



**Mémoire présenté pour la validation de la Formation
« Certificat d'Expertise Actuarielle »
de l'Institut du Risk Management
et l'admission à l'Institut des actuaires
le 28/04/2023**

Par : **Emmanuel THOUZEAU**

Titre : **Le risque incapacité d'un portefeuille prévoyance : construction de lois d'entrée, analyse de l'hétérogénéité par régression de Cox et étude du maintien**

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

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

Membres présents du jury de l'Institut des
actuaires :

Membres présents du jury de l'Institut du Risk
Management :

Secrétariat :

Bibliothèque :

Entreprise : PREVOIR VIE

Nom : SETTTON Franck

Signature et Cachet :

Franck Settton

PRÉVOIR-VIE
GROUPE PRÉVOIR
19 rue d'Aumale
75306 PARIS Cedex 09
Tél. 01 53 20 30 40

Directeur de mémoire en entreprise :

Nom : RODRIGUES Henri-Pierre

Signature : *Henri-Pierre Rodrigues*

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

Henri-Pierre Rodrigues

Signature(s) du candidat(s)

Emmanuel Thouzeau

RESUME

MOT CLES : Prévoyance, incapacité, estimateur, Kaplan-Meier, lissage, Whittaker-Henderson, optimisation, validation croisée, Cox, maintien, BCAC, référence externe, Brass.

Un arrêt de travail entraîne une perte de revenu dans de nombreuses situations. La perte est davantage marquée pour les Travailleurs Non-Salariés (TNS) pour lesquels les mécanismes d'indemnisation sont moins protecteurs. Une Interruption Temporaire de Travail (ITT) de plusieurs semaines est donc susceptible d'entraîner des conséquences plus ou moins lourdes en fonction du revenu des travailleurs.

Les garanties en cas d'incapacité prévoient une compensation de nature indemnitaire ou forfaitaire pour compenser au moins partiellement la perte de revenus. Elles peuvent être souscrites en tant que garanties obligatoires d'un contrat, ou comme des garanties optionnelles (par exemple en assurance emprunteur) et sont assorties d'une franchise. La franchise est habituellement choisie par l'assuré lors de la souscription et son niveau varie en fonction des populations et des types de contrats (90 jours sur des contrats en couverture d'ADP, 40 à 90 jours sur des contrats vendus à des salariés, 15 jours sur des contrats à destination des TNS).

Le Groupe Prévoir, assureur de personnes depuis 1910, intervient historiquement auprès des foyers populaires (particuliers ou professionnels), par l'intermédiaire d'un réseau debout. Le Groupe a entamé depuis 2012 un virage stratégique vers l'assurance emprunteur, d'abord par l'intermédiaire de sa marque blanche. Puis cette trajectoire de développement s'est concrétisée par des acquisitions importantes dans l'univers du courtage. La plupart des contrats de prévoyance commercialisés ou en run-off ont donc des garanties d'Incapacité Temporaire de Travail. Or les provisions de primes et de sinistres peuvent atteindre des niveaux importants sur ce type de garanties. L'objectif principal de ce mémoire est donc de fournir des lois d'expérience qui seront utilisées pour les travaux de provisionnement et de tarification. Cette étude du risque d'incapacité porte sur plusieurs portefeuilles de contrats, dont certains en run-off. L'ensemble des travaux présentés sont systématiquement soumis à la contrainte de la rareté des données disponibles. Cette rareté peut être absolue, comme c'est souvent le cas sur des risques comme l'invalidité. Mais elle peut aussi être relative, c'est-à-dire créée par la nécessité de segmentation de l'analyse du risque.

Le premier enjeu de l'étude revient à retraiter les données sinistres. Un même assuré peut disposer de plusieurs garanties ITT dans plusieurs contrats. Dans cette configuration, il convient en premier lieu de s'assurer que chaque sinistre n'est comptabilisé qu'une seule fois et sur sa durée totale. Un contrôle de cohérence entre les dates de survenances et d'indemnisations, relativement aux niveaux de franchise, est également nécessaire.

Le second enjeu est de construire les lois d'entrée en incapacité. Ces travaux doivent être entrepris sous le prisme de l'analyse de survie, c'est-à-dire en exprimant la problématique ainsi : « à chaque âge, quelle est la probabilité de ne pas observer l'évènement entrer en incapacité ? » Ces travaux doivent tenir compte de l'incomplétude des observations. Les données sont dites censurées (à droite) lorsque l'individu sort de la cohorte avant la fin de la période étudiée (résiliation entre deux âges entiers par exemple ou fin de l'étude). Elles sont tronquées lorsque l'individu entre dans la cohorte après le début de la période étudiée (souscription entre deux âges entiers par exemple). L'estimateur de Kaplan-Meier est adapté à cette configuration. Il s'agit d'un estimateur de la fonction de survie.

Une autre forme de troncature est introduite par les franchises. Seuls les sinistres dont la durée excède la période de franchise seront connus. Afin de ne pas introduire de biais, des lois propres à chaque niveau de franchise peuvent être établies. Enfin, les lois d'entrée en incapacité doivent être continues et régulières. Elles ne doivent pas présenter d'évolutions brusques ou de décrochages. Les travaux de lissage sont réalisés à partir de la méthode non-paramétrique de Whittaker-Henderson. Cette méthode nécessite un calibrage qui est paramétré par validation croisée.

Le troisième enjeu est de définir des coefficients de majoration ou d'abattement à appliquer à une population de référence. Cette problématique est particulièrement présente dans le cas de segmentations insuffisamment représentées dans les données pour établir une loi sur l'ensemble des âges (CSP, consommation de tabac etc.). Le modèle de Cox à risques proportionnels répond à cette problématique. Ce modèle permet de décomposer le risque d'entrée en incapacité entre un risque de base, dépendant du temps et identique pour tous les individus, et un risque ne dépendant que des covariables et indépendant du temps.

Enfin, une étude du maintien en incapacité est réalisée à des fins de comparaison avec la table réglementaire du BCAC. Cette étude repose également sur l'estimateur de Kaplan-Meier mais est limitée par la quantité de données disponibles. La lecture de la table du BCAC indique que la distribution des sinistres en fonction de leur durée décroît très rapidement. Dès lors, il apparaît nécessaire de disposer d'un nombre important de sinistres pour produire des taux de maintien significatifs au-delà des premiers mois d'indemnisation. Afin de se soustraire au moins partiellement de cette problématique, un modèle à référence externe de Brass est proposé. Les résultats obtenus indiquent que l'utilisation de la table réglementaire du BCAC conduit à un sur-provisionnement dans le cas des deux portefeuilles étudiés.

ABSTRACT

KEY WORDS: Provident insurance, incapacity, estimator, Kaplan-Meier, smoothing, Whittaker-Henderson, optimization, cross-validation, Cox, maintenance, BCAC, external reference, Brass.

A work stoppage leads to a loss of income. The loss is more significant for self-employed workers as their compensation mechanisms are less protective. A temporary work interruption of several weeks is therefore likely to have serious consequences on the total income of the workers.

In case of short-term disability, certain guarantees aim to compensate for part of this loss of income. These guarantees come with a deductible and can be a mandatory part of the contract or an optional one (e.g., mortgage insurance). The deductible is usually chosen by the insured when subscribing. Its level will vary based on the populations and contract types.

Since 1910, The Groupe Prévoir is an insurance provider that mainly targets working-class households (individuals or professionals), through a standing network. Most provident insurance contracts marketed or in run-off have incapacity guarantees. This study of the risk of incapacity covers several contract portfolios, some of which are in run-off. All the works presented are subject to the problem of the relative scarcity of available data.

The first challenge of this study is to reprocess the claims data. The same subscriber can have multiple incapacity guarantees in several contracts. In this configuration, it is necessary to start with ensuring that each claim is counted only once and over its total duration. A consistency check between the dates of occurrence and compensation, with respect to deductible levels, is also necessary.

The second challenge is to build the laws allowing the subscriber to get a work stoppage. This work must be undertaken under the prism of survival analysis, i.e., by expressing the problem as follows: At each age, what is the probability of not observing the event “entering incapacity”? This work must consider the incompleteness of the observations. The data is said to be censored (on the right) when the individual leaves the cohort before the end of the studied period (e.g.: termination between two even ages or end of the study). They are truncated when the individual enters the cohort after the start of the studied period (subscription between two even ages, for example). The Kaplan-Meier estimator is suitable for this configuration. It is an estimator of the survival function.

Another form of truncation is introduced by deductible levels. Only claims with durations exceeding the deductible period will be known. In order not to introduce bias, law specific to each level of deductible are established. Finally, the law of entry into short-term disability must be continuous and regular. They must not present sudden changes or drop-outs. Smoothing work is carried out using the non-parametric Whittaker-Henderson method. This method requires a calibration which is fitted by cross-validation.

The third challenge is to define the increase or reduction coefficients to be applied to a reference population. This problem is particularly present when segmentations are insufficiently represented in the data to establish a law for all ages (socio-professional category, tobacco consumption, etc.). The Cox proportional hazards model will answer this problem. This model will allow to break down the risk of incapacity into a basic risk, which depends on time and is identical for all individuals, and a risk which depends only on the covariates and is independent of time.

Finally, a study of the durations is carried out for the purpose of comparison with the regulatory table of the BCAC. This study also relies on the Kaplan-Meier estimator but is limited by the amount of data available. Reading the BCAC table indicates that the distribution of claims according to their duration is decreasing very rapidly. Therefore, it appears necessary to have many claims to produce significant duration statistics beyond the first months of compensation. To avoid at least partially this problem, an external reference model of Brass is proposed. The proposed results indicate that the use of the BCAC regulatory table leads to over-provisioning in the case of the two portfolios studied.

NOTE DE SYNTHÈSE

INTRODUCTION

Un arrêt de travail entraîne une perte de revenu dans de nombreuses situations. Cette perte est davantage marquée pour les Travailleurs Non-Salariés (TNS) pour lesquels les mécanismes d'indemnisation sont moins protecteurs. Une Interruption Temporaire de Travail (ITT) de plusieurs semaines est donc susceptible d'entraîner des conséquences plus ou moins lourdes en fonction du niveau de revenus des travailleurs. Pour cette raison, des acteurs privés de la protection sociale (assureurs, mutuelles, institutions de prévoyance) proposent des garanties complémentaires. Ces garanties prévoient une indemnisation au moins partielle de la perte de revenus liée à l'arrêt de travail.

Deux composantes participent au risque d'arrêt de travail. La première relève de la probabilité pour un assuré de connaître une période d'incapacité. Cette composante est exprimée sous la forme d'une fréquence généralement qualifiée de taux d'entrée en incapacité ou de taux d'incidence. La seconde composante tient à la durée de la période d'incapacité. Elle est généralement exprimée par une espérance de durée résiduelle en incapacité ou sous la forme d'une probabilité de maintien dans cet état à chaque pas d'ancienneté choisi. **L'examen de ces deux composantes s'inscrit donc dans une démarche de modélisation statistique des phénomènes de durées, également mentionnée sous le terme d'analyse de survie.**

L'objectif principal de cette étude est de fournir des lois d'expérience d'entrée et de maintien en incapacité. Cette démarche est réalisée d'une part pour le calcul des provisions, qui peuvent représenter des niveaux importants sur l'arrêt de travail. Et d'autre part pour les travaux de tarification, particulièrement exigeants compte tenu du dynamisme et de l'intensité concurrentielle du marché sur ces garanties.

PRESENTATION DES DONNÉES

Cette étude a pour objet l'estimation des lois de probabilité qui régissent les taux d'entrée en incapacité ainsi que l'estimation des durées de maintien sur un portefeuille de 250 000 contrats (ce qui représente environ 1,3 millions d'années d'exposition). Les données sont partitionnées en trois portefeuilles : la moitié concerne des contrats d'assurance emprunteur. L'autre moitié regroupe des contrats individuels de prévoyance à adhésion facultative, destinés aux salariés d'une part, et aux travailleurs non-salariés (TNS) d'autre part.

La population étudiée sur les portefeuilles de prévoyance « classiques » sont principalement composées de foyers populaires. A l'origine composée majoritairement d'hommes ouvriers, cette population a suivi une importante mutation depuis le déclin de ce type d'emplois en France dans les années 2000 pour devenir plus hétérogène et mixte aujourd'hui. La population étudiée sur le portefeuille d'assurance emprunteur est en revanche principalement composée de cadres, en raison des contraintes financières liées à l'accès à la propriété.

Compte tenu de cette forte hétérogénéité des assurés, il serait nécessaire de construire des lois segmentées. Mais le volume de données étant limité, de telles segmentation se traduirait par une forte volatilité des probabilités estimées. Afin de se soustraire à cette limite, des régressions de Cox sont réalisées pour évaluer l'effet de certaines variables sur le risque incapacité.

ELABORATION DES LOIS D'ENTRÉE EN INCAPACITÉ

L'élaboration des taux bruts d'entrée en incapacité consiste, en corolaire, à estimer la fonction de survie $S(t)$ des individus en capacité de mener leur activité professionnelle. Le choix de l'estimateur de la fonction de survie est contraint par l'impossibilité pour l'assureur d'observer intégralement l'ensemble des durées de survie compte tenu des troncatures et des censures.

Dans ces conditions, l'estimateur de Kaplan-Meier est retenu dans la mesure où il est réputé être l'unique estimateur cohérent de la fonction de survie. L'estimateur de la fonction de survie est donné ci-après avec d_j et n_j , respectivement le nombre d'événements d'intérêt et l'exposition au risque à la date j :

$$\hat{S}(t) = \prod_{T_j \leq t} \left(1 - \frac{d_j}{n_j}\right), \forall j \in [x, x+1[.$$

Comme l'atteste la **Figure I** ci-contre, à moins de disposer d'une quantité très importante de données à tout âge, les taux bruts issus du calcul précédent sont erratiques. C'est-à-dire que leur allure est chaotique et présente des sauts entre des tranches d'âge adjacentes. Ce phénomène est particulièrement marqué sur les âges extrêmes, pour lesquels la population sous risque est moins importante. Il est ainsi difficile de les lire et de les interpréter sous cette forme. Ces taux sont surtout parfaitement ajustés aux données d'apprentissage. Il est donc peu probable qu'ils s'ajustent correctement à la sinistralité des futurs assurés. **Un travail de lissage est donc nécessaire afin de généraliser les taux d'entrée en incapacité.**

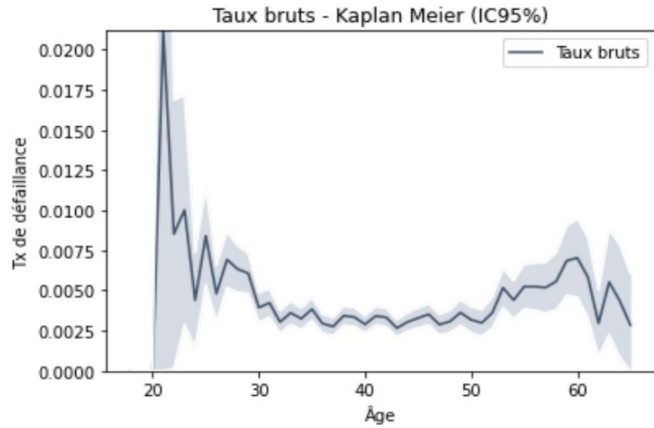


Figure I - Taux bruts d'entrée en incapacité

La méthode de lissage de Whittaker-Henderson cherche à minimiser la somme M d'une mesure F de fidélité aux données et d'une mesure S de régularité. Un coefficient h agit comme une constante de normalisation permettant de calibrer l'influence de chacune des deux mesures sur le lissage :

$$M = F + hS,$$

La fidélité est mesurée par la somme des écarts quadratiques pondérés par des poids (w_x) entre les taux bruts $\hat{q}_x$ et les taux lissés g_x :

$$F = \sum_{x=x_{inf}}^{x=x_{sup}} w_x (g_x - \hat{q}_x)^2.$$

La régularité est mesurée par la somme des carrés des différences avant d'ordre z des taux lissés :

$$S = \sum_{x=x_{inf}}^{x=x_{sup}-z} (\Delta^z g_x)^2 = \sum_{x=x_{inf}}^{x=x_{sup}-z} \left(\sum_{k=0}^z (-1)^{z-k} \binom{z}{k} g_{x+k} \right)^2.$$

En pratique, il convient de disposer de la forme matricielle de ces expressions. Le vecteur des taux lissés est noté g tel que $g = (g_x)_{x_0 \leq x \leq x_0+n}$, le vecteur des taux bruts est noté $\hat{q}$ tel que $\hat{q} = (\hat{q}_x)_{x_0 \leq x \leq x_0+n}$, la matrice diagonale des poids est notée W telle que $W = \text{diag}(w_x)_{x_0 \leq x \leq x_0+n}$.

L'expression du critère de fidélité F devient donc $F = {}^t(g - \hat{q})W(g - \hat{q})$.

Une matrice K_z , composée des coefficients binomiaux signés d'ordre z est introduite telle que $\Delta^z g = K_z g$. L'expression matricielle du critère de régularité S devient donc :

$$S = {}^t(K_z g)(K_z g).$$

Et donc $M = {}^t(g - \hat{q})W(g - \hat{q}) + h^t(K_z g)(K_z g)$. Son minimum est donné pour le vecteur des taux lissés g tel que :

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

Si z est généralement fixé à 2, il existe un consensus sur l'absence de solution idéale dans le choix de h . **Pour autant, afin de se soustraire au choix arbitraire de l'allure de la courbe, une approche par validation croisée est proposée.** Cette méthode consiste à découper le jeu d'entraînement en K sous-ensembles de même taille, appelés plis. Les plis sont à tour de rôle utilisés comme ensemble de validation tandis que les $K - 1$ plis sont utilisés comme ensemble d'entraînement. Cette méthode permet à toutes les observations d'être utilisées à la fois pour l'entraînement et pour la validation. Et chaque observation est utilisée exactement une seule fois pour la validation. L'avantage de l'utilisation d'une telle stratégie de calibrage est de produire une loi plus généraliste puisqu'elle est testée sur des données n'ayant pas servi à l'élaboration de son lissage. Elle doit donc permettre de mieux prédire la sinistralité des futurs assurés en évitant les effets de surapprentissage.

Les paramètres testés sont le degré de lissage z , la constante de normalisation h et l'utilisation, ou non, de la matrice des poids. Ainsi, pour chaque valeur testée des paramètres, une nouvelle loi d'entrée en incapacité est calculée puis lissée sur chaque jeu d'entraînement. Le dernier pli est utilisé pour mesurer la performance du modèle. Dans le cas du lissage, la performance du modèle peut être mesurée par un ratio de mortalité standardisé (ratio SMR). Cependant, un ratio SMR proche de 100% calculé globalement peut masquer des compensations entre les âges. Le choix se porte donc sur la statistique de test d'homogénéité du χ^2 . Il s'agit d'additionner les écarts quadratiques relatifs en chacun des points de la courbe :

$$X = \sum_{x_{\min}}^{x_{\max}} \frac{(d_x - \hat{d}_x)^2}{\hat{d}_x}.$$

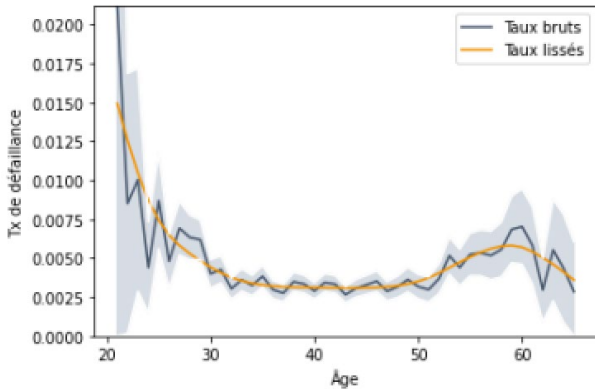


Figure II - Taux lissés pour $z=2$ et $h=23$

Le nombre de degrés de liberté n'a pas d'importance ici, dans le sens où il ne s'agit pas de déterminer une *p-valeur* mais de déterminer les taux lissés les plus adaptés aux échantillons de validation, c'est à dire présentant la statistique de test la plus faible. Cette méthode aboutit à fixer $z = 2$, $h = 23$ et à ne pas utiliser la matrice des poids. Les taux lissés sont présentés dans la **Figure II** ci-contre. Ces derniers présentent un aspect beaucoup plus régulier.

ETUDE DE L'HETEROGENEITE DU PORTEFEUILLE

Les taux d'entrée en incapacité précédemment présentés sont calibrés sur l'intégralité des portefeuilles. Ils traduisent le comportement moyen de nombreux individus hétérogènes. Cette simplification est indispensable pour disposer de suffisamment de données. Toutefois, afin de réaliser des travaux de tarification, des calculs de provisions, ou simplement de mieux comprendre les forces sous-jacentes, il est nécessaire d'affiner ces lois en évaluant l'impact de variables exogènes sur le risque. Plus précisément, **l'objectif est de disposer de coefficients de majoration de ou de minoration du risque pour chacune de ces variables. La méthode statistique se porte sur la régression de Cox à hasard proportionnel.** Le modèle de Cox repose sur la décomposition du risque de défaillance en deux composantes :

- Un risque de base noté $\lambda_0(t)$, dépendant du temps t et commun à tous les individus (i.e. indépendant des caractéristiques des individus).
- Un risque additionnel $\exp(\beta^T Z_i)$ ne dépendant pas du temps mais des caractéristiques des individus. β correspond au vecteur des p paramètres et Z_i à la valeur des p covariables pour l'individu i .

La fonction de hasard pour l'individu i est ainsi donné par $\lambda_i(t|Z_i, \beta) = \exp(\beta^T Z_i) \lambda_0(t)$. Aucune hypothèse n'est faite sur la distribution de $\lambda_0(t)$ qui correspond à la fonction de hasard, inconnue, pour $z = 0$. En ce sens, le modèle de Cox est dit semi-paramétrique.

Dans ces conditions, le rapport des fonctions de hasard (*risk ratio*) entre deux individus est indépendant du temps :

$$\frac{\lambda_1(t|Z_1, \beta)}{\lambda_2(t|Z_2, \beta)} = \frac{\exp(\beta^T Z_1)}{\exp(\beta^T Z_2)} = \exp(\beta^T (Z_1 - Z_2)).$$

Le modèle de Cox reposant donc sur une proportionnalité stable des risques à travers le temps, il convient de s'assurer que cette hypothèse est respectée. Il est en particulier utile de tracer les résidus de Schönefeld standardisés. Concrètement, il s'agit de calculer, pour chaque événement, la différence entre la valeur de la variable de l'individu concerné et la moyenne pondérée de cette variable pour l'ensemble des individus à risque. Si l'hypothèse de hasard proportionnel est vérifiée, les résidus doivent être distribués de façon homogène au cours du temps. Therneau et Grambsch (1994) ont établi que :

$$E[S_{kj}] + \hat{\beta}_j \approx \beta_j(t_k),$$

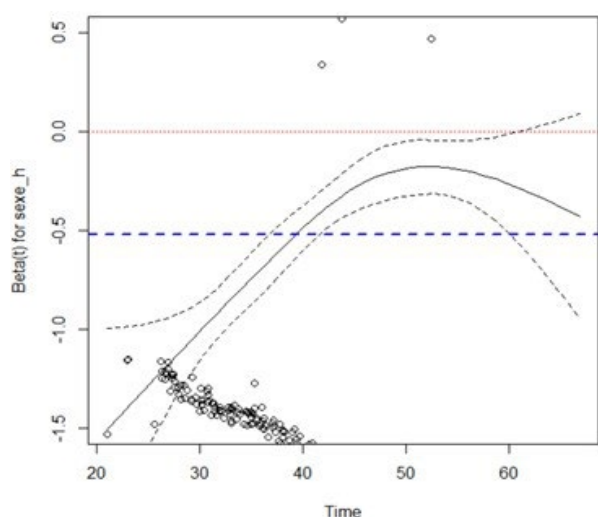


Figure III - Résidus de Schönefeld pour la variable sexe

où S_{kj} est le résidu standardisé pour la variable j au moment de l'événement k et $\hat{\beta}_j$ le coefficient indépendant du temps issu du modèle de Cox. Il est dès lors possible d'utiliser la représentation graphique de $S_{kj} + \hat{\beta}_j$ en fonction du temps (après lissage) pour révéler la forme fonctionnelle de $\beta(t)$ pour chacune des variables. La **Figure III** ci-contre permet d'apprécier cette représentation pour la variable sexe sur le portefeuille emprunteur. La ligne bleu correspond au coefficient de la régression pour un homme comparativement à une femme. Le tracé de $\beta(t)$ n'a pas l'apparence d'une droite horizontale, ce qui indique que l'effet de cette variable n'est pas stable dans le temps. Dès lors, le coefficient issu de la régression est un coefficient moyen.

Afin de gérer la dépendance au temps de l'effet de certaines variables, il est possible de partitionner le temps sur des tranches d'âges. Chaque individu est dupliqué sur chaque intervalle de temps correspondant au découpage retenu. La duplication des assurés sur les différents intervalles est sans impact sur la loi de survie puisqu'en tout instant la population sous risque est identique au modèle d'origine (sorties par censure). Il est ainsi possible de créer des fonctions en escalier afin de disposer d'un coefficient sur chaque intervalle de temps sélectionné (hasard proportionnel par morceaux), tel que présenté dans la **Figure IV**.

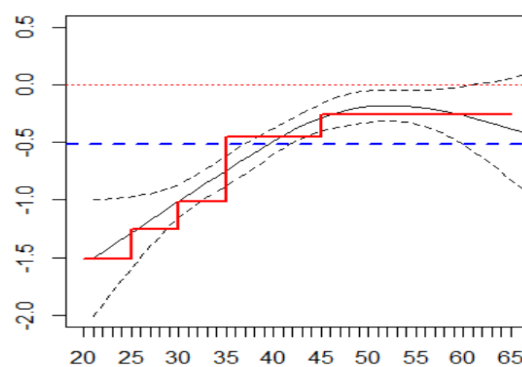


Figure IV - Fonction en escalier issue du modèle partitionné.

Les coefficients issus de la régression ