu$ stockera pour chaque simulation l'état du bénéficiaire, les valeurs contenues dans cette matrice seront T, I, et S respectivement pour travail, incapacité, et sortie.

Remarque : les âges retenus pour l'obtention des différentes probabilités étant exprimés en jour, il est donc nécessaire de convertir les unités mensuelles ou annuelles des tables obtenues précédemment en jour. Ceci est réalisé par interpolation linéaire pour les lois de maintien en incapacité et de maintien en non indemnisation, et par répartition uniforme des retours sur le mois pour la loi de retour au travail.

2. L'algorithme

Les étapes qui vont suivre seront réalisées nb_simu fois.

Initialisation

Nous déterminons tout d'abord l'état du bénéficiaire au début de la simulation. Si le bénéficiaire possède une franchise f non nulle, alors nécessairement il est dans l'état au travail au début de la simulation. En effet, s'il était en incapacité, cet engagement serait à associer à l'année $N-1$ et non à l'année N actuellement étudiée. Cependant, si le bénéficiaire possède une franchise f nulle, alors nous devons simuler son état à l'initialisation. Étant nécessairement au travail la veille du $01/01/N$ (pour les mêmes raisons que celles énoncées au début de ce paragraphe), nous utilisons la loi de maintien en non indemnisation pour obtenir son état au 1er jour de la simulation. Soit r une simulation de $R \sim \mathcal{U}([0,1])$. Si $r < p_{NI,f}\big(max(age-1;\alpha)\big)$ alors le bénéficiaire est au travail à l'initialisation, sinon il est en incapacité à l'initialisation.

Remarque : le maximum retenu dans la probabilité de maintien en non indemnisation permet d'éviter de sortir de la table de maintien en non indemnisation. En effet, comme présenté précédemment, cette loi est construite pour des âges compris entre α et β. Donc dans le cas d'un bénéficiaire ayant l'âge α, nous devons conserver α et non $\alpha - 1$ qui n'existe pas dans la table. Le minimum présent dans la formule de age_{incap} du paragraphe suivant a le même objectif.

Si l'individu débute dans l'état au travail, nous initialisons $debut\_travail = 1$. Sinon, dans le cas où l'individu début dans l'état d'incapacité, nous initialisons $age_{incap} = min(age+1;\beta)$ et $anc = 1$.

Corps de l'algorithme

Le corps de l'algorithme est une boucle qui sera répétée pour chaque t entre 1 et $duree\_maximum$. Les tâches à accomplir vont permettre de déterminer l'état du bénéficiaire en $t+1$ en fonction de son état observé en t.

Si le bénéficiaire est au travail en t

Si en t le bénéficiaire est au travail depuis une durée inférieure à f alors il reste au travail. En effet, tant que le bénéficiaire n'a pas atteint sa durée de franchise, il ne peut pas sortir de l'état de non indemnisation (sa probabilité de rester dans l'état est égale à 1). Cette condition peut s'écrire : $t < debut\_travail + f$.

Sinon, nous devons étudier si le bénéficiaire entrera en incapacité à la période suivante ou s'il restera en non indemnisation. Pour cela, nous simulons tout d'abord une réalisation u de $U \sim \mathcal{U}([0,1])$. Puis :

- si $u < p_{NI,f}\big(min(age + t; \beta)\big)$ alors le bénéficiaire restera au travail en $t + 1$;
- sinon il entrera en incapacité en $t + 1$. Dans ce cas nous définissons son âge de passage en incapacité : $age_{incap} = min(age + t + 1; \beta)$ et initialisons son ancienneté dans l'arrêt : $anc = 1$.

Si le bénéficiaire est en incapacité en t

Si le bénéficiaire est en incapacité en t, nous différencions trois cas après simulation de v une réalisation de $V \sim \mathcal{U}([0,1])$.

- Si $v < p_{IC}\big(age_{incap}; anc\big)$ alors le bénéficiaire reste en incapacité en $t + 1$. Dans ce cas nous incrémentons l'ancienneté en incapacité anc de 1 jour.
- Si $p_{IC}\big(age_{incap}; anc\big) \leq v < p_{IC}\big(age_{incap}; anc\big) + p_{retour}\big(age_{incap}; anc\big)$ alors le bénéficiaire sort de sa période d'incapacité et sera donc au travail en $t + 1$. Dans ce cas, nous redéfinissons la date de début de travail comme : $debut\_travail = t + 1$.
- Et enfin, si $p_{IC}\big(age_{incap}; anc\big) + p_{retour}\big(age_{incap}; anc\big) \leq v$ alors le bénéficiaire sort du portefeuille en $t + 1$. Dans ce cas, cette simulation se termine et nous passons à la suivante.

3. *Résultats*

Une fois l'état du bénéficiaire d'âge age et de franchise f obtenu pour chaque jour de chaque année parmi les nb_simu années simulées, nous obtenons un tableau de la forme suivante :

Jours	Simulation n°1	Simulation n°2	Simulation n°3	…	Simulation n°nb_simu
1	T	T	T	…	T
2	T	T	T	…	T
3	T	I	T	…	T
4	T	I	T	…	T
5	T	I	I	…	T
⋮	⋮	⋮	⋮	…	⋮
$365 + f$	T	S	T	…	I

Figure 24 – Exemple d'une sortie de l'algorithme de simulation

Maintenant que nous connaissons pour toute simulation l'état du bénéficiaire en tout temps, nous pouvons calculer pour chacune de ces simulations la VAP des engagements de l'assureur et ainsi en moyennant l'ensemble de ces valeurs, obtenir la prime pure.

B. Formules de tarification

Comme indiqué dans les paragraphes précédents, nous cherchons à calculer la VAP des engagements de l'assureur pour chacune des simulations réalisées précédemment. Puis, en calculant la moyenne de ces multiples VAP nous obtiendrons la prime pure d'incapacité pour le bénéficiaire d'âge age et la franchise f.

Nous allons détailler les formules pour une simulation uniquement. Ensuite il faudra appliquer cette méthode à chacune des simulations puis utiliser la formule d'agrégation qui sera présentée à la fin de cette partie afin d'obtenir la prime pure finale notée Π.

Dans les parties qui vont suivre, i représentera le taux d'actualisation et $v = \frac{1}{1+i}$.

Le calcul de la VAP d'une simulation se partage en deux éléments :

$$VAP = VAP_{\leq 1an} + VAP_{>1an}$$

où :

- $VAP_{\leq 1an}$ est la part des engagements de l'assureur relatifs aux périodes d'incapacité survenues pendant l'année de la simulation N et à verser courant N ;
- $VAP_{>1an}$ est la part des engagements de l'assureur relatifs aux périodes d'incapacité survenues pendant l'année de la simulation N mais à verser après N.

Il faut préciser que $VAP_{>1an} \neq 0$ si et seulement si le bénéficiaire est toujours en incapacité au dernier jour de la simulation, c'est-à-dire au jour $365 + f$.

Pour un engagement de 1€ par jour en incapacité, nous posons $\forall t \in [\![1; 365 + f]\!]$, $X_t \in \{T, I, S\}$ l'état du bénéficiaire à la date t. La prime pure $VAP_{\leq 1an}$ s'exprime alors :

$$VAP_{\leq 1an} = \sum_{t=1}^{365+f} \mathbb{I}_{X_t=I} * v^{\frac{t}{365}}$$

A la différence de $VAP_{\leq 1an}$, quelques calculs sont nécessaires pour d'obtenir $VAP_{>1an}$. Afin de déterminer ces engagements dus après l'année de simulation, nous devons estimer l'espérance de maintien en incapacité pour le bénéficiaire observé. Pour cela, nous utilisons la valeur du barème de provisionnement d'incapacité en cours. Nous présentons donc ci-après la formule de construction de ce barème ainsi que les paramètres à retenir.

Le barème d'incapacité en cours s'obtient depuis la table de maintien en incapacité. Pour chaque âge x en jour et ancienneté anc en jour nous obtenons le barème par la formule suivante :

$$a_{IC}(x; anc) = \frac{1}{L_{IC}(x; anc)} * \sum_{j=anc}^{1095} \left[\frac{1}{2} * \left(L_{IC}(x; anc) + L_{IC}(x; anc+1) * v^{\frac{1}{365}}\right) * v^{\frac{j-anc}{365}}\right]$$

Maintenant que nous avons notre barème, nous devons déterminer quel âge d'entrée en incapacité et quelle ancienneté dans l'arrêt retenir pour notre bénéficiaire concerné. Pour cela, nous calculons tout d'abord le nombre de jour nb qu'a duré cette dernière période d'incapacité au dernier jour de la simulation :

$$duree = \sum_{t=1}^{365+f} \left(\prod_{k=t}^{365+f} \mathbb{I}_{\{X_k=I\}} \right)$$

Une fois cette durée connue, nous pouvons calculer l'âge de début d'incapacité x' et l'ancienneté au dernier jour de la simulation anc' :

- $x' = age + 365 - duree$
- $anc' = f + duree$

Ainsi, pour un engagement de 1€ par jour en incapacité,

$$VAP_{>1an} = \mathbb{I}_{X_{365+f}=I} * a_{IC}(x'; anc') * v^{\frac{365+f}{365}}$$

Maintenant que nous connaissons VAP pour chacune des simulations, nous pouvons calculer Π par la formule suivante :

$$\Pi \approx \frac{1}{nb\_simu} * \sum_{k=1}^{nb\_simu} VAP(k)$$

où $VAP(k)$ représente la VAP de la $k^{\text{ème}}$ simulation.

Grâce à cette formule, nous pouvons pour un bénéficiaire avec une franchise f fixée, déterminer un tarif par âge. Il est possible que malgré un grand nombre de simulations, la courbe représentant les différentes primes pures par âge comporte des irrégularités. Dans ce cas, nous appliquerons la méthode de lissage par moyennes mobiles d'ordre $2d + 1$ suivante :

$$\Pi_{lisse}(x) = \frac{1}{2d+1} \sum_{k=x-d}^{x+d} \Pi(k)$$

où $\Pi(x)$ représente la prime pure du bénéficiaire d'âge x. Le choix de d est déterminé en fonction des données étudiées de manière à ne pas modifier la tendance de la courbe tout en la rendant plus régulière.

C. Déclinaison de la tarification par facteur discriminant

Dans l'ensemble des parties précédentes, nous avons construit notre tarification de manière générale, c'est-à-dire sur l'ensemble de la population. Dans la pratique, l'assureur cherchera à segmenter son portefeuille en sous populations afin d'obtenir une prime pure plus proche de son risque. Afin de réaliser une tarification segmentée, la méthode sera identique au cas général mais la base de données utilisée sera modifiée.

Par exemple, si un assureur possède une population que nous pouvons diviser selon deux paramètres :

- le premier est $chiffre$ et prend les valeurs 1 et 2 ;
- le second est $lettre$ et prend les valeurs a et b.

Dans ce cas, l'assureur devra fractionner sa base de données en quatre bases :

- une première base où $chiffre = 1$ et $lettre = a$;
- une seconde base où $chiffre = 1$ et $lettre = b$;
- une troisième base où $chiffre = 2$ et $lettre = a$;
- une quatrième et dernière base où $chiffre = 2$ et $lettre = b$.

Puis pour chacune de ces bases l'assureur appliquera la méthode globale développée précédemment.

Il est nécessaire de porter une attention particulière aux expositions dans chacune des classes. Comme nous l'avons souligné précédemment l'estimateur de Kaplan Meier est très sensible aux problèmes d'expositions, et décliner la tarification selon plusieurs variables réduit fortement le volume de la base de données utilisée.

Résumé du chapitre sur l'implémentation de la tarification

Afin d'obtenir la prime pure pour un bénéficiaire d'âge x et de franchise f, nous commençons par simuler à de nombreuses reprises l'évolution de ce bénéficiaire entre les trois états : travail, incapacité, et sortie au cours d'une année. Cette simulation est réalisée à l'aide d'un algorithme présenté dans la partie A.

A la suite de cette phase de simulation, nous possédons nb_simu années au sein desquelles nous connaissons l'état du bénéficiaire en tout temps. Nous pouvons donc obtenir la valeur actuelle probable des engagements de l'assureur sur chacune de ces années. Cette dernière étant divisée en deux parties :

- d'un côté les engagements de l'assureur relatifs aux périodes d'incapacité survenues pendant l'année de la simulation N et à verser courant N ;
- de l'autre côté les engagements de l'assureur relatifs aux périodes d'incapacité survenues pendant l'année de la simulation N mais à verser après N.

Finalement, en calculant la moyenne de l'ensemble des valeurs actuelles probables obtenues pour le bénéficiaire d'âge x et de franchise f, nous approximons la prime pure pour un engagement de l'assureur de 1€ par jours passé en incapacité.

IV. Tarification par simulation en pratique

Dans cette partie nous mettrons en pratique la méthode de simulation présentée précédemment à l'aide du logiciel R[11]. Comme pour la méthode par MLG, nous ne présenterons les résultats que globalement afin de ne pas nuire à la confidentialité des données.

A. Construction des lois

En utilisant l'estimateur de Kaplan Meier nous avons construit la table de maintien en incapacité brute suivante :

[11] Les packages utilisés sont :
- SPINU V. et al. : *lubridate* *(version 1.7.4)*
- WICKHAM H. : *stringr* *(version 1.3.1)*
- DOWLE M. et al. : *data.table* *(version 1.11.4)*
- WICKHAM H. et al. : *dplyr* *(version 0.7.5)*
- HOTHORN T. et al. : *lmtest* *(version 0.9-36)*

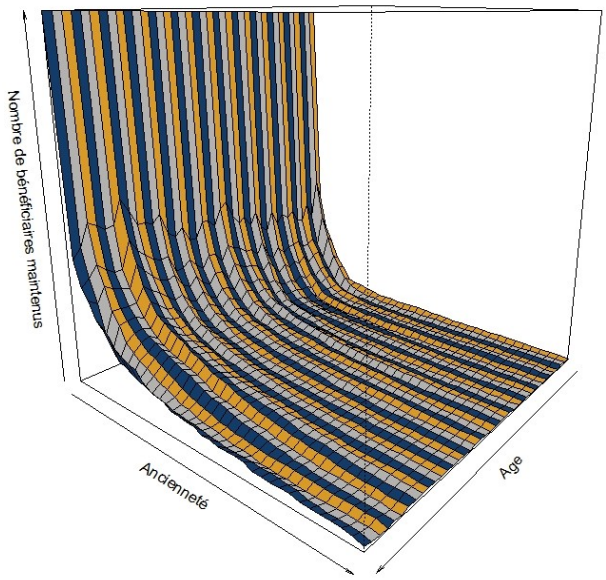


Figure 25 – Table maintien en incapacité brute

Avant de lisser cette table au comportement erratique, nous allons étudier la variance de l'estimation obtenue ainsi que les différents intervalles de confiance par âge. Grâce à l'estimateur de la variance de Greenwood, nous pouvons exclure les âges extrêmes non suffisamment exposés. De ce fait, nous conservons les âges à l'entrée de 23 à 62 ans. Les deux graphiques suivants présentent à droite la variance de l'estimateur de Kaplan Meier pour les âges retenus, et à gauche le logarithme de cette variance afin d'éviter les problèmes d'échelle dus aux très grandes valeurs.

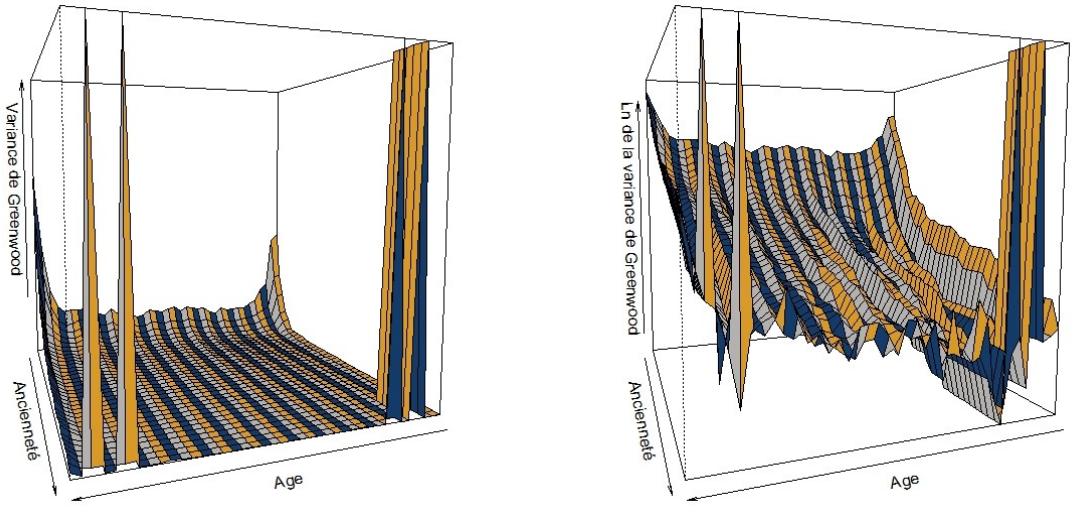


Figure 26 – Variance de l'estimation par Kaplan Meier du maintien en incapacité

Sur la plage d'âges conservés nous repérons neuf valeurs pour lesquelles nous ne possédons pas suffisamment de données et donc pour lesquelles l'estimation par Kaplan Meier est fortement biaisée. Cependant, n'étant pas nombreuses, ces valeurs seront corrigées par le lissage. Les autres valeurs ne présentent pas une variance trop élevées, nous pouvons donc avoir confiance en l'estimation de Kaplan Meier pour la loi de maintien en incapacité. Ceci est confirmé par la construction de quelques intervalles de confiances de Rothman. Ci-dessous, de gauche à droite, les intervalles de confiance pour les âges à l'entrée : 23, 30, 40, 50, 58, et 62.

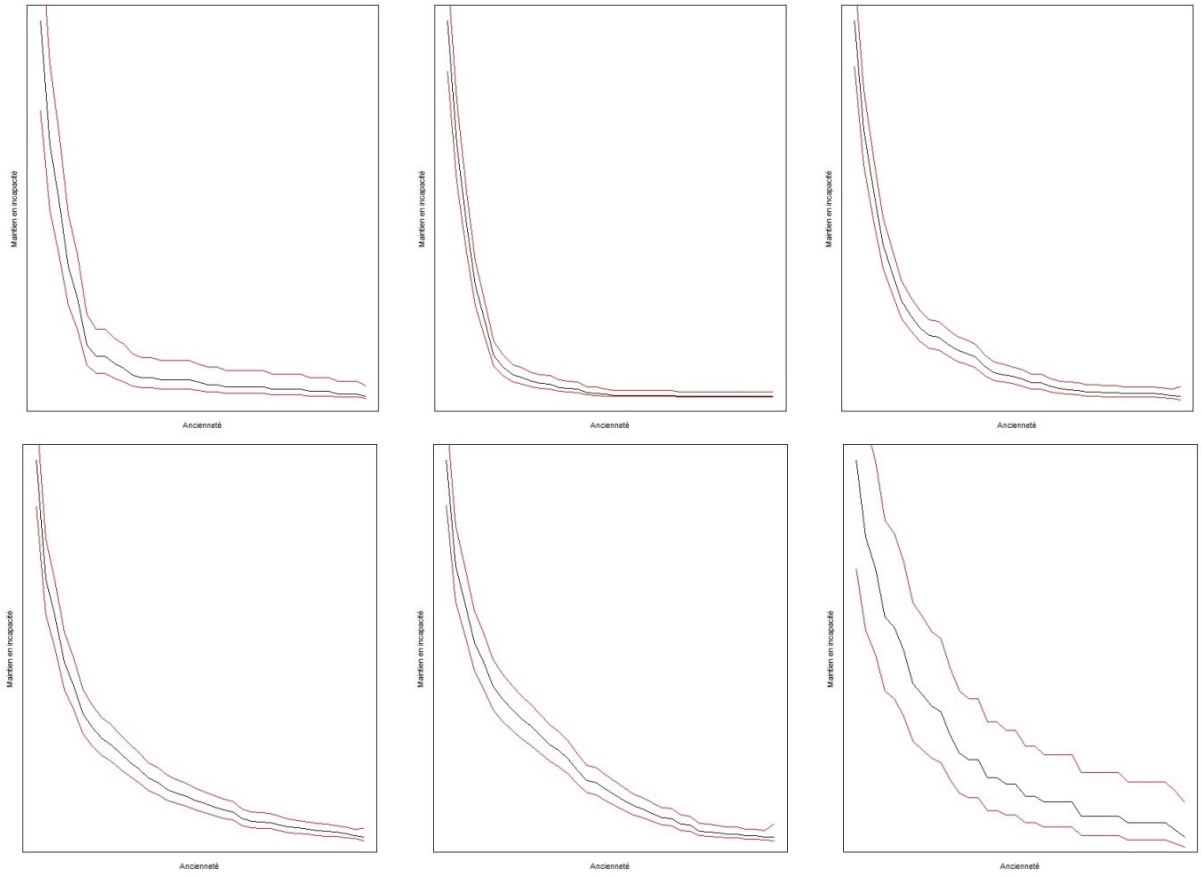


Figure 27 – Intervalles de confiance du maintien en incapacité

Cette table de maintien en incapacité est comme nous le voyons ci-dessus erratique. Nous allons donc la lisser en appliquant la méthode de Whittaker-Henderson en dimension deux aux taux de sortie d'incapacité. La fonction utilisée pour ce lissage (présentée en ANNEXE 4) est celle proposée par PLANCHET F. sur le site www.ressources-actuarielles.net.

Après de nombreux tests de diverses combinaisons de paramètres, nous retenons les paramètres suivants :

- le poids donné à la régularité verticale $\alpha = 16$;
- le poids donné à la régularité horizontale $\beta = 16$;
- l'ordre vertical $OV = 3$;
- l'ordre horizontal $OH = 3$;
- les poids de la matrice de poids W sont l'exposition de chacun des âges (identiques pour l'ensemble des anciennetés en incapacité pour un âge fixé).

Ces paramètres sont ceux qui permettent d'obtenir le meilleur lissage tout en restant proche de la distribution empirique vis-à-vis du test du Chi2.

Le lissage nous permet de passer des taux de sortie d'incapacité bruts (à gauche) aux taux lissés (à droite).

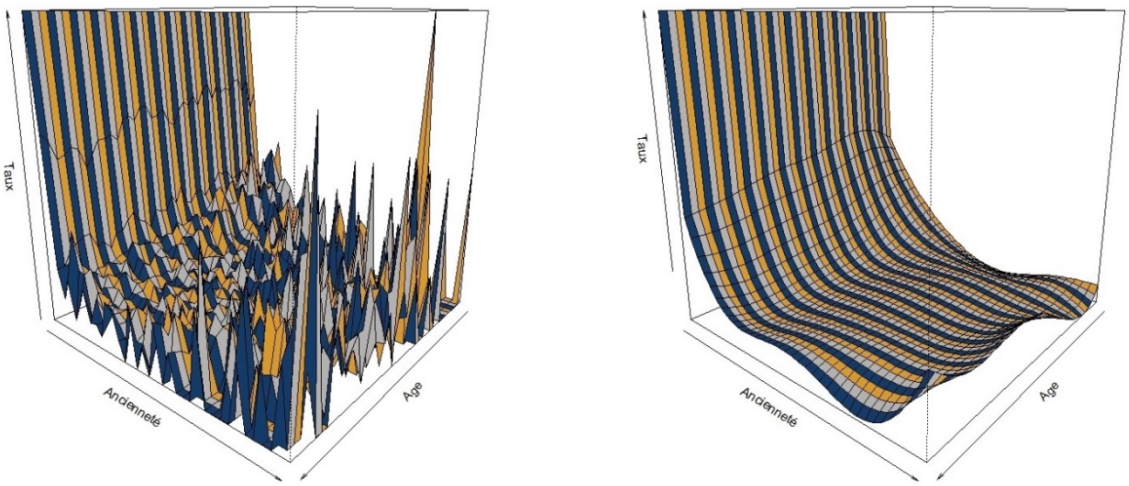


Figure 28 – Taux de sortie d'incapacité bruts et lissés

Finalement, depuis les taux lissés précédents nous obtenons la table de maintien en incapacité suivante :

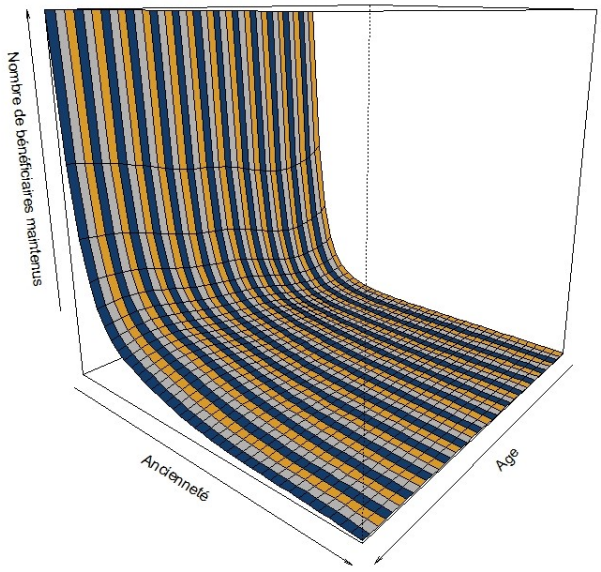


Figure 29 – Table maintien en incapacité lissée

Cette table a finalement été back-testée afin de vérifier son adéquation à la population étudiée. Afin de réaliser ce test, nous comparons à chaque fin d'années, les durées empiriques de maintien en incapacité avec les estimations obtenues depuis notre table empirique. La date de censure non informative étant en octobre 2016, nous pouvons réaliser ce test pour les années de 2009 à 2015. Voici les écarts relatifs entre les valeurs empiriques et les valeurs théoriques avec notre table de maintien en incapacité lissée :

Année	2009	2010	2011	2012	2013	2014	2015
Écart relatif	10%	7%	1%	2%	2%	2%	-2%

Figure 30 – Backtest de la table de maintien en incapacité lissée

Nous repérons une légère dérive de la population au fil des années étudiées. Cependant dans son ensemble la table est fidèle à la population.

Ensuite, le même processus a été appliqué pour l'obtention de la table de retour au travail. La table lissée finalement obtenue est présentée ci-dessous (la table initiale ainsi que les taux bruts et lissés sont disponibles en ANNEXE 5).

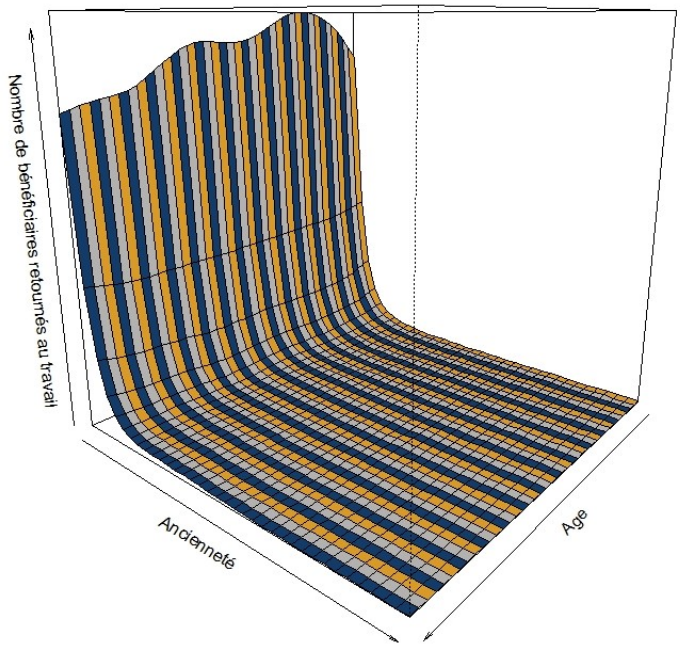


Figure 31 – Table de retour au travail lissée

Enfin, nous avons construit les lois de maintien en non indemnisation par franchise. Nous ne présenterons ici que les lois obtenues depuis les franchises les plus exposées. De plus, en ce qui concerne cette loi de maintien en non indemnisation, la population particulière contenue dans notre base de données a un comportement différentes du reste de la population. De ce fait, nous la traiterons à part.

Les lois de maintien en non indemnisation brutes obtenues par Kaplan Meier pour la population générale sont les suivantes :

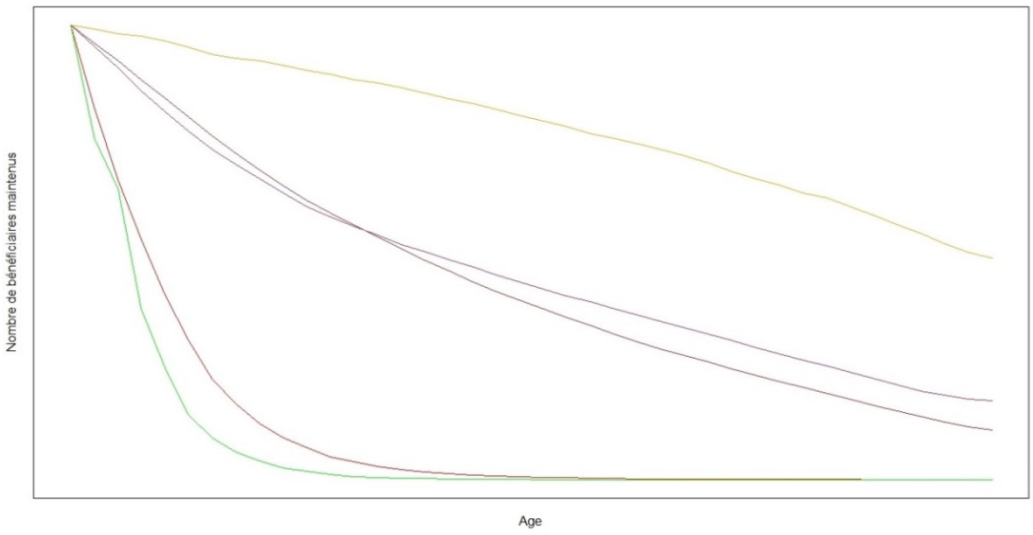


Figure 32 – Lois de maintien en non indemnisation brutes

Remarque : les courbes rouge et verte sont des franchises courtes, les courbes marron et violette sont des franchises longues, la courbe orange est une franchise très longue.

Et voici la loi de maintien en non indemnisation brute obtenue pour la population particulière :

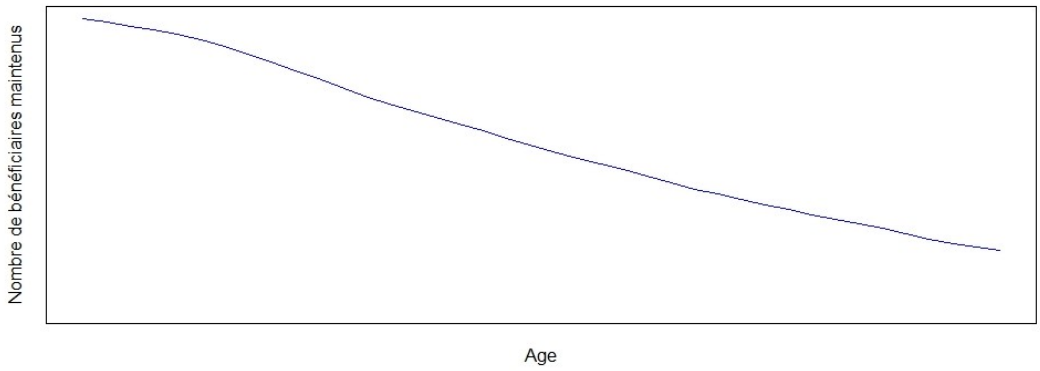


Figure 33 – Loi de maintien en non indemnisation brute pour la population spécifique

Sur le premier graphique représentant les lois de maintien en non indemnisation pour notre population générale, nous repérons deux incohérences.

- La franchise retenue pour obtenir la loi rouge est inférieure à celle retenue pour construire la loi verte, donc nous nous attendons à observer plus de maintien en non indemnisation dans le second cas que dans le premier. Or la courbe verte est majoritairement sous la courbe rouge, c'est donc le comportement opposé qui est observé.
- La même situation s'observe pour les courbes marron et violette alors que l'écart entre les franchises est beaucoup plus important. Ces courbes ne devraient pas se croiser.

Ces incohérences proviennent d'un manque de données pour la construction de ces lois. En effet, comme nous l'avons vu dans la partie théorique, ces lois se construisent par franchise. Ainsi, nous devons isoler les données relative à la franchise étudiée afin de construire la loi de maintien en non indemnisation qui lui sera associée. Or, au niveau de la base des prestations, nous possédons environ 30000 périodes d'incapacité et notre base contient de très nombreuses franchises. A titre d'exemple, la franchise retenue pour la construction de la loi de maintien en non indemnisation pour la population spécifique est la plus exposée et représente environ 30% des données de prestation. De ce fait, l'exposition par franchise pour construire les lois restantes est beaucoup plus faible.

Pour ne pas clôturer la méthode ici et afin de montrer l'application du processus de tarification dans son ensemble, nous continuerons le développement avec seulement deux lois de maintien en non indemnisation :

- la loi de maintien en non indemnisation retenue pour la population spécifique ;
- et la loi de maintien en non indemnisation retenue pour la population générale représentée par la courbe marron sur le graphique précédent.

Ces lois sont celles obtenues pour les franchises les plus exposées représentant respectivement 30% et 20% de la base des prestations.

Les variances de Greenwood obtenues pour l'estimateur de Kaplan Meier de ces lois sont les suivantes (à gauche la population spécifique avec une franchise très courte, à droite la population générale avec une franchise longue) :

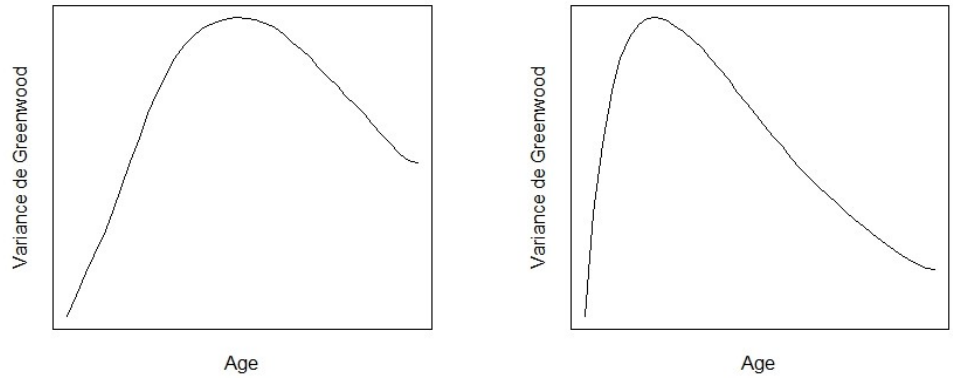


Figure 34 – Variance de l'estimation par Kaplan Meier du maintien en non indemnisation

Remarque : l'exclusion des franchises non suffisamment exposées pour la population générale nous empêche de structurer la loi de maintien en non indemnisation par franchise.

Ainsi, nous obtenons les intervalles de confiance suivants :

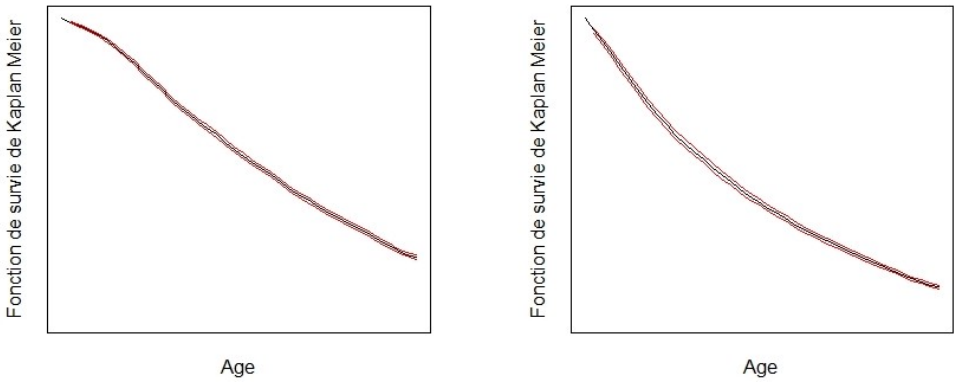


Figure 35 – Intervalles de confiance du maintien en non indemnisation

Ces deux lois sont ensuite lissées par lissage polynomial. Nous obtenons ainsi les deux lois de maintien en non indemnisation lissées suivantes.

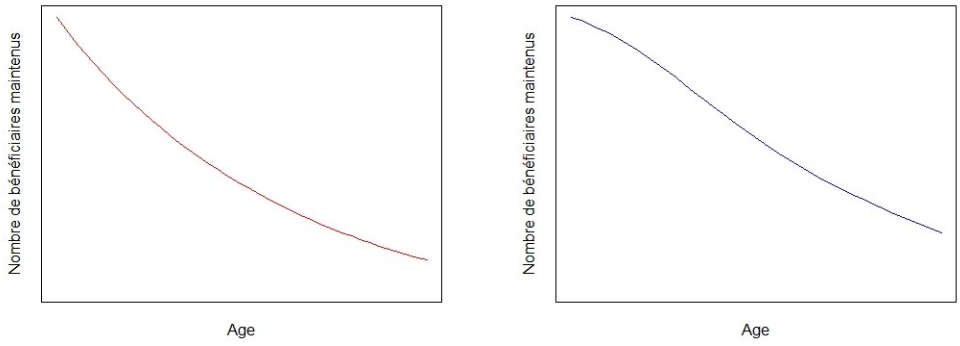


Figure 36 – Lois de maintien en non indemnisation lisses retenues

B. Implémentation de la tarification

Maintenant que nous possédons les lois lissées requises pour notre méthode de simulation, nous pouvons obtenir les approximations des primes pures. Nous réalisons 100 000 simulations par

franchise et obtenons les tarifs bruts en fonction de l'âge suivants. Tout d'abord le tarif pour la population générale, puis celui pour la population spécifique.

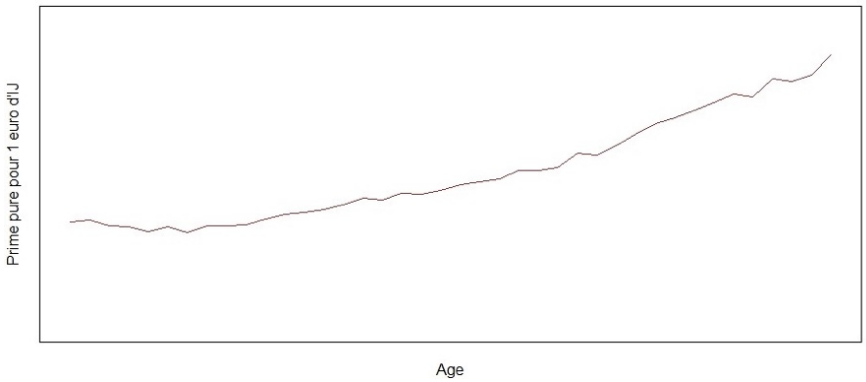


Figure 37 – Primes pures brutes pour la population générale

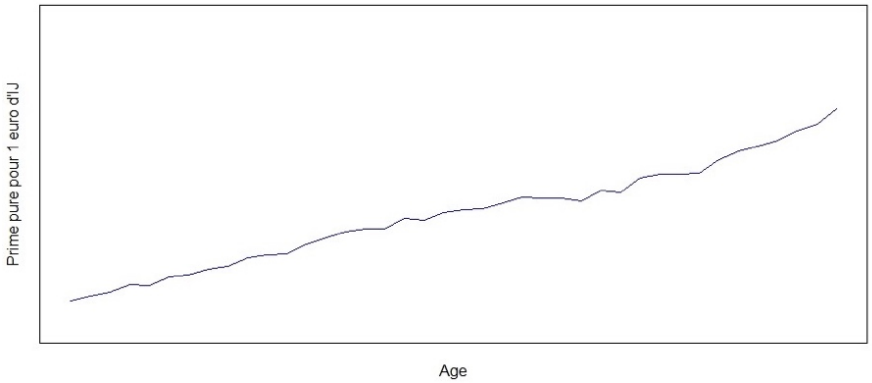


Figure 38 – Primes pures brutes pour la population spécifique

Ces primes pures ayant un comportement erratique, nous appliquons la méthode de lissage par moyenne mobile de degré 2 à nos primes pures brutes. Nous obtenons ainsi les primes pures lissées suivantes : à gauche la franchise longue pour la population classique et à droite la franchise courte pour la population spécifique.

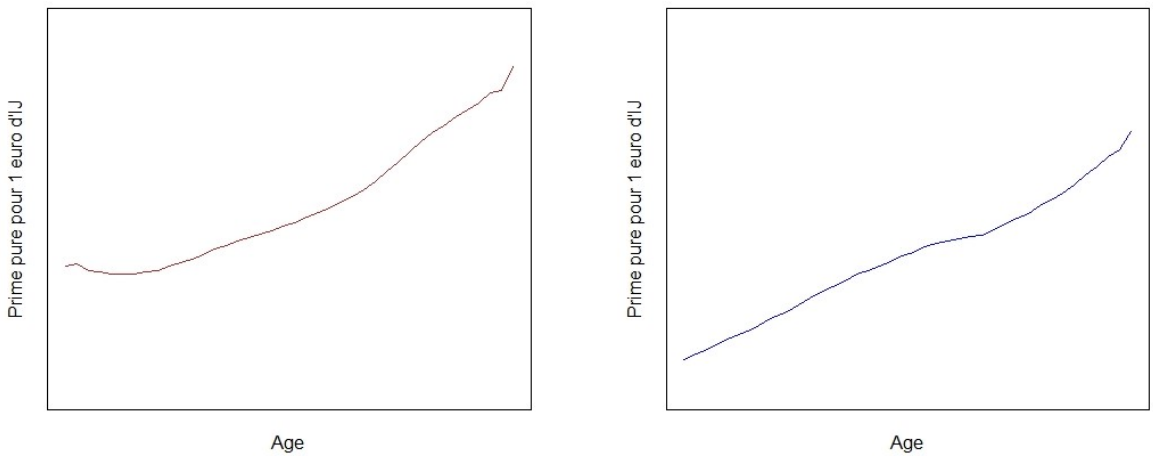


Figure 39 – Primes pures lissées

Résumé du chapitre sur la méthode de tarification par simulation en pratique

Dans cette partie nous avons appliqué la méthode de tarification du risque incapacité par simulation à notre base collective. Tout d'abord, nous avons construit la loi de maintien en incapacité et la loi de retour au travail. Ces lois ont été lissées et la loi de maintien en incapacité spécifiquement back-testée. Elles sont donc aptes à être utilisée pour notre méthode de tarification. Cependant, nous n'avons pas été en mesure de construire la loi de maintien en non indemnisation sur l'ensemble du portefeuille. En effet, nous ne possédions pas suffisamment de prestations pour estimer de manière fiable les lois de maintien en non indemnisation par franchise. Ceci a été mis en avant par des comportements incohérents entre les différentes lois de maintien en non indemnisation construites précédemment.

Néanmoins, pour ne pas avorter totalement le processus de tarification, nous avons souhaité continuer en ne retenant que la franchise associée à la population spécifique et la franchise la plus exposée associée à la population générale. Ces deux franchises représentent respectivement 30% et 20% des prestations sont les plus exposées de notre base de données. Ainsi, pour ces deux franchises, nous avons appliqué l'algorithme de simulation et les formules de tarification afin d'obtenir les primes pures pour une garantie de 1€ d'IJ versée par jour en incapacité.

Partie 4 – Ajout de la déclaration sociale nominative à la modélisation

Dans cette partie nous présenterons ce que la DSN peut apporter à la tarification du risque incapacité et plus particulièrement aux méthodes de tarification présentées précédemment. Nous précisons que cette partie sera entièrement théorique. En effet, notre base de données regroupe des sinistres de 2009 à 2016, hors la DSN n'est en application que depuis le milieu de l'année 2015, de ce fait nous ne pouvions pas récupérer les données relatives à nos assurés.

I. Intérêt pour les organismes assureurs

En plus de faciliter et sécuriser le traitement des données, la DSN a un véritable intérêt pour les organismes complémentaires couvrant le risque incapacité. Nous verrons que la DSN permet une meilleure gestion de l'évolution des contrats au cours du temps ainsi qu'une amélioration du processus de tarification.

A. Intérêt pour la gestion et l'évolution des contrats

La DSN peut devenir un outil simplifiant grandement la gestion des couvertures d'incapacité en ce qui concerne les changements de franchise. Supposons qu'un assureur ait un contrat couvrant l'incapacité avec une franchise unique fixée à 60 jours et qu'il souhaite proposer – toutes choses égales par ailleurs – un nouveau contrat avec une franchise à 30 jours. Ce changement de franchise impliquera une augmentation de ses prestations provenant de deux causes différentes.

La première cause d'augmentation des prestations provient des sinistres déjà indemnisés : ces sinistres seront indemnisés pendant 30 jours de plus. L'assureur est en mesure de déterminer exactement ce surcoût car ce dernier connait le montant des IJ versées à ses bénéficiaires et le nombre exact de jours supplémentaires à indemniser.

La seconde cause d'augmentation des prestations est plus compliquée à déterminer. Du fait de la troncature gauche des données, l'assureur n'a généralement pas connaissance des sinistres durant moins de la durée de franchise. De ce fait, lors du passage de 60 à 30 jours, de nombreux sinistres précédemment non indemnisés deviendront des sinistres indemnisés. La détermination de cette augmentation du coût est actuellement compliquée et généralement basée sur des statistiques. La DSN améliorera grandement la précision de cette estimation. En effet, la survenance d'un arrêt de travail et la fin d'un arrêt de travail font partie des évènements à déclarer dans le contexte de la DSN. De ce fait, l'assureur est en mesure de connaitre l'ensemble des arrêts survenus sur sa population. Plus précisément, avec l'exemple développé ci-dessus, il connaitra, pour son contrat actuel ayant une franchise de 60 jours, l'ensemble des sinistres durant entre 30 et 59 jours, sinistres qu'il n'observait pas jusque-là. Donc, il pourra très précisément analyser l'augmentation de coût engendrée par cette réduction de franchise. Et si sa population est suffisamment grande, il sera en mesure de déterminer en moyenne combien cette réduction de franchise lui coutera.

B. Intérêt pour la tarification des contrats

La bonne utilisation des données de la DSN par un assureur peut lui permettre d'améliorer fortement le processus de tarification. Dans un premier temps via l'augmentation du volume de données utilisées, et dans un second temps en permettant un perfectionnement des méthodes de tarification.

Comme nous l'avons vu dans les parties précédentes, le manque de données est un problème majeur du processus de tarification. Ce problème se rencontre particulière dans le cas de l'incapacité car, à cause de la troncature gauche des données, l'assureur n'a accès qu'à une partie des sinistres. Plus la durée de franchise sera longue, plus le nombre de sinistres auxquels a accès l'assureur sera faible.

Cependant, une bonne utilisation des données de la DSN permettra de résoudre ce problème. En effet, si ces données sont utilisées, l'assureur sera en mesure de connaitre l'ensemble des sinistres survenus et non plus seulement ceux dépassant la durée de franchise, c'est à dire l'ensemble des sinistres de sa population. En d'autres termes, cela lui permettra de **supprimer la troncature gauche des données due à la présence de franchise**. En effet, la déclaration évènementielle de l'arrêt de travail est à produire dès que l'arrêt de travail est connu, donc les données de la DSN contiennent tous les arrêts de travail de la population assurée quelle que soit la franchise. Ceci nous permettra de refondre quelques éléments des modélisations précédentes.

Remarque : la déclaration de la survenance d'un arrêt de travail doit être faite dans les cinq jours suivants l'arrêt. Donc nous aurions pu penser que la suppression de la troncature gauche des données n'est pas rigoureusement appliquée car certains sinistres auraient pu survenir sans être déclarés à la DSN durant cet intervalle de cinq jours. Cependant, nous rappelons que la date de censure n'est pas la date d'exportation des données. En effet, afin que la censure soit non informative, nous retenons généralement comme date de censure la date d'extraction moins plusieurs mois afin d'éviter tout biais. Les sinistres en cours à cette date seront considérés censurés, et les sinistres survenus après cette date ne seront pas retenus. De ce fait, à la date de censure non informative, nous avons bien connaissance de l'ensemble déclarations d'arrêt de la DSN. Donc la suppression de la troncature gauche des données due à la franchise est rigoureusement obtenue.

La seule troncature qui restera dans notre base de données sera celle causée par la date de début d'observation des données. Cependant, cette dernière ne peut être supprimée de la modélisation car il existe nécessairement une date de début d'observation.

Résumé du chapitre sur l'intérêt pour les organismes assureurs

L'analyse des données de la DSN présente un fort intérêt pour les organismes assureurs. En effet, l'employeur déclare via la DSN l'ensemble de ses arrêts de travail dès leur survenance. De ce fait, l'assureur est en mesure de connaitre l'ensemble des sinistres survenus sur sa population et non plus seulement ceux durant plus de la durée de franchise. En d'autres termes, l'assureur peut avec une bonne analyse de ses données de DSN supprimer la troncature gauche des données due à la présence de franchises.

Cette analyse des données de la DSN lui permettra tout d'abord de connaitre exactement le coût d'une réduction de franchise car il connaitra dans ce cas précisément le volume de sinistres qui ne sont actuellement pas couverts mais qui le deviendront à la suite du changement de franchise.

Cependant, le véritable point d'intérêt de ces nouvelles données est qu'elles vont permettre de refondre les processus de modélisation du risque incapacité.

II. Déclaration sociale nominative d'un arrêt de travail

Dans les généralités de ce mémoire nous avons présenté la DSN dans son ensemble. Au sein de cette partie nous nous concentrerons sur le fonctionnement des déclarations relatives à l'arrêt de travail. Nous présenterons tout d'abord la procédure de signalement de la survenance d'un arrêt de travail, puis celle de la clôture de cet arrêt.

A. Signalement de la survenance

La déclaration de la survenance d'un arrêt de travail doit être envoyée dans les cinq jours qui suivent la connaissance de l'arrêt par l'employeur, et ce pour tout arrêt durant au moins un jour. Le processus varie légèrement entre un employeur pratiquant la subrogation et un employeur ne la pratiquant pas. Cependant, dans tous les cas, la déclaration d'un arrêt de travail se présente comme dans l'exemple ci-dessous provenant du site www.dsn-info.fr. Il s'agit de la déclaration d'arrêt de travail d'un salarié en arrêt maladie du 19 au 30 juillet (date de fin prévisionnelle).

Arrêt de travail - S21.G00.60		
S21.G00.60.001	Motif de l'arrêt	01 - maladie
S21.G00.60.002	Date du dernier jour travaillé	18/07/2016
S21.G00.60.003	Date de fin prévisionnelle	30/07/2016
S21.G00.60.004	Subrogation	02 - non
S21.G00.60.010	Date de la reprise	- (*)
S21.G00.60.011	Motif de la reprise	- (*)

Figure 40 – Exemple de déclaration de survenance d'un arrêt dans la DSN

Tout d'abord, il est nécessaire de préciser que l'ensemble des éléments renseignés dans la DSN ont un identifiant unique. Ces différents éléments sont séparés en blocs. Un bloc est identifié par les deux chiffres XX après le préfixe S21.G00. Par exemple :

- S21.G00.30 représente le bloc 30 qui regroupe les informations sur le salarié ;
- S21.G00.40 représente le bloc 40 qui regroupe les informations sur le contrat de travail du salarié ;
- Etc.

Le bloc qui nous intéressera par la suite est le bloc 60 qui regroupe les données sur les arrêts de travail. Ce bloc se fractionne en multiples sous blocs :

- S21.G00.60.001 : motif de l'arrêt ;
- S21.G00.60.002 : date du dernier jour travaillé ;
- S21.G00.60.003 : date de fin prévisionnelle ;
- S21.G00.60.004 : subrogation ;
- S21.G00.60.005 : date de début de subrogation ;
- S21.G00.60.006 : date de fin de subrogation ;
- S21.G00.60.007 : IBAN ;
- S21.G00.60.010 : date de la reprise du travail ;
- S21.G00.60.011 : motif de la reprise du travail ;
- S21.G00.60.012 : date de l'accident ou de la première constatation.

Ainsi, dès que l'employeur a connaissance d'un arrêt durant au moins un jour, il devra renseigner le motif de l'arrêt, la date du dernier jour travaillé, la date de fin prévisionnelle, et enfin les informations concernant la subrogation si subrogation il y a. Les différents motifs d'arrêt pour la DSN sont les suivants :

- 01 – maladie ;
- 02 – maternité / adoption ;
- 03 – paternité / accueil de l'enfant ;
- 04 – congé suite à un accident de trajet ;
- 05 – congé suite à maladie professionnelle ;
- 06 – congé suite à accident de travail ou de servie ;
- 07 – femme enceinte dispensée de travail.

Dans le contexte de ce mémoire, nous ne retiendrons que les motifs 01, 04, 05, et 06 représentant pour 01 un arrêt de travail VP et pour les autres indicateurs un arrêt de travail AT/MP.

Remarque : pour la DSN, le prolongement d'un arrêt de travail n'est pas géré comme un nouvel arrêt de travail mais comme une modification de la date de fin prévisionnelle de l'arrêt en cours. Ainsi chaque arrêt fait bien l'objet d'une seule fiche au sein de la DSN.

B. Signalement de la clôture

Afin de signaler la date de fin de l'arrêt de travail, l'employeur doit renseigner dans les données de la DSN la date de reprise du travail. Cette informations ne fait pas l'objet d'une déclaration évènementielle spécifique et est à renseigner dans la DSN mensuelle classique. Cependant, si la reprise du travail est anticipée par rapport à la date de fin prévisionnelle, l'employeur devra réaliser une déclaration évènementielle.

Parmi les motifs de reprise du travail, la DSN liste :

- la reprise normale ;
- la reprise en temps partiel pour cause de thérapie ;
- et enfin la reprise en temps partiel pour raison personnelle.

Résumé du chapitre sur la déclaration sociale nominative d'un arrêt de travail

La déclaration d'un arrêt de travail au sein de la DSN se réalise en deux temps. Premièrement une déclaration de la survenance d'un arrêt avec la définition du motif de cet arrêt, de la date de début de l'arrêt, et de la date prévisionnelle de fin de l'arrêt. Et deuxièmement une déclaration de la déclaration de la date de retour au travail définitive dans la DSN mensuelle suivant la reprise d'activité.

Les données de la DSN contiennent donc les informations de l'assuré, la cause de l'arrêt de travail (que nous pouvons fractionner en VP, AT/MP, maternité/paternité), la date de début de l'arrêt, et enfin la date de fin de l'arrêt.

III. Amélioration des modèles de tarification

L'utilisation des données de la DSN permettra d'améliorer fortement les deux modélisations précédentes. En effet, si nous rassemblons les éléments présentés ci-dessus, grâce aux données de la DSN, l'assureur détiendra pour chaque sinistre survenu et durant au moins un jour :

- les informations sur l'assuré sinistré ;
- la cause de l'arrêt ;
- la date de début de l'arrêt ;
- et la date de fin d'arrêt.

De plus, à l'aide des informations sur l'assuré, l'assureur sera en mesure de remonter dans sa base de données personnelle à l'ensemble des informations relatives au contrat de cet assuré. Ainsi, grâce à la DSN, un assureur de contrats de prévoyance collective est en mesure d'avoir une **très bonne connaissance tête par tête de son portefeuille**.

Dans les deux parties suivantes nous analyserons comment les améliorations présentées précédemment vont – grâce aux nouvelles données présentées ci-dessus – permettre d'affiner la modélisation par MLG et la modélisation par simulation.

A. Amélioration de la modélisation par modèle linéaire généralisé

Lors de la mise en pratique de la tarification du risque incapacité par MLG, nous nous sommes heurtés à une forte hétérogénéité des données. En effet, nos modèles étaient bons pour prédire les arrêts de la majeure partie de notre population, mais rencontraient des difficultés pour les cas plus atypiques (les queues de distribution). Grâce à l'augmentation de la volumétrie des données obtenues par la DSN, il est probable que ces difficultés soient fortement réduites. Cependant, le principal intérêt de la DSN pour la tarification par MLG ne se trouve pas ici, mais dans la gestion de la troncature des données. En effet, comme observé dans la partie sur les MLG, cette méthode de tarification ne tient pas compte des troncatures et censures.

Le cas des censures ne sera pas amélioré par la DSN car à la date d'observation, même avec l'utilisation de ces nouvelles données, il y aura des sinistres en cours. Cependant les censures ne sont pas un réel problème grâce à l'extrapolation des durées restantes en arrêt via la table de maintien en incapacité règlementaire.

En revanche, les données de la DSN permettront de corriger le biais introduit par la troncature gauche due à la présence de franchises. En effet, la modélisation par MLG ne permet pas de prendre en compte la présence de cette troncature. Afin de contourner ce problème, nous retenons donc la durée de franchise comme variable explicative pour modéliser une durée en indemnisation par franchise. Or ceci introduit un biais (présenté dans la partie sur la tarification par MLG en pratique) car pour de ne pas avoir trop de modalités nous créons des classes de franchises qui vont induire une hétérogénéité non négligeable au sein de la variable à expliquer. L'utilisation des données de la DSN permettra de corriger ce problème car il ne sera plus nécessaire de retenir la franchise dans la modélisation par MLG.

En effet, grâce aux données de la DSN, nous sommes en mesure de connaitre le nombre exact de jours en incapacité. Ainsi, la modélisation deviendra beaucoup plus précise. Plutôt que modéliser un taux d'entrée en indemnisation, nous pourrons modéliser un taux d'entrée en arrêt, plus précis car se basant sur beaucoup plus de données et ne dépendant plus des franchises. De même, nous modéliserons le nombre de jours passés en arrêt plutôt que le nombre de jours indemnisés. Il suffira alors de soustraire la franchise pour obtenir le nombre de jours indemnisés depuis le nombre de jours d'arrêts.

A titre d'exemple nous reprendrons la seconde modélisation par MLG. L'objectif était de modéliser deux variables aléatoires p_a et n_a où p_a représentait la probabilité d'entrée en indemnisation durant l'année et n_a représente le nombre moyen de jours indemnisés au titre des arrêts survenus dans l'année sachant que l'individu est sinistré. Dorénavant, grâce aux données de la DSN, nous chercherons à modéliser :

- p'_a la **probabilité d'entrée en arrêt durant l'année**;
- et n'_a **le nombre de jours passés en arrêt au titre des arrêts survenus dans l'année sachant que l'individu est sinistré**.

Finalement, pour obtenir le nombre de jours indemnisés, nous appliquerons la formule suivante aux résultats du modèle expliquant la variable n'_a :

$$n_a = max(n'_a - f_a\ ; 0)$$

où f_a représente la franchise du bénéficiaire a.

B. Amélioration de la modélisation par simulation

La construction des lois de maintien en non indemnisation de la méthode de tarification par simulation présentée précédemment a dû être avortée à cause d'un fort manque de données. L'utilisation des données de la DSN nous aurait permis d'éviter cette difficulté.

En effet, comme indiqué précédemment, une bonne utilisation des données de la DSN permet de supprimer la troncature gauche des données due aux franchises. De ce fait, nous pourrons remplacer les multiples lois de maintien en non indemnisation par une unique loi de maintien au travail qui sera construite sur beaucoup plus de données que la première puisque nous n'aurons dans ce cas plus besoin de la distinguer par franchise. En effet, la loi de maintien en non indemnisation devait être construite par franchise puisque la franchise était inclue dans la durée modélisée par celle loi, mais ce n'est plus le cas dans la loi de maintien au travail qui sera construite sur l'ensemble des données.

Aussi l'utilisation des données de la DSN nous permettra de construire de manière plus précise la loi de maintien en incapacité et la loi de retour au travail car nous considèrerons les sinistres depuis leur origine (hors sinistres en cours à la date de début d'observation des données).

Nous présenterons dans les deux parties suivantes tout d'abord la modification de la construction des lois de maintien en incapacité et de retour au travail, puis la transformation de la loi de maintien en non indemnisation en loi de maintien au travail, et enfin la refonte de l'algorithme de simulation ainsi que des formules de tarifications.

1. Loi de maintien en incapacité et loi de retour au travail

Grâce aux données de la DSN, nous aurons besoin de moins d'éléments que précédemment. Voici les variables indispensables à l'obtention de la nouvelle loi de maintien en incapacité et de la nouvelle loi de retour au travail.

- A propos de l'individu :
  - la date de naissance ddn ;
- à propos du contrat :
  - les dates de début et de fin de garantie $t1$ et $t2$;
  - la durée maximale d'indemnisation $dmax$;
- à propos du sinistre :
  - les dates de début et de fin d'incapacité $pjAT$ et $djAT$ obtenues avec la DSN ;
- à propos de la gestion des données :

- la date de censure non informative cs ;
- la date de troncature tc.

Nous conservons les spécifications des notations des parties précédentes :

- un indice fera référence à un individu ;
- un exposant fera référence à un sinistre.

Soit V_x le portefeuille regroupant l'ensemble des sinistres qui vont être utilisés pour la construction de la loi de maintien en incapacité pour l'âge x en année. Ce vecteur est de dimension $n \times 1$ où n représente le nombre d'incapacités survenues en base pour un bénéficiaire d'âge x.

$$V_x = \begin{pmatrix} \{e_1; z_1; \delta_1\} \\ \vdots \\ \{e_n; z_n; \delta_n\} \end{pmatrix}$$

Les différents V_x sont supposés indépendants mais ne sont pas identiquement distribués.

Remarque : nous ne pouvons pas supprimer totalement le e_i *à cause de la troncature initiale des données.*

Pour chaque sinistre $k \in [\![1; n]\!]$, les variables contenues dans V_x sont :

- e^k la durée en nombre de jours dans l'état d'incapacité au début des observations :
$$e^k = max(tc - pjAT^k; 0)$$
- z^k la durée en nombre de jours dans l'état d'incapacité à la fin des observations :
$$z^k = min(djAT^k; cs) - pjAT^k$$
- δ^k l'indicateur des sorties non censurées, c'est-à-dire les sorties pour motif de passage en invalidité, de décès, de résiliation du contrat, de retraite, ou de retour au travail :
$$\delta^k = \begin{cases} 1 \; si \; fin \; de \; l'incapacité \\ 0 \; sinon \end{cases}$$

Avec ces nouvelles définitions des e_i et z_i, nous allons reprendre les exemples d'observation des variables pour Kaplan Meier que nous avons présentés lors de la construction de la loi de maintien en incapacité initiale (FIGURE 21).

Nous observons ci-dessous que les données de la DSN permettent de mieux utiliser les informations des divers sinistres. En effet, seules les dates de troncature et de censures nous font perdre de l'information. Tous les sinistres survenus après la date de troncature seront utilisés dès leur survenance, et non plus dès la fin de la franchise comme précédemment. Et nous voyons avec l'exemple $k6$ qu'un sinistre non indemnisé entre maintenant dans la modélisation.

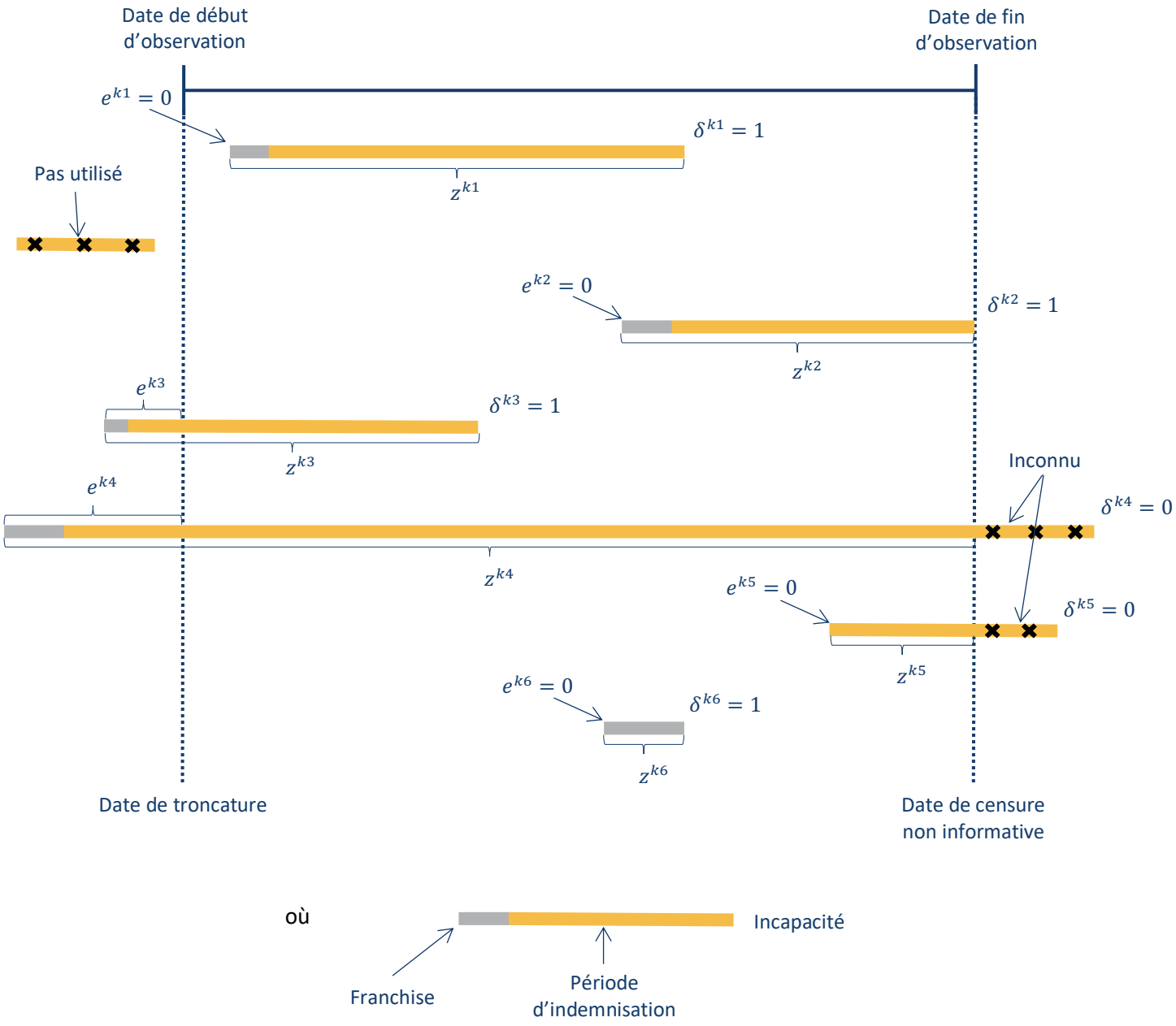


Figure 41 – Exemple variables utilisées pour Kaplan Meier avec la DSN

La suite et fin de la construction de la loi de maintien en incapacité et la construction de la loi de retour au travail sont inchangées.

2. *Loi de maintien au travail*

Comme précédemment nous utiliserons les deux bases de données : la base des prestations et la base des bénéficiaires, mais avec l'utilisation des données de la DSN nous n'aurons plus besoin de construire une loi par franchise, nous aurons une unique loi de maintien au travail.

Comme pour la construction des deux nouvelles lois précédentes, la construction de la loi de maintien au travail demandera moins d'éléments que la loi de maintien en non indemnisation. Il faudra nécessairement connaitre les éléments suivants pour chaque bénéficiaire i :

- à propos de l'individu :
  - la date de naissance ddn_i ;
- à propos du contrat :
  - les dates de début et de fin de garantie $t1_i$ et $t2_i$;
  - la durée maximale d'indemnisation $dmax_i$;
- à propos de chaque sinistre (k est ici l'indice des différents sinistres du bénéficiaire i) :
  - dates de début et de fin d'incapacité $pjAT_i^k$ et $djAT_i^k$ obtenues avec la DSN ;
- à propos de la gestion des données :
  - la date de censure non informative cs ;
  - la date de troncature tc.

La période d'observation de chaque bénéficiaire i est la même que précédemment :

$$[\tau 1_i; \tau 2_i] = \begin{cases} [max(tc; t1_i); min(t2_i; cs)] \; si \; [tc; cs] \cap [t1_i; t2_i] \neq \emptyset \\ \emptyset \; sinon \end{cases}$$

Afin de calculer les périodes au travail, la méthode est la même que pour les périodes de non indemnisation. Il faudra identifier chacune des périodes d'incapacité de chaque bénéficiaire (franchise comprise). Le calcul s'effectuera alors bénéficiaire par bénéficiaire. Et de nouveau nous retenons une plage d'âge $[\alpha; \beta]$ pour la définition de la table.

Notons V_x le portefeuille contenant l'ensemble des données qui seront utilisées pour la construction de la loi de maintien au travail pour l'âge x en jour quelle que soit la franchise :

$$V_x = \begin{pmatrix} \{x_1^{min}; x_1^{max}; \delta_1\} \\ \vdots \\ \{x_n^{min}; x_n^{max}; \delta_n\} \end{pmatrix}$$

où :

- x_i^{min} et x_i^{max} sont l'âge de l'individu respectivement au début et à la fin de la période au travail
- δ_i l'indicatrice des sorties non censurées, c'est-à-dire des sorties de l'état au travail engendrées par un début d'arrêt de travail :

$$\delta_i = \begin{cases} 0 \; si \; sortie \; censurée \\ 1 \; sinon \end{cases}$$

Les vecteurs V_x sont toujours supposés indépendants mais non identiquement distribués.

La suite de la construction de la loi de maintien au travail est identique à la construction de la loi de maintien en non indemnisation. Finalement, nous obtiendrons une unique loi de maintien au travail de la forme suivante :

Age (en jour)	Maintien au travail
α	10 000
$\alpha + 1$	⋮
⋮	⋮

Figure 42 – Exemple table maintien au travail

3. Algorithme de simulation

Les paramètres de la simulation ainsi que la durée de simulation sont inchangés.

Au niveau des variables de l'algorithme, nous dénoterons deux modifications. La première modification concerne $p_{NI,f}(x)$. Cette probabilité est remplacée par $p_T(x) = \frac{L_T(x+1)}{L_T(x)}$ la probabilité qu'un bénéficiaire d'âge x reste au travail le jour suivant (où $L_T(x)$ est le nombre de bénéficiaires maintenus au travail à l'âge x). La seconde modification concerne la matrice de stockage des états du bénéficiaire des multiples simulations. Dans le cas présent elle prendra les valeurs : T, F, I, et S respectivement pour travail, franchise, incapacité et sortie. En effet, grâce à la DSN nous connaissons les sinistres dès leur survenance, nous devons donc différencier les jours d'incapacité indemnisés des jours d'incapacité non indemnisés du fait de la franchise. Dans la suite, F représentera donc l'état de franchise et I représentera les jours d'incapacité indemnisés. De ce fait, l'état I de cet algorithme est le même que celui de l'algorithme précédent, alors que l'état travail T précédent regroupant et les périodes de franchise et les périodes au travail a été fractionné en T et F.

Contrairement aux variables retenues, l'algorithme subira de nombreuses modifications. Ces dernières concernent le corps de l'algorithme qui sera donc entièrement réécrit ci-dessous (l'initialisation restant inchangée ne sera pas rappelée).

Corps de l'algorithme

Le corps de l'algorithme est une boucle qui sera répétée pour chaque t entre 1 et $duree\_maximum$. Les tâches à accomplir vont permettre de déterminer l'état du bénéficiaire en $t+1$ en fonction de son état observé en t.

Si le bénéficiaire est au travail en t

Si en t le bénéficiaire est au travail, nous devons déterminer s'il restera ou non au travail le jour suivant. Pour cela, nous simulons tout d'abord une réalisation u de $U \sim \mathcal{U}([0,1])$. Puis :

- si $u < p_T\big(min(age + t; \beta)\big)$ alors le bénéficiaire restera au travail en $t+1$;
- sinon il sortira de l'état au travail en $t+1$. Dans ce cas nous définissons son âge de passage en incapacité : $age_{incap} = min(age + t + 1; \beta)$ et initialisons son ancienneté dans l'arrêt : $anc = 1$. Si le bénéficiaire possède une franchise f nulle, alors son état sera I, sinon il sera en période de franchise et son état sera donc F.

Si le bénéficiaire est en franchise en t

Nous devons déterminer si le bénéficiaire retourne au travail, reste dans l'état de franchise, entre dans l'état incapacité (l'incapacité indemnisée), ou sort du portefeuille. Pour cela, nous allons simuler une réalisation v de $V \sim \mathcal{U}([0,1])$.

- Si $v < p_{IC}\big(age_{incap}; anc\big)$ alors le bénéficiaire reste en incapacité en $t+1$. Dans ce cas nous incrémentons l'ancienneté en incapacité anc de 1 jour. Après cette incrémentation, si $anc > f$ alors le bénéficiaire sera dans l'état I en $t+1$, sinon il restera dans l'état F en $t+1$.

- Si $p_{IC}(age_{incap}; anc) \leq v < p_{IC}(age_{incap}; anc) + p_{retour}(age_{incap}; anc)$ alors le bénéficiaire sort de sa période d'incapacité et sera donc au travail en $t+1$. Dans ce cas, nous redéfinissons la date de début de travail comme : $debut\_travail = t+1$.
- Et enfin si $p_{IC}(age_{incap}; anc) + p_{retour}(age_{incap}; anc) \leq v$ alors le bénéficiaire sort du portefeuille en $t+1$. Dans ce cas, cette simulation se termine et nous passons à la suivante.

Si le bénéficiaire est en incapacité en t

Si le bénéficiaire est en incapacité en t, nous différencions trois cas après simulation de v une réalisation de $V \sim \mathcal{U}([0,1])$.

- Si $v < p_{IC}(age_{incap}; anc)$ alors le bénéficiaire reste en incapacité en $t+1$. Dans ce cas nous incrémentons l'ancienneté en incapacité anc de 1 jour.
- Si $p_{IC}(age_{incap}; anc) \leq v < p_{IC}(age_{incap}; anc) + p_{retour}(age_{incap}; anc)$ alors le bénéficiaire sort de sa période d'incapacité et sera donc au travail en $t+1$. Dans ce cas, nous redéfinissons la date de début de travail comme : $debut\_travail = t+1$.
- Et enfin si $p_{IC}(age_{incap}; anc) + p_{retour}(age_{incap}; anc) \leq v$ alors le bénéficiaire sort du portefeuille en $t+1$. Dans ce cas, cette simulation se termine et nous passons à la suivante.

4. Formules de tarification

La seule modification concernant les formules de tarification concerne l'obtention de la durée dans l'arrêt. En effet, nous devons rattacher la période de franchise à la période d'incapacité indemnisée. Donc $duree$ devient ici :

$$duree = \sum_{t=1}^{365+f} \left(\prod_{k=t}^{365+f} \left(\mathbb{I}_{\{X_k = I\}} + \mathbb{I}_{\{X_k = F\}} \right) \right)$$

Les autres formules étant inchangées, nous pouvons avec l'ensemble des étapes précédentes obtenir notre prime pure simulée en utilisant les données de la DSN.

Résumé du chapitre sur l'amélioration des modèles de tarification

La suppression de la troncature gauche des données grâce à la DSN permet de fortement améliorer les deux modélisations présentées précédemment.

La DSN permettra tout d'abord de corriger l'absence de gestion des troncatures par la modélisation par GLM. Nous pourrons ainsi modéliser un taux d'entrée en arrêt plutôt qu'un taux d'entrée en indemnisation et un nombre de jours passés en arrêt et non plus en indemnisation. Puis, pour la méthode de tarification par simulation, la DSN nous permettra de construire une loi de maintien au travail plutôt qu'une loi de maintien en non indemnisation. Cette loi de maintien au travail sera beaucoup plus précise car il ne sera plus nécessaire de la fractionner par franchise, elle sera donc construite sur l'ensemble de la base de données. Aussi, la DSN nous permettra de construire les loi de maintien en incapacité et de retour au travail plus précises car basées sur plus de données. Les évolutions des méthodologies permettant la construction ces nouvelles lois sont présentées dans cette partie, ainsi que le nouvel algorithme de simulation et les nouvelles formules de tarification en découlant.

Conclusion

Nous avons adapté les méthodes de tarification individuelle par MLG et par simulation à notre base de données collective. Théoriquement, cette transition ne devrait poser aucun problèmes, cependant en pratique nous avons rencontré plusieurs difficultés.

La méthode de tarification par MLG s'est heurtée à une trop forte hétérogénéité des données. La modélisation du taux d'entrée en indemnisation était correcte, mais les modèles de comptage des nombres de jours indemnisés ne permettaient pas une bonne prédiction des queues de distribution. En effet, les queues de distribution empiriques des nombres de jours indemnisés étaient trop épaisses pour les lois classiques utilisées dans les MLG. La méthode de tarification par simulation s'est heurtée à un important manque d'exposition. Les lois de maintien en incapacité et de retour au travail – construites sur l'ensemble de la population – ont bien été obtenues. Cependant, la loi de maintien en non indemnisation – construite par franchise – n'a pu être complètement déterminée car notre base de données présentait de trop nombreuses franchises trop peu exposées. Nous n'avons obtenu une loi de maintien en non indemnisation viable que pour les deux franchises les plus exposées.

Cependant, les résultats incomplets des deux méthodes précédentes nous ont permis de repérer des problèmes qui dans un futur proche pourront être corrigés grâce aux données de la DSN. En effet, une bonne utilisation de la DSN permettra à un assureur de supprimer la troncature gauche des données. Il sera alors possible pour les MLG de modéliser les taux d'entrée en arrêt de travail et le nombre de jours passés en arrêt, variables basées sur beaucoup plus de données, ce qui réduira l'hétérogénéité de ces dernières. Et pour la méthode par simulation, il sera possible de remplacer les multiples lois de maintien en non indemnisation par une unique loi de maintien au travail, loi construite sur l'ensemble de la base de données, ce qui supprimera le problème de faible exposition.

Nous ne pouvions obtenir les données de la DSN car ce système administratif n'est en place que depuis mi-2015 alors que nos données s'étendaient de 2009 à 2016. Mais il est fortement probable que si nous les avions eues, nous aurions pu pousser à terme ces modélisations.

Bibliographie

Mémoires de l'Institut des Actuaires

- LEBOSSE C. (2009) : *Construction de barèmes de tarification d'arrêt de travail par une méthode de Simulation – Application à un portefeuille de Travailleurs Non-Salariés*
- LEFRANC C. (2013) : *Provisionnement des garanties incapacité et invalidité et problématiques associées*
- PAUGAM M. (2013) : *Mise en place d'indicateurs de pilotage d'une Institution de Prévoyance*
- PELLICIER J. (2010) : *Étude des facteurs discriminants en arrêt de travail pour les travailleurs non-salariés – Application à la tarification*

Ouvrages et documents

- BOCQUAIRE E. et al. (2015) : *Les grands principes de l'actuariat*, L'Argus Éditions
- CHARPENTIER A. (2011) : *Statistique de l'assurance*, HAL
- CHARPENTIER A. (2013) : *Actuariat IARD – modèles linéaires généralisés*, UQAM
- JUILLARD M., PLANCHET F. (2010) : *Tables d'expériences*, Winter
- MASIELLO E. (2016 – 2017) : *Modèles linéaires généralisés*, Support de cours ISFA
- MCCULLAGH P., NELDER J.A. (1991) : *Generalized Linear Models*, Chapman and Hall
- PLANCHET F., THEROND P. (2011) : *Modélisation statistique des phénomènes de durée – Application actuarielles*, Economica
- PLANCHET F. (2017) : *Modèles de durée*, Support de cours ISFA
- ROUVIERE L. (non précisé) : *Régression logistique avec R*, Université Rennes 2

- Auteur inconnu (2017) : Pour une bonne gestion des arrêts de travail dans la DSN, l'Assurance Maladie

Sites internet

- Site du cabinet ACTUARIS : www.actuaris.fr
- Site de l'Assurance Maladie : www.ameli.fr
- Site de l'Argus de l'assurance : www.argusdelassurance.com
- Site de la déclaration sociale nominative : www.dsn-info.fr
- Portail de l'Économie, des Finances, de l'Action et des Comptes publics : www.economie.gouv.fr
- Site du ministère de l'intérieur : www.interieur.gouv.fr
- Site du service public de la diffusion du droit : www.legifrance.gouv.fr
- Site de Mr Frédéric Planchet : www.ressources-actuarielles.net

Liste des figures

Liste des acronymes

- AGIRC – Association Générale des Institutions de Retraite des Cadres
- AIC – *Akaike Information Criterion* (critère d'information d'Akaike)
- ANC – Autorité des Normes Comptables
- ARRCO – Association pour le Régime de Retraite Complémentaire des salariés
- AT – Arrêt de Travail
- AT/MP – Accident du Travail, Maladie Professionnelle
- BCAC – Bureau Commun des Assurances Collectives
- BIC – *Bayes Information Criterion* (critère d'information de Bayes)
- CCN – Convention Collective Nationale
- CPAM – Caisse Primaire d'Assurance Maladie
- DSN – Déclaration Sociale Nominative
- EMV – Estimateur du Maximum de Vraisemblance
- ETP – Équivalent Temps Plein
- IJ – Indemnité(s) Journalière(s)
- MGDC – Maintien des Garanties Décès
- MLG – Modèles Linéaires Généralisés
- MSA – Mutualité Sociale Agricole
- NAF – Nomenclature d'Activité Française
- PASS – Plafond Annuel de la Sécurité Sociale
- PMSS – Plafond Mensuel de la Sécurité Sociale
- SIREN – Système d'Identification du Répertoire des Entreprises
- SIRET – Système d'Identification du Répertoire des Établissements
- SMIC – Salaire Minimum de Croissance
- URSSAF – Unions de Recouvrements des cotisations de Sécurité Sociale et d'Allocations Familiales
- VAP – Valeur Actuelle Probable
- VP – Vie Privée
- ZIM – *Zero-Inflated Model* (modèle avec surreprésentation de zéros)
- ZINB – *Zero Inflated Negative Binomial* (modèle binomial négatif avec surreprésentation des zéros)
- ZIP – *Zero Inflated Poisson* (modèle de Poisson avec surreprésentation des zéros)

Annexes

Annexe 1 – Retraitements des bases de données

Divers retraitements ont été effectués sur les bases de données. Dans cette partie de l'annexe, nous présenterons ces retraitements dans la limite de la confidentialité des données.

Tests de cohérence

Les premiers retraitements effectués ont été des tests de cohérences sur les personnes morales présentes en base (par personne morale nous parlons des entreprises). Comme expliqué dans la partie de présentation générale des données, nous possédons une base collective regroupant à la fois les contrats individualisés sinistrés et les contrats individualisés non sinistrés. Avec ces tests, nous voulions vérifier que pour chaque personne morale nous possédions bien et les non sinistrés et les sinistrés. En effet, nous redoutions la présence des seuls contrats individualisés sinistrés pour certaines personnes morales.

Tout d'abord, nous avons étudié l'Équivalent Temps Plein *(ETP)* afin d'exclure les personnes morales non cohérentes. L'ETP permet d'évaluer l'effectif d'une personne morale de façon homogène par rapport au temps travaillé. Ainsi, nous avons exclu les personnes morales avec un ETP trop faible (inférieur à un certain seuil). Il a fallu fractionner ce traitement entre la population classique présente en base et la population plus spécifique. A la suite de ce premier traitement, 10% des personnes morales sont placées en anomalie.

En parallèle nous avons étudié la sinistralité des différentes personnes morales. En suivant la même logique que le test précédent, si pour une personne morale nous ne possédons que les contrats individualisés sinistrés, le taux de sinistralité sera anormalement élevé. Ainsi, nous avons exclu les personnes morales avec un taux de sinistralité trop élevé (supérieur à un certain seuil). Ce traitement a été fractionné en fonction de la population concernée, et de la taille de l'entreprise (il n'est pas possible de raisonner de la même manière entre une entreprise ayant un unique salarié et une entreprise de 50 salariés). Ceci place environ 10% des personnes morales en anomalie.

Enfin, en croisant les personnes morales en anomalie des deux tests précédents, nous conservons environ 90% des personnes morales initialement présentes en base.

Agrégation des lignes

A ce stade de l'étude, notre base des bénéficiaires possède encore un trop grand nombre de lignes. Ces retraitements ont permis de réduire fortement le volume de cette base.

Nos données initiales comprenaient de nombreux champs dates :

- les dates associées à l'affiliation de la personne physique ;
- les dates associées à l'affiliation de la personne morale ;
- les dates associées à la vie du produit ;
- les dates associées à la vie du contrat ;
- etc.

Nous avons réalisé un travail de regroupement de ces dates en une simple date de début d'appartenance et date de fin d'appartenance pour chacun des contrats individualisé. Ceci couplé à la suppression de quelques variables non utiles pour l'étude nous a permis de réduire drastiquement le volume de la base de données (de quelques millions de lignes à 500000 lignes).

Suppression des anomalies

Après cette première série de tests de cohérence, nous avons effectué quelques tests de détection des anomalies de saisie. Ces tests sont résumés dans les tableaux suivants.

N°	Test	% base des bénéficiaires	% base des prestations
1	Identifiant de base non renseigné	0,00 %	0,00 %
2	Date naissance non renseignée	0,00 %	0,00 %
3	Date survenance du sinistre non renseignée	0,00 %	0,00 %
4	Date fin d'incapacité non renseignée	~ 0 %	0,04 %
5	Age au début de l'appartenance < 16 ans ou > 70 ans	0,55 %	0,05 %
6	Age à l'entrée en incapacité < 16 ans ou > 70 ans	~ 0 %	0,03 %
7	Franchise > 1095 jours	0,00 %	0,00 %
8	Taux indemnisation > 100%	0,00%	0,00%
9	Taux indemnisation nuls	0,52%	0,00%
10	Plusieurs dates de naissance pour une même personne physique	0,00%	0,00%
11	Date au début de l'appartenance < date naissance	0,00%	0,00%

Ces anomalies nous font supprimer au global que 0,5% des personnes physiques présentes en base.

Préparation des bases pour l'étude

Ensuite, nous effectuerons plusieurs corrections afin d'obtenir des données propres pour l'application des modèles et l'utilisation du logiciel R. Ainsi, nous avons :

- remplacé les valeurs vides et les NAs ;
- défini les types des différentes variables ;
- harmonisé les variables ayant des liens les unes avec les autres ;
- regroupé quelques lignes de la base des prestations ;
- supprimé les sinistres relatifs à des périodes d'invalidité tout en conservant les dates de passage en invalidité pour les lignes d'incapacité associées ;
- supprimé les lignes avec un montant indemnisé nul (moins de 50 lignes).

Enfin, nous avons identifié les sinistres relatifs à de l'AT/MP des sinistres relatifs de la VP. Ainsi, nous avons identifié quelle franchise était applicable à quel indemnisation.

Suite à ces longs retraitements, nous avons pu commencer la modélisation.

Annexe 2 – Estimation des paramètres du MLG par la méthode du maximum de vraisemblance

Nous devons estimer la valeur de β à partir des variables explicatives. Nous appliquons pour cela la méthode du maximum de vraisemblance sur l'expression suivante :

$$g_n(\mathbb{E}[Y_i]) = \beta_0 + \beta_1 * X_{1i} + \cdots + \beta_p * X_{pi}$$

En reprenant les notations précédentes, la vraisemblance s'écrit :

$$L(y;\theta;\phi) = \prod_{i=1}^{n} f(y_i|\theta_i;\phi)$$
$$= exp\left(\sum_{i=1}^{n} \frac{\left(y_i * \theta_i - b(\theta_i)\right) * \omega_i}{\phi} + \sum_{i=1}^{n} c(y_i;\phi)\right)$$

Nous rappelons aussi que :

- $\mu_i = b'(\theta_i)$
- $\eta_i = g(\mu_i)$

Donc :

$$\theta_i = (b')^{-1}(\mu_i) = ((b')^{-1} \circ g^{-1})(\eta_i)$$
$$= ((b')^{-1} \circ g^{-1})\left(\beta_0 + \beta_1 * X_{1i} + \cdots + \beta_p * X_{pi}\right)$$

L'expression de θ en fonction de β : $L(y;\theta;\phi) = L(y;\theta(\beta);\phi)$ nous permet de terminer le calcul.

Dans la suite de cette partie, les Estimateurs du Maximum de Vraisemblance *(EMV)* sont notés :

- ${}^t\left(\widehat{\beta_0};\widehat{\beta_1};\dots;\widehat{\beta_p}\right)$ pour ${}^t\left(\beta_0;\beta_1;\dots;\beta_p\right)$
- $\hat{\phi}$ pour ϕ

Ceci nous donnera au final :

$$\widehat{y_i} = g^{-1}\left(\widehat{\beta_0} + \widehat{\beta_1} * X_{1i} + \cdots + \widehat{\beta_p} * X_{pi}\right)$$

La log-vraisemblance est donc :

$$l(y;\theta(\beta);\phi) = \sum_{i=1}^{n} ln\left(f(y_i|\theta_i;\phi)\right) = \sum_{i=1}^{n} \frac{\left(y_i * \theta_i - b(\theta_i)\right) * \omega_i}{\phi} + \sum_{i=1}^{n} c(y_i;\phi)$$

Les équations de vraisemblance sont :

$$\frac{\partial l(y_i;\theta(\beta);\phi)}{\partial \beta_j} = 0 \; ; j \in [\![0;p]\!]$$

où :

$$\frac{\partial ln\left(f(y_i|\theta_i;\phi)\right)}{\partial \beta_j} = \frac{\partial ln\left(f(y_i|\theta_i;\phi)\right)}{\partial \theta_i} * \frac{\partial \theta_i}{\partial \mu_i} * \frac{\partial \mu_i}{\partial \beta_j}$$

avec :

$$\frac{\partial ln\left(f(y_i|\theta_i;\phi)\right)}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{\frac{\phi}{\omega_i}} = \frac{y_i - \mu_i}{\frac{\phi}{\omega_i}}$$
$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b''(\theta_i)} \; car \; \mu_i = b'(\theta_i)$$
$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} * x_{ji}$$

Finalement, nous obtenons :

$$\frac{\partial ln\big(f(y_i|\theta_i;\phi)\big)}{\partial \beta_j} = \frac{(y_i-\mu_i)*\frac{\partial \mu_i}{\partial \eta_i}*x_{ji}}{\frac{\phi}{\omega_i}*b''(\theta_i)}$$

Et enfin puisque $\frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{g\prime(\mu_i)}$ nous obtenons :

$$\sum_{i=1}^{n} \frac{\omega_i*(y_i-\mu_i)*x_{ji}}{b''(\theta_i)*g'(\mu_i)} = 0\ ;\ j \in [\![0;p]\!]$$

Les équations qui viennent d'être obtenues ne possèdent pas de solution explicite. Il faudra utiliser une méthode de résolution numérique telle que la méthode de Newton-Raphson ou la méthode de Fisher-Scoring par exemple.

Le paramètre de dispersion ϕ est estimé par l'estimateur de Pearson :

$$\hat{\phi} = \frac{{}^t(y-\hat{\mu})*\mathrm{I}_n(\hat{\mu})*(y-\hat{\mu})}{n-p-1} = \frac{\mathcal{X}^2_{n-p-1;1-\alpha}}{n-p-1}$$

où $\mathcal{X}^2_{n-p-1;1-\alpha}$ est le quantile d'ordre $1-\alpha$ d'une loi du Chi 2 à $n-p-1$ degrés de liberté.

Annexe 3 – Table d'espérance résiduelle en incapacité utilisée pour les MLG

Table obtenue avec le logiciel addactis® PM Expert® depuis la table réglementaire de maintien en incapacité du BCAC

En ordonnée se trouve l'âge de l'assuré à la survenance de l'arrêt

En abscisse se trouve la durée de l'arrêt en mois

Les valeurs présentes dans la table correspondent à la durée restante de l'arrêt en mois

Age \ mois	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
16	1,61	3,42	4,26	5,23	5,96	6,82	8,18	9,81	10,34	10,82	11,65	12,32	12,27	12,9	13,16
17	1,61	3,42	4,26	5,23	5,96	6,82	8,18	9,81	10,34	10,82	11,65	12,32	12,27	12,9	13,16
18	1,61	3,42	4,26	5,23	5,96	6,82	8,18	9,81	10,34	10,82	11,65	12,32	12,27	12,9	13,16
19	1,61	3,42	4,26	5,23	5,96	6,82	8,18	9,81	10,34	10,82	11,65	12,32	12,27	12,9	13,16
20	1,64	3,38	4,07	4,93	5,52	6,33	7,52	9,28	9,85	10,13	10,86	11,78	11,94	12,78	13,01
21	1,76	3,61	4,28	5,12	5,74	6,57	7,86	9,83	10,4	10,7	11,25	11,99	12,23	12,8	12,8
22	1,87	3,82	4,49	5,27	5,92	6,75	8,09	9,9	10,4	10,83	11,28	11,84	11,99	12,42	12,38
23	1,99	4,08	4,83	5,63	6,34	7,26	8,64	10,5	11,19	11,59	11,97	12,38	12,38	12,68	12,5
24	2,09	4,33	5,14	6	6,78	7,8	9,21	10,93	11,7	12,05	12,37	12,73	12,76	12,99	12,72
25	2,17	4,5	5,39	6,28	7,1	8,11	9,52	11,21	12,08	12,47	12,75	12,86	12,8	13	12,81
26	2,2	4,52	5,48	6,33	7,15	8,21	9,49	10,98	11,87	12,29	12,57	12,67	12,57	12,74	12,64
27	2,21	4,56	5,67	6,62	7,42	8,46	9,66	11,14	12,14	12,49	12,76	12,77	12,76	12,88	12,84
28	2,23	4,53	5,69	6,68	7,47	8,46	9,56	10,92	11,72	12,01	12,3	12,3	12,33	12,49	12,41
29	2,24	4,52	5,73	6,69	7,49	8,46	9,51	10,72	11,38	11,64	12,01	12,06	12,04	12,16	11,98
30	2,35	4,68	5,99	6,99	7,77	8,79	9,85	10,82	11,32	11,57	11,92	11,99	12	12,1	11,84
31	2,46	4,87	6,33	7,47	8,26	9,24	10,29	11,22	11,62	11,78	12,02	12,15	12,15	12,33	12,12
32	2,53	4,97	6,43	7,61	8,43	9,47	10,46	11,19	11,59	11,76	11,9	12,07	12,03	12,14	12,05
33	2,59	4,97	6,44	7,6	8,46	9,41	10,47	11,11	11,57	11,77	11,86	12	11,93	12,06	12,11
34	2,7	5,06	6,65	7,89	8,75	9,73	10,73	11,42	11,93	12,09	12,13	12,2	12,16	12,24	12,26
35	2,86	5,36	7,04	8,35	9,29	10,2	11,14	11,77	12,28	12,36	12,39	12,48	12,55	12,5	12,59
36	2,96	5,53	7,22	8,54	9,57	10,46	11,33	11,82	12,3	12,36	12,41	12,5	12,53	12,42	12,43
37	3,17	5,82	7,58	8,88	9,96	10,78	11,64	12,11	12,55	12,54	12,57	12,68	12,67	12,64	12,56
38	3,43	6,22	8,02	9,35	10,42	11,18	12,02	12,52	12,95	12,82	12,86	12,91	12,9	12,85	12,66
39	3,64	6,53	8,35	9,72	10,82	11,55	12,39	12,89	13,25	13,16	13,21	13,25	13,2	13,12	12,96
40	3,76	6,55	8,37	9,77	10,85	11,69	12,52	13,01	13,35	13,31	13,3	13,29	13,11	13,09	12,91
41	4	6,82	8,61	10,02	11,11	12,03	12,89	13,42	13,68	13,73	13,71	13,61	13,49	13,47	13,21
42	4,17	7	8,88	10,26	11,27	12,19	13,01	13,51	13,67	13,76	13,77	13,59	13,5	13,45	13,14
43	4,36	7,2	9,13	10,48	11,48	12,47	13,15	13,54	13,7	13,81	13,77	13,57	13,44	13,39	13,03
44	4,65	7,54	9,42	10,73	11,7	12,66	13,25	13,61	13,68	13,76	13,66	13,38	13,26	13,18	12,81
45	4,87	7,99	9,9	11,12	12,05	12,93	13,48	13,7	13,75	13,78	13,71	13,41	13,23	13,08	12,68
46	5,07	8,22	10,15	11,36	12,24	12,93	13,43	13,68	13,74	13,7	13,57	13,28	13,05	12,77	12,42
47	5,29	8,52	10,41	11,6	12,43	13,07	13,45	13,69	13,74	13,69	13,52	13,26	12,96	12,63	12,3
48	5,51	8,96	10,9	12,04	12,81	13,39	13,69	13,92	13,89	13,82	13,65	13,35	13,01	12,66	12,33
49	5,72	9,29	11,25	12,26	12,95	13,48	13,75	13,88	13,81	13,74	13,52	13,2	12,82	12,41	12,03
50	5,83	9,53	11,51	12,47	13,15	13,59	13,84	13,95	13,79	13,68	13,41	13,05	12,7	12,26	11,87
51	5,73	9,3	11,62	12,67	13,33	13,81	14,01	14,03	13,86	13,73	13,49	13,11	12,76	12,37	11,99
52	5,66	9,11	11,96	13,13	13,82	14,33	14,6	14,6	14,39	14,22	13,98	13,62	13,29	12,96	12,53
53	5,5	8,72	12,22	13,51	14,27	14,8	15,13	15,12	14,9	14,72	14,45	14,12	13,79	13,49	13,06
54	5,5	8,72	12,48	13,82	14,55	15,06	15,37	15,28	15	14,78	14,51	14,15	13,83	13,57	13,12
55	5,44	8,57	12,78	14,22	14,97	15,49	15,8	15,65	15,32	15,05	14,77	14,4	14,11	13,88	13,42
56	5,38	8,42	13,11	14,65	15,41	15,93	16,25	16,04	15,66	15,34	15,05	14,68	14,4	14,21	13,73
57	5,32	8,27	13,46	15,12	15,89	16,41	16,74	16,45	16,01	15,64	15,35	14,96	14,69	14,56	14,05
58	5,26	8,12	13,84	15,64	16,42	16,93	17,27	16,88	16,38	15,96	15,64	15,26	15	14,92	14,39
59	5,19	7,97	14,26	16,21	16,99	17,49	17,83	17,35	16,77	16,29	15,96	15,56	15,32	15,29	14,73

15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
13,04	12,65	12,61	13,03	12,55	12,06	11,44	10,94	10,32	9,81	9,3	8,3	7,51	6,7	6,07	5,07	4,28	3,4	2,44	1,58	0,64
13,04	12,65	12,61	13,03	12,55	12,06	11,44	10,94	10,32	9,81	9,3	8,3	7,51	6,7	6,07	5,07	4,28	3,4	2,44	1,58	0,64
13,04	12,65	12,61	13,03	12,55	12,06	11,44	10,94	10,32	9,81	9,3	8,3	7,51	6,7	6,07	5,07	4,28	3,4	2,44	1,58	0,64
13,04	12,65	12,61	13,03	12,55	12,06	11,44	10,94	10,32	9,81	9,3	8,3	7,51	6,7	6,07	5,07	4,28	3,4	2,44	1,58	0,64
12,94	12,7	12,75	13,04	12,59	12	11,13	10,51	9,76	9,11	8,69	7,92	7,35	6,46	5,94	4,94	4,24	3,31	2,41	1,52	0,65
12,63	12,41	12,42	12,76	12,08	11,4	10,81	10,22	9,72	9,22	8,72	8,11	7,6	6,79	6,05	5,05	4,24	3,41	2,45	1,55	0,64
12,42	12,16	12,02	12,11	11,83	11,29	10,74	10,19	9,82	9,28	8,93	8,3	7,47	6,63	6,01	5,07	4,25	3,35	2,47	1,58	0,68
12,37	12,3	12,16	11,92	11,57	11,16	10,67	9,96	9,39	8,8	8,51	8,01	7,22	6,48	5,91	5,15	4,3	3,39	2,49	1,59	0,69
12,59	12,47	12,15	11,79	11,39	10,94	10,42	9,72	9,12	8,52	8,08	7,71	6,93	6,36	5,78	5,13	4,22	3,37	2,46	1,59	0,7
12,65	12,39	12,1	11,79	11,44	10,93	10,43	9,75	9,23	8,59	8,05	7,67	6,87	6,35	5,72	5,03	4,23	3,37	2,45	1,61	0,69
12,55	12,27	11,85	11,5	11,17	10,81	10,29	9,71	9,13	8,43	7,92	7,52	6,81	6,24	5,65	5	4,25	3,35	2,43	1,59	0,69
12,69	12,39	11,92	11,59	11,24	10,8	10,2	9,7	9,1	8,43	7,76	7,35	6,57	6,13	5,56	4,98	4,25	3,36	2,44	1,57	0,66
12,39	12,06	11,61	11,42	11,06	10,61	10,06	9,55	9	8,34	7,61	7,11	6,37	6,05	5,49	4,87	4,19	3,33	2,44	1,56	0,64
11,95	11,63	11,23	11,05	10,76	10,29	9,77	9,31	8,79	8,23	7,7	7,08	6,44	5,97	5,51	4,72	4,2	3,39	2,45	1,53	0,6
11,73	11,52	11,14	10,83	10,58	10,18	9,76	9,34	8,79	8,28	7,77	7,11	6,49	5,98	5,47	4,79	4,12	3,35	2,46	1,55	0,61
11,98	11,67	11,47	11,03	10,75	10,3	9,9	9,38	8,95	8,48	7,8	7,11	6,55	6,02	5,49	4,73	4,08	3,33	2,45	1,55	0,61
11,89	11,64	11,47	11	10,77	10,3	9,99	9,34	8,95	8,46	7,81	7,14	6,6	6,02	5,5	4,73	4,06	3,33	2,42	1,55	0,6
11,94	11,65	11,38	10,92	10,66	10,35	9,95	9,36	9	8,5	7,88	7,17	6,62	5,96	5,42	4,6	3,93	3,29	2,4	1,5	0,58
12,06	11,87	11,69	11,29	11	10,7	10,31	9,75	9,41	8,78	8,03	7,22	6,59	6,01	5,45	4,67	4	3,3	2,41	1,5	0,58
12,41	12,2	11,98	11,56	11,07	10,72	10,16	9,62	9,24	8,71	8,04	7,28	6,6	5,97	5,45	4,64	4,04	3,28	2,39	1,49	0,58
12,18	12	11,8	11,47	11,05	10,69	10,07	9,56	9,11	8,63	8,04	7,28	6,52	5,93	5,36	4,68	4,09	3,3	2,4	1,47	0,57
12,21	11,96	11,81	11,51	11,06	10,74	10,1	9,65	9,17	8,74	8,16	7,42	6,63	6,05	5,39	4,71	4,09	3,27	2,38	1,49	0,55
12,36	12,1	11,94	11,64	11,13	10,64	10,07	9,62	9,09	8,76	8,15	7,39	6,65	5,99	5,4	4,75	4,13	3,27	2,4	1,51	0,56
12,59	12,27	12,04	11,67	11,16	10,68	10,05	9,54	9,01	8,64	8,1	7,35	6,67	5,95	5,35	4,76	4,06	3,22	2,37	1,52	0,58
12,51	12,26	11,98	11,64	11,15	10,65	10,11	9,65	9,11	8,6	8,01	7,26	6,62	5,9	5,3	4,74	4,04	3,22	2,36	1,5	0,55
12,76	12,51	12,19	11,71	11,29	10,78	10,24	9,74	9,14	8,6	7,95	7,3	6,67	5,97	5,36	4,77	4	3,17	2,33	1,48	0,56
12,72	12,49	12,08	11,52	11,13	10,65	10,07	9,54	9	8,4	7,81	7,23	6,65	5,96	5,34	4,75	3,99	3,16	2,35	1,49	0,58
12,63	12,34	11,86	11,32	10,98	10,52	9,96	9,54	9,07	8,47	7,93	7,42	6,79	6,08	5,43	4,8	3,99	3,16	2,33	1,48	0,58
12,51	12,2	11,71	11,15	10,82	10,27	9,73	9,35	8,92	8,35	7,82	7,34	6,72	6,05	5,33	4,69	3,93	3,15	2,37	1,49	0,57
12,41	12,05	11,65	11,12	10,7	10,23	9,71	9,31	8,94	8,34	7,82	7,36	6,71	6,09	5,41	4,69	3,89	3,13	2,37	1,5	0,58
12,17	11,82	11,37	10,91	10,38	9,9	9,46	9,06	8,7	8,19	7,76	7,24	6,74	6,07	5,44	4,66	3,92	3,14	2,39	1,51	0,57
12,03	11,7	11,34	10,92	10,39	9,94	9,58	9,2	8,77	8,26	7,8	7,24	6,67	6,05	5,42	4,63	3,9	3,14	2,37	1,49	0,56
11,98	11,71	11,43	10,98	10,43	9,96	9,55	9,11	8,64	8,12	7,68	7,05	6,58	5,95	5,36	4,62	3,87	3,09	2,38	1,51	0,56
11,71	11,42	11,15	10,74	10,24	9,82	9,42	9,01	8,53	8,01	7,57	6,92	6,52	5,9	5,37	4,6	3,87	3,08	2,36	1,5	0,56
11,6	11,32	11,04	10,69	10,25	9,74	9,35	8,97	8,46	8,06	7,65	7,04	6,64	6,02	5,45	4,68	3,94	3,13	2,37	1,5	0,55
11,69	11,48	11,24	10,89	10,49	9,95	9,53	9,12	8,65	8,19	7,72	7,11	6,61	6,07	5,52	4,74	3,97	3,17	2,38	1,51	0,57
12,24	11,97	11,69	11,32	10,89	10,29	9,85	9,39	8,94	8,43	7,97	7,32	6,81	6,2	5,64	4,83	4,06	3,25	2,43	1,56	0,62
12,79	12,49	12,21	11,84	11,44	10,84	10,4	9,92	9,44	8,89	8,36	7,7	7,1	6,49	5,86	5,01	4,23	3,39	2,52	1,62	0,68
12,85	12,59	12,31	11,95	11,59	10,91	10,45	9,95	9,48	8,95	8,41	7,76	7,12	6,53	5,91	5,05	4,26	3,41	2,52	1,62	0,68
13,16	12,89	12,61	12,26	11,91	11,18	10,7	10,17	9,72	9,16	8,59	7,93	7,24	6,64	6	5,13	4,33	3,47	2,54	1,65	0,7
13,47	13,21	12,92	12,56	12,25	11,45	10,95	10,4	9,94	9,37	8,77	8,11	7,36	6,76	6,1	5,22	4,39	3,53	2,57	1,67	0,72
13,8	13,52	13,23	12,88	12,59	11,72	11,2	10,62	10,17	9,58	8,94	8,27	7,46	6,87	6,19	5,29	4,46	3,58	2,59	1,69	0,73
14,13	13,86	13,55	13,2	12,95	12	11,46	10,84	10,39	9,78	9,1	8,43	7,57	6,98	6,28	5,36	4,51	3,62	2,61	1,7	0,75
14,48	14,21	13,89	13,54	13,31	12,28	11,72	11,08	10,62	9,97	9,28	8,58	7,66	7,08	6,36	5,42	4,57	3,67	2,63	1,72	0,76

Annexe 4 – Lissage de Whittaker-Henderson en dimension deux (par PLANCHET F.)

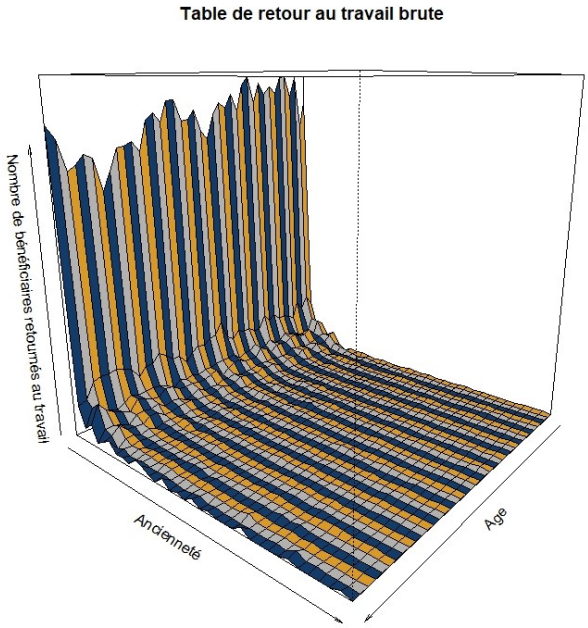
```
W_H=function(Tbrut,Poids,Ordre_vertical,Ordre_horizontal,alpha,beta)
{
  #Déclaration de la taille des variables#

  q=ncol(Tbrut)
  p=nrow(Tbrut)

  U=matrix(0,q*p)
  v=matrix(0,Ordre_vertical+1)
  h=matrix(0,Ordre_horizontal+1)
  Kv=matrix(0,(p-Ordre_vertical)*q,q*p)
  Kh=matrix(0,p*(q-Ordre_horizontal),q*p)
  M=matrix(0,q*p,q*p)
  W=matrix(0,q*p,q*p)
  qlisse=matrix(0,p,q)
  Tlisse=matrix(0,p,q)
  Ecart=matrix(0,p,q)

  #Transformation de la matrice de taux bruts en vecteur et construction de la matrice des poids#

  for (j in 1:q){
    for (i in 1:p){
      U[(i-1)*q+j,1]=Tbrut[i,j]
      W[q*(i-1)+j,q*(i-1)+j]=Poids[i,j]
    }
  }

  #Construction matrice kv#
  for (k in 0:Ordre_vertical)
  {
    v[(k+1),1]=(-1)^(Ordre_vertical-k)*factorial(Ordre_vertical)/(factorial(k)*factorial(Ordre_vertical-k))
  }

  for (j in 1:q)
  {
    for (z in 1:(p-Ordre_vertical))
    {
      for (i in 1:(Ordre_vertical+1))
      {
        Kv[z+(j-1)*(p-Ordre_vertical),j+(q)*(i-1)+(z-1)*(q)]=v[i,1]
      }
    }
  }

  #Construction matrice Kh#

  for (k in 0:Ordre_horizontal)
  {
    h[(k+1),1]=(-1)^(Ordre_horizontal-k)*factorial(Ordre_horizontal)/(factorial(Ordre_horizontal-k)*factorial(k))
  }

  for (i in 1:p)
  {
    for (j in 1:(q-Ordre_horizontal))
    {
      for (z in 1:(Ordre_horizontal+1))
      {
        Kh[j+(i-1)*(q-Ordre_horizontal),z+(j-1)+(i-1)*(q)]=h[z,1]
      }
    }
  }
  #Calcul des taux lissés#

  M=W+alpha*t(Kv)%*%Kv+beta*t(Kh)%*%Kh
  qlisse=solve(M)%*%W%*%U

  for (j in 1:q)
  {
    for (i in 1:p)
    {
      Tlisse[i,j]=qlisse[(i-1)*q+j,1]
      Ecart[i,j]=ifelse(is.finite(Tlisse[i,j]/Tbrut[i,j]),Tlisse[i,j]/Tbrut[i,j],0)
    }
  }
  #Stockage et exportation des résultats
  Tlisse<<-Tlisse
  Ecart<<-Ecart

}
```

D'après le site www.ressources-actuarielles.net.

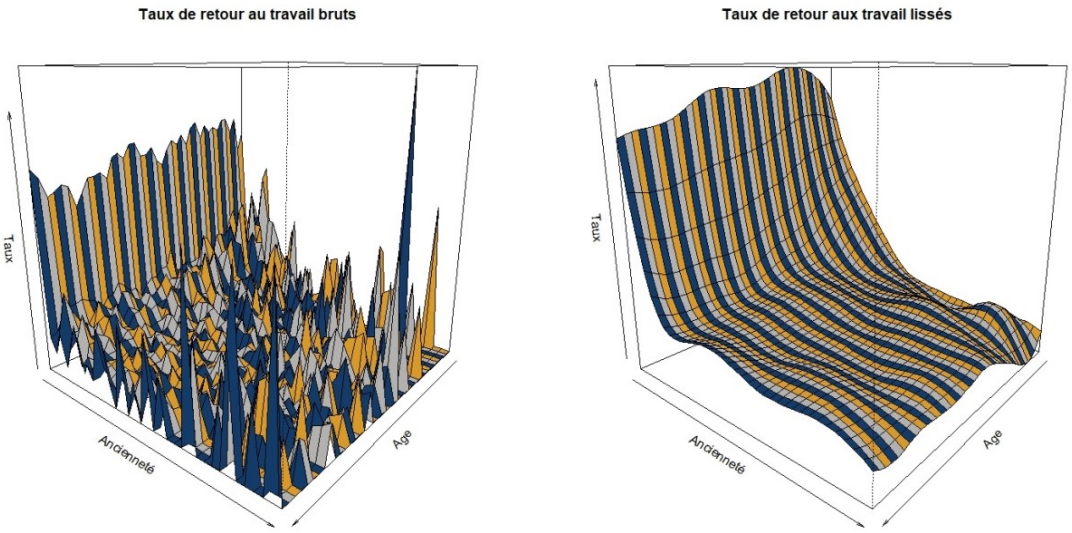
Remarque : la variable « Ecart » n'était pas présente dans la fonction initiale, nous l'avons rajoutée.

Annexe 5 – Lissage de la loi de retour au travail

La table de retour au travail brute obtenue par l'estimateur de Kaplan Meier est la suivante :



Afin de lisser cette table erratique, nous allons appliquer la méthode de lissage de Whittaker-Henderson en dimension 2 aux taux de retour. Nous passons ainsi des taux de retour bruts à gauche aux taux de retour lissés à droites dans les graphiques ci-dessous.



Après plusieurs tests, les paramètres retenus pour le lissage sont les suivants :

- le poids donné à la régularité verticale $\alpha = 2$;
- le poids donné à la régularité horizontale $\beta = 2$;
- l'ordre vertical $OV = 3$;
- l'ordre horizontal $OH = 3$;
- la matrice de poids de chacune des observations W est l'exposition de chacun des âges.

Nous obtenons ainsi la table de retour au travail lissée présentée dans le corps du mémoire.