Mémoire présenté le :

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

Par : Maylis HÉLIOT

Titre : Tarification du risque invalidité en prévoyance collective

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

Laure OLIÉ

Membres présents du jury de l'ISFA

Pierre RIBEREAU

Entreprise :

Nom : Groupama Gan Vie

Signature :

Directeur de mémoire en entreprise :

Nom : Mathilde LOMBARD

Signature :

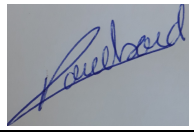
Invité :

Nom :

Signature :

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

Signature du responsable entreprise



Signature du candidat

Résumé

Mots-clés : tarification, invalidité, tables d'expérience, Kaplan-Meier, ajustement par référence externe, régression logistique

Actuellement, les assureurs tarifent généralement le risque invalidité à partir des barèmes fournis par le *Bureau Commun des Assurances Collectives* (BCAC). Ces derniers ne reflètent pas forcément le risque propre à une compagnie d'assurance. C'est pourquoi nous avons cherché à proposer une méthode de tarification du risque invalidité à partir de nos données internes.

En amont des différentes études, une réflexion sur la construction du tarif a été nécessaire. L'incidence et le maintien sont les grandes composantes du tarif. En revanche, il est aussi important de tenir compte des catégories d'invalidité, qui influent sur le montant des prestations. Notre tarif repose sur l'hypothèse d'une répartition constante des assurés dans les trois catégories.

Dans un premier temps, cette étude a exigé un long travail sur les données. En effet, la reconstitution de l'historique des sinistres a demandé le recours à plusieurs sources et les données ont dû être fiabilisées.

Par la suite, la modélisation du maintien en invalidité s'est révélée délicate. L'effectif disponible ne permettait pas d'obtenir des résultats robustes à tous les âges par l'estimateur de Kaplan-Meier. L'ajustement par référence externe a finalement permis d'obtenir des résultats fiables en utilisant la structure de la table du BCAC. Grâce au positionnement par rapport à cette table de référence, la spécificité de notre risque a été conservée. Les différences observées entre nos fonctions de survie et celles du BCAC valident l'utilité des tables d'expérience.

En revanche, la segmentation du maintien a été très compliquée. En effet, le volume de données ne permettait pas d'utiliser les méthodes traditionnelles. Si la quantification des effets n'a pas été un succès, cette étude a révélé l'influence de plusieurs variables (âge, secteur d'activité, département). Le maintien étant majoritairement très élevé et étant difficile à expliquer à partir des variables explicatives, il n'a pas été segmenté pour construire le tarif.

L'incidence de l'*invalidité indirecte* (faisant suite à l'incapacité) a pu être étudiée grâce à un précédent mémoire modélisant l'incidence en incapacité à partir des données de la *Déclaration Sociale Nominative* (DSN). Il restait donc à modéliser le taux de passage en invalidité, estimé à l'aide d'une régression logistique. Comme l'effectif sous risque n'est pas connu pour l'*invalidité directe*, elle a été évaluée par une fraction de l'incidence en *invalidité indirecte*.

Enfin, le tarificateur a été créé sur Excel et nous avons commencé à comparer les barèmes construits aux barèmes actuels.

Abstract

Key words : pricing, disability, experience table, Kaplan-Meier, referential adjustment, logistic regression

Nowadays, insurers usually price stoppage thanks to the pricing schedule delivered by the *Bureau Commun des Assurances Collectives* (BCAC). The pricing model does not always reflect the specific risk of an insurance company. That is the reason why we wished to put forward a pricing method, based on our internal data.

Upstream of the different studies, thinking on the way to build a new pricing model has been needed. Incidence and maintenance are the major components of the price. Nevertheless, it is also important to take account of disability categories, which influence benefit amounts. Our pricing model proceeds on the assumption of a constant distribution in the three disability categories.

Firstly, this study required an important work on data. Indeed, we had to use different sources in order to piece together the history of claims and we needed to increase reliability of data.

Then, modeling the maintenance of disability has been tricky. The small amount of data did not provide robust results for all the ages with the Kaplan-Meier estimator. Referential adjustment finally gave us reliable results, by using the structure of the BCAC table. By positioning our results in relation to this reference table, the specificity of our risk has been preserved. Differences between our survival functions and those of the BCAC show that experience tables are useful.

However, the segmentation of maintenance has been really difficult. Indeed, the amount of data did not enable us to use traditional methods. If quantifying impacts has not been a success, this study has revealed the impact of several variables (age, line of business, department). As maintenance is mostly high and difficult to explain with variables, it has not been segmented in order to build the price.

Incidence in *indirect disability* (following incapacity) has been studied thanks to another memoir of the team, which had modeled incapacity incidence basing on the *Déclaration Sociale Nominative* (DSN). It remained to model the transition rate to disability, estimated with a logistic regression. As we did not know how many people were insured, the *direct disability* has been measured as a fraction of the incidence in *indirect disability*.

Finally, the pricing tool has been created on Excel and we have started comparing the built prices to the current ones.

Remerciements

Je tiens à remercier vivement mes deux premières tutrices pour leur aide et leur sincère bienveillance : Marie DALONGEVILLE et Justine GRUEL. C'est elles m'ont permis de commencer à poser le sujet.

Je souhaite aussi adresser ma gratitude à Amandine NSEKE, qui m'a encadrée le plus longtemps, pour le temps qu'elle m'a dédié, la clarté de ses explications et ses conseils avisés.

J'ai également à coeur de remercier Mathilde LOMBARD pour m'avoir aidée à finir cette étude. Un grand merci pour ses précieux conseils et ses explications qui m'ont permis de mieux appréhender le risque invalidité.

D'autre part, mes remerciements vont aussi à l'ensemble de l'équipe de la Direction Technique, en particulier Vincent LAUDOU, Adel KOBTAN et Manuel PARMENTIER, pour leurs idées qui m'ont permis d'avancer dans cette étude ainsi que pour leur bienveillance. Je souhaite aussi remercier Julie RAMU pour son aide, son soutien et sa convivialité.

Je tiens enfin à adresser ma gratitude à mon tuteur pédagogique Aurélien COULOUMY, pour l'aide qu'il m'a apportée lors de difficultés mais aussi pour ses conseils.

Sommaire

Introduction

En prévoyance collective, les assureurs tarifient très souvent leur risque invalidité à partir des barèmes fournis par le *Bureau Commun des Assurances Collectives* (BCAC). Ces barèmes ont été construits à partir des données de sept assureurs. Or les assurés d'un organisme n'ont pas forcément les mêmes caractéristiques et le risque porté n'est donc pas le même. Cela engendre donc un risque de sous-tarification, si l'entreprise fait face à de mauvais risques, ou de sur-tarification dans le cas inverse. De surcroît, les barèmes, relativement anciens, pourraient ne pas refléter le risque actuel. Le taux d'absentéisme a d'ailleurs connu une forte hausse avant 2019. De même, les taux d'actualisation ont baissé, ce qui influence le tarif.

Les barèmes du BCAC reposent sur un taux d'entrée dans le risque (incidence) et sur des lois de maintien. Certains assureurs ont fait le choix de construire leurs propres tables de maintien à partir de leurs données internes afin de mieux gérer leur risque. Ces tables doivent être certifiées pour pouvoir être utilisées pour le provisionnement. Quant à l'incidence, elle était jusqu'ici difficile à modéliser, comme l'effectif assuré n'était pas connu en prévoyance collective. Mais l'arrivée de la *Déclaration Sociale Nominative* (DSN) a permis de l'estimer. L'incidence en incapacité a fait l'objet d'un précédent mémoire de l'équipe, dont les résultats seront utilisés dans ce mémoire.

L'objectif de ce mémoire est de proposer une méthode de tarification du risque invalidité, à partir de nos données internes. Pour cela, nous nous questionnerons sur les étapes permettant la construction du tarif ainsi que sur leur mise en oeuvre.

Tout d'abord, il convient de présenter le risque invalidité et le régime français. Les principes de tarification ainsi que les garanties à tarifer y seront aussi exposés.
Une deuxième partie traitera du choix du périmètre d'étude, de la construction de la base de données à partir de diverses sources et de la fiabilisation des bases. Nous y présenterons également les principales statistiques des bases construites.
Puis la modélisation du maintien en fonction de l'âge et de l'ancienneté, à l'aide de l'estimateur de Kaplan-Meier et d'un ajustement par référence externe, sera expliquée. Nous décrirons aussi la segmentation du maintien, qui ne pouvait pas être effectuée à l'aide des méthodes usuelles en raison du volume de données.
Ensuite, c'est l'incidence, deuxième grande composante du tarif, qui sera modélisée à l'aide d'une régression logistique et de précédents résultats sur l'incidence en incapacité.
Enfin, nous expliquerons la démarche de construction du tarif et appliquerons notre méthode.

1 Cadre du mémoire

Afin de tarifer l'invalidité, il est important de bien comprendre le risque, le régime de prévoyance et les grands principes de tarification.

1.1 Le risque invalidité

Le risque invalidité est un risque spécifique en arrêt de travail, qui est divisé en trois catégories.

1.1.1 Définition

Il existe deux types d'arrêts de travail : l'incapacité et l'invalidité.

L'incapacité temporaire est une incapacité physique d'exercer une activité professionnelle, qui dure au maximum 3 ans. Elle doit être constatée par un médecin. Les principales causes de sortie sont la reprise du travail, le passage en invalidité et le décès.

L'invalidité fait généralement suite à l'incapacité. Selon l'article L341-1 du code de la Sécurité Sociale,
"L'assuré a droit à une pension d'invalidité lorsqu'il présente une invalidité réduisant dans des proportions déterminées sa capacité de travail ou de gain, c'est-à-dire le mettant **hors d'état de se procurer un salaire supérieur à une fraction de la rémunération** soumise à cotisations et contributions sociales qu'il percevait dans la profession qu'il exerçait avant la date de l'interruption de travail suivie d'invalidité ou la date de la constatation médicale de l'invalidité."

Si la capacité de travail ou de gain de l'assuré est réduite de 2/3, il entre en invalidité. L'invalidité est le plus souvent causée par une maladie ou un accident d'origine non professionnelle.

Quant à l'incapacité permanente, elle a pour origine un accident du travail ou une maladie professionnelle. Un taux d'incapacité permanente est fixé et ouvre droit à une indemnité forfaitaire s'il est inférieur à 10% ou à une rente d'incapacité le cas échéant.

Les principales causes de sortie de l'invalidité sont la retraite et le décès. La reprise d'activité est très rare.

1.1.2 Les trois catégories d'invalidité

Il existe trois catégories d'invalidité. C'est le médecin-conseil de la caisse primaire d'assurance maladie (CPAM) qui détermine la catégorie d'invalidité de l'assuré. Cette catégorie permet de

déterminer le niveau de prestation de la Sécurité Sociale.

Selon l'article L341-4 du code de la Sécurité Sociale, les invalides de première catégorie sont capables d'exercer une activité rémunérée.
La deuxième catégorie correspond aux individus qui sont incapables d'exercer une profession.
Quant à la troisième catégorie, elle regroupe les invalides incapables d'exercer une profession et qui doivent avoir recours à l'assistance d'une tierce personne pour effectuer les actes ordinaires de la vie.

Néanmoins, les invalides de $2^{ème}$ et $3^{ème}$ catégories peuvent parfois être aptes à travailler. C'est le médecin du travail qui détermine l'inaptitude ou l'aptitude (dans des conditions fixées) au travail.

1.1.3 Contexte actuel du risque invalidité

Afin de ne pas biaiser l'analyse en raison du Covid-19, les derniers chiffres présentés datent de 2019.

Alors que le taux d'absentéisme a connu une forte hausse ces dernières années (augmentation de 8% entre 2018 et 2019) [1], il s'est stabilisé entre 2018 et 2019. En moyenne, on dénombre 18,7 jours d'absence par salarié (18,6 en 2018). Cependant, cette constance globale masque des évolutions différentes selon les secteurs. En effet, si certains secteurs connaissant auparavant une augmentation de leur taux d'absentéisme ont réussi à inverser la tendance (commerce, industrie BTP et santé), le taux d'absentéisme des services a augmenté de 9%.

En revanche, le nombre de nouvelles invalidités était à la baisse entre 2012 et 2016, stable entre 2017 et 2018 avant de connaitre une hausse de près de 2% en 2019 [2].

Les prestations versées au titre des arrêts de travail sont très importantes. L'invalidité est le premier poste de dépenses de l'Assurance Maladie – Risques professionnels. En 2019, ses prestations s'élevaient à 4,35 milliards d'euros.

1.2 Le régime français de prévoyance

Il existe trois degrés de couverture du risque arrêt de travail :
- les garanties de la Sécurité Sociale
- la couverture assurée par l'entreprise (obligation de maintien de salaire par l'employeur)
- les garanties complémentaires versées par les assureurs

1.2.1 Régime de la Sécurité Sociale

La Sécurité Sociale est un ensemble d'institutions dont le but est de protéger les personnes des conséquences financières des risques sociaux. Elle a été fondée par les ordonnances des 4 et 19 octobre 1945 promulguées par le gouvernement du Général de Gaulle afin de fusionner toutes les anciennes assurances (maladie, retraite...). Il existe trois grands types de régime :

1. D'après le $11^{ème}$ et le $12^{ème}$ baromètres de l'absentéisme et de l'engagement d'Ayming
2. D'après le Rapport annuel de 2019 de l'Assurance Maladie-Risques Professionnels

— le régime général, majoritaire (80% de la population), applicable aux salariés et travailleurs assimilés à des salariés ;
— le régime social des indépendants (RSI), correspondant aux travailleurs non salariés non agricoles (artisans, commerçants et professions libérales).
— la Mutualité Sociale Agricole, applicable aux exploitants et salariés agricoles et aux secteurs rattachés à l'agriculture, comme l'industrie agroalimentaire.

Des régimes spéciaux (comme les Caisses de prévoyance et de retraite SNCF, CPR SNCF) subsistent.

La Sécurité Sociale est principalement financée par les cotisations sociales (à environ 60%) et les impôts (Contribution sociale généralisée (CSG) à 20%, autres impôts à 13%).

Elle est divisée en quatre branches :
— la branche "Maladie" (maladie, maternité, paternité, invalidité, décès) ;
— la branche "Accidents du travail et Maladies professionnelles" ;
— la branche "Vieillesse et veuvage"(retraite) ;
— la branche "Famille" (handicap, logement, RSA...).

❖ *Ouverture des droits*

Un individu ne peut être indemnisé que s'il était immatriculé au moins un an avant l'arrêt de travail suite à l'invalidité ou lors de la constatation de l'invalidité par le médecin conseil de la caisse d'Assurance Maladie. De plus, il doit justifier qu'au cours des 12 mois précédant l'arrêt de travail pour invalidité ou la constatation de l'invalidité, il avait soit effectué au moins 600 heures de travail salarié, soit cotisé un salaire au moins égal à 2030 fois le SMIC horaire.

❖ *Durée de service de la rente*

La pension est suspendue en tout ou partie en cas de reprise du travail. Elle cesse lors de la retraite pour être remplacée par la pension de vieillesse allouée en cas d'inaptitude au travail. L'âge de départ à la retraite a été modifié par la Loi portant réforme des retraites de 2010, passant progressivement de 60 à 62 ans.

❖ *Couverture*

La pension d'invalidité est calculée sur la base de la moyenne des dix meilleurs salaires annuels bruts plafonnés à la tranche A. Ces salaires sont revalorisés à la date de calcul.

Lorsque l'invalidité fait suite à un **accident ou une maladie non professionnelle**, le taux de couverture dépend de la catégorie d'invalidité :
— Pour la $1^{ère}$ catégorie, 30% du salaire annuel moyen ;
— Pour la $2^{ème}$ catégorie, 50% du salaire annuel moyen ;
— Pour la $3^{ème}$ catégorie, 50% du salaire annuel moyen, majoré par l'allocation tierce personne qui s'élevait à 1121,92 euros par mois en 2020.

De plus, la pension d'invalidité doit être supérieure au montant de l'allocation aux vieux travailleurs salariés.

Lorsque l'origine est professionnelle et que la prestation est sous forme de rente (cf 1.1.1 Définition), celle-ci est égale au salaire annuel multiplié par le taux d'invalidité, majorée par une prestation pour tierce personne le cas échéant. Dans le cas de l'incapacité permanente, aucune condition n'existe pour l'ouverture des droits. Si le taux d'incapacité est inférieur à 10%, un capital est versé. En 2021, un taux de 1% ouvre droit à 418,96 euros tandis que pour un taux de 9%, le capital est de 4188,62 euros.

1.2.2 Maintien de salaire par l'employeur et régime complémentaire

❖ *Obligation de maintien de salaire par l'employeur*

L'employeur, sous certaines conditions, doit maintenir le salaire de ses employés en arrêt de travail (loi de mensualisation, améliorée par l'ANI, cf **1.2.3 Cadre législatif**). Il peut financer ce maintien de salaire sur sa propre trésorerie ou souscrire à un contrat de prévoyance.

❖ *Régime complémentaire*

Au régime de la Sécurité Sociale et au maintien de salaire par l'employeur s'ajoute le régime complémentaire. Les prestations de la Sécurité Sociale sont souvent insuffisantes, ce qui incite à souscrire à un contrat de prévoyance complémentaire. Les individus peuvent eux-même souscrire à un contrat (prévoyance individuelle) ou une entreprise souscrit un contrat pour le compte de ses employés (prévoyance collective).

En cas d'incapacité, l'assureur verse les indemnités journalières à l'entreprise et non directement au salarié. L'assureur se réfère à la position de la Sécurité Sociale pour la reconnaissance de l'incapacité.

En cas d'invalidité, la rente est cette fois-ci directement versée au salarié et l'assureur ne se réfère pas toujours à la Sécurité Sociale pour la reconnaissance de l'invalidité.
En cas de contrats collectifs, le montant de la rente perçue peut être fixe ou s'exprimer comme un pourcentage du dernier salaire de l'assuré. Très souvent, le montant dépend de la catégorie d'invalidité.

Selon le baromètre CTIP/Crédoc [3] de 2020, 60% des salariés déclarent qu'ils disposent d'une garantie invalidité dans leur entreprise.

1.2.3 Cadre législatif

❖ *Loi de mensualisation*

En cas d'arrêt de travail, l'employeur est tenu de compléter les versements de la Sécurité Sociale si le salarié a au moins un an d'ancienneté.

3. Centre technique des institutions de prévoyance/ Centre de recherche pour l'étude et l'observation des conditions de vie

Ce maintien de salaire n'intervient qu'après une franchise de 7 jours, sauf s'il s'agit d'une maladie professionnelle ou d'un accident de travail (auquel cas aucune franchise n'existe).

L'employeur doit compléter les prestations de la Sécurité Sociale à hauteur de 90% puis de 66,66% du salaire brut qu'il aurait touché. Les durées de maintien de salaire varient selon l'ancienneté du salarié. Les conventions collectives fixent l'obligation de maintien (ancienneté minimale, montant, durée), qui doit être au moins aussi favorable que l'obligation minimale :

Ancienneté	Durée d'indemnisation	
	90 %	66.67 %
De 1 à 6 ans	30 jours	30 jours
De 6 à 11 ans	40 jours	40 jours
De 11 à 16 ans	50 jours	50 jours
De 16 à 21 ans	60 jours	60 jours
De 21 à 26 ans	70 jours	70 jours
De 26 à 31 ans	80 jours	80 jours
Au-delà de 31 ans	90 jours	90 jours

FIGURE 1.1 – Obligation minimale de maintien de salaire

❖ *Loi Evin (31/12/1989)*

La loi Evin est la première loi spécifique de prévoyance.

L'article 2 de la loi Evin interdit aux assureurs d'exclure du champ d'application des contrats une maladie antérieure à la signature du contrat.

L'article 7 impose en cas de résiliation le maintien par le premier assureur des prestations au niveau atteint (incapacité, invalidité, rente éducation et rente de conjoint). Elle oblige la constitution de provisions.

Remarque :

Si une entreprise change d'assureur et qu'un salarié en incapacité avant la résiliation passe en invalidité après celle-ci, c'est à l'assureur résilié de verser la rente d'invalidité. En effet, on considère que le risque invalidité est né au moment de l'entrée en incapacité. Néanmoins, il existe quelques cas de jurisprudence où le premier assureur a pu prouver que l'invalidité ne faisait pas suite à l'incapacité.

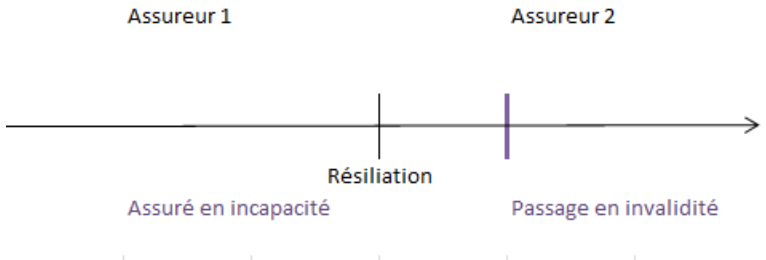


FIGURE 1.2 – Passage en invalidité après la résiliation

La loi Evin impose aussi aux assureurs de proposer aux anciens salariés la poursuite de la couverture santé. Par ailleurs, elle interdit à l'employeur d'imposer unilatéralement de cotiser à un régime de prévoyance. Parmi les autres principales obligations de la loi Evin, on retrouve la remise d'une notice résumant les garanties, la consultation du Comité d'entreprise et la remise par l'assureur d'un rapport sur les résultats du contrat.

- ❖ *Loi du 08 août 1994*

La loi du 08 août 1994 s'inscrit dans la continuité de la loi Evin. Elle impose à l'employeur d'organiser la revalorisation des rentes en cas de changement d'assureur. Souvent c'est le nouvel assureur qui prendra en charge la poursuite des revalorisations, en contrepartie d'une prime versée par l'entreprise.

D'autre part, elle oblige le maintien des garanties décès pour les personnes en arrêt de travail au moment de la résiliation du contrat.

- ❖ *Accord National Interprofessionnel (ANI)*

L'article 14 de l'ANI du 11 janvier 2008 permet le maintien des garanties existantes dans l'entreprise pour les anciens salariés bénéficiant de l'assurance chômage (hors démission et licenciement pour faute lourde).

1.3 Tarification

Nous présenterons les méthodes de tarification, les tables de référence et les garanties à tarifer.

1.3.1 Méthode de tarification

La tarification de l'arrêt de travail n'est **pas encadrée réglementairement**.

En pratique, le tarif est obtenu à partir du maintien, de l'incidence, du coût (niveau des garanties) et de correctifs (liés au sexe et à la catégorie socioprofessionnelle par exemple).

Le tarif est déterminé à partir d'un ensemble de paramètres, tels que l'âge moyen, le salaire moyen, le collège ou encore la part d'hommes. La loi française interdit de mettre en place un tarif différent pour un homme et une femme toutes choses égales par ailleurs.

1.3.2 Tables de maintien et de passage d'incapacité à invalidité

❖ *Les tables du Bureau commun d'assurances des collectives (BCAC)*

Le Bureau commun d'assurances des collectives (BCAC) a produit des tables de maintien en incapacité et en invalidité et une table de passage d'incapacité à invalidité. Elles ont été établies pour la première fois en 1993, puis ont été mises à jour en 2010 suite à la réforme des retraites. En l'absence de tables d'expérience, ces tables sont utilisées pour le provisionnement par les assureurs. Le BCAC a aussi remis à jour ses tables en 2013 mais elles n'ont à ce jour pas été homologuées pour le provisionnement.

La table de maintien en incapacité du BCAC est une table bidimensionnelle faisant intervenir l'âge de l'assuré à l'entrée en incapacité et l'ancienneté en incapacité. L'ancienneté est ici exprimée en mois. C'est le nombre d'individus maintenus en incapacité, rapporté à un effectif de 10000 qui est renseigné.

La table de maintien en invalidité dépend aussi de l'âge d'entrée (en invalidité) et de l'ancienneté, ici exprimée en années.

Enfin, la table de passage de l'incapacité à l'invalidité est toujours à deux dimensions, l'âge d'entrée et l'ancienneté. Les probabilités déduites des nombres de passages de la table ne sont pas des probabilités conditionnelles au maintien en incapacité : le maintien en incapacité est déjà inclus.

❖ *Tables d'expérience*

Certains assureurs choisissent de faire leurs propres tables de maintien ou de passage. Ces tables ne peuvent ensuite être utilisées pour le provisionnement que si elle ont été certifiées. De plus, un suivi annuel est mis en place.

Les tables d'expérience permettent d'étudier le maintien ou le passage sur la population assurée. La population étudiée par le BCAC ne présente pas les mêmes caractéristiques que celle-ci. Ainsi les tables d'expérience permettent de mieux évaluer le risque propre à l'assureur et donc d'éviter une sur-tarification ou sous-tarification et un sur-provisionnement ou un sous-provisionnement.

1.3.3 Présentation des garanties

Les garanties de GGVIE dépendent du taux d'invalidité (noté n), défini comme le taux d'incapacité professionnelle.

Si le taux d'invalidité est supérieur à 66% (invalidité totale), le montant de la rente est celui exprimé aux conditions particulières. Celui-ci est soit exprimé en complément des prestations de la Sécurité Sociale, soit sous-déductions de celles-ci.

Si le taux d'invalidité est compris entre 33 et 66%, le montant de la rente est égal à celui de l'invalidité totale, multiplié par n/66.

Enfin, si le taux est inférieur à 33%, aucune prestation n'est versée.

La rente est versée après une période de franchise égale à celle prévue en cas d'incapacité temporaire. Néanmoins, lorsque l'invalidité fait suite à de l'incapacité, le délai de franchise n'est pas réappliqué.

2 Traitement des données

Cette partie abordera la construction des bases d'étude et les premières statistiques descriptives.

2.1 Construction des bases de données

La construction des bases de données s'est appuyée sur plusieurs sources. Puis les étapes de retraitement ont permis de fiabiliser les bases.

2.1.1 Présentation des principales sources

Pour construire les tables d'étude de l'incidence et du maintien, nous avons utilisé plusieurs tables, qui sont présentées ci-dessous.

❖ *Historique des sinistres BDD_AT - archivé depuis fin 2018*

Afin d'étudier l'invalidité, nous disposions tout d'abord de la table BDD_AT recensant tous les arrêts de travail en cours et clos, ainsi que leurs principales caractéristiques (date de survenance, annuité, nature de la prestation, identité du bénéficiaire...). Cette table présente la dernière vision du sinistre : si l'individu passe de l'incapacité à l'invalidité ou change de catégorie d'invalidité, une seule ligne indiquant la dernière nature et la date de début correspondante est présente. Cette table est archivée tous les mois depuis fin 2018.

❖ *Historique des Provisions Mathématiques*

Les tables de Provisions Mathématiques ont permis d'élargir l'historique. En effet, les Provisions Mathématiques annuelles (archives de 2006 à 2020) et trimestrielles (archives de 2010 à 2020) contiennent notamment la date de début d'invalidité (et non celle d'un éventuel changement de catégorie) pour les sinistres déjà passés en invalidité. D'autre part, une autre table de Provisions Mathématiques archivée de 2012 à 2017 précise les catégories d'invalidité.

❖ *Historique des flux comptables*

L'historique des flux comptables de 2001 à 2020 a également été utilisé.

❖ *Autres tables utiles pour l'ajout des caractéristiques de l'assuré et de l'entreprise*

Enfin, d'autres tables recensant les caractéristiques des contrats, des assurés et des entreprises

ont permis l'ajout des caractéristiques suivantes : sexe, catégorie socioprofessionnelle, code APE et département de l'entreprise.

2.1.2 Périmètre d'observation

L'étude porte sur les arrêts de travail en gestion interne et déléguée en excluant les accidents (dont la durée peut excéder 3 ans). Les sinistres repris à la concurrence ne font également pas partie du périmètre d'étude puisqu'ils n'entrent pas en jeu dans le tarif.

La période d'observation devait permettre une volumétrie suffisante tout en évitant la prise en compte de sinistres trop anciens. En effet, le maintien peut se déformer au cours du temps. De surcroît, l'année 2020 a connu un nombre très important d'arrêts de travail à cause du covid. Il a donc été choisi d'arrêter la période d'observation fin 2019. Enfin, la catégorie d'invalidité peut beaucoup influer sur les prestations, donc il est pertinent d'en tenir compte. Or, nous ne disposions de l'historique qu'à partir de 2012. C'est pourquoi nous avons choisi la période d'observation de 2012 à 2019. A titre de comparaison, le BCAC a travaillé sur une période d'observation de 7 ans.

2.1.3 Méthodologie de construction des bases

Cette section décrit les étapes de construction des bases.

2.1.3.1 Extraction de la dernière vision des sinistres

La première étape a consisté à récupérer tous les sinistres du périmètre dans la table des sinistres de 2019. Nous obtenions ainsi la dernière vision du sinistre à la fin de la période d'observation.

2.1.3.2 Reconstitution de l'historique des sinistres

La date de début d'invalidité et les périodes dans les catégories d'invalidité ont ensuite été reconstituées.

❖ *Extraction de la date de début d'invalidité*

Il était primordial de retrouver la date de début d'invalidité pour étudier le maintien. Celle-ci a été déterminée grâce à l'historique des tables des Provisions Mathématiques.

D'autre part, la catégorie d'invalidité influant sur les prestations, nous nous sommes intéressés aux catégories d'invalidité du sinistre et à leurs périodes respectives. Pour cela, nous disposions de l'historique des Provisions Mathématiques de 2012 à 2017 (vision en fin d'année) puis de l'historique de la BDD_AT de 2018 à 2019 (vision mensuelle, cf 2.1.1 Présentation des principales sources).

❖ *Détermination des périodes des catégories d'invalidité*

L'historique des sinistres nous fournit toutes les catégories observées chaque mois de 2018 à 2019 pour tous les sinistres, ainsi que les dates du début de ces catégories. Cependant, s'il y a eu un changement de catégorie entre le début de l'invalidité et la catégorie observée en 2018, c'est-à-dire si la date de début de la première catégorie retrouvée dans BDD_AT ne coïncide pas avec la date

de début d'invalidité, nous n'avons pas d'information.

Quant à l'historique des Provisions Mathématiques, il indique les catégories d'invalidité à la fin de l'année de 2012 à 2017. Ainsi, il est possible d'en déduire l'année des éventuels changements de catégorie. Donc cet historique rajoute de l'information lorsqu'un changement de catégorie a eu lieu entre 2012 et 2018. Cependant, contrairement à la table BDD_AT seule l'année de changement pourra être renseignée et si un changement de catégorie a lieu avant 2012, la première catégorie ne peut être connue. Dans un peu plus de 4% des cas, les tables des Provisions Mathématiques (PM) apportaient une part d'information supplémentaire pour retracer l'historique du sinistre. Néanmoins, l'information disponible n'étant qu'annuelle, bien que peu restrictives, des hypothèses ont dû être posées :

— si le sinistre est retrouvé dans les PM l'année même du début de l'invalidité et que la catégorie correspondante est égale à la première catégorie retrouvée dans BDD_AT (qui n'est pas la catégorie d'arrivée), on suppose qu'un seul changement a eu lieu cette même année : la catégorie d'arrivée n'est pas connue mais l'on connaît la $2^{ème}$ catégorie et la date exacte grâce à la BDD_AT. L'hypothèse d'un seul changement dans l'année n'est pas contraignante puisque dans la base finale, seuls 15 sinistres ont connu 2 changements tout au long de leur déroulement.

— si le sinistre est de même retrouvé dans les PM l'année du début de l'invalidité mais que la catégorie correspondante n'est pas égale à la première catégorie retrouvée dans BDD_AT (qui n'est pas la catégorie d'arrivée), on suppose cette fois qu'il s'agit bien de la catégorie d'arrivée en invalidité. Sinon, cela signifierait que le sinistre a connu au moins 2 changements de catégorie au cours de son déroulement, ce qui est très rare.

— si la dernière catégorie présente dans les PM ne coïncide pas avec la première catégorie retrouvée dans BDD_AT (changement pas encore renseigné dans les dernières PM), alors la $2^{ème}$ catégorie (dernière catégorie des PM) est bien connue mais seule l'année de changement est renseignée. Dans ce cas, on émet l'hypothèse que le changement a lieu au milieu de l'année de changement. Cependant, seuls 3 cas sont concernés.

2.1.3.3 Ajout des montants et des périodes de prestations

Afin de vérifier l'impact de la catégorie d'invalidité sur le montant des prestations, nous avons rajouté les dates et montants des prestations de 2001 à 2020 en séparant les flux relatifs à l'incapacité de ceux liés à l'invalidité.

2.1.3.4 Ajout des variables explicatives

La table contenait déjà les principales caractéristiques du sinistre et du bénéficiaire, contenues dans BDD_AT. Le sexe, la catégorie socioprofessionnelle, le code APE et le département de l'entreprise ont été rajoutés.

2.1.4 Principaux retraitements

Les principaux retraitements concernent l'identification des sinistres, la date de début d'invalidité, le sexe et la catégorie socioprofessionnelle.

❖ *Création d'une clé unique par sinistre*

Les sinistres sont identifiés par un numéro. Cependant, il arrive que ce numéro ne suive pas le sinistre tout au long de sa vie. Par exemple, cela arrive parfois lorsqu'un sinistre en gestion déléguée passe en invalidité en gestion interne. La gestion des sinistres d'invalidité se fait toujours en interne. Il convenait donc de créer une clé unique pour ne pas les compter plusieurs fois lors de l'étude de l'incidence et pour prendre en compte la bonne durée de maintien. La clé retenue est constituée du numéro du contrat, d'un identifiant assuré, de la date de naissance et de la date de survenance.

❖ *Fiabilisation de la date de début d'invalidité*

Lors d'un changement de catégorie d'invalidité ou de numéro de sinistre, il arrive que la date de début d'invalidité soit modifiée. La date retenue est donc le minimum des dates trouvées.

La cohérence de la date de début d'invalidité a été testée en vérifiant que celle-ci était bien postérieure à la date de survenance et antérieure à la date de début de la première catégorie d'invalidité retrouvée dans BDD_AT. De plus, la différence entre la date de début d'invalidité et la date de survenance ne devait pas excéder 3 ans en raison de la limite de durée de l'incapacité.

Lors d'incohérences ou d'absence d'information dans les PM, si la date présente dans BDD_AT assurait un maintien en incapacité de moins de 3 ans et était cohérente avec les flux comptables (date de début d'invalidité jouxtant la fin de la période des prestations d'incapacité), alors c'est cette date qui était retenue.

❖ *Retraitement du sexe*

Lorsque le sexe n'était pas renseigné pour un individu et que son prénom n'était associé qu'à un seul sexe dans la base des individus, ce sexe a été retenu pour l'individu.

❖ *Retraitement de la catégorie socioprofessionnelle*

Lorsque le code de la catégorie bénéficiaire n'avait pas pu être retrouvé ou si celui-ci ne permettait pas de distinguer les cadres des non-cadres, la catégorie socioprofessionnelle a été approximée grâce au salaire : si le salaire brut était inférieur à la tranche 1 du Plafond Annuel de la Sécurité Sociale de l'année de survenance, alors on estimait qu'il s'agissait d'un non-cadre.

2.1.5 Présentation des tables créées

Les invalidités non consécutives à de l'incapacité (franchise non fiable) ainsi que les sinistres pour lesquels la date de début d'invalidité présente une anomalie ou est absente ne peuvent pas être pris en compte pour l'étude sur le maintien. Deux tables ont donc été créées : *AT_MAINTIEN* et

AT_INCIDENCE.

Voici les principales variables contenues dans *AT_INCIDENCE* :

- Clé identifiant le sinistre
- Montant des prestations de 2001 à 2020
- Sexe
- CSP
- Code APE
- Département de l'entreprise

La table *AT_MAINTIEN* contient également les variables suivantes :
- Date de début d'invalidité
- Périodes dans les catégories d'invalidité et annuités correspondantes

2.2 Statistiques descriptives

Les statistiques descriptives permettent de découvrir la composition des bases d'études.

2.2.1 Statistiques sur la base *AT_INCIDENCE*

La base AT_INCIDENCE permettra d'étudier la probabilité de passage de l'incapacité à l'invalidité. La période et le périmètre d'étude seront déterminés ultérieurement. La base totale comprend tous les arrêts en cours de 2012 à 2019, tandis que le BCAC observait les années 2005-2011 pour construire ses tables de passage.

Le tableau suivant résume les premières statistiques de nos données internes :

Statistiques	
Nombre d'arrêts observés	272 095
Âge moyen	42,2
Proportion de femmes	54,5%
Proportion de non cadres	84,8%

FIGURE 2.1 – Composition de la table *AT_INCIDENCE*

Le secteur d'activité est très concentré sur quelques sections. En effet, en regroupant les sections qui ont un poids inférieur à 5% dans la catégorie "Autre", on obtient la répartition suivante :

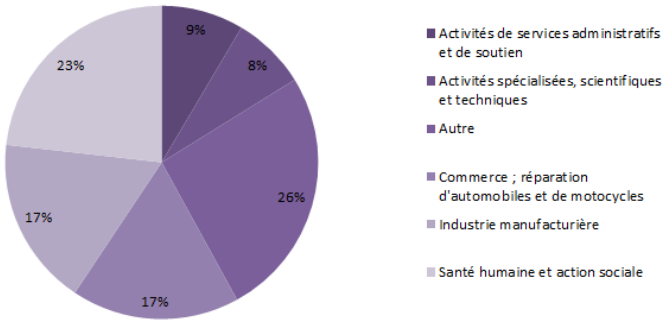


FIGURE 2.2 – Répartition dans les sections de secteur d'activité pour la table *AT_INCIDENCE*

On remarque que les sections regroupées dans la catégorie "Autres" représentent environ un quart de la base. Les secteurs de santé, commerce et industrie sont quant à eux très représentés.

Au contraire, les observations ne sont pas concentrées sur un département. Seuls deux départements ont un poids supérieur à 5% mais limité tout de même : le 75 (9%) et le 92 (7%). Ces départements ont une population relativement importante.

Par ailleurs, le poids des valeurs manquantes est relativement faible pour la plupart des variables mais représente un peu plus de 5% pour le secteur d'activité :

Variable	Sexe	CSP	Département	Secteur d'activité
Poids valeurs manquantes	2,3%	0,3%	0,4%	5,9%

FIGURE 2.3 – Pourcentage de valeurs manquantes pour chaque variable

2.2.2 Statistiques sur la base *AT_MAINTIEN*

2.2.2.1 Composition de la base

La base *AT_MAINTIEN* permettra l'étude du maintien. Les premières statistiques ont été comparées avec celles des données utilisées par le BCAC [1] :

Statistiques	Données internes	BCAC
Nombre d'arrêts observés	14 406	63 148
Âge moyen d'entrée en invalidité	49 ans	49,5 ans
Part d'hommes	51,6%	41%
Proportion de non cadres	84,8%	

FIGURE 2.4 – Composition de la table *AT_MAINTIEN*

D'une part, on remarque que l'on dispose de beaucoup moins de données que le BCAC (moins de 30% du volume du BCAC). Ainsi, l'étude sur le maintien devrait être délicate. En revanche, Claire Elias [2], qui avait également travaillé sur le maintien en invalidité (pour l'assurance individuelle)

1. Source : "Présentation et comparaison des nouvelles tables BCAC", Prim'Act-25/06/2014

lors de son mémoire, était parvenue à des résultats alors qu'elle disposait de moins de données : sa base contenait 6724 contrats alors que dans nos données internes, 9151 contrats ont des prestations en invalidité.

Si l'âge moyen de nos données est proche de celui des données du BCAC, on peut noter que notre portefeuille est composé majoritairement d'hommes, à l'inverse du BCAC. Ainsi, il est intéressant de faire une loi d'expérience, qui sera plus proche de notre risque.

De nouveau, la base est très concentrée sur les secteurs d'activité du commerce et de l'industrie :

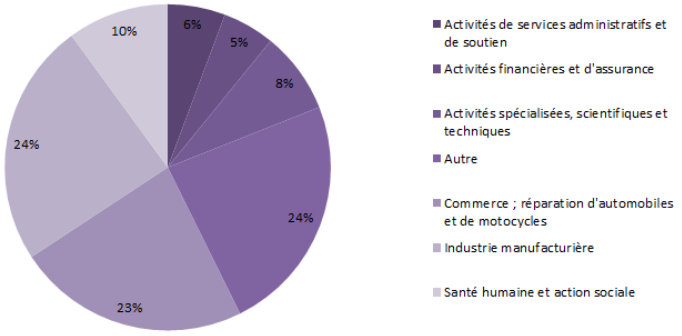


FIGURE 2.5 – Répartition dans les sections de secteur d'activité pour la table *AT_MAINTIEN*

Seuls 3 départements ont un poids supérieur à 5% : le 13 (5%) et de nouveau le 75 (9%) et le 92 (8%).

Le poids des valeurs manquantes avoisine 0% pour toutes les variables, hormis le secteur d'activité pour lequel il est de 5,9%.

2.2.2.2 Statistiques du maintien

❖ *Proportion de données censurées*

Le concept de censure sera expliqué plus précisément dans l'explication de la modélisation du maintien (cf 3.1.2 Troncatures et censures). Les données censurées (ici à droite) sont les données pour lesquelles nous n'avons qu'une information partielle sur la durée de maintien : nous savons qu'elle est supérieure à une certaine valeur.
88,9% des sinistres sont censurés. Le BCAC avait également beaucoup de données censurées puisque seuls 22% de ses sinistres étaient clos.

❖ *Durée moyenne des sinistres non censurés*

La durée moyenne des sinistres non censurés est de 6,48 ans.

❖ *Motif de sortie*

Nous avons étudié les causes de sortie pour les sinistres clos et non censurés (cf 3.1.2 Troncatures et censures) :

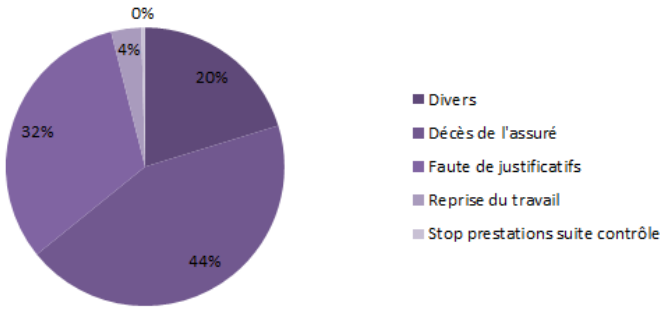


FIGURE 2.6 – Motif de sortie de l'invalidité

Ainsi 3 causes de fin de sinistres sont majoritaires :

- le décès ;
- la faute de justificatifs (lorsque l'assuré ne renvoie pas les justificatifs permettant la poursuite des prestations). Cette situation se produit en cas de décès ou encore lorsque l'assuré reprend le travail [2]. Comme la $2^{ème}$ cause est très rare (56 cas observés dans notre étude), il semble que ce motif soit principalement lié au décès, qui comptabiliserait donc environ 70% des cas ;
- les causes diverses : il s'agit des fraudes, décisions juridiques... L'ensemble de ces causes constitue tout de même 20% des motifs de fin.

Remarque : La retraite n'est pas prise en compte ici puisqu'elle est considérée comme une censure (cf 3.1.2 Troncatures et censures).

2. Les cas de retraite devraient majoritairement avoir été exclus, cf 3.1.2 Troncatures et censures

3 Modélisation du maintien en invalidité

Maintenant que les bases ont été constituées, il nous est possible d'étudier le maintien en invalidité. Le maintien sera d'abord étudié en fonction de l'âge d'entrée et de l'ancienneté, puis nous étudierons l'effet des caractéristiques de la population.

3.1 Présentation des modèles

Cette section vise à présenter les principales notions et notations ainsi que les méthodes envisagées pour l'étude du maintien.

3.1.1 Fonction de survie et conventions

Notre objectif est d'estimer le maintien en invalidité. Pour cela, nous pouvons introduire la fonction de survie :

$$S(t) = P(T \geq t)$$

où T est une variable aléatoire positive.

Par la suite, T correspondra à la durée de maintien en invalidité, exprimée en année.

Remarque : Ici, nous travaillons avec la version continue à gauche de la fonction de survie. La version continue à droite définit quant à elle la fonction de survie ainsi :

$$S(t) = P(T > t)$$

3.1.2 Troncatures et censures

❖ *Définitions*

L'information de la durée n'est pas toujours complète.

La durée est censurée à droite si l'on ne connaît pas la durée mais que l'on sait que celle-ci est supérieure à une certaine valeur : la fin n'est pas observée.
Elle est censurée à gauche si l'on sait que la durée est inférieure à une certaine valeur : l'événement de fin s'est déjà produit.

Une observation est tronquée si elle n'est observable que conditionnellement à un événement. Elle est tronquée à gauche si elle n'est observable que si elle est supérieure à un certain seuil.

Elle est tronquée à droite si au contraire elle n'est observable que si elle est inférieure à un seuil.

❖ *Exemple :*

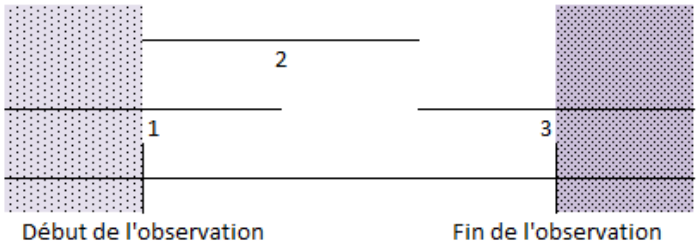


FIGURE 3.1 – Schéma explicatif des troncatures et censures

L'individu 1 est tronqué à gauche : si la fin de son arrêt avait été inférieure au début de l'observation, il n'aurait pas été observé.
L'individu 2 n'est ni tronqué ni censuré : l'information est complète.
L'individu 3 est censuré à droite : on ne connaît pas la durée de son arrêt mais on sait que la fin est postérieure à la fin de l'observation.

❖ *Censures de notre étude*

Les principales causes de fin de l'invalidité sont le décès, la reprise du travail et la retraite (cf 1.1.1 Le risque invalidité). Cependant, la retraite sera considérée comme une censure. En effet, l'effet des garanties (et donc l'arrêt du versement des prestations à l'âge de départ à la retraite) sera pris en compte dans les calculs, donc ce n'est pas la peine de l'intégrer dans les tables : les individus sortis pour un motif de retraite ne sont plus observés, et sont donc considérés comme censurés. De plus, en cas de changement de l'âge de départ à la retraite, les tables ne seront pas affectées. Enfin, les tables du BCAC ne présentent pas de cassure au moment de la retraite (sorties massives), donc le BCAC a lui aussi considéré la retraite comme une censure.

Les sinistres censurés sont :
- les sinistres en cours fin 2019
- les sinistres clos pour un motif censurant (cf tableau ci-dessous).

Le tableau suivant récapitule les motifs de clôture et indique s'il s'agit d'une censure :

Cause de clôture	**Censure**
Reprise du travail	0
Décès de l'assuré	0
Stop prestations suite contrôle	0
Divers (fraude, décision juridique...)	0
Retraite	1
Fin de garantie	1
Maternité	1
Faute de justificatifs* avant 59 ans	0
Faute de justificatifs après 59 ans	1

❖ *Faute de justificatifs** :

Lorsqu'un sinistre est clos pour cause de retraite, le motif est en général "Retraite" dans le système de gestion. Ainsi, nous suspections les fautes de justificatifs d'être des décès ou des reprises d'activité. Elles étaient donc toutes initialement considérées comme de vraies sorties. Néanmoins, les probabilités de transition obtenues ont remis en question cette approche.
La probabilité de transition correspond à la probabilité de se maintenir en invalidité l'année N sachant que l'on s'était maintenu l'année N-1. Voici le graphique que nous obtenions pour les âges 55, 56, 57 et 58 ans (âges ayant les plus grands effectifs) :

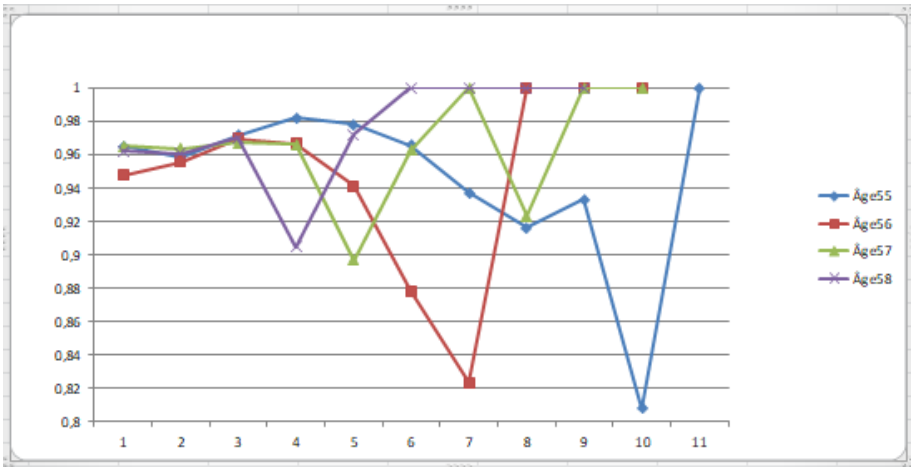


FIGURE 3.2 – Probabilités de transition obtenues en considérant les "Fautes de justificatif" comme des sorties

On remarque une baisse importante aux âges correspondant à la retraite. Ainsi, on peut supposer que les fins de justificatifs correspondent en partie à des départs à la retraite.

Ces fautes de justificatifs représentent 12% des sinistres clos. En les supprimant, le maintien semblait très surestimé (en comparant avec le BCAC). En effet, l'échantillon enlevé contenait principalement de vraies sorties et la base étudiée n'était donc plus représentative.

Finalement, nous avons choisi de considérer les fautes de justificatifs comme de vraies sorties avant 59 ans puis comme des censures après 59 ans (prise en compte du changement progressif d'âge de départ à la retraite) par prudence. Cependant, nous risquons ainsi de surestimer le maintien. C'est pourquoi nous faisons le choix de calculer le maintien jusqu'à 61,5 ans pour le tarif et non 62 ans.

❖ *Troncatures de notre étude*

Les données tronquées à gauche de notre étude sont les sinistres dont la date de début d'invalidité est antérieure au 01/01/2012.

3.1.3 Estimateur de Kaplan-Meier

3.1.3.1 Justification de la sélection de l'estimateur

L'estimateur de Kaplan-Meier est un estimateur non paramétrique de la fonction de survie. Il permet donc de ne pas faire d'hypothèse sur la distribution. Les deux principaux estimateurs non paramétriques des modèles de durée sont l'estimateur de Nelson-Aalen du taux de hasard cumulé [1] (qui conduit à l'estimateur de Harrington-Flemming de la fonction de survie [2]) et l'estimateur de Kaplan-Meier. Cependant, ces deux estimateurs surestiment la durée, mais le biais est moins important avec l'estimateur de Kaplan-Meier [11] (cf Annexe 1 : Comparaison des biais de Kaplan-Meier et Harrington-Flemming). C'est pourquoi nous avons retenu l'estimateur de Kaplan-Meier.

3.1.3.2 Présentation de la méthode

L'estimateur de Kaplan-Meier repose sur la définition de la probabilité conditionnelle. En notant s et t deux entiers tels que $t \geq s$,

$$P(T > t) = P(T > t|T > s) \times P(T > s)$$

Par la suite, $P(T > s)$ peut être décomposé de la même manière et ainsi de suite.

Notons $T_{(1)}, ..., T_{(N)}$ les instants où les sorties et les censures se produisent, ordonnés par ordre croissant.
Par le raisonnement précédent,

$$\begin{aligned} \hat{S(t)} &= \prod_{T_{(i)}<t} P(T > T_{(i)}|T > T_{(i-1)}) \\ &= \prod_{T_{(i)}<t} (1 - \hat{q}_i) \end{aligned}$$

avec $\hat{q}_i$, la probabilité de sortir de l'état d'invalidité.

Remarque : Dans la version continue à gauche (cf 2.1.1 Fonction de survie et conventions), on fait le produit sur $T_{(i)} < t$. Si l'on avait travaillé avec la version continue à droite, cela aurait été sur $T_{(i)} \leq t$.

Il reste à exprimer $\hat{q}_i$. Notons :
- d_i, le nombre de décès entre $]T_{(i-1)}; T_{(i)}]$,
- r_i, l'effectif sous risque en $T_{(i-1)}$.

Le nombre de sorties en $T_{(i)}$ suit une loi binomiale $B(r_i, q_i)$. La vraisemblance s'écrit donc :

$$L(q_1, ..., q_n) = \prod_{i=1}^{n} \binom{d_i}{r_i} q_i^{d_i} (1 - q_i)^{r_i - d_i}$$

1. La fonction de hasard cumulé est définie par $H(t) = \int_0^t \frac{dS(u)}{S(u-)}$, soit dans le cas continu par : $\int_0^t h(u)du$ où $h(t) = -\frac{d}{dt} ln(S(t))$
2. $\hat{S}_{HF}(t)) = exp(-\hat{H}_{NA}(t))$

Pour trouver le maximum de vraisemblance, on résoud pour i=1,...,n :

$$\frac{\partial ln(L)}{\partial q_i} = 0$$

On trouve : $\hat{q}_i = \frac{d_i}{r_i}$

Ainsi,

$$\hat{S(t)} = \prod_{T_{(i)}<t} \left(1 - \frac{d_i}{r_i}\right)$$

Remarque :

Par convention, on suppose que si un décès survient à la même période qu'une censure ou une troncature, le décès précède les autres évènements.

3.1.3.3 Propriétés

L'estimateur de Kaplan-Meier présente beaucoup de bonnes propriétés :

- Il est convergent [3] :
$$\forall \epsilon > 0, \lim_{n \to +\infty} P(|\hat{S}_x(t) - S_x(t)| > \epsilon) = 0$$
- Il est asymptotiquement gaussien [4] ;
- C'est l'unique estimateur cohérent de la fonction de survie : la probabilité de décéder au-delà de t est la somme de la probabilité d'être à risque à t et de celle d'avoir été censuré avant t ;
- C'est un estimateur du maximum de vraisemblance généralisé.

Cependant, il est biaisé positivement.

❖ *Variance de l'estimateur*

Cette partie présente l'estimateur de la variance de Greenwood.

Par définition de la fonction de survie de Kaplan-Meier :

$$ln(\hat{(}\hat{S}(t)) = \sum_{T_{(i)}<t} ln\left(1 - \frac{d_i}{r_i}\right) = \sum_{T_{(i)}<t} ln(1 - q_i)$$

Or, la loi de $r_i\hat{p}_i$ est binomiale de paramètres (r_i,p_i) et par la méthode Delta, $V(f(X)) \approx \left(\frac{df}{dx}(E(X))\right)^2 V(X)$.

3. Si la fonction de survie et la distribution de censures n'ont pas de discontinuité commune
4. à la même condition que la propriété précédente

Donc :

$$\begin{aligned} V(ln(\hat{p}_i)) = V(ln(r_i\hat{p}_i)) &\approx V(\hat{p}_i)\left(\frac{d}{dp}ln(\hat{p}_i)\right)^2 \\ &= \frac{\hat{q}_i(1-\hat{q}_i)}{r_i} \times \frac{1}{(1-\hat{q}_i)^2} \\ &= \frac{\hat{q}_i}{r_i(1-\hat{q}_i)} \end{aligned}$$

Par indépendance des variables $ln(1-\hat{p}_i)$,

$$\hat{V}(ln(\hat{S}(t))) = \sum_{T_{(i)}<t} \frac{\hat{q}_i}{r_i(1-\hat{q}_i)} = \sum_{T_{(i)}<t} \frac{d_i}{r_i(r_i-d_i)}$$

En appliquant de nouveau la méthode Delta à la fonction exponentielle cette fois-ci :

$$\hat{V}(\hat{S}(t)) \approx \hat{S}(t)^2\gamma(t)^2$$

avec $\gamma(t)^2 = \sum_{T_{(i)}<t} \frac{d_i}{r_i(r_i-d_i)}$

Or l'estimateur de Kaplan-Meier est asymptotiquement normal, donc on peut estimer l'intervalle de confiance de la fonction de survie en T_i dont les bornes sont :

$$S_i \times (1 \pm u_{1-\frac{\alpha}{2}}\gamma(T_{(i)}))$$

avec $u_{1-\frac{\alpha}{2}}$ le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi Normale.

3.1.4 Estimateur de Turnbull

L'estimateur de Turnbull est l'estimateur utilisé par le BCAC. Il s'agit d'une version discrétisée de Kaplan-Meier [4] : cet estimateur ne fait plus d'estimation en temps continu mais par période.

Notons :

- m le nombre de périodes
- s_i la probabilité de sortie pendant la $i^{ème}$ période ;
- n_i le nombre de sorties pendant la $i^{ème}$ période ;
- t_i le nombre d'individus tronqués à la $i^{ème}$ période ;
- c_i le nombre d'individus censurés à la $i^{ème}$ période ;
- N_0 le nombre d'assurés non tronqués
- $S(j)$, la probabilité de se maintenir au-delà de la $j^{ème}$ période.

La fonction de survie peut être exprimée à partir des s_i :

$$S(j) = 1 - \sum_{i=1}^{j} s_i$$

Les équations du maximum de vraisemblance ne permettent pas d'obtenir une solution numérique pour les s_i. Turnbull a proposé en 1976 le « Self-Consistent algorithm », un algorithme itératif

qui apporte une solution approchée. Nous présenterons ici la forme opérationnelle de l'estimateur, version simplifiée de la formule originale.

L'algorithme est le suivant :

1. Initialisation des s_j^0, pour j=1,...,m+1, de façon à ce qu'ils définissent une distribution de probabilité

2. Itération jusqu'au critère d'arrêt :

$$s_j^{h+1} = \frac{n_j + \left(\frac{c_1}{S(1)} + ... + \frac{c_m}{S(m)}\right) \times s_j^h}{N_0 + \frac{t_1}{S(1)} + ... + \frac{t_m}{S(m)}}, \text{ pour j=1,...,m+1}$$

Le calcul de s_j^{h+1} fait apparaître au numérateur le nombre de vraies sorties n_j, mais aussi une estimation des sorties non observées à cause des censures. De même, le dénominateur correspond au nombre réel d'individus observés en 0, augmenté d'une estimation du nombre de personnes non observées du fait des troncatures.

Un critère d'arrêt doit être choisi. L'algorithme peut par exemple être stoppé lorsque la différence entre s_j^{h+1} et s_j^h est inférieure à un seuil.

3.1.5 Ajustement par référence externe

Le modèle de Brass est un modèle fréquemment utilisé pour les tables de mortalité lorsque l'effectif est faible [11]. **Son utilisation dans cette étude sera justifiée ultérieurement (cf 3.2.1.1 Choix de la méthode).** Pour essayer de structurer correctement la fonction de survie malgré tout, on peut utiliser la structure d'une table de référence et positionner la fonction de survie par rapport à celle de référence.

Ce modèle consiste à faire la régression des logits des taux de sortie observés par les logits des taux de sortie d'une table de référence :

$$ln\left(\frac{q_{xt}}{1-q_{xt}}\right) = a \times ln\left(\frac{q_{xt}^{rf}}{1-q_{xt}^{rf}}\right) + b$$

Les paramètres peuvent être obtenus par la méthode des moindres carrés.

De manière heuristique, le taux $\frac{q_{xt}}{1-q_{xt}}$ représente le rapport entre la probabilité de sortie de l'invalidité et celle de se maintenir. Ce rapport étant positif, il est assez naturel de l'exprimer sous la forme d'une exponentielle.

Sur $]0; \frac{1}{2}[$, la fonction logistique est concave donc par l'inégalité de Jensen [5], les logits des taux de sortie seront sous-estimés. Or, l'inverse de la fonction logistique est croissante donc les taux de sortie seront eux aussi sous-estimés. Ainsi, le maintien sera au contraire surestimé. Cette approche est donc prudente.

5. Si f est concave, alors $E(f(X)) \geq f(E(X))$, donc $E(logit(q_{xt})) \geq logit(q_{xt})$

3.1.6 Modèles intégrant des variables explicatives

3.1.6.1 Présentation des modèles

❖ *Présentation sommaire des méthodes usuelles (non applicables dans notre étude)*

En modèle de durée, la prise en compte des variables explicatives se fait le plus souvent grâce au modèle de Cox ou au modèle additif d'Aalen [11]. **Ces modèles seront présentés très brièvement puisqu'ils ne seront pas appliqués dans notre étude** (cf 3.2.2.1 Choix de la méthode).

Définissons tout d'abord la fonction de hasard h :

$$h(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} = -\frac{d}{dt} ln(S(t))$$

Le modèle de Cox est un modèle à hasard proportionnel, c'est-à-dire que l'on cherche à exprimer la fonction de survie d'une sous-population en fonction d'une fonction de survie de base. En notant z le vecteur des variables explicatives et θ celui des paramètres à estimer, on recherche une relation de la forme :

$$h(x|z, \theta) = exp(-z'\theta)h_0(x)$$

avec h_0 la fonction de hasard de base. Le modèle de Cox correspond au cas où la fonction de hasard est inconnue, c'est un modèle semi-paramétrique.

En pratique, l'hypothèse de hasard proportionnel n'est pas toujours vérifiée et les paramètres peuvent dépendre du temps. On peut alors se tourner vers le modèle additif d'Aalen. Ce modèle est non paramétrique et suppose que la fonction de survie est de la forme suivante :

$$h(t) = X^T \beta(t)$$

❖ *Modèles linéaires généralisés*

Les modèles linéaires généralisés sont très souvent appliqués pour étudier l'impact de variables sur le coût et la fréquence moyens mais ne sont généralement pas utilisés en modèle de durée. **La sélection de ce modèle ainsi que le contexte d'utilisation seront expliqués ultérieurement (cf 3.2.2.1 Choix de la méthode)**.

Le modèle linéaire généralisé (GLM) comprend trois composantes : la variable Y à expliquer, les variables explicatives X_i et la fonction de lien g.

➤ *Variable à expliquer*

Le GLM repose sur l'hypothèse que la variable à expliquer Y suit une loi appartenant à la famille exponentielle.

Une loi appartient à la famille exponentielle si sa densité $f_{\theta,\Phi}$ s'exprime ainsi :

$$f_{\theta,\Phi}(x) = exp\left(\frac{x\theta - b(\theta)}{a(\Phi)} + c(x, \Phi)\right)$$

où :

— a et c sont des fonctions dérivables ;
— b est trois fois dérivable ;
— b' est inversible ;
— $\theta \in \mathbb{R}$ est le paramètre naturel et $\Phi \in \mathbb{R}$ est le paramètre de dispersion.

Les lois gamma, normale et exponentielle font par exemple partie de la famille exponentielle.

Si une variable aléatoire X suit une loi appartenant à la famille exponentielle,

$$E(X) = b'(\theta) \; et \; V(X) = b''(\theta)\Phi$$

➤ *Les variables explicatives*

On cherche à établir une relation entre la variable à expliquer Y et une combinaison linéaire des variables explicatives X_i.

➤ *La fonction de lien g*

Dans un modèle linéaire généralisé, on ne cherche pas directement la relation de la combinaison linéaire des variables explicatives avec l'espérance de Y $E(Y|X)$, mais avec $g(E(Y|X))$:

$$g(E(Y|X)) = X^T\beta$$

où β est la matrice des coefficients à estimer.

En notant μ l'espérance, la fonction de lien canonique g_c est telle que $g_c(\mu) = \theta$.

➤ *Estimation des paramètres*

Dans le cas de la famille exponentielle, la log-vraisemblance s'écrit :

$$ln(L(\theta_1, ..., \theta_n, \Phi, y_1, ..., y_n)) = \sum_{i=1}^{n} \left(\frac{y_i\theta_i - b(\theta_i)}{a(\Phi)} + c(y_i, \Phi) \right)$$

En dérivant par rapport aux paramètres β_j, on obtient les équations du score :

$$\sum_{i=1}^{n} \frac{\partial lnL_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \eta_i} \times \frac{y_i - \mu_i}{V(Y_i)} X_{ij} = 0$$

avec $\mu_i = E(Y_i)$ et $\eta_i = g(\mu_i)$

Ces équations ne peuvent pas être résolues analytiquement. Les solutions peuvent être approchées par l'algorithme de Newton-Raphson.

3.1.6.2 Tests et évaluation du modèle

❖ *Tests de significativité*

Il existe trois tests de significativité : le test de Wald, le test du rapport de vraisemblance et le test du score. L'hypothèse nulle H_0 est la nullité du coefficient.

La statistique de Wald est :

$$S_W = \frac{\hat{\beta}_j^{\,2}}{V(\hat{\beta})}$$

Sous H_0, S_W suit une loi du χ^2 à un degré de liberté.

Le test du rapport de vraisemblance compare deux modèles emboîtés. Notons $\hat{L}$ et $\widetilde{L}$ les vraisemblances des modèles sans et avec contrainte. La statistique du test est :

$$S_r = 2 \times ln\left(\frac{\hat{L}}{\widetilde{L}}\right) = 2\left(ln(\hat{L}) - ln(\widetilde{L})\right)$$

Dans le cas où seul un coefficient est ajouté, cette statistique suit sous H_0 une loi du χ^2 à un degré de liberté.

Quant au test du score, il est basé sur la dérivée de $ln(L)$, appelée score et notée l'. La statistique est :

$$S_s = \left(l'(\hat{\beta})\right)^t \left[V(l'(\hat{\beta})\right]^{-1} l'(\hat{\beta})$$

Sous H_0, S_s suit une loi du χ^2 à un degré de liberté.

❖ *Qualité d'ajustement*

➤ *Statistique de Pearson*

La statistique de Pearson est définie ainsi :

$$\chi^2 = \sum_{i=1}^{n} \frac{(Y_i - \hat{Y}_i)^2}{V(\hat{Y}_i)}$$

Sous l'hypothèse nulle d'une bonne adéquation du modèle, la statistique de Pearson standardisée ($\chi^2* = \frac{\chi^2}{\Phi}$) suit une loi du χ^2 à $n - p$ degrés de liberté, où p représente le nombre de variables explicatives.

➤ *Déviance*

L'idée de ce test est de comparer le modèle avec un modèle dans lequel il y a autant de paramètres que d'observations : le modèle saturé. Dans ce modèle, $E(Y_i|X_i) = Y_i$.

La déviance est :

$$D = 2 \times \Phi \times (ln(L_{satur}) - ln(L))$$

Sous l'hypothèse nulle d'une bonne adéquation du modèle, la déviance standardisée ($D* = \frac{D}{\Phi}$) suit une loi du χ^2 à $n - p$ degrés de liberté.

De manière empirique, l'ajustement est bon si $\frac{D}{n-p-1}$ est proche de 1.

➤ *Analyse des résidus*

Les résidus lignes sont définis ainsi : $\hat{r_i} = y_i - \hat{\mu_i}$. Ils présentent l'inconvénient de ne pas avoir la même variance et sont donc difficilement interprétables. C'est pourquoi les résidus de Pearson standardisés ont été introduits :

$$r_{Pi} = \frac{\sqrt{w_i}(y_i - \mu_i)}{\sqrt{V(\mu_i)}}$$

On remarque que : $E(r_{Pi}) = 0$ et $V(r_{Pi}) = \frac{w_i V(Y_i)}{V(\mu_i)} = \Phi$.

Graphiquement, les résidus devraient être répartis aléatoirement autour de 0.

Remarque : D'autres types de résidus ont été définis, comme les résidus de déviance.

3.2 Application

Maintenant que les modèles retenus ont été présentés, nous pouvons choisir une méthode puis l'appliquer.

3.2.1 Etude du maintien en fonction de l'âge et de l'ancienneté

Commençons par l'étude du maintien en deux dimensions.

3.2.1.1 Choix de la méthode

❖ *Choix de la méthode de calcul des taux bruts*

Deux méthodes ont été préalablement retenues : les méthodes de Kaplan-Meier et Turnbull.

La méthode de Turnbull présente l'avantage d'être applicable même si on ne connaît pas les dates exactes des événements. Pour une petite part des sinistres, nous n'avons pas pu remonter jusqu'à la date de début d'invalidité. Cependant, cela demanderait d'étudier la survie sur des périodes de 3 ans (période maximale avant le passage), ce qui ne serait pas pertinent pour le calcul du tarif. Néanmoins, le calcul par période peut être moins lourd que celui de Kaplan-Meier (à nuancer en raison des itérations). Mais le problème ne se pose pas ici puisque nous avons un volume de données relativement faible.

D'autre part, l'estimateur de Kaplan-Meier est en temps continu et est plus proche des données individuelles. Il est donc plus précis que l'estimateur de Turnbull. C'est pourquoi nous avons retenu l'estimateur de Kaplan-Meier.

❖ *Modèle à référence externe*

Le volume de données dont nous disposons est assez faible pour certains âges et anciennetés, notamment aux âges très jeunes :

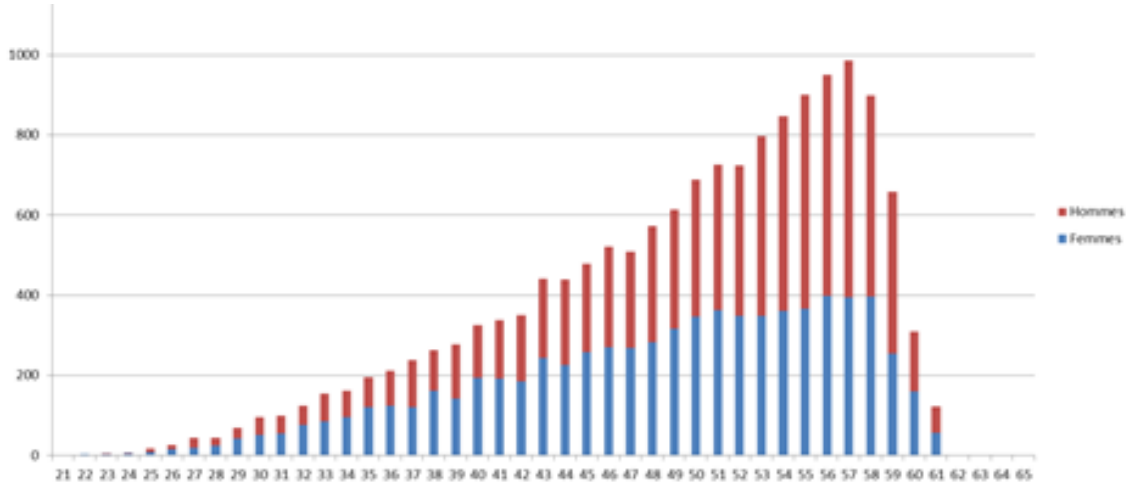


FIGURE 3.3 – Âge à l'entrée en invalidité

C'est pourquoi nous avons songé à appliquer le modèle de Brass pour fiabiliser la structure de la fonction de survie. Ce modèle est d'habitude utilisé pour construire des tables de mortalité en présence de peu de données. Dans notre étude, si la principale cause de sortie est bien le décès, ce n'est pas l'unique motif. Néanmoins, cette cause est très largement majoritaire (cf 2.2.2.2 Statistiques du maintien). De plus, l'hypothèse selon laquelle les logits des taux de sortie observés peuvent s'exprimer grâce à la régression des logits des taux de sortie d'une table de référence sera validée par la suite (cf 3.2.1.3 Ajustement par référence externe).

Remarque : **Cette étape de fiabilisation de la fonction de survie permet aussi de lisser les taux bruts**.

3.2.1.2 Calcul des taux bruts par l'estimateur de Kaplan-Meier

Dans un premier temps, les taux bruts ont été calculés par l'estimateur de Kaplan-Meier. La fonction de survie a été estimée pour chaque âge d'entrée en invalidité. Des travaux ont cherché à généraliser l'estimateur de Kaplan-Meier en dimension 2. Cependant, selon F. Planchet et P. Thérond, l'estimation en une dimension reste la méthode la plus robuste et la perte d'information due à la non prise en compte de la loi conjointe selon les deux dimensions est faible.

La fonction de survie a été calculée pour chaque âge. En présence de peu de données, il est courant de regrouper les âges par classes lorsqu'on étudie l'incapacité. En revanche, cela s'avère plus compliqué pour l'invalidité. En effet, les bornes de la durée dépendent de l'âge d'entrée en invalidité : la période de versement de la pension est souvent limitée au départ à la retraite.

Comme la fonction LIFETEST de SAS, qui permet de calculer les taux par l'estimateur de Kaplan-Meier, ne prend pas en compte les troncatures à gauche, la méthode a été implémentée (cf Annexe 1 : Code Kaplan-Meier). De plus, les intervalles de confiance ont été calculés grâce à l'estimateur de la variance de Greenwood, présenté précédemment.

Cependant, comme le volume de données est faible pour certains âges (notamment les plus jeunes), les intervalles de confiance étaient larges. Voici par exemple la courbe des taux bruts obtenus pour l'âge de 57 ans, qui correspond à l'âge pour lequel on disposait du volume de données le plus important :

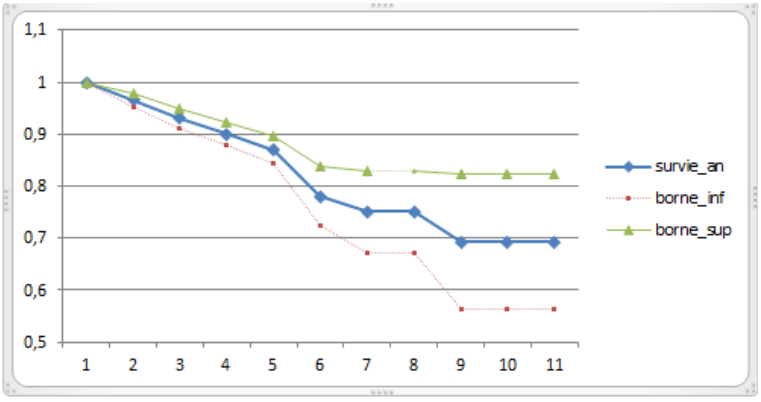


FIGURE 3.4 – Taux bruts pour l'âge de 57 ans obtenus par l'estimateur de Kaplan-Meier

Même à 57 ans, les intervalles de confiance sont larges pour les anciennetés élevées.

Par ailleurs, on constate que la fonction de survie est relativement élevée, même pour les anciennetés élevées. Cela suggère que beaucoup d'assurés restent invalides jusqu'à leur départ à la retraite.

3.2.1.3 Ajustement par référence externe

Afin de pallier le manque de données, le modèle de Brass a été appliqué, en utilisant la table du BCAC comme référence externe.

❖ *Présentation de la démarche*

Pour rappel, cette méthode consiste à faire une régression des logits des taux de sortie observés par les logits des taux de sortie du BCAC :

$$ln\left(\frac{q_{xt}}{1-q_{xt}}\right) = a \times ln\left(\frac{q_{xt}^{BCAC}}{1-q_{xt}^{BCAC}}\right) + b$$

avec $q_{xt} = P_x(T < t+1 | T \geq t)$ [6] $= 1 - \frac{S_x(t+1)}{S_x(t)}$

Le premier nuage de points représentant les logits observés en fonction des logits du BCAC était assez étalé et ne semblait pas décrire une droite. Néanmoins, en traçant les points par classe d'âge, on se rapprochait bien d'une relation linéaire. C'est pourquoi nous avons appliqué la variante du modèle dans laquelle les paramètres dépendent de l'âge :

$$lg_x(t) = a_x \times lg_x^{BCAC}(t) + b_x$$

Les âges ont été regroupés en 4 classes afin de conserver un volume suffisant.

D'autre part, certains points ont été calculés à partir d'une exposition très faible, donc leur estimation est moins robuste. Afin de ne pas accorder trop de poids à ces points, la fonction de perte utilisée lors du calcul des coefficients a et b a été pondérée par l'exposition :

$$\phi(a,b) = \sum_{x,t} E_{xt} \times (lg_{xt}(a,b) - \hat{lg_{xt}})^2$$

avec E_{xt} l'exposition à l'âge x pour l'ancienneté t.

Pour un assuré, l'exposition sur $[t-1, t[$ a été calculée ainsi :

$$E_{xt} = max\,(0\ ;\ min(T,t) - max(seuil\_troncature, t-1)\ )$$

où $seuil\_troncature$ correspond à la durée en invalidité avant le début de l'observation.

6. Dans le version continue à gauche introduite initialement

❖ *Optimisation des classes d'âges*

Plusieurs découpages en classes d'âges ont été testés, en fixant la classe 24-34 ans afin de disposer d'un effectif suffisant.

Pour chaque découpage, on a procédé aux étapes suivantes :

- Pour chaque classe :
  - Régression des logits grâce à la procédure *REG* sur SAS
  - Calcul de la somme des carrés des résidus pondérés par l'exposition (SCRP)
- Somme des SCRP de chaque classe du découpage

Finalement, on a retenu le découpage minimisant la somme des SCRP : 24-34, 35-46, 47-54, 55-63.

❖ *Calcul de la fonction de survie*

La régression permet d'estimer $y = logit(q_x(t))$.
On en déduit $q_x(t)$ en appliquant l'inverse f de la fonction logistique :

$$f(x) = \frac{exp(y)}{1 + exp(y)}$$

Enfin, par définition : $S_x(t+1) = S_x(t)(1 - q_x(t))$.

3.2.1.4 Validation du modèle et comparaison avec le BCAC

❖ *Validation du modèle*

Afin de valider le modèle, nous avons tracé les logits observés, les logits prédits ainsi que les intervalles de confiance à 95%. Voici par exemple le graphique obtenu pour la classe d'âges 47-54 ans :

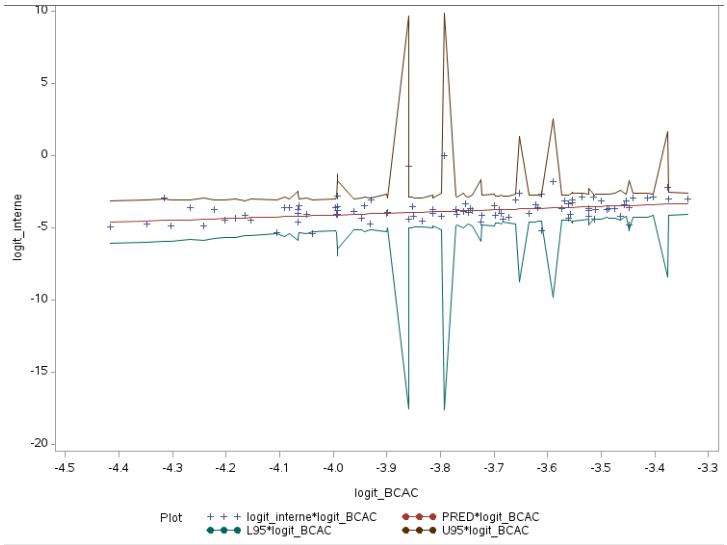


FIGURE 3.5 – Régression des logits des taux de sortie bruts sur les logits des taux de sortie du BCAC pour les 47-54 ans

La très grande majorité des points est située dans l'intervalle de confiance à 95%, ce qui permet de valider le modèle. De plus, les points sont situés pour la plupart très proches de la droite et les points les plus éloignés sont ceux pour lesquels l'exposition est faible et l'intervalle de confiance est donc élevé.

❖ *Comparaison de la fonction de survie avec celle du BCAC*

La fonction de survie obtenue est inférieure à celle du BCAC pour les âges très jeunes mais supérieure pour les âges plus élevés. Les graphiques ci-dessous représentent les taux de survie bruts (en rouge), la fonction de survie ajustée (en bleu) et celle du BCAC (en vert) :

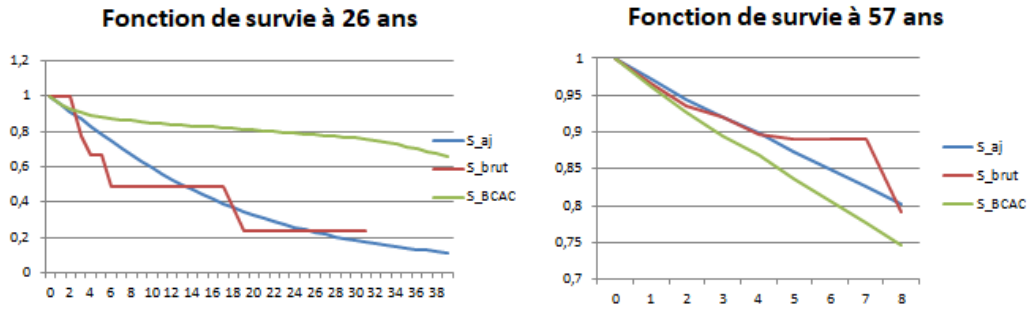


FIGURE 3.6 – Comparaison des fonctions de survie obtenues avec celles du BCAC

La fonction de survie à 26 ans est bien plus faible que celle du BCAC. Comme les âges très jeunes sont moins représentés, il est important de se demander si la survie reste fiable. D'une part, dans le nuage de points représentant les logits observés en fonction des logits du BCAC (cf Annexe 3 : Spécificité de la classe d'âges 24-34 ans), la majorité des points de la classe d'âges se distinguait bien des autres classes. D'autre part, les poids ont permis d'accorder plus d'importance aux points calculés avec les plus gros effectifs et la classe a été fixée à 11 âges afin de calculer les paramètres avec un effectif suffisant. Il semble donc bien se dégager un effet différent pour les âges jeunes.

Par ailleurs, il est intéressant de se demander pourquoi nous observons plus de sorties aux âges très jeunes. Voici les causes de sortie pour cette tranche d'âges :

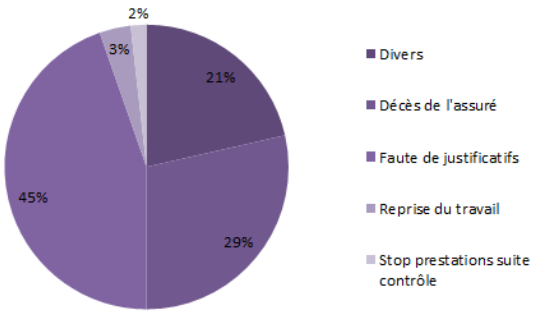


FIGURE 3.7 – Causes de sortie chez les moins de 34 ans

Si l'on compare avec les causes de sorties de l'ensemble de la population (cf 2.2.2.2 Statistiques du maintien), on remarque que le décès est un peu moins représenté et que la faute de justificatif l'ait

au contraire plus. Néanmoins, comme précédemment, la faute de justificatif est sans doute liée soit au décès, soit à la reprise du travail.

Par ailleurs, comme c'est le maintien à l'âge moyen de la population assurée qui sera utilisé dans le tarif, cette différence aura rarement un impact sur le tarif.

Au contraire, la fonction de survie obtenue est plus élevée aux âges plus élevés que celle du BCAC. Néanmoins, il faut garder à l'esprit que celle-ci a pu être surestimée à cause des fautes de justificatifs (cf 3.1.2 Troncatures et censures). Mais le maintien ne sera pris en compte dans le tarif que jusqu'à 61,5 ans pour compenser une éventuelle surestimation due aux fautes de justificatifs.

3.2.2 Segmentation du maintien

Dans le cadre de la tarification, il est important de prendre en compte l'effet des variables explicatives : l'âge, le sexe, la CSP, le secteur d'activité et le département de l'entreprise.

3.2.2.1 Choix de la méthode

La prise en compte des variables explicatives en modèle de durée se fait très souvent grâce au modèle de Cox ou au modèle additif d'Aalen. Ces modèles expriment la fonction de hasard en fonction des variables (cf 3.1.6 Modèles intégrant des variables explicatives). Cependant, l'étude de la fonction de survie en fonction de l'âge et de l'ancienneté seuls s'est déjà révélée difficile en raison du volume de données. Il ne semblait donc pas pertinent d'ajouter des variables.

D'autre part, pour le tarif, nous n'avions pas besoin de toute la fonction de hasard, mais seulement de l'espérance du maintien. Les fonctions de hasard et de survie permettent de prendre en compte toutes les données, y compris celles censurées, mais requièrent un volume suffisant pour obtenir des résultats fiables. Les intervalles de confiance obtenus en dimension 2 étaient déjà relativement larges (cf 3.2.1.2 Calcul des taux bruts par l'estimateur de Kaplan-Meier). A l'inverse, la durée (qui n'est plus une fonction) nécessite moins de données. Cependant, elle n'est connue que pour les sinistres clos non censurés et les durées des sinistres non observés à cause des troncatures ne sont pas connues.

Nous avons donc choisi d'étudier la *durée estimée* :

— Pour les sinistres non censurés, la durée estimée sera la durée réelle.
— Pour les sinistres censurés, on ajoute la durée résiduelle moyenne connaissant le maintien actuel, grâce à la table du BCAC (calcul présenté ci-dessous).

Cette durée estimée sera limitée à 62 ans. En effet, si l'on tenait compte des départs tardifs liés aux garanties viagères dans le calcul de l'espérance, on surestimerait la durée limitée au départ à la retraite.

L'histogramme de la durée estimée se rapprochant des lois classiques de durée (cf 3.2.2.3 Analyses préliminaires et démarche), il a été possible d'étudier l'impact des variables explicatives grâce à un modèle linéaire généralisé.

Néanmoins, l'objectif était d'étudier l'impact des variables explicatives, donc il fallait veiller à conserver l'influence des variables, malgré un prolongement uniforme. C'est pourquoi nous avons commencé par des tests de sensibilité afin de déterminer le périmètre d'étude.

❖ *Calcul de l'espérance résiduelle de maintien*

La durée résiduelle moyenne est estimée à l'aide de la table du BCAC, qui indique les fonctions de survie par année. Ainsi, c'est l'espérance du nombre d'années restantes en invalidité qui a été calculée.

Par définition de l'espérance conditionnelle, en notant T_k la durée résiduelle :

$$E(T_k) = \frac{E((T-k)1_{T\geq k})}{P(T \geq k)}$$

avec :

$$E((T-k)1_{T\geq k}) = \sum_{t=k+1}^{+\infty} S(k)$$

En effet, si $T \geq k$:

$$\sum_{t=k+1}^{+\infty} S(t) = \sum_{t=k+1}^{+\infty} E(1_{T\geq t}) = E\left(\sum_{t=k+1}^{+\infty} 1_{T\geq t}\right) = E\left(\sum_{t=k+1}^{T} 1\right) = E(T-k)$$

3.2.2.2 Détermination du périmètre d'étude

Le prolongement de la durée des sinistres censurés pouvait biaiser l'influence des variables. En effet, les tables du BCAC n'intègrent pas de variables explicatives, donc le prolongement était uniforme pour tous les individus. Un prolongement assez long pouvait modifier l'impact des variables. Une première idée a donc été de ne conserver que les lignes pour lesquelles la durée résiduelle estimée représente au maximum x% de la durée totale estimée. Puis des tests de sensibilité ont été réalisés à la fois sur le maintien moyen et sur l'influence des variables explicatives.

- ❖ *Tests d'influence sur la durée moyenne*

La durée moyenne en année (par âge) a été calculée en fonction du seuil x :

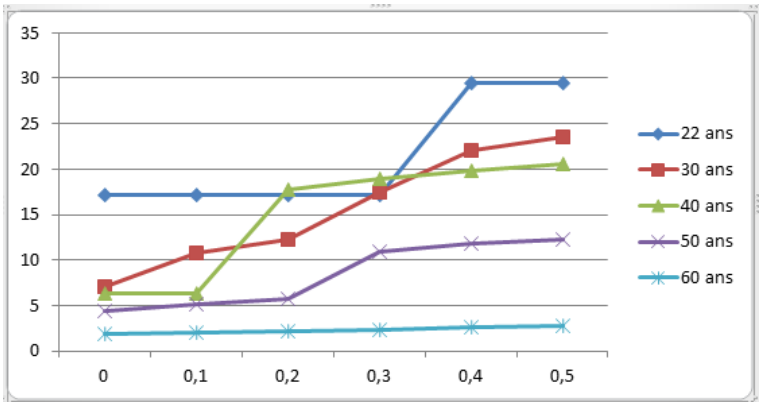


FIGURE 3.8 – Durée moyenne en fonction du seuil

On remarque que si l'on prend un seuil trop petit (qui permet de limiter le biais de l'impact des variables), la durée moyenne est diminuée. En ajoutant un seuil, on élimine des sinistres longs, ce qui risque de biaiser la durée globale.

- ❖ *Tests d'influence sur l'impact des variables explicatives :*

L'influence des variables explicatives devait aussi être conservée. Pour l'âge de 56 ans (897 données), le rapport entre la durée moyenne chez les femmes et chez les hommes a été calculé en fonction du seuil :

seuil	Quotient durée F/H
0	0,986144482
0,1	0,982444214
0,2	0,882174389
0,3	0,871005463
0,4	1,057278275
0,5	1,066203517
0,6	1,051821945
0,7	1,042405772
0,8	1,037777247
0,9	1,034713403
1	1,033537995

FIGURE 3.9 – Rapport des durées moyennes des femmes et des hommes en fonction du seuil

On remarque que selon le seuil, la durée moyenne est plus importante chez les femmes ou chez les hommes. Ainsi, le seuil risque de biaiser l'influence des variables.

Si dans une segmentation donnée, le maintien connu est plus avancé, la durée totale estimée sera plus importante, sans que ce soit lié aux variables explicatives.

Afin de ne pas biaiser l'impact des variables explicatives, l'impact des variables a été étudié seulement sur les sinistres non censurés et ceux clos censurés après 59 ans (retraite, faute de justificatifs avec suspicion de retraite...). Seul l'impact des variables explicatives après 59 ans serait effacé. Nous conservions ainsi 6358 lignes.

Cependant, en ne conservant que les sinistres observés clos et donc possiblement plus courts en moyenne, la durée serait biaisée. Nous avons supposé que l'impact des variables explicatives était le même sur les sinistres clos que sur les sinistres totaux et devions ajouter un coefficient multiplicatif pour retrouver la durée moyenne globale. Cette durée moyenne globale a été calculée à partir de nos données internes grâce au modèle de Brass.

Néanmoins, l'impact des variables explicatives sur les sinistres observés et clos (surreprésentation des sinistres courts probable) n'est pas forcément le même que sur les sinistres globaux. Afin de vérifier cette hypothèse, nous avons calculé les taux bruts de survie pour plusieurs sous-populations grâce à l'estimateur de Kaplan-Meier, pour les âges au volume suffisant. Les effets obtenus ont ensuite été comparés aux résultats du GLM.

3.2.2.3 Analyses préliminaires et démarche

❖ *Effet de l'âge*

En moyenne, la durée est décroissante en fonction de l'âge d'entrée en invalidité, ce qui est cohérent : la durée maximale (limitée par le départ à la retraite) diminue. L'âge n'a pas le même effet

aux âges très jeunes (avant 34 ans) qu'au reste des âges. Bien que l'effectif soit faible aux âges jeunes, l'histogramme de la *durée relative* (rapport entre la durée estimée et la durée maximale) ci-dessous montre un nombre élevé de sinistres très courts. Comme nous n'étudions que les sinistres clos, cette proportion de sinistres courts est à nuancer, mais est en accord avec la fonction de survie trouvée précédemment.

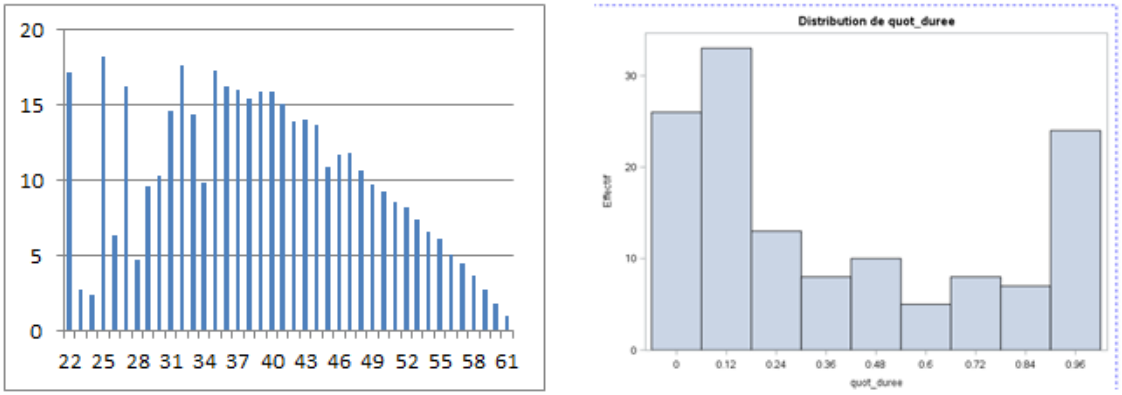


FIGURE 3.10 – Durée moyenne en fonction de l'âge (à gauche) et histogramme de la durée relative pour les moins de 34 ans (à droite)

C'est pourquoi nous avons décidé d'étudier dans un premier temps la durée estimée des plus de 34 ans, puis nous souhaitions faire un modèle très simple pour les moins de 34 ans en raison du petit volume de données (modèle à une variable explicative).

D'autre part, il était de nouveau impossible de faire des classes d'âge, comme les bornes dépendent de l'âge.

❖ *Segmentation du secteur d'activité et du département*

Le code APE et le département possèdent de nombreuses modalités. Il convenait donc de créer des classes avant de procéder au GLM, afin de ne pas estimer trop de coefficients. Ces variables ont été discrétisées à l'aide d'arbres de décision (cf 4.1.2.1 Présentation de l'arbre CART), grâce à la procédure *HPSPLIT* sur SAS. Afin de choisir le nombre optimal de feuilles de l'arbre, on trace l'erreur de prédiction en fonction du nombre de feuilles de l'arbre :

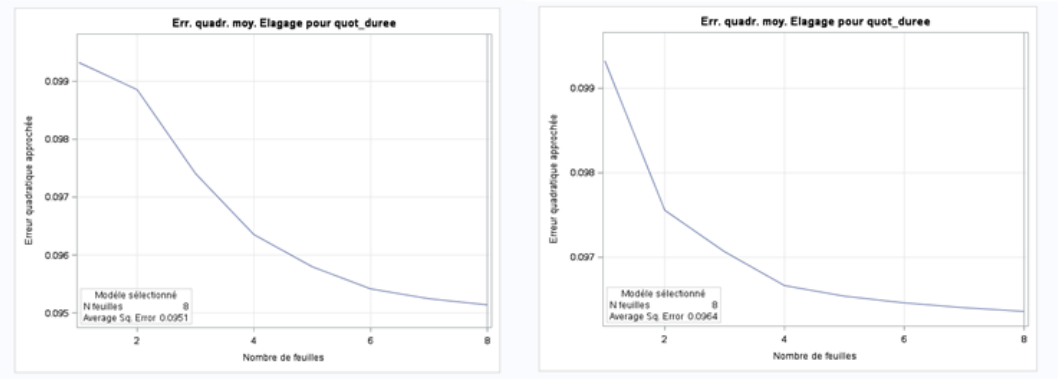


FIGURE 3.11 – Evolution de l'erreur de prédiction de la durée relative en fonction du nombre de feuilles pour le secteur d'activité (à gauche) et pour le département (à droite)

La procédure SAS fonctionne par validation croisée et renvoie le nombre de feuilles optimal. Cependant, certaines classes formées étaient très peu représentées. Nous avons donc choisi de ne pas ajouter les feuilles qui ne diminuaient pas sensiblement l'erreur. Ainsi, six classes de secteurs d'activité et quatre classes de départements ont été formées.

La composition des classes n'est pas précisée par souci de confidentialité mais des exemples pourront être donnés lors de l'interprétation finale.

❖ *Choix de la loi et de la fonction de lien*

L'adéquation des lois traditionnelles (appartenant à la famille exponentielle) utilisées en modèle de durée a été testée. Voici les histogrammes obtenus :

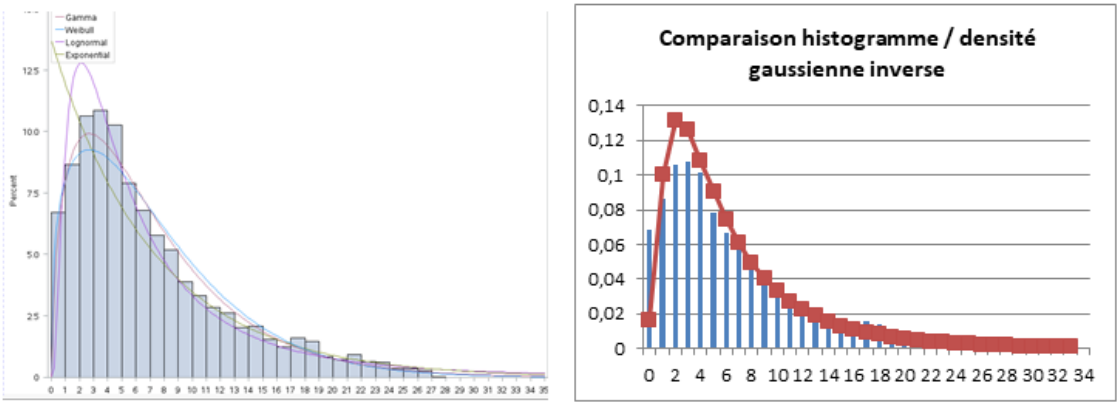


FIGURE 3.12 – Adéquation des lois classiques de durée

L'adéquation n'est pas parfaite. Les tests de Kolmogorov-Smirnov [7] donnent des p-value inférieures à 0,05 donc on rejette l'hypothèse d'égalité des fonctions de répartition. Bien qu'imparfaite, la loi gamma semble convenir, donc nous avons testé un GLM log-gamma. Une amélioration ultérieure

7. Test dont l'hypothèse nulle est l'égalité de la fonction de répartition des données et de celle de la loi testée

permettra d'ailleurs d'obtenir une meilleure adéquation (cf 3.2.2.4 Inadéquation du modèle).

❖ *Analyse des corrélations*

Les variables sont un peu corrélées mais les VCramer (présentés en **Annexe 4 : Présentation du V de Cramér**) restent inférieurs à 0,7.

❖ *Définition de l'individu de référence*

La modalité la plus présente de chaque variable a été choisie comme référence.

❖ *Création de la base d'échantillon et de la base de validation*

La base a été séparée en un échantillon de validation représentant les 2/3 de la base et un échantillon test.

❖ *Sélection des variables et regroupement des modalités*

Pour sélectionner les variables et regrouper les modalités, nous avons procédé ainsi (sélection descendante, "backward selection") :

— le modèle a été établi en sélectionnant l'ensemble des variables ;
— le modèle a été réestimé en regroupant la modalité la moins significative avec la modalité ayant le paramètre estimé le plus proche ;
— l'étape précédente a été renouvelée jusqu'à ce que l'ensemble des variables soit significatif.

Remarque : Il existe d'autres méthodes de sélection des variables. Parmi les plus courantes, la méthode ascendante ("forward selection") procède à l'inverse en commençant avec une seule variable et en introduisant une à une des variables explicatives. L'inconvénient majeur de ces deux premières méthodes est de ne pas pouvoir respectivement réintroduire une variable éliminée et éliminer une variable ajoutée. En effet, en raison des corrélations entre les variables, le caractère significatif de certaines variables peut varier selon les autres variables présentes. La méthode Stepwise est une méthode proche de la méthode ascendante, mais elle permet en plus à chaque étape d'éliminer une variable introduite précédemment qui est devenue non significative.

La procédure automatique stepwise ne renvoie pas toujours la sélection de variables optimale, en raison des multicorrélations. De plus, la méthode descendante permet de prendre en compte toutes les variables. C'est pourquoi cette méthode a été retenue.

3.2.2.4 Inadéquation du modèle

Les résultats n'ont pas permis de valider le modèle.

La moyenne de la valeur absolue des écarts (MAE) et la racine de la moyenne du carré des écarts (RMSE) ont été calculés sur l'échantillon test :

RMSE	4,05791838
MAE	2,75169009

FIGURE 3.13 – MAE et RMSE calculées sur l'échantillon test

Les écarts sont considérables : en moyenne, la valeur absolue des écarts est de 2,75 ans.

D'autre part, les résidus devraient être répartis aléatoirement autour de 0. Ce n'est pas le cas :

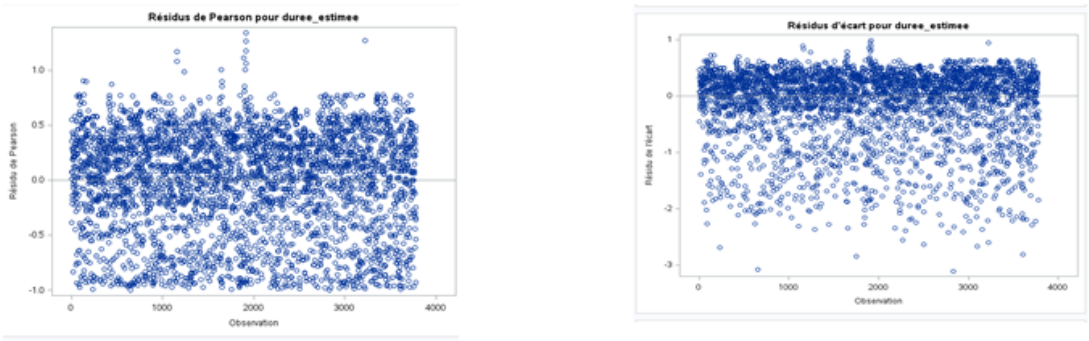


FIGURE 3.14 – Résidus

On remarque qu'une grande part des résidus sont positifs : le modèle sous-estime très souvent la durée. Il n'est donc pas prudent.

Enfin, le test basé sur la déviance rejette l'hypothèse d'une bonne adéquation. De manière empirique, le rapport $\frac{D}{n-p-1}$ est très éloigné de 1, ce qui annonçait cette inadéquation.

Nous avons donc cherché à comprendre pourquoi le modèle n'était pas adéquat.

❖ *Recherche de la cause de l'inadéquation du modèle*

L'une des premières hypothèses a été l'hypothèse de loi gamma (cf 3.2.2.4 Analyses préliminaires et démarche). En effet, l'adéquation semblait suffisante mais n'était pas parfaite. En n'étudiant que les âges supérieurs à 45 ans, l'adéquation est bien meilleure :

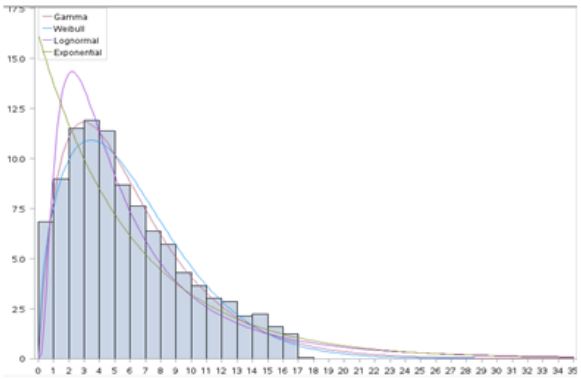


FIGURE 3.15 – Adéquation des lois pour les plus de 45 ans

La loi gamma épouse bien la distribution empirique. En revanche, en relançant le GLM sur cette partie de la population, les résultats restaient insatisfaisants. Cela ne semblait donc pas être la cause principale de l'inadéquation du modèle.

L'une des autres causes possibles est la dispersion de la variable durée. Les durées relatives[8] observées pour les invalides du secteur de l'industrie manufacturière sont représentées ci-dessous à gauche en fonction de l'âge d'entrée :

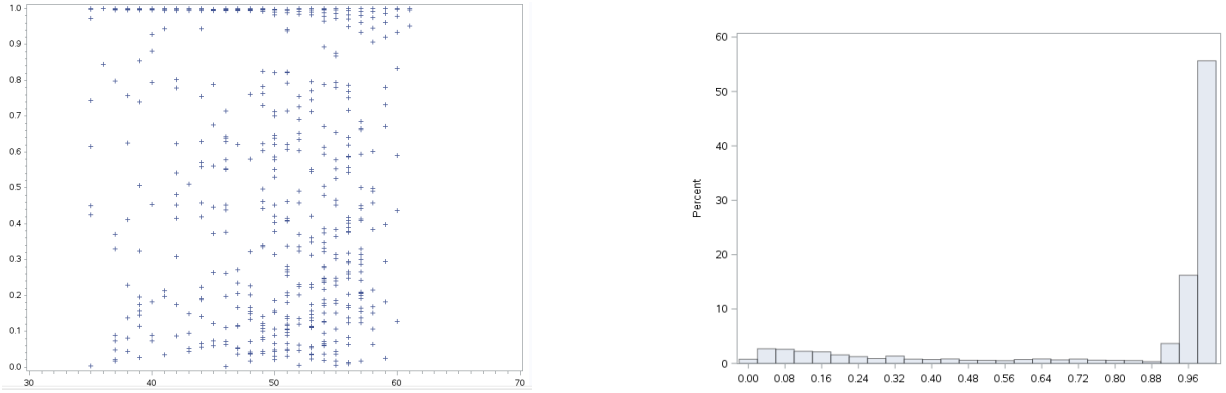


FIGURE 3.16 – Dispersion de la durée relative des invalides du secteur de l'industrie manufacturière (à gauche) et histogramme de la durée relative (à droite)

On remarque que pour un même âge, les durées relatives observées sont très dispersées, malgré un même secteur d'activité. Cependant, la majorité des durées relatives est très proche de 1 (cf Figure 21 à droite). Le reste des durées relatives est assez étalé avec toutefois un petit amas de sinistres très courts.

Ces graphiques expliquent l'allure des résidus observés précédemment : les durées sont assez dispersées mais majoritairement élevées. Les sinistres courts, moins nombreux, viennent baisser la moyenne. C'est pourquoi nous observons un grand nombre de résidus positifs.

Cette surdispersion n'est pas si surprenante. En effet, la cause principale de sortie de l'invalidité est le décès. Or, il est difficile d'exprimer la durée de vie en fonction de variables explicatives.

Comme le maintien est difficilement explicable par des variables et qu'il est la plupart du temps très élevé, nous avons choisi de **privilégier la segmentation de l'incidence**. Il est malgré tout intéressant d'analyser l'impact des variables donné par le GLM.

3.2.2.5 Comparaison des résultats avec ceux observés par Kaplan-Meier

❖ *Présentation de la démarche*

Le modèle GLM a malgré tout estimé l'impact des variables, en révélant la significativé de certaines. Comme il n'ajuste pas bien les données et que nous n'avons étudié que les sinistres non censurés et clos non censurés après 59 ans, il est intéressant de vérifier si l'influence des variables

8. cf III. 2.2.1 Choix de la méthode

n'a pas été biaisée.

Pour cela, les taux bruts de survie ont été estimés par Kaplan-Meier sur certaines sous-populations. En revanche, les âges sur lesquels les taux ont été calculés devaient avoir un effectif suffisant pour que les résultats soient fiables. C'est pourquoi nous nous sommes concentrés sur les âges 56-58 qui ont les effectifs les plus élevés. De plus, le secteur d'activité et le département n'ont été divisés qu'en deux classes afin de garder un effectif suffisant. Elles ont été formées de la même manière que précédemment.

❖ *Impact des variables révélés par le GLM*

Les impacts des variables révélés par le GLM sont assez surprenants : en effet, le sexe n'est pas significatif. Or, comme la principale cause de sortie est le décès, on aurait pu penser que le sexe serait significatif. De même, la CSP n'est pas significative selon le GLM. Il est surprenant que les cadres ne se maintiennent pas davantage que les non cadres. L'âge, le secteur d'activité et le département sont en revanche bien significatifs.

❖ *Comparaison avec les résultats de Kaplan-Meier*

A 57 ans, les graphiques obtenus (**cf Annexe 5 : Analyse de l'impact des variables sur les taux bruts de survie estimés par Kaplan-Meier**) montrent que, comme le prédisait le GLM :

— la CSP ne semble pas significative ;
— le département et le secteur d'activité semblent influents ;

Cependant, ces résultats sont à prendre avec précaution. En effet, à la différence du GLM, on ne compare pas le maintien **toutes choses égales par ailleurs**. Ainsi, on pourrait penser au vu de la Figure 31 (**cf Annexe 5 : Analyse de l'impact des variables sur les taux bruts de survie estimés par Kaplan-Meier**) que les femmes se maintiennent plus que les hommes toutes choses égales par ailleurs. Néanmoins, en étudiant cette sous-population, on constate que 83% des femmes entrées à 57 ans font partie du secteur d'activité 2. Le volume de données ne permet pas d'étudier l'influence des variables toutes choses égales par ailleurs. A 58 ans, nous disposons néanmoins de suffisamment de données pour étudier l'impact du sexe sur le maintien pour les assurés du secteur d'activité 2 :

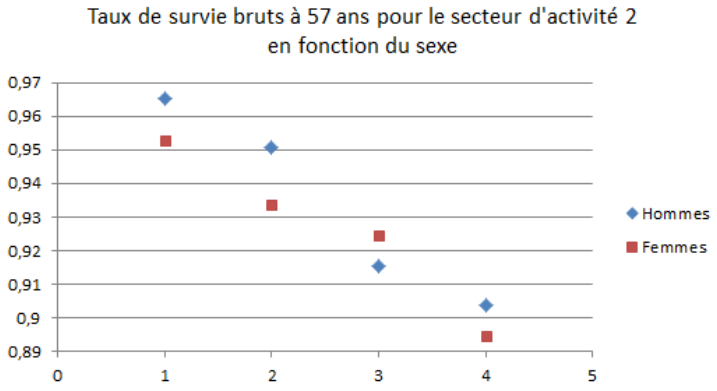


FIGURE 3.17 – Influence du sexe sur le maintien à 58 ans pour le secteur d'activité 2

Ainsi le sexe paraît peu influent sur la durée moyenne. La moyenne du maintien est d'environ 3,7 pour les deux sexes.

En conclusion, l'impact n'a été testé qu'à certains âges et pas toutes choses égales par ailleurs mais les taux bruts semblent bien confirmer l'impact des variables révélé par le GLM.

4 Modélisation de l'incidence en invalidité indirecte

Une fois le maintien modélisé, il reste à étudier la deuxième grande composante du tarif : l'incidence en invalidité.

Très souvent, l'invalidité fait suite à de l'incapacité. Appelons ce type d'invalidité l'invalidité indirecte, par opposition à l'invalidité directe.

4.1 Présentation des modèles

Cette section vise à présenter des modèles permettant d'étudier l'incidence en invalidité indirecte : la régression logistique et les arbres de classification.

4.1.1 Régression logistique

4.1.1.1 Présentation de la méthode

La régression logistique est un cas particulier de modèle linéaire généralisé (cf 3.1.6 Modèles intégrant des variables explicatives) dans lequel la variable à modéliser Y est une variable binaire prenant les valeurs 0 ou 1. Or, en reprenant les notations introduites précédemment :

$$E(Y|X) = 1 \times P(Y = 1|X) + 0 \times P(Y = 0|X) = P(Y = 1|X)$$

Ainsi, la relation cherchée est de la forme suivante :

$$g(P(Y = 1|X)) = X^T\beta$$

❖ *Fonctions de lien*

Trois fonctions de lien peuvent être utilisées en régression logistique :

— le lien logit : $g(y) = ln\left(\frac{y}{1-y}\right)$
— le lien probit : $g(y) = \phi^{-1}(y)$ où ϕ est la fonction de répartition de la loi normale $N(0,1)$
— le lien log-log : $g(y) = log(-log(1-y))$

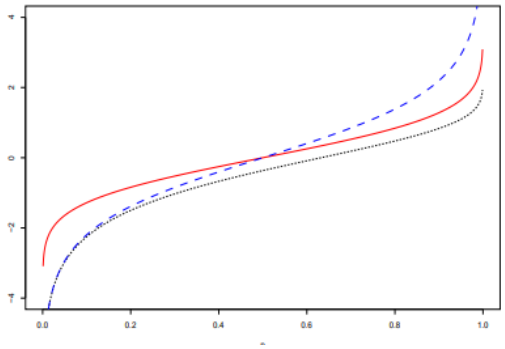


FIGURE 4.1 – Fonctions de lien : logit (bleu), probit(rouge) et log-log (noir)

Si la probabilité à modéliser est asymétrique, la fonction log-log est plus adapée. Par ailleurs, la fonction logit est souvent préférée car elle est plus simple à manipuler.

❖ *Estimation des paramètres*

Comme $Y_i|X_i$ suit une loi de Bernoulli de paramètre $\pi(X_i)$, la vraisemblance vaut :

$$\begin{aligned} L(y_1, ..., y_n) &= \prod_{i=1}^{n} P(Y_i = y_i | X = X_i) \\ &= \prod_{i=1}^{n} \pi(X_i)^{y_i} (1 - \pi(X_i))^{1-y_i} \end{aligned}$$

Donc :

$$ln(L(y_1, ..., y_n)) = \sum_{i=1}^{n} (y_i ln(\pi(X_i)) + (1 - y_i) ln(1 - \pi(X_i)))$$

La résolution des équations $\frac{d}{d\beta_i} lnL(y_1, ..., y_n) = 0$ est réalisée à l'aide de méthodes numériques telles que celle de Newton-Raphson.

4.1.1.2 Adéquation et évaluation des modèles

❖ *Test d'adéquation d'Hosmer-Lemeshow*

Le test d'Hosmer-Lemeshow permet d'évaleur la qualité d'ajustement du modèle. La statistique est basée sur l'écart entre les prédictions et les observations. Pour réaliser ce test, les observations sont réparties en g groupes.

Notons :
- n_i le nombre d'observations dans le $i^{ème}$ groupe ;
- π_i la moyenne des probabilités prédites dans le $i^{ème}$ groupe ;
- o_i le nombre de positifs (passages dans notre étude) réellement observés dans le $i^{ème}$ groupe.

$$\chi^2_{HL} = \sum_{i=1}^{g} \frac{(n_i \pi_i - o_i)^2}{n_i \pi_i (1 - \pi_i)}$$

L'hypothèse nulle (H_0) correspond à un bon ajustement du modèle.
Si l'on réalise le test sur l'échantillon d'apprentissage, la statistique suit sous H_0 une loi du χ^2 à $(g-2)$ degrés de libertés. Donc lorsque la p-value est supérieur à un seuil fixé (en général 5%), on accepte H_0.
Si le test est réalisé sur un échantillon de validation, comme aucun paramètre n'a été estimé à partir de ces données, le degré de liberté n'est plus le même : sous H_0, la statistique suit une loi du χ^2 à g degrés de libertés [1].

❖ *Diagramme de fiabilité*

Le diagramme de fiabilité repose également sur la comparaison des probabilités prédites aux observations mais est un outil plus visuel. Pour le réaliser, on estime les probabilités sur l'échantillon à partir de notre modèle. Puis l'échantillon est à nouveau divisé en groupes, formés à partir des quantiles de la probabilité prédite. Pour chaque groupe, la moyenne des probabilités prédites et la proportion de positifs observés sont calculées. Enfin, on trace les points correspondant à la moyenne prédite en fonction de la proportion observée.

❖ *Comparaison avec le modèle trivial : les pseudos R^2*

Les pseudos R^2 sont basés sur la comparaison des vraisemblances du modèle trivial constitué uniquement de la constance et du modèle construit composé de l'ensemble des variables. Il existe plusieurs formules. Nous présenterons celle du R^2 de McFadden qui est plus adaptée aux modèles logistiques d'après Menard (2000) puisqu'elle ne dépend pas de la proportion de positifs dans la base.

$$R^2 = 1 - \frac{log(L_T)}{log(L_M)}$$

avec L_T et L_M les vraisemblances du modèle trivial et du modèle étudié.

❖ *Tableau de contingence*

Le tableau de contingence est un tableau à double entrée confrontant les prédictions et les observations :

	Y_préd=0	Y_préd=1
Y_obs=0	VN	FP
Y_obs=1	FN	VP

FIGURE 4.2 – Tableau de contingence

Il est constitué des :

— vrais négatifs (VN) : les observations négatives bien prédites négatives ;
— faux positifs (FP) : les observations négatives mais prédites positives ;

1. cf D.W. Hosmer, S. Lemeshow, Applied Logistic Regression, Second Edition, Wiley, 2000, p 204-205

— faux négatifs (FN) : les observations positives mais prédites négatives ;
— vrais positifs (VP) : les observations positives bien prédites positives.

La régression logistique permet d'estimer la probabilité de l'événement. Donc pour compléter ce tableau de contingence, il faut auparavant fixer un seuil limite au-dessus duquel on prédit l'événement. Cet outil dépend ainsi du choix du seuil.

Divers indicateurs peuvent être calculés à partir du tableau de contingence : le taux de bonnes prédictions et de mauvaises prédictions, la sensibilité et la spécificité.
La sensibilité correspond à la proportion de vrais positifs parmi tous les individus observés positifs :

$$sensibilité = \frac{VP}{VP + FN}$$

Quant à la spécificité, il s'agit de la proportion de vrais négatifs parmi les observations négatives :

$$spécificité = \frac{VN}{VN + FP}$$

❖ *Courbe ROC*

Comme expliqué ci-dessus, le tableau de contingence dépend du seuil de probabilité choisi. L'idée de la courbe ROC est de faire varier ce seuil et regarder l'évolution de la sensibilité et de la spécificité. La courbe représente la sensibilité en fonction du "complémentaire" de la spécificité ($1 - spécificité$). Plus la courbe s'éloigne de la bissectrice, plus le modèle est performant.

❖ *Area Under the Curve (AUC) et coefficient de Gini*

L'AUC correspond à l'aire sous la courbe ROC. Elle est comprise entre 0,5 et 1. C'est un indicateur de la qualité de prédiction du modèle : si elle est proche de 0,5, le modèle est peu informatif alors que proche de 1, la discrimination est élevée.

Le coefficient de Gini est :

$$Gini = 2\ AUC - 1$$

Donc plus le Gini est proche de 1, plus le modèle est performant.

4.1.1.3 Présentation des rapports de côte (*odds ratio)*

La côte (*odd*) correspond au rapport entre la probabilité de passage et de non passage. En reprenant les notations précédentes :

$$C(X_i) = \frac{\pi(X_i)}{1 - \pi(X_i)}$$

Le rapport de côte (*odds ratio)* vaut donc :

$$R(X_i, X_i') = \frac{C(X_i)}{C(X_i')} = \frac{\pi(X_i)(1 - \pi(X_i'))}{(1 - \pi(X_i))\pi(X_i')}$$

Cette expression se simplifie lorsque la fonction de lien est la fonction logistique. En effet, dans ce cas :

$$ln\left(\frac{\pi(X_i)}{1 - \pi(X_i)}\right) = X^T\beta$$

Donc : $C(X_i) = \exp(X^T\beta)$ et $R(X_i, X_i') = \exp(X_i^T\beta - X_i'^T\beta)$

4.1.2 Arbre de classification

4.1.2.1 Présentation de l'arbre CART

De même que la régression logistique, des méthodes de machine learning telles que les arbres de décision permettent aussi de modéliser la survenance d'événements. Les arbres présentent l'avantage d'être non paramétriques, de capter des relations non linéaires et de ne nécessiter aucune hypothèse. Leurs avantages et inconvénients seront développés ultérieurement lors du choix de la méthode.(cf 4.2.2 Choix de la méthode).

Les arbres de décision sont une suite de tests sur les variables explicatives aboutissant à une prédiction de la variable à expliquer. Deux types d'arbres sont à distinguer : les arbres de régression pour lesquels la variable à prédire est quantitative et les arbres de classification, pour lesquels elle est qualitative. Ce sont donc ces derniers qui retiendront notre attention.

Un arbre est constitué :
- d'une racine : point initial contenant l'ensemble de la population à segmenter ;
- de branches : tests qui permettent de segmenter la population ;
- de noeuds : sous-échantillons de la population formés après une branche
- de feuilles : sous-échantillons à l'extrémité de l'arbre associés à une prédiction

Dans l'arbre CART, chaque test admet une réponse binaire (segmentation de la population en deux) et ne repose que sur une seule variable explicative.

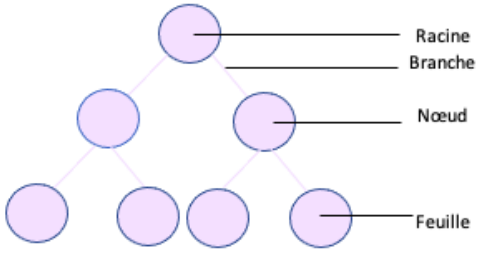


FIGURE 4.3 – Schéma représentant un arbre CART

4.1.2.2 Création de l'arbre maximal

La première étape est la création de l'arbre maximal, qui correspond à l'arbre obtenu sans se soucier du surapprentissage (apprentissage des spécificités de la base d'apprentissage).

❖ *Choix des critères de segmentation*

A chaque noeud, le test optimal doit être choisi afin de minimiser l'hétérogénéité (ou l'**impureté**) dans les deux noeuds fils.

Notons :

- N_1 (resp. N_2) : le noeud fils gauche (resp. droit) ;
- p_1 : la proportion de la population assignée au noeud fils gauche ;
- V : la variable explicative utilisée dans cette segmentation ;
- c : le critère appliqué à V pour la segmentation (seuil si V est quantitative et partage en deux groupes de modalités sinon) ;
- $impuret(V, c, N_i)$: la fonction d'impureté appliquée au noeud N_i après segmentation à l'aide de la variable V et du critère c.

La segmentation optimale consiste à sélectionner la variable explicative V et le critère c qui minimisent l'impureté. En notant $(V, c)^*$ la combinaison optimale,

$$(V, c)^* = \underset{(V,c)}{\arg\min} \left[p_1 \times impureté(V, c, N_1) + (1 - p_1) \times impureté(V, c, N_2) \right]$$

La fonction d'impureté est positive et doit :

- être nulle si une seule modalité de la variable à expliquer est représentée dans la population (population homogène) ;
- être maximale si les différentes modalités sont équireprésentées.

En pratique, la fonction d'impureté la plus couramment utilisée est celle basée sur la concentration de Gini.
En notant K le nombre de classes de la variable à expliquer et p_k la proportion de la population de la classe k dans le noeud considéré, l'hétérogénéité basée sur la concentration de Gini vaut :

$$impureté_{Gini} = \sum_{k=1}^{K} p_k(1 - p_k)$$

Dans notre étude, seules deux classes sont présentes : 0 pour l'absence de passage en invalidité et 1 pour le passage.

D'autres critères peuvent aussi être utilisés pour définir l'hétérogénéité, tels que la fonction d'entropie :

$$entropie(p) = -p\ log(p), \text{ en posant par convention } 0\ log(0) = 0.$$

❖ *Critère d'arrêt et affectation de l'estimation*

L'algorithme ne produit plus de noeuds fils dans les cas suivants :

- toute la population du noeud est de même modalité (noeud homogène) ;
- le nombre d'observations dans le noeud est inférieur à un seuil fixé.

Pour chaque feuille, la modalité attribuée est généralement la plus représentée [2].

2. en l'absence de coûts de mauvais classements, sans se placer dans un cadre Bayésien

4.1.2.3 Elagage et choix de l'arbre optimal

L'élagage consiste à extraire un sous-arbre de l'arbre maximal pour réduire l'erreur due au surapprentissage (cf Annexe 6 : Présentation sommaire du surapprentissage). Comme il n'est pas possible de tester tous les sous-arbres de l'arbre maximal, Breiman et col ont proposé en 1984 de construire une suite de sous-arbres emboîtés et de sélectionner l'arbre optimal parmi cette suite. Il s'agit donc d'un optimum local.

❖ *Construction de la séquence d'arbres emboîtés*

La construction de cette suite d'arbres repose sur la pénalisation de la complexité de l'arbre. Pour un sous-arbre A, notons K_A le nombre de feuilles et E_A le taux d'erreur dans le sous-arbre (moyenne pondérée des taux de mauvaise classification des feuilles). Pour un paramètre de complexité α, la fonction coût-complexité vaut :

$$R(A) = E_A + \alpha K_A$$

Lorsque le paramètre de complexité est fixé, le sous-arbre optimal est celui qui minimise la fonction coût-complexité.

Pour construire la séquence d'arbres, le paramètre de complexité est dans un premier temps nul. Dans ce cas, c'est l'arbre maximal (à K feuilles) qui est optimal. Puis en augmentant la valeur de α, l'une des divisions de l'arbre maximal ne permet plus une réduction de l'erreur suffisante pour compenser le terme de complexité. Les deux feuilles sont donc regroupées et l'on obtient le premier sous-arbre associé à une deuxième valeur de α. Puis on augmente à nouveau afin de construire un autre sous-arbre, associé à une nouvelle valeur de α. En itérant ce procédé, on obtient une suite d'arbres emboîtés et une suite croissante de valeurs de α, $(\alpha_k)_{k=1,\dots,K}$.

❖ *Identification de l'arbre optimal*

Il reste à trouver l'arbre optimal parmi la suite de sous-arbres construite précédemment. Pour cela, si le volume de données a permis la formation d'un échantillon de validation, nous recherchons le sous-arbre qui minimise l'erreur de prédiction sur l'échantillon de validation.

Lorsque le volume de données est insuffisant ou que l'on veut gagner en robustesse, il est possible d'utiliser une méthode de **validation croisée**. Dans ce cas, la population est divisée en V segments. A chaque itération, l'un de ces segments servira à la validation tandis que les autres permettront l'apprentissage et donc la construction de l'arbre maximal de ces sous-arbres.

La suite de sous-arbres est construite pour la suite de paramètres de complexité $(\alpha_k)_{k=1,\dots,K}$ construite initialement. Puis l'erreur de prédiction est calculée sur l'échantillon de validation pour chaque valeur de α.

Les erreurs obtenues à chaque itération sur les V échantillons sont ensuite sommées pour chaque α. Le paramètre de complexité optimal est celui qui minimise la somme des erreurs. Enfin, l'arbre retenu est celui construit à partir de l'ensemble de l'échantillon et associé au paramètre de complexité optimal.

4.2 Application

Maintenant que les méthodes retenues ont été expliquées, il convient de choisir une démarche d'étude et de l'appliquer.

4.2.1 Démarche et détermination de la variable à modéliser

En prévoyance collective, l'effectif d'assurés n'est pas connu. Néanmoins, un précédent mémoire avait étudié l'incidence en incapacité grâce à la Déclaration Sociale Nominative (DSN) [3]. Il nous est donc possible d'étudier le taux de passage des sinistres en invalidité, puis d'en déduire la probabilité d'entrer en invalidité indirecte.

D'autre part, afin d'étudier le taux de passage en invalidité, nous disposons d'une table dans laquelle chaque ligne correspond à un épisode d'incapacité qui donne lieu ou non à un passage. Cependant, les franchises de ces épisodes d'incapacité diffèrent et il est important de prendre en compte ces troncatures. Si ce n'est pas le cas, les incapacités courtes non observées ne sont pas considérées et le taux de passage est surestimé. Néanmoins, dans la base, la variable de la franchise n'est pas fiable. Or, les franchises sont toutes inférieures à 120 jours. C'est pourquoi nous avons décidé de ne modéliser le taux de passage que pour les épisodes d'incapacité de durée supérieure à 120 jours. Quant aux passages ayant lieu avant 120 jours d'incapacité, ils seront considérés comme de l'invalidité directe au sens large et seront étudiés par la suite.

Comme la probabilité de passage dans les trois ans suivant de le début de l'incapacité est déjà relativement faible, le passage ne sera pas étudié mensuellement, comme dans les tables du BCAC. En effet, la tarification ne nécessite pas le même niveau de précision que le provisionnement. Ainsi nous modéliserons le taux de passage en invalidité dans les 3 ans sachant que l'incapacité dure plus de 120 jours.

4.2.2 Choix de la méthode

Une première possibilité est de recourir, comme le BCAC [4], à un modèle de durée. En effet, nous pourrions considérer la durée avant le passage et étudier la fonction de survie à l'aide de l'estimateur de Kaplan-Meier. L'estimateur de Hoem, qui exprime la probabilité de passage comme le rapport entre le nombre de passages et l'exposition est une alternative possible. Ces méthodes permettent de prendre en compte tous les sinistres y compris ceux censurés (épisodes d'incapacité non clos n'ayant pas donné lieu à un passage à la fin de la période d'observation). Néanmoins, le volume d'épisodes d'incapacité est relativement important et supprimer ces données n'engendre ici pas de biais : pour chaque année de survenance étudiée, toutes les informations de passage ou d'absence de passage sont connues (cf Périmètre d'étude ci-dessous). De surcroît, comme expliqué précédemment, dans le cadre de la tarification, il n'est pas nécessaire de connaître la période exacte de passage mais la segmentation est plus importante. C'est pourquoi cette approche n'a pas été retenue.

3. Déclaration en ligne mensuelle qui transmet des informations sur chaque salarié
4. cf http://www.ressources-actuarielles.net/EXT/ISFA/fp-isfa.nsf/34a14c286dfb0903c1256ffd00502d73/a0d8fe4a9807886ac1257d11002a0be9/$FILE/presentation.pdf, p 32

La probabilité de passage en invalidité peut aussi être modélisée grâce à des méthodes de modélisation de survenance d'évènements, en l'absence de troncatures (cf 4.2.1 Démarche et détermination de la variable à modéliser) et de censures (cf Périmètre d'étude ci-dessous). Il existe principalement deux types de modèles : la régression logistique et les méthodes de machine learning telles que les arbres de décision. D'autres modèles plus poussés de machine learning tels que les forêts aléatoires auraient pu être retenus mais nous souhaitions un compromis entre simplicité et performance de prédiction.

Les arbres de décision présentent l'avantage de capter des relations non linéaires entre la variable à étudier et les variables explicatives. De plus, la segmentation des données n'a plus à être réalisée au préalable et gagne en précision. En effet, la segmentation préliminaire au GLM n'est réalisée qu'en bivarié. Malgré tout, le GLM présente plusieurs avantages. Il produit des coefficients associés à chaque variable ou modalité, ce qui permet d'analyser les impacts globaux. Cela s'avère plus difficile avec les arbres de décision, dans lesquels l'effet des variables est testé séquentiellement et non simultanément. Enfin, les arbres de décision sont très sensibles à l'échantillon d'apprentissage et sont donc moins robustes. Donc c'est la régression logistique qui a été retenue.

❖ *Périmètre d'étude*

Le périmètre d'étude contient originellement les années 2012-2019. En revanche, afin de connaitre tous les passages, nous nous limiterons aux épisodes d'incapacité dont l'année de survenance est comprise entre 2012 et 2016. En effet, les passages ont lieu au plus tard trois ans après le début de l'incapacité.

4.2.3 Hypothèses du modèle et expression de la probabilité d'invalidité indirecte

❖ *Hypothèse d'indépendance des observations*

Le GLM repose sur l'hypothèse d'indépendance des observations, ce qui permet d'exprimer la vraisemblance comme un produit. Néanmoins, certains assurés ont plusieurs épisodes d'incapacité et l'hypothèse d'indépendance pourrait dont être remise en question. Ainsi, d'une part les arrêts associés à un même individu peuvent avoir la même cause. D'autre part, si un assuré a par exemple deux arrêts, au plus un arrêt donne lieu à un passage.

Afin de pallier ce problème, une solution aurait été de modifier la structure de la base en binarisant les informations [5] :

— soit créer pour chaque assuré une ligne pour chaque année de 2013 à 2016 et ajouter les variables d'incidence de l'incapacité et de l'invalidité des années N-1 et N
— soit ne conserver que la dernière ligne pour chaque assuré et créer les variables d'incidence N, N-1, N-2...

Cette proposition permettrait de rendre les lignes indépendantes. En revanche, l'objectif final est la construction du tarif et l'historique des sinistres passés n'est pas utilisé en prévoyance collective.

5. solution proposée par le tuteur pédagogique

Une alternative a donc été envisagée pour limiter la dépendance entre les lignes : pour chaque assuré, regrouper les sinistres de même année de survenance sur une seule ligne et indiquer si l'un des épisodes a conduit à un passage. Cependant, cela aurait demandé d'étudier la probabilité d'avoir au moins un épisode d'incapacité dans l'année. De plus, peu d'assurés ont plusieurs arrêts de plus de 120 jours dans l'année et il resterait des assurés ayant plusieurs arrêts mais sur des années différentes. C'est pourquoi cette solution n'a pas été retenue non plus.

Enfin, il existe une extension du GLM qui tient compte de la dépendance entre les lignes d'un même individu mais repose sur l'hypothèse d'indépendance entre les individus : le GEE (Generalized Estimating Equations) [6]. La procédure GEE existe sur SAS.

Cependant, nous avons préféré commencer par un modèle simple comme l'hypothèse d'indépendance n'est pas aberrante. En effet, le problème est limité puisque seuls les arrêts de durée supérieure à 120 jours sont étudiés. 99,31% des assurés présents dans la base n'ont qu'un épisode de plus de 120 jours dans l'année. De plus, un premier arrêt peut être dû à une jambe cassée et un deuxième à une maladie grave. Enfin, les tests de validation permettront par la suite de valider ou non le modèle.

❖ *Hypothèse de covariance entre le nombre d'épisodes d'incapacité annuel et le taux de passage négligeable*

La probabilité de passer en invalidité est calculée à partir de l'estimation du nombre d'épisodes d'incapacité et du taux de passage (cf 4.2.1 Démarche et détermination de la variable à modéliser). En notant P_{II} la probabilité d'entrer en invalidité indirecte, τ le taux de passage et $N_{INC}*$ le nombre annuel d'épisodes d'incapacité de durée supérieure à 120 jours,

$$\begin{aligned} P_{II} = E\left(\frac{1_{passage}}{N_{INC}*} \times \ N_{INC}*\right) &= E\left(\tau \times N_{INC}*\right) \\ &= E(\tau) \times E(N_{INC}*) + cov(\tau, N_{INC}*) \end{aligned}$$

Or,

$$cov(\tau, N_{INC}*) = E\left((\tau - E(\tau)) \times N_{INC}*\right)$$

L'espérance du taux est calculée à partir de la base contenant les épisodes d'incapacité (cf 4.2.1 Démarche et détermination de la variable à modéliser) et correspond donc au cas $N_{INC}* \geq 1$. Posons pour $N_{INC}* = 0$, $\tau = E(\tau | N_{INC}* \geq 1)$ de façon à ce que $E(\tau) = E(\tau | N_{INC}* \geq 1)$ tout en conservant la relation $1_{passage} = \tau \times N_{INC}* = 0$.

Afin de prendre en compte le cas $N_{INC}* = 0$ (non présent dans la base), on conditionne par rapport à $N_{INC}*$. Comme le terme associé au cas $N_{INC}* = 0$ est nul et qu'on observe au plus 2 sinistres de plus de 120 jours dans l'année, on obtient :

$$cov(\tau, N_{INC}*) = \sum_{i=1}^{2} E\left((\tau - E(\tau)) \times N_{INC}* | N_{INC}* = i\right) \times P(N_{INC}* = i)$$

6. cf http://www.ressources-actuarielles.net/EXT/ISFA/1226-02.nsf/0/92af0ab1dd6a4dccc1257e11002ddd1b/$FILE/Memoire_Actuaire_Ahmed_Tidiane_DIOMANDE.pdf

A partir de notre base de données, il est possible d'estimer $P((N_{INC}* = i|N_{INC}* \geq 1)$.

Or, en notant N_{INC} le nombre annuel d'épisodes d'incapacité modélisé dans un précédent mémoire pour une franchise de 3 jours,

$$\begin{aligned} P(N_{INC}* = i) &= P((N_{INC}* = i|N_{INC}* \geq 1) \times P(N_{INC}* \geq 1) \\ &\leq P((N_{INC}* = i|N_{INC}* \geq 1) \times P(N_{INC} \geq 1) \end{aligned}$$

De plus, $P(N_{INC} \geq 1)$ a été estimé dans un précédent mémoire sur l'incidence en incapacité.

Ainsi, $cov(\tau, N_{INC}*) \leq 1,7 \times 10^{-6}$.

Elle est donc négligeable devant le premier terme de la probabilité d'incidence indirecte (de l'ordre de 10^{-3}).

❖ *Calcul de la probabilité d'invalidité indirecte*

En négligeant la covariance,

$$P_{II} \approx E(\tau) \times E(N_{INC}*)$$

avec :

$$E(N_{INC}*) = E\left(\sum_{i=1}^{N_{INC}} 1_{T_i>120}\right)$$

En supposant que les durées T_i suivent la même loi et en notant S_{INC} la fonction de survie du maintien en incapacité,

$$E(N_{INC}*) = E(N_{INC} \times 1_{T>120}) = E(N_{INC}) \times \frac{S_{INC}(120)}{S_{INC}(3)} + cov(N_{INC}, 1_{T>120})$$

La covariance n'est pas calculable à partir de notre base puisque N_{INC} dépend de la franchise qui n'est pas fiable. Néanmoins, comme la probabilité d'avoir un sinistre est inférieure à 10%, la covariance sera relativement faible, comme précédemment. De plus, plus la durée est importante plus le nombre de sinistres maximal dans l'année diminue. La covariance devrait donc être négative et la simplification reste prudente.
Donc,

$$P_{II} \approx E(\tau) \times E(N_{INC}) \frac{S_{INC}(120)}{S_{INC}(3)}$$

4.2.4 Comparaison du taux de passage par âge avec celui du BCAC

Comme seuls les arrêts de plus de 120 jours ont été conservés (cf 4.2.1 Démarche et détermination de la variable à modéliser), les taux du BCAC ont été calculés en additionnant les passages ayant lieu au cours du $5^{ème}$ mois ou ultérieurement. Cette somme est rapportée à l'effectif correspondant au maintien au-delà du $5^{ème}$ mois.

Cette comparaison toutes survenances confondues a mis en lumière une importante différence entre les taux :

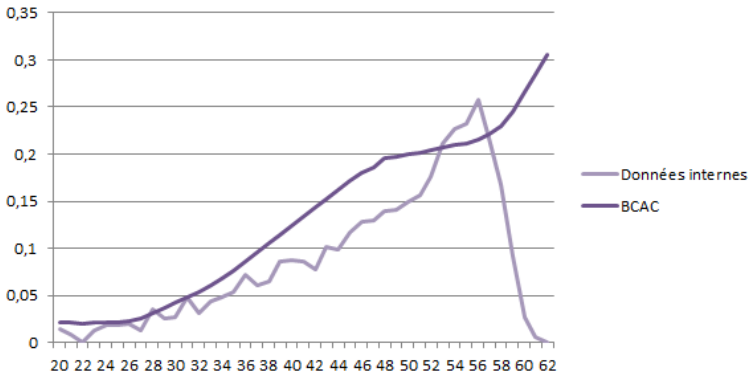


FIGURE 4.4 – Comparaison des taux bruts internes de passage aux taux lissés du BCAC en fonction de l'âge d'entrée en incapacité

D'une part, on remarque que les taux internes sont majoritairement très inférieurs à ceux du BCAC. Cependant, il est rassurant de noter que dans son mémoire, Tom Leurent observait un ratio de 51% entre ses taux et ceux du BCAC de 1993 pour la prévoyance individuelle. De plus, le BCAC constatait lui-même une surévaluation des taux de passage [7] :

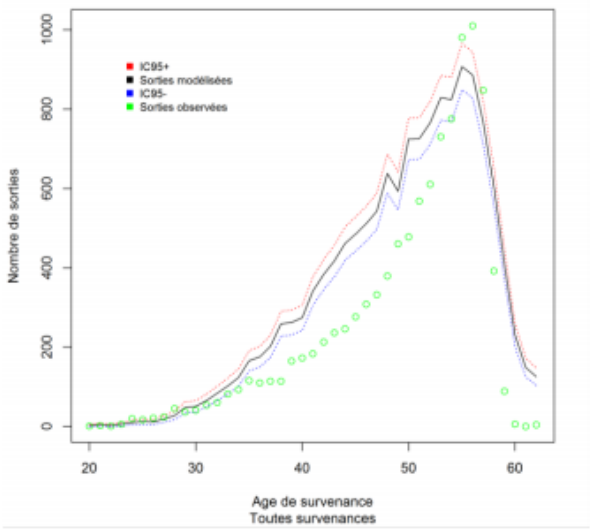


FIGURE 4.5 – Comparaison entre les nombres de passage (toutes survenances confondues) observé et modélisé du BCAC

D'autre part, on constate une décroissance du taux de passage à partir d'un certain âge (cf Figure 4.2) alors que le taux lissé du BCAC poursuit sa croissance. En revanche, cette tendance n'est pas présente au niveau des taux bruts (cf Figure 4.3). Cette décroissance a lieu à partir de 57 ans, soit 3 ans avant l'âge de départ à la retraite anticipé. Le BCAC a sans doute voulu gommer l'effet des garanties dues à la retraite. Afin que notre modèle reste valable lors d'un changement d'âge de départ à la retraite, nous avons également choisi de pas modéliser cette chute.

7. Source : "Pr esentation et comparaison des nouvelles tables BCAC", Prim'Act-25/06/2014, p33

4.2.5 Analyses préliminaires et démarche

❖ *Effet de l'âge*

Comme expliqué ci-dessus, nous avons choisi de ne pas modéliser la chute du taux de passage au-delà de 56 ans due à la retraite anticipée. C'est pourquoi seuls les individus d'au plus 56 ans ont été étudiés pour réaliser notre modèle. Afin de modéliser une poursuite similaire de l'augmentation du taux au-delà de 56 ans, l'âge est traité en variable continue. Ainsi, nous obtenons un coefficient propre à l'âge qui sera appliqué de la même manière aux âges élevés. Cependant, pour ne pas surestimer le taux de passage à ces âges, il conviendra d'appliquer un abattement (à modifier en cas de changement d'âge de départ à la retraite). Néanmoins, cet abattement sera rarement nécessaire. En effet, c'est souvent l'incidence correspondant à l'âge moyen de la population à assurer [8] qui est utilisée pour le tarif. Or, l'âge moyen ne devrait pas souvent excéder 56 ans. Il arrive parfois cependant que l'on tarifie sur des populations détaillées.

❖ *Segmentation du secteur d'activité et du département*

Comme le secteur d'activité et le département présentent de nombreuses modalités, ces variables ont été discrétisées à l'aide d'arbres de décision, grâce à la procédure *HPSPLIT* sur SAS. Afin de choisir le nombre optimal de feuilles de l'arbre, on trace l'erreur de prédiction en fonction du nombre de feuilles de l'arbre :

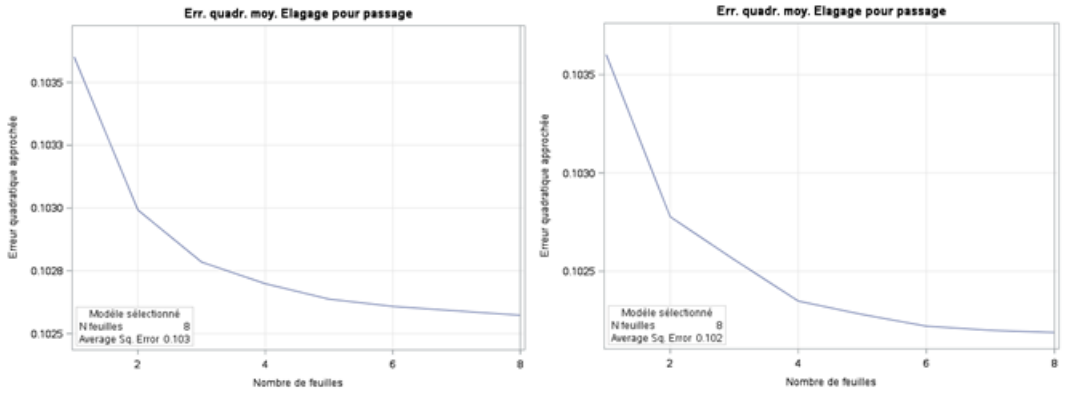


FIGURE 4.6 – Evolution de l'erreur de prédiction en fonction du nombre de feuilles pour le secteur d'activité (à gauche) et pour le département (à droite)

La procédure SAS fonctionne par validation croisée et renvoie le nombre de feuilles optimal. Cependant, certaines classes formées étaient très peu représentées. Nous avons donc choisi de ne pas ajouter les feuilles qui ne diminuaient pas sensiblement l'erreur. Ainsi, cinq classes de secteurs d'activité et six classes de départements ont été formées.

La composition des classes n'est pas précisée par souci de confidentialité mais des exemples seront donnés lors de l'interprétation finale.

8. souvent le seul âge connu en prévoyance collective et pas d'âge actuariel pour ce tarif

❖ *Choix de la fonction de lien*

Trois fonctions de lien sont possibles : les fonctions logit, probit et log-log. Les fonctions logit et probit sont relativement proches mais la fonction logit est plus facile à manipuler. Quant à la fonction log-log, elle est plus adaptée aux probabilités asymétriques. C'est pourquoi nous avons retenu les fonctions logit et log-log. La statistique du test d'Hosmer-Lemeshow, qui permet d'évaluer la qualité d'ajustement des modèles sera ensuite calculée pour chacune de ces fonctions afin de les départager.

❖ *Analyse des corrélations*

Les variables (âge discrétisé, sexe, CSP, département,secteur d'activité) sont un peu corrélées mais les VCramer restent inférieurs à 0,7.

❖ *Définition de l'individu de référence*

La modalité la plus présente de chaque variable a été choisie comme référence.

❖ *Création de la base d'échantillon et de la base de validation*

La base a été séparée en un échantillon de validation représentant les 2/3 de la base et un échantillon test.

❖ *Sélection des variables et regroupement des modalités*

Pour sélectionner les variables et regrouper les modalités, nous avons procédé ainsi ("backward selection") :

— le modèle a été établi en sélectionnant l'ensemble des variables ;
— le modèle a été réestimé en regroupant la modalité la moins significative avec la modalité ayant le paramètre estimé le plus proche ;
— l'étape précédente a été renouvelée jusqu'à ce que l'ensemble des variables soit significatif.

4.2.6 Choix et qualité d'adéquation du modèle

❖ *Choix du modèle*

Comme expliqué ci-dessus, les modèles résultant des fonctions de lien logit et log-log sont comparés. Le modèle minimisant l'écart entre les nombres de passages prédit et observé sur les bases d'apprentissage et de validation est retenu.

Voici les statistiques du test d'Hosmer-Lemeshow obtenues sur chacun des échantillons pour les deux fonctions de lien :

Lien \ Base	Apprentissage	Validation
Log-log	3,8437	9,39
Logit	3,5194	8,68

FIGURE 4.7 – Comparaison des statistiques du test d'Hosmer-Lemeshow pour les deux fonctions de lien

L'adéquation est donc meilleure pour la fonction de lien **logit** à la fois sur la base d'apprentissage et de validation.

❖ *Test d'Hosmer-Lemeshow*

Les nombres de passage issus du modèle sont comparés à ceux observés sur les bases d'apprentissage et de validation pour la fonction de lien logit :

Base apprentissage

effectif	cluster	nb_reel	nb_pred
2548	0	43	43,28
2552	1	73	81,62
2549	2	125	116,19
2568	3	167	162,22
2546	4	225	214,71
2563	5	281	280,12
2475	6	335	343,61
2555	7	430	442,07
2537	8	531	539,94
2625	9	786	772,24

Khi-2	DF	P-value
3,5194	8	0,8977

Base validation

effectif	cluster	nb_reel	nb_pred	moy_pred	
1313	0	19	22,4027527	0,01706226	0,5258155
1313	1	52	42,308935	0,0322231	2,29369501
1314	2	57	60,193279	0,04580919	0,17753767
1342	3	94	84,531737	0,06298937	1,13181727
1301	4	125	108,846291	0,08366356	2,61623014
1303	5	128	140,187469	0,10758823	1,18727835
1310	6	178	180,360344	0,13767965	0,03582127
1283	7	211	221,936709	0,17298263	0,65167267
1368	8	291	294,734903	0,21544949	0,06032624
1296	9	387	387,794616	0,29922424	0,00232346
				Khi2	8,68251756

Khi-2	DF	P-value
8,68251756	10	0,5625

FIGURE 4.8 – Test d'Hosmer-Lemeshow

Rappelons que sur la base de validation, la statistique de test est la même mais sous H0, elle suit une loi du Khi-2 à 10 degrés de liberté et non plus 8 (cf 4.1.1.2 Adéquation et évaluation des modèles).
Les nombres prédits sont proches des nombres observés. La p-value étant supérieure à 0,05, l'adéquation est bonne.

❖ *Diagramme de fiabilité*

Afin de tester le caractère prédictif du modèle, les observations ont été classées en 10 groupes sur lesquels le taux moyen de passage prédit et le taux observé ont été calculés pour tracer le diagramme de fiabilité :

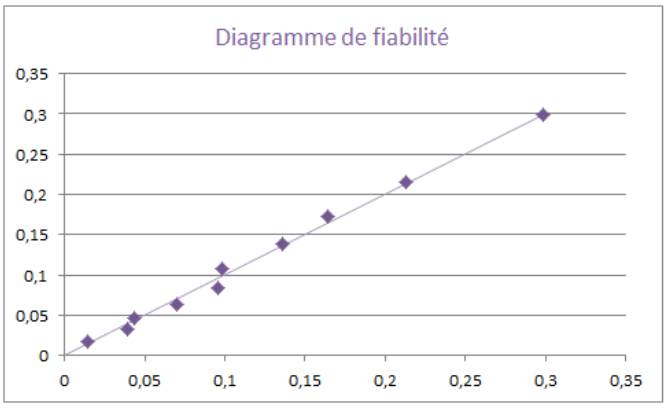


FIGURE 4.9 – Diagramme de fiabilité

L'ensemble des points est proche de la première bissectrice : les taux prédits sont proches des taux réels. La qualité prédictive du modèle est donc bonne.

❖ *R^2 de McFadden, AUC et indice de Gini*

Le R^2 de McFadden, l'AUC et l'indice de Gini ont été calculés sur la base d'apprentissage :

R^2 McFadden	AUC	Gini
0,098	0,728	0,456

FIGURE 4.10 – R^2 de McFadden, AUC et indice de Gini de la base d'apprentissage

Si le R^2 de McFadden est relativement faible, il montre tout de même que le modèle obtenu est plus performant que le modèle trivial composé uniquement de la constante. D'autre part, l'AUC est compris entre 0,7 et 0,8 donc la discrimination est bonne. Enfin, l'indice de Gini est relativement élevé, ce qui confirme que la qualité du modèle est assez bonne.

4.2.7 Interprétation des résultats

Il est intéressant de noter le niveau d'influence et l'impact des variables.

❖ *Degré d'influence des variables et cohérence des résultats*

Afin d'étudier le degré d'influence de chaque variable, les *odds ratio* ont été calculés en fixant les autres variables. Pour rappel :

$$R(X_i, X_i') = \exp(X_i^T \beta - X_i'^T \beta)$$

Un avantage des odds ratios est que les termes correspondant aux autres variables se simplifient, ce qui permet d'étudier l'impact de la variable à tout âge.

Voici les odds ratios obtenus :

Effet	OddsRatio
age_deb_inc	1,090519272
classe_dpt_2 1 vs 0	1,361735639
classe_dpt_2 2 vs 0	0,096114428
classe_dpt_2 3 vs 0	0,428235439
classe_dpt_2 4 vs 0	1,639823469
classe_dpt_2 5 vs 0	2,618151389
classe_ape_2 1 vs 0	1,767428577
classe_ape_2 2 vs 0	1,226637115
classe_ape_2 3 vs 0	0,125714287
CSP C vs NC	0,890138014

FIGURE 4.11 – Rapports de côte

Plus le rapport de côte s'éloigne de 1, plus l'impact de la variable est important. Le département est très influent : toutes choses égales par ailleurs, pour un individu dont le département est en classe 5, le rapport entre la probabilité de passer en invalidité et celle de ne pas passer est 2,6 fois plus élevé que celui d'un individu dont le département est en classe 0 !

Les variables les plus influentes sont l'âge, le département et le secteur d'activité. Le sexe n'est pas significatif mais il l'était dans le cas de l'incidence en incapacité. Ainsi, s'il ne joue pas dans le passage même, il interviendra tout de même dans l'incidence en *invalidité indirecte*. Quant à la catégorie socioprofessionnelle, son impact est limité.

Les résultats semblent **cohérents**. En effet, le passage en invalidité peut s'apparenter à un maintien long en incapacité. Or, dans un précédent mémoire traitant du maintien en incapacité, le sexe et la CSP étaient significatifs mais avaient un impact très faible, et n'avaient donc pas été retenus.

❖ *Influence des variables*

Pour chaque variable d'étude, les taux de passage ont été tracés en fonction de l'âge d'entrée en incapacité en distinguant les classes et en fixant les autres variables aux références.

Par souci de confidentialité, la composition des classes ne sera pas révélée. Seuls quelques exemples seront cités.

➤ *Catégorie socioprofessionnelle (CSP)*

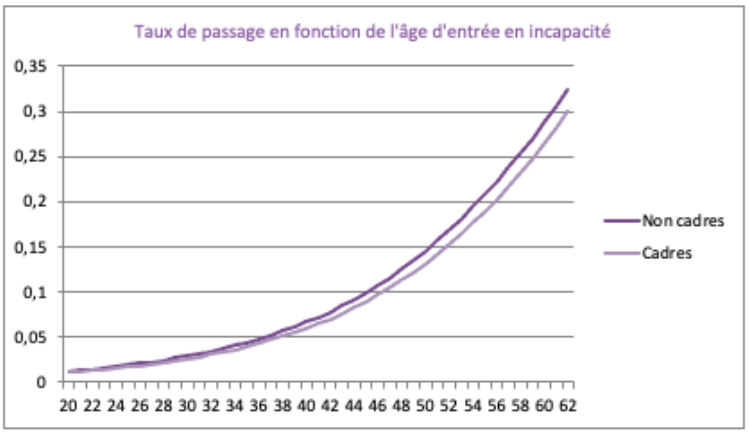


FIGURE 4.12 – Influence de la CSP sur le taux de passage en fonction de l'âge d'entrée en incapacité

D'une part, on remarque que la catégorie socioprofessionnelle a un impact limité. Le rapport de côte était proche de 1, donc c'est cohérent. D'autre part, aux âges plus élevés, le taux de passage est assez important. Cependant, il ne faut pas oublier que nous étudions le taux de passage pour un maintien en incapacité de plus de 120 jours.

Remarque : On peut noter qu'à la différence d'un modèle linéaire, pour le modèle logistique avec fonction de lien logit, la différence observée entre les taux des deux classes varie en fonction de l'âge.

➤ *Secteur d'activité*

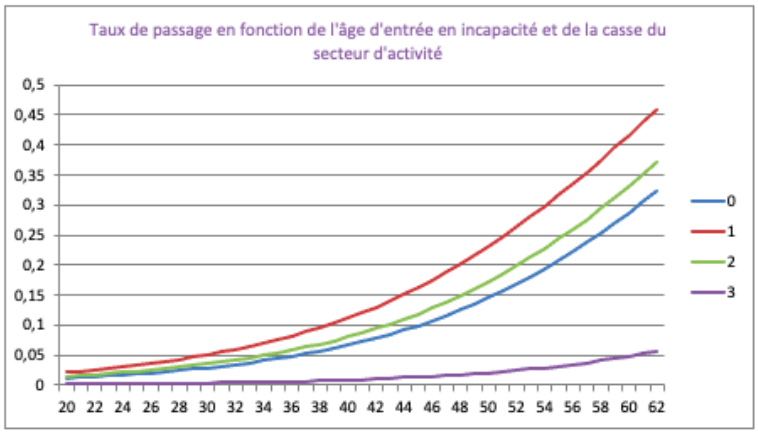


FIGURE 4.13 – Influence du secteur d'activité sur le taux de passage en fonction de l'âge d'entrée en incapacité

Le secteur d'activité a une influence notable sur le taux de passage en invalidité au-delà de 120 jours d'incapacité. En effet, à 43 ans (âge moyen en entreprise), le taux de passage du profil de référence dans la fabrication de textiles (classe 1) est 12 fois supérieur à celui relatif à l'adminis-

tration publique (classe 3). Cela pourrait être dû à la pénibilité du travail.

➤ *Département de l'entreprise*

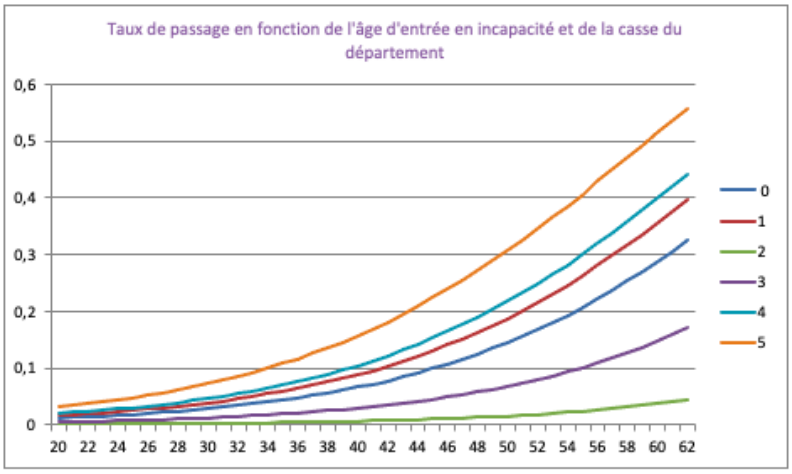


FIGURE 4.14 – Influence du département de l'entreprise sur le taux de passage en fonction de l'âge d'entrée en incapacité

Le département a également un impact très important sur le taux de passage. Par exemple, à 43 ans (âge moyen en entreprise), le taux de passage du profil de référence pour la Meurthe-et-Moselle et pour le Lot (classe 5) est 2,3 fois supérieur à celui de Paris et des Hauts de Seine (classe 0). Au contraire, pour le profil de référence, les départements d'outre-mer et l'Indre (classe 2) ont un taux de passage très faible, 9,6 fois inférieur à celui de Paris. Ces différences pourraient notamment être dues à la répartition de la population (âge, sexe, statut).

5 Construction des barèmes tarifaires

Les deux principales composantes du tarif ont été modélisées. Une fois les derniers éléments étudiés, il sera possible de présenter une méthode de tarification. Enfin, celle-ci sera appliquée et nos barèmes seront comparés aux actuels.

5.1 Etudes préliminaires

Le maintien et l'incidence ont été estimés au préalable, mais d'autres composantes du tarif restent à étudier.

5.1.1 Evolution temporelle de la proportion d'invalides dans les 3 catégories d'invalidité

Les garanties dépendent de la catégorie d'invalidité. Il faut donc en tenir compte dans le tarif. C'est pourquoi nous avons étudié la répartition dans les 3 catégories d'invalidité au 31/12 de 2012 à 2019.

Cependant, l'information de la première catégorie d'invalidité n'est pas toujours disponible.
Concrètement, parfois la date de début d'invalidité de l'individu, la date de son changement de catégorie et sa $2^{ème}$ catégorie sont connues, mais la $1^{ère}$ ne l'est pas (2.1.3.2 Reconstitution de l'historique des sinistres). Dans ce cas, nous avons émis les hypothèses suivantes, selon la durée entre le début d'invalidité renseigné et la date de changement de catégorie (durée D) :

- Si la $2^{ème}$ catégorie est la catégorie 1, l'individu sera supposé arrivé en catégorie 1 lorsque la durée D est inférieure à 3 mois et en catégorie 2 sinon.
- Sinon, l'individu sera supposé arrivé directement dans sa $2^{ème}$ catégorie lorsque la durée D est inférieure à 3 mois et dans la catégorie inférieure sinon.

Si la durée D est supérieure à 3 mois, on suppose qu'il y a réellement eu un changement. Dans ce cas, on fait l'hypothèse que la catégorie d'arrivée était la catégorie inférieure ou la catégorie 2 dans le premier cas. En effet, on observe très rarement un changement d'une catégorie supérieure vers une catégorie inférieure (2 cas dans toute la base) et aucun passage de la catégorie 3 vers la catégorie 1.

Voici les répartitions obtenues :

Année / Catégorie	2012	2013	2014	2015	2016	2017	2018	2019
Catégorie 1	20,5%	20,7%	21,6%	21,6%	22,2%	22,3%	22,4%	22,2%
Catégorie 2	77,3%	77,2%	76,4%	76,5%	76,0%	76,0%	76,0%	76,1%
Catégorie 3	2,3%	2,1%	2,0%	1,9%	1,9%	1,7%	1,6%	1,6%

FIGURE 5.1 – Répartition dans les 3 catégories d'invalidité de 2012 à 2019

Cette répartition est relativement proche de celle de la Sécurité Sociale :

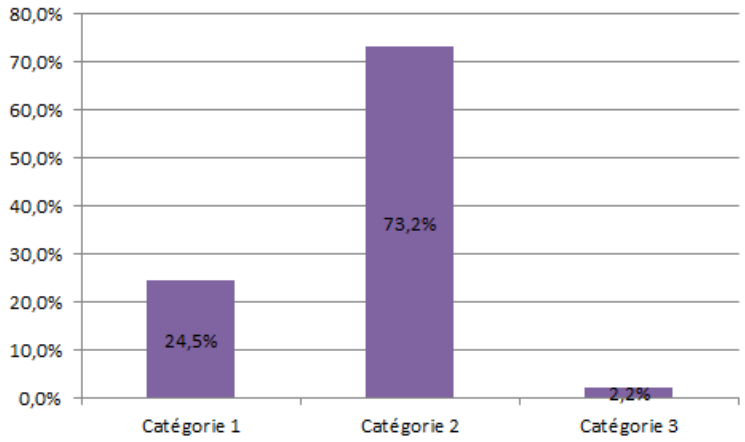


FIGURE 5.2 – Répartition des invalides du régime général dans les 3 catégories en 2015 (Source : Mémoire de Pierre Morlon, page 16)

En revanche, nous comptions plus d'individus en $2^{ème}$ catégorie et moins en $1^{ère}$ catégorie.

D'autre part, la proportion des catégories 2 et 3 est en légère baisse, au contraire de celle de la catégorie 1. Néanmoins, au vu de ces résultats, une hypothèse de répartition constante dans les 3 catégories n'est pas aberrante. L'évolution de la répartition pourra être étudiée par la suite.

5.1.2 Durée en incapacité avant le passage en invalidité

La probabilité de passer en invalidité dans les 3 ans suivant le début d'incapacité a été préalablement calculée. Cependant, le maintien dépend de l'âge d'entrée en invalidité et l'année de passage influe sur l'actualisation. C'est pourquoi nous avons étudié la durée en incapacité avant le passage en invalidité, qui permet d'avoir l'information de l'année de passage en invalidité. Ainsi, il sera possible de calculer les taux de passage pour la $1^{ère}$, la $2^{ème}$ et la $3^{ème}$ années.

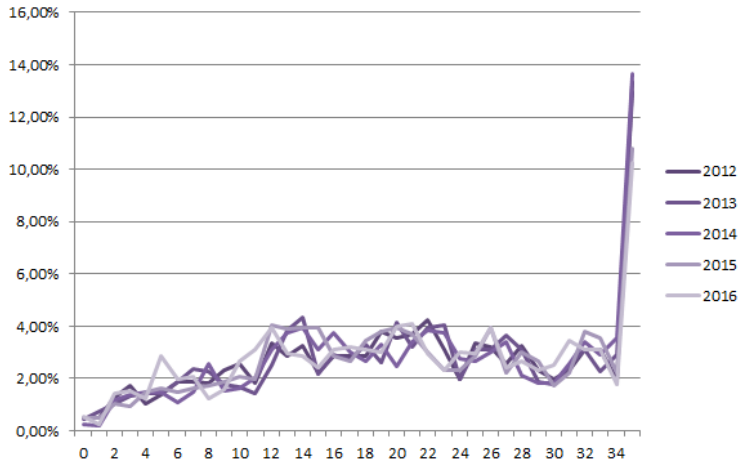


FIGURE 5.3 – Répartition des passages en invalidité en fonction de la durée en incapacité (en mois) avant le passage

On constate que la répartition des passages au cours des 3 ans reste à peu près constante. De plus, le pic au $35^{ème}$ mois montre qu'une part importante des passages a lieu au bout des trois années d'incapacité. Néanmoins, on aurait pu s'attendre à ce que cette part soit plus élevée : entre 85 et 90% des passages ont lieu avant la durée maximale d'incapacité.

En moyennant les parts de 2012 à 2019, voici la part de passages au cours de la $1^{ère}$, $2^{ème}$ et $3^{ème}$ années :

Année de passage	Taux
$1^{ère}$ année	18,05%
$2^{ème}$ année	39,63%
$3^{ème}$ année	42,32%

FIGURE 5.4 – Répartition des passages au cours des années

5.1.3 Prise en compte de l'invalidité directe

❖ *Méthode*

L'incidence de l'invalidité découlant de l'incapacité a pu être estimée, grâce à un mémoire précédent qui avait traité de l'incidence en incapacité. En revanche, pour l'invalidité directe, nous ignorons l'effectif sous risque. Pour tenir malgré tout compte de l'invalidité directe, nous avons estimé le rapport entre le nombre d'invalidités directes et le nombre de passages en invalidité (dans les 3 ans suivant le début de l'incapacité), sur plusieurs années de survenance. Ainsi nous pouvions exprimer la probabilité d'entrer en invalidité directe comme une fraction de la probabilité d'invalidité indirecte.

Par ailleurs, nous émettons l'hypothèse que la loi de maintien des invalidités directes est la même que celle des invalidités indirectes.

❖ *Estimation de la probabilité d'entrer en invalidité directe dans l'année*

Le rapport entre l'incidence en invalidité directe et l'incidence en invalidité indirecte a été calculé pour plusieurs années de survenance :

nb_direct	nb_indirect	taux_direct
96	1129	0,085031
71	1119	0,06344951
65	1124	0,05782918
69	1112	0,06205036
94	1145	0,08209607

Moyenne	0,07009122

FIGURE 5.5 – Rapport entre l'incidence en invalidité directe et l'incidence en invalidité indirecte

Ce rapport est assez fluctuant, de moyenne 7%. Comme les sinistres dont la date de survenance (d'incapacité) est postérieure à 2016 peuvent passer en invalidité au bout de 3 ans, on ne connaît pas tous les passages après 2016. Donc le taux n'a pas pu être estimé pour les années 2017-2019. Néanmoins, au vu du nombre d'invalidités directes de ces années, le taux moyen paraît prudent.

5.2 Méthode de la tarification

Les éléments constitutifs du tarif ayant été modélisés, il ne reste plus qu'à construire le tarif.

5.2.1 Hypothèses

❖ *Constance de la proportion d'invalides dans les 3 catégories*

En amont des différentes études, une réflexion sur la construction du tarif a été nécessaire. L'incidence et le maintien en invalidité sont les grandes composantes du tarif. En revanche, les prestations dépendent fortement de la catégorie d'invalidité. De plus, les assurés peuvent changer de catégorie. Une solution possible aurait donc été d'étudier de modéliser les changements de catégorie et d'étudier les lois de maintien pour chaque catégorie d'invalidité. Néanmoins, la quantité de données n'aurait pas permis d'étudier le maintien pour chaque catégorie et le modèle aurait été très complexe.

D'autre part, une étude statistique a révélé que la répartition des assurés dans les trois catégories d'invalidité est relativement constante. Notre tarif repose sur cette hypothèse de répartition constante dans les 3 catégories. La constance de la répartition dans les catégories assure la mutualisation du risque parmi les assurés.

❖ *Prise en compte du maintien global, sans distinguer les catégories d'invalidité*

Le maintien a été étudié toutes catégories confondues, il s'agit donc du maintien global. Au vu du volume de données, il aurait été difficile de modéliser le maintien pour chaque catégorie séparément.

Mais la constance de la répartition dans les catégories d'invalidité permet la mutualisation du maintien.

5.2.2 Calcul du tarif

Le tarif du collège correspond à la moyenne des tarifs sur les différentes sous-populations (femmes par exemple), pondérée par les poids de ces catégories.

Pour la **sous-population considérée**, introduisons tout d'abord les notations suivantes :

- P_{II}, la probabilité d'entrer en invalidité indirecte dans les trois ans suivant le début de l'incapacité estimée précédemment :

$$P_{II} \approx E(\tau) \times E(N_{INC}) \frac{S_{INC}(120)}{S_{INC}(3)}$$

- P_{ID}, la probabilité d'entrer en invalidité directe dans l'année ;
- P_{cat_i} pour i=1,2,3, la part d'invalides de catégorie i déterminée précédemment ;
- P_{pass_i} pour i=1,2,3, la probabilité de passer en invalidité la $i^{ème}$ année après le début de l'incapacité sachant que l'assuré passe en invalidité (indirecte) ;
- v, le facteur d'actualisation.
- $tarif_{II}$, la part du tarif incombant à l'invalidité indirecte ;
- $tarif_{ID}$, la part du tarif propre à l'invalidité directe ;
- S_x, la fonction de survie définie précédemment pour un âge d'entrée en invalidité égal à x ;
- k_{ID}, le rapport entre l'incidence en invalidité directe et celui en invalidité indirecte (dans les 3 ans) déterminé précédemment.

Le tarif correspondant au risque de l'invalidité est constitué du tarif propre à l'invalidité directe et de celui propre à l'invalidité indirecte.

❖ *Tarif propre à l'invalidité indirecte*

Pour calculer ses barèmes, le BCAC a multiplié le taux d'entrée dans le risque par le coût moyen pour 100 euros de prestation annuelle (décomposition non disponible). [1]

De plus, les prestations dépendent de la catégorie d'invalidité. Ainsi :

$$Tarif_I = P_{II} \times (P_{cat_1} \times E(coût_{cat_1}) + P_{cat_2} \times E(coût_{cat_2}) + P_{cat_3} \times E(coût_{cat_3}))$$

Le coût moyen de chaque catégorie reste à exprimer. Il s'exprime comme une rente annuelle versée par mois échu. Pour rappel, on prend en compte le maintien jusqu'à 61,5 ans (cf 3.1.2 Troncatures et censures).

Si le risque d'invalidité naît bien des incapacités de l'année et fait donc partie des engagements futurs de l'assureur [2], l'invalidité indirecte peut débuter dans les trois ans qui suivent. Or, l'année

1. Source : Actuaris, https ://docplayer.fr/20553228-Tech-prevoyance-risque-incapacite-invalidite-projet-de-nouvelles-tables-bcac-infotech-32-introduction.html

2. La prime est égale à la valeur actuelle probable des engagements futurs de l'assureur

de passage intervient dans l'actualisation et dans le maintien (qui dépend de l'âge d'entrée en invalidité). Il faut donc tenir compte de l'année de passage.

Notons, pour un âge d'entrée en incapacité x et une année de passage i :

$$R_{x,i} = \frac{1}{12} \times \sum_{t=1}^{(61,5-(x+i-1))\times 12} \left(v^{\frac{t}{12}+(i-1)} \times S_{x+i-1}\left(\frac{t}{12}\right)\right)$$

$R_{x,1}$ correspond à une rente unitaire versée par mois échu pour un assuré d'âge x en début d'incapacité et passé en invalidité dès la $1^{ère}$ année.
$R_{x,2}$ correspond à la rente d'un assuré entré en incapacité à l'âge x mais passé en invalidité la $2^{ème}$ année (alors âgé de $x+1$).
La probabilité de survie mensuelle peut être approchée par interpolation linéaire.

On en déduit que pour un assuré entré en incapacité à l'âge x :

$$E(coût_{cat_i}) = (P_{pass_1} \times R_{x,1} + P_{pass_2} \times R_{x,2} + P_{pass_3} \times R_{x,3}) \times garanties_{cat_i}$$

Notons $R_{pass,x} = P_{pass_1} \times R_{x,1} + P_{pass_2} \times R_{x,2} + P_{pass_3} \times R_{x,3}$

Comme on a calculé un maintien global et une durée en incapacité avant le passage toutes catégories confondues, on trouve :

$$Tarif_{II} = P_{II} \times R_{pass,x} \times (P_{cat_1} \times garanties_{cat_1} + P_{cat_2} \times garanties_{cat_2} + P_{cat_3} \times garanties_{cat_3})$$

❖ *Tarif propre à l'invalidité directe*

L'invalidité directe n'est prise en charge que suite à une période de franchise, identique à celle de l'incapacité. Au contraire, l'invalidité indirecte donne lieu à des prestations d'invalidité directement. L'incapacité temporaire a dans ce cas déjà permis son écoulement.

En présence de franchise, les arrêts n'ayant pas dépassé celle-ci ne sont pas connus. Comme il y a différentes franchises dans la base et qu'elles ne sont pas connues (variable non fiable), l'incidence est légérement modifiée. Cependant, la probabilité de maintien au-delà de 90 jours est très proche de 1 (estimation des taux bruts par Kaplan-Meier), donc cet aspect peut être négligé.

Supposons que la probabilité de se maintenir au-delà de la franchise est de 1 et notons f la franchise exprimée en trimestre. Notons :

$$R_{f,x} = \frac{1}{12} \times \sum_{t=1}^{(61,5-x)\times 12} \left(v^{\frac{t}{12}} \times S_x\left(\frac{t+f}{12}\right)\right)$$

Remarque : En pratique, la franchise sera négligée afin de simplifier la tarification et d'avoir un unique barème pour toutes les franchises. Cette simplification a peu d'impact sur le tarif.

Par un raisonnement identique à celui développé pour l'invalidité indirecte :

$$Tarif_{ID} = P_{ID} \times R_{f,x} \times (P_{cat_1} \times garanties_{cat_1} + P_{cat_2} \times garanties_{cat_2} + P_{cat_3} \times garanties_{cat_3})$$

avec $P_{ID} \approx k_{ID} \times P_{II}$

Finalement,

$$Tarif = P_{II} \times (P_{cat_1} \times garanties_{cat_1} + P_{cat_2} \times garanties_{cat_2} + P_{cat_3} \times garanties_{cat_3}) \times (k_{ID} \times R_{f,x} + R_{pass,x})$$

Cette méthode reste à appliquer afin d'établir notre tarif. Ensuite, il sera pertinent de le comparer avec le tarif actuel.

5.3 Application

Maintenant que notre méthode de tarification a été définie, il nous est possible d'établir nos barèmes.

5.3.1 Résultats issus du tarificateur

Le tarificateur a été construit sur Excel, à partir des résultats des précédents mémoires sur l'incidence et le maintien en incapacité. Les formules présentées précédemment ont permis d'établir les barèmes d'invalidité.

Le taux d'actualisation n'est réglementé que dans le cadre du provisionnement. Dans ce cas, il ne peut être supérieur à 75% du taux moyen d'emprunt d'Etat (TME, taux de rendement sur le marché secondaire des emprunts d'État à taux fixe supérieurs à 7 ans) moyen des 24 derniers mois. Cependant, comme ce taux moyen est devenu négatif, l'article 2 du règlement N° 2020-11 du 22 decembre 2020 permet dans ce cas de retenir un taux d'actualisation inférieur ou égal à 0. Dans notre étude de tarification, nous avons choisi un taux nul.

D'autre part, le tarif est proportionnel à la moyenne des garanties servies pondérée par les proportions dans les trois catégories d'invalidité. Or, les garanties dépendent des contrats. Il est possible d'obtenir un barème $Barème_{100\%}$ indépendant de cette moyenne, qui correspond à une garantie à 100% peu importe la catégorie. Afin d'avoir une meilleure idée des taux de cotisation, nous avons également calculé un **barème simplifié**. Pour cela, quelques hypothèses simplificatrices ont été posées :

— les garanties servies pour la catégorie 1 sont supposées nulles.
En effet les assureurs interviennent peu pour cette catégorie.

— les garanties servies pour les catégories 2 et 3 sont supposées égales.
Les garanties sont semblables.

Ainsi, $Barème_{simplifié} = Barème_{100\%} \times (P_{cat_2} + P_{cat_3})$.

❖ *Influence des variables*

Pour évaluer l'impact de chaque variable, nous avons fait varier les modalités de façon à minimiser et maximiser le risque, tout en fixant les autres variables à la référence :

Variable	Modalités risque faible / référence / fort	Rapport profil faible/référence	Rapport profil fort/référence
Secteur d'activité	65 / 74 / 03	-66%	9%
Département	97 / 64 / 54	-96%	115%
Sexe	/ H / F		31%
CSP	C / NC /	-57%	
Age	25 / 39 / 55	-92%	55%

FIGURE 5.6 – Influence des variables dans le tarif

Tout d'abord, on remarque de gros écarts entre les différents profils. Néanmoins, cela pourrait être dû au nombre important de classes (combinaison des classes formées au cours des trois études). En effet, les classes extrêmes sont alors peu représentées et très éloignées du profil de référence. Ainsi, si ces résultats prouvent bien l'impact important de chaque variable, les écarts vont diminuer après l'harmonisation des classes (cf 5.4. Limites de l'approche et prolongement de l'étude).

Les différences observées pour le profil de référence entre les secteurs d'assurance (65) et de pêche et aquaculture (03) pourraient s'expliquer par la pénibilité du travail. D'autre part, on remarque que le sexe a un impact relativement important, malgré sa non-significativité dans le passage de l'incapacité à l'invalidité. Cela s'explique par son influence sur l'incidence en incapacité. Pour le profil de référence, le barème des femmes est 31% supérieur à celui des hommes.

5.3.2 Comparaison des barèmes construits avec les barèmes actuels

Il est important de comparer nos barèmes avec ceux utilisés actuellement, afin de vérifier leur cohérence. En revanche, des différences devraient apparaître, comme les barèmes actuels ne sont pas très récents. Pour pouvoir comparer les barèmes (à 100% de garantie), le taux d'actualisation a été modifié et les frais et la revalorisation ont été pris en compte.

Les barèmes actuels ne dépendent que de l'âge moyen, de la catégorie soci-professionnelle et du sexe (uniquement pour les non cadres). Ainsi, la comparaison n'est pas immédiate. Les barèmes ont donc été comparés pour plusieurs profils : un profil proche de la référence et un profil pour lequel le risque d'incidence en incapacité est légèrement inférieur. Voici les résultats obtenus, pour un âge moyen de 39 ans :

Population	BCAC	Profil référence	Rapport	Profil plus faible	Rapport
Cadres	0,88	1,21	1,36	0,52	0,59
Non cadres H	1,10	2,80	2,54	0,96	0,87
Non cadres F	1,73	3,66	2,11	1,25	0,72

FIGURE 5.7 – Comparaisons de nos barèmes à 100% à ceux du BCAC selon le profil

Pour le profil de référence, les barèmes construits sont plus élevés que ceux du BCAC mais pour un profil légèrement plus faible, ils sont inférieurs. Le barème moyen modélisé pour les cadres dépend de la répartition de la population. Il serait intéressant de comparer la moyenne des barèmes modélisés sur un portefeuille à celui du BCAC.

Les effets modélisés des variables ont été comparés à ceux employés actuellement, en fixant le profil à la référence.

L'impact de l'âge n'est pas le même. En effet, alors que pour le même profil, notre barème était supérieur à celui du BCAC à 39 ans, il est cette fois inférieur à 30 ans :

	30 ans		
Population	BCAC	Modèle	Rapport
C	0,65	0,37	0,57
NC H	0,81	1,05	1,30
NC F	1,43	1,37	0,96

FIGURE 5.8 – Comparaison de nos barèmes à 100% à ceux du BCAC pour le profil de référence à 30 ans

D'autre part, pour les profils étudiés, le correctif de la CSP est plus important dans nos barèmes alors que celui du sexe est plus faible :

Correctif CSP

	BCAC	Profil référence	Profil plus faible
39 ans	1,25	2,32	1,84
30 ans	1,25	2,84	

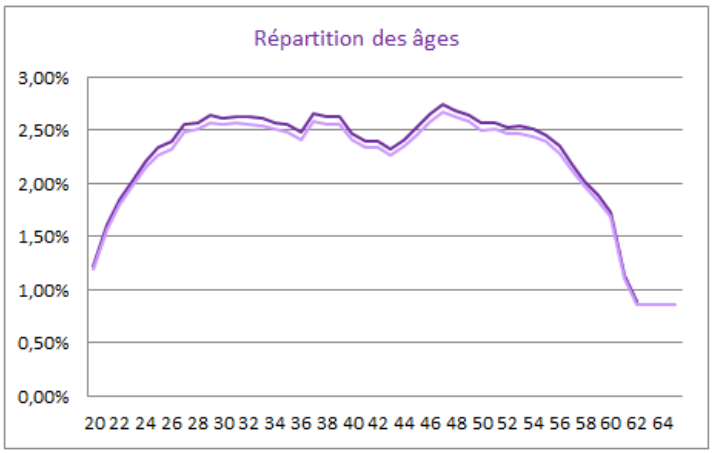
Correctif sexe

	BCAC	Profil référence	Profil plus faible
39 ans	1,57	1,31	1,31
30 ans	1,76	1,31	

FIGURE 5.9 – Comparaisons des correctifs CSP et sexe de nos barèmes à ceux du BCAC

5.3.3 Effet d'une réforme des retraites

Le changement d'âge de départ à la retraite joue sur la durée de la rente mais aussi sur l'âge moyen de la population. Pour estimer l'augmentation de l'âge moyen, nous nous sommes basés une précédente étude des effectifs d'assurés à chaque âge. Les effectifs des âges avant 62 ans ont été conservés et les effectifs au-delà sont supposés égaux à l'effectif correspondant à 62 ans. Cette hypothèse forte nous permettra de déduire l'augmentation maximale de l'âge moyen. Voici la courbe représentant la répartition des assurés par âge :



Age moyen actuel	Age moyen simulé
40,7	41,3
	0,6

FIGURE 5.10 – Répartition des âges et modification de l'âge moyen

Dans notre modèle de tarification, l'âge moyen est entier. Donc en simulant l'effet de la retraite en conservant d'une part l'âge moyen et en l'augmentant d'un an d'autre part, il sera possible de

déduire un encadrement de l'augmentation du barème. La modification de l'âge moyen influe à la fois sur l'incidence et sur le maintien.

L'effet est modélisé pour un profil de référence et pour un profil légèrement moins risqué, pour un allongement à 64 et à 65 ans [3] :

	Profil référence		Profil plus faible	
	Ecart relatif 64/62 ans	Ecart relatif 65/62 ans	Ecart relatif 64/62 ans	Ecart relatif 65/62 ans
Age	7,00%	10,35%	7,00%	10,35%
Age +1	11,22%	14,87%	11,98%	15,66%

FIGURE 5.11 – Effet d'une réforme des retraites sur le tarif

Ainsi, pour un départ décalé à 64 ans, l'augmentation du tarif pour ces profils serait comprise entre 7 et 12%. Pour un changement à 65 ans, elle serait comprise entre 10 et 16%.

5.4 Limites de l'approche et prolongement de l'étude

❖ *Limites de l'approche*

Lors de la modélisation de l'incidence indirecte, nous avons pris en compte les sinistres survenus entre 2012 et 2016 et avons recensé les passages connus fin 2019. Cependant, il se peut que des invalides de catégorie 1 n'aient pas déclaré leur passage s'ils savaient qu'ils ne seraient pas indemnisés. Or ces invalides silencieux peuvent ensuite changer de catégorie. Néanmoins, ce risque est limité puisque les éventuelles invalidités silencieuses survenues en 2012 avaient jusqu'à fin 2019 pour être déclarées. D'autre part, notre base montre que seuls entre 10 et 20% d'invalides changent de catégorie. Seul un intervalle peut être déterminé comme nous ne connaissons pas toujours le début de l'invalidité.

Par ailleurs, les invalides de catégorie 2 peuvent travailler, ce qui diminue leurs prestations. Il serait donc intéressant d'en tenir compte.

❖ *Prochaines étapes*

Nos barèmes d'invalidité ont été construits à partir des précédentes études sur le maintien et l'incidence en incapacité. Or, pour chacune de ces études, la population a été segmentée. Ainsi, la classe des assurés résulte de la combinaison de leur classe d'incidence en incapacité, de maintien en incapacité et de passage en invalidité. Le nombre de classes pour le barème est donc important. Pour éviter une sursegmentation de la population, une étape d'harmonisation des classes sera nécessaire.

Par ailleurs, il conviendra de poursuivre la comparaison des barèmes obtenus avec ceux utilisés actuellement afin de vérifier leur cohérence.

3. Comme pour le départ à 62 ans, l'âge est limité respectivement à 63,5 et 64,5 ans, cf 3.2.1.4 Validation du modèle et comparaison avec le BCAC

Enfin, pour obtenir le tarif final, la portabilité, l'exonération et la revalorisation devront être pris en compte. La portabilité permet à un salarié quittant l'entreprise de bénéficier d'une couverture pendant au maximum un an sans cotisations, sous conditions. Quant à l'exonération, elle correspond à l'absence de cotisation du salarié en arrêt de travail, qui bénéficie toujours de l'assurance décès.

Conclusion

Ce mémoire visait à proposer une méthode de tarification du risque invalidité.

Nos barèmes reposent principalement sur le maintien, l'incidence et la proportion d'invalides dans les trois catégories d'invalidité.

En premier lieu, un long travail sur les données a été nécessaire afin de reconstituer l'historique des sinistres et de fiabiliser les données. Les statistiques effectuées sur la base du maintien ont révélé une composition différente du portefeuille par rapport à celui étudié par le BCAC.

Par la suite, la modélisation du maintien en invalidité s'est révélée délicate. L'effectif disponible ne permettait pas d'obtenir des résultats robustes à tous les âges. En effet, l'incidence en incapacité aux âges très jeunes est faible. Il n'a pas non plus été possible de regrouper les âges par classes. L'ajustement par référence externe a finalement permis de fiabiliser la structure de notre table tout en conservant la spécificité de notre risque. Nos fonctions de survie sont inférieures à celles du BCAC aux âges jeunes mais supérieures aux âges plus élevés.

Le maintien n'a pas été segmenté pour construire le tarif. En effet, la quantification des effets des caractéristiques de la population assurée sur le maintien n'a pas pu aboutir en raison du volume de données. Néanmoins, cette étude a tout de même montré des observations intéressantes. Les sorties de l'état d'invalidité se révèlent très erratiques et il est difficile d'exprimer la durée de maintien à l'aide de variables explicatives. L'âge, le secteur d'activité et le département ont été estimés influents par deux méthodes différentes.

L'incidence, deuxième grande composante du tarif, a pu être estimée grâce à un précédent mémoire sur l'incidence en incapacité, qui s'appuyait sur la DSN. Pour estimer l'*invalidité indirecte*, le taux de passage de l'incapacité à l'invalidité a été modélisé grâce à une régression logistique. Il en ressort que les caractéristiques les plus déterminantes sont l'âge, le secteur d'activité et le département alors que le sexe n'est pas significatif. Quant à l'incidence en *invalidité directe* (en incluant les passages avant 120 jours d'incapacité), elle représente environ 7% de l'incidence en *invalidité indirecte*.

Le tarificateur construit sur Excel a permis de constater une forte influence de toutes les variables (âge, secteur d'activité, département, CSP et sexe). Le sexe, non significatif pour le passage en invalidité, l'est néanmoins dans le tarif en raison de son influence sur l'incidence en incapacité. Toutes choses égales par ailleurs, le tarif est plus élevé pour les femmes que pour les hommes, et pour les non-cadres que pour les cadres.

La principale limite de notre étude est la non prise en compte d'une éventuelle activité des invalides de catégorie 2. Cela diminuerait ainsi nos tarifs. Par ailleurs, la finalisation de l'étude demanderait

l'harmonisation des classes créées lors des différentes études (incidence et maintien en incapacité, et taux de passage en invalidité). De plus, il convient de poursuivre la comparaison des barèmes construits aux barèmes actuels. Enfin, la portabilité, l'exonération et la revalorisation doivent être prises en compte.

Table des figures

Bibliographie

[1] DJOFFON O. (2017) Modélisation de la survenance d'un sinistre dans le cas d'une asymétrie des classes et utilisation dans le cadre d'un modèle interne partiel. Mémoire, ISFA.

[2] ELIAS C. (2016) Construction d'une table d'expérience sur le maintien en invalidité. Mémoire, ISFA.

[3] FAVRE-BEGUET M. (2021). Protection Sociale. Cours, ISFA.

[4] FETOUI N. (2015) Impact de l'utilisation des tables d'expérience sur le provisionnement en prévoyance. Mémoire, Université Paris Dauphine.

[5] HUBER-CAROL C. (1994) "Durées de survie tronquées et censurées". Journal de la société statistique de Paris, tome 135, n° 4 , 3-23.
http://www.numdam.org/article/JSFS_1994__135_4_3_0.pdf

[6] LEURENT T. (2010) Construction de tables d'expérience des risques incapacité et invalidité. Mémoire, DUAS.

[7] MORLON P. (2017) Construction d'une loi de changement de catégorie d'invalidité et étude de l'impact sur les provisions mathématiques. Mémoire, DUAS.

[8] MTAR M. (2017) Gestion du risque incapacité : Etude de l'impact des tables d'expérience de maintien en incapacité et de passage en invalidité. Mémoire, ISUP.

[9] https://www.ayming.fr/insights/barometres-livres-blancs/ (baromètre de l'absentéisme), site consulté le 2 août 2021

[10] https://eric.univ-lyon2.fr/~ricco/cours/cours/pratique_regression_logistique.pdf (cours sur la régression logistique), site consulté le 29 juillet 2021

[11] http://www.ressources-actuarielles.net/ (site de Frédéric Planchet, modèles de durée), site consulté en octobre 2020

Annexes

Annexe 1 : Comparaison des biais de Kaplan-Meier et Harrington-Flemming

La fonction de hasard cumulée de Nelson-Aalen se calcule ainsi :

$$\hat{H}_{NA}(t) = \sum_{i,T_i \leq t} \frac{d_i}{r_i}$$

avec d_i le nombre de sorties en T_i et r_i l'exposition au risque.
La fonction de survie de Harrington-Flemming découle de la fonction de hasard de Nelson-Aalen :

$$\hat{S}_{HF}(t) = exp(\hat{H}_{NA}(t))$$

La fonction de survie de Kaplan-Meier est :

$$\hat{S(t)} = \prod_{T_{(i)} \leq t} \left(1 - \frac{d_i}{r_i}\right)$$

Donc :

$$ln\hat{S}_{KM}(t) - ln\hat{S}_{HF}(t) = \sum_{T_{(i)} \leq t} \left(ln\left(1 - \frac{d_i}{r_i}\right) + \frac{d_i}{r_i}\right)$$

Or, la fonction f telle que $f(x) = ln(1-x) + x$ est négative, donc $\hat{S}_{KM}(t) \leq \hat{S}_{HF}(t)$.

Annexe 2 : Code Kaplan-Meier

```
%macro km(perimetre) ;

/*Dénombrement des troncatures par pas de temps*/
proc sql ;
create table entree as
select age_deb_inv,seuil_tronc as
temps,count(seuil_tronc) as trc
from mylib.at_maintien(where=(&perimetre))
group by age_deb_inv,seuil_tronc ;
quit ;

/*Dénombrement des sorties par pas de temps*/
proc sql ;
create table sortie as
select age_deb_inv,duree as temps,count(duree) as s
from mylib.at_maintien(where=(&perimetre))
where censure=0
group by age_deb_inv,duree ;
quit ;

/*Dénombrement des censures par pas de temps*/
proc sql ;
create table censure as
select age_deb_inv,duree as temps,count(duree) as c
from mylib.at_maintien(where=&perimetre))
where censure=1
group by age_deb_inv,duree ;
quit ;

/*Estimation de la fonction de survie par l'estimateur de Kaplan-Meier*/

data km ;

merge entree sortie censure ;
by age_deb_inv temps ;

/*remplacement des valeurs manquantes par des 0 pour pouvoir sommer*/
if trc=. then trc=0 ;if s=. then s=0 ;if c=. then c=0 ;

/*événements en t-1*/
trc_prec=lag(trc) ;s_prec=lag(s) ;c_prec=lag(c) ;

/*ajout de l'exposition*/
retain exposition ;
if temps=0 then exposition=0 ; /*en t=0*/
```

```
else exposition=exposition+trc_prec-s_prec-c_prec;

/*ajout de S(t)*/
retain survie;
if temps=0 then survie=1;
else survie=survie*(1-s/exposition);

tps_an=floor(temps/365.25);

/*Ajout de S_IC (composante de l'IC)*/
retain S_IC;
if temps=0 then S_IC=0;
else S_IC=S_IC+s/(exposition*(exposition-s));

run;

data km_an(where=(survie_an ne .) keep=age_deb_inv tps_an survie_an borne_inf borne_sup
);
set km;
by age_deb_inv tps_an;
if temps=0 then do;survie_an=1; borne_inf=1;borne_sup=1;end;
if last.tps_an then
do;survie_an=survie;tps_an=tps_an+1;/*dernier jour précédent le changement d'année*/
borne_inf=survie*(1-1.96*sqrt(S_IC));borne_sup=survie*(1+1.96*sqrt(S_IC));end;
run;

%mend;
```

Annexe 3 : Spécificité de la classe d'âges 24-34 ans

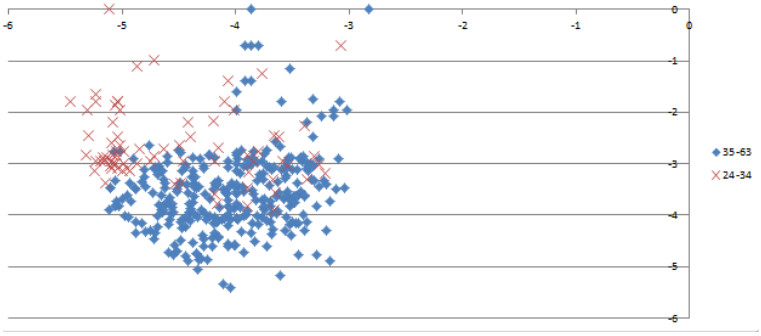


FIGURE 5.12 – Logits observés en fonction des logits du BCAC

Annexe 4 : Présentation du V de Cramér

Le V de Cramér permet de mesurer la liaison entre deux variable quantitatives et est basé sur la statistique du test du χ^2 de Pearson.

Soient X et Y deux variables qualitatives à respectivement K_1 et K_2 modalités et n le nombre d'observations.
Notons n_{k_1,k_2} l'effectif de la cellule correspondant aux modalités k_1 et k_2 du tableau de contingence de X et Y et $n_{k_1,.}$ (resp. $n_{.,k_2}$)l'effectif de la ligne (resp. colonne) associée à la modalité k_1 (resp. k_2).

$$\chi^2 = \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \frac{\left(n_{k_1,k_2} - \frac{n_{k_1,.} n_{.,k_2}}{n}\right)^2}{\frac{n_{k_1,.} n_{.,k_2}}{n}}$$

Le V de Cramér est alors défini ainsi :

$$V = \sqrt{\frac{\chi^2}{n \times min(K_1 - 1, K_2 - 1)}}$$

Le dénominateur de la fraction représente le maximum de la statistique du χ^2. Ainsi, le V de Cramér est le rapport entre la statistique du χ^2 et la valeur maximale possible. C'est pour cela que cette statistique est considérée plus fiable que la statistique du χ^2, qui est très sensible à la taille de l'échantillon.

Annexe 5 : Analyse de l'impact des variables sur les taux bruts de survie estimés par Kaplan-Meier

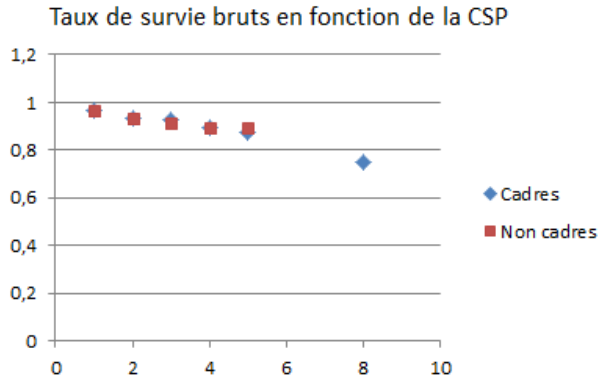


FIGURE 5.13 – Impact peu significatif de la CSP sur le maintien à 57 ans à première vue

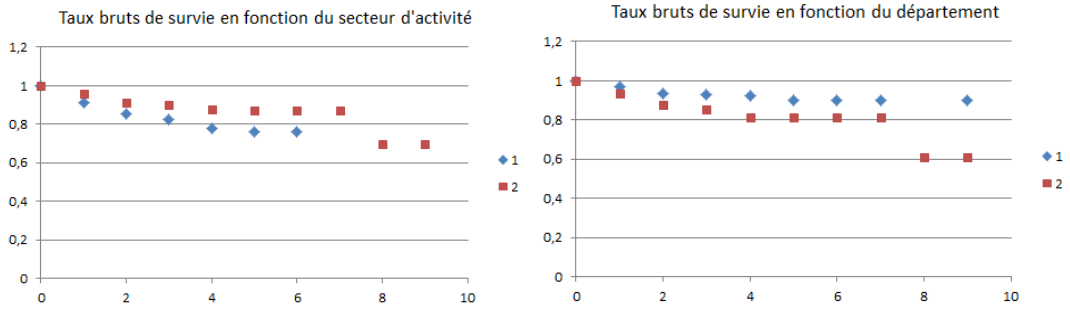


FIGURE 5.14 – Impact du secteur d'activité et du département sur le maintien à 57 ans

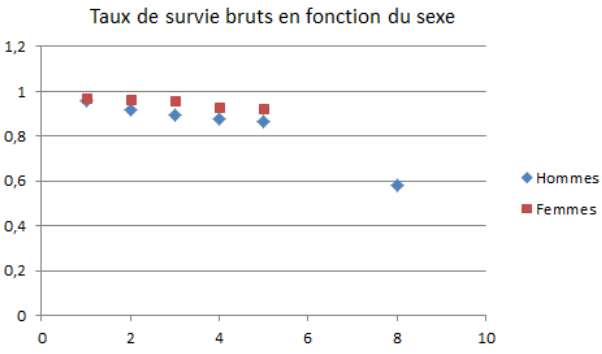


FIGURE 5.15 – Impact du sexe sur le maintien à 57 ans à première vue

Annexe 6 : Présentation sommaire du surapprentissage

Lorsque le nombre d'itérations augmente, l'erreur de prédiction diminue sur l'échantillon d'apprentissage. En revanche, au-delà d'un certain nombre de feuilles, l'erreur sur l'échantillon de validation augmente : on observe du surapprentissage. Le graphique suivant [4] illustre ce phénomène :

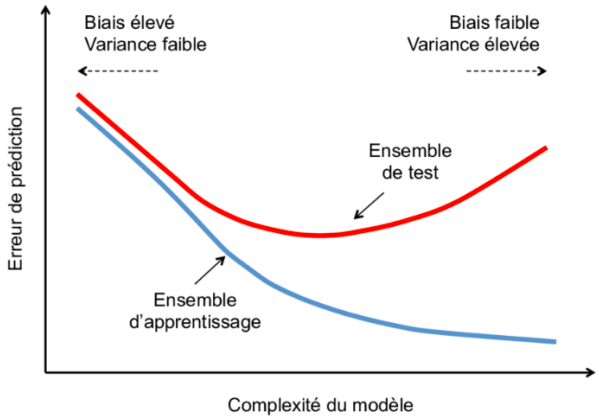


FIGURE 5.16 – Graphique représentant la notion de surapprentissage

4. Source : https://www.researchgate.net/publication/292151058_Inference_de_reseaux_d'interaction_proteine-proteine_par_apprentissage_statistique#pf36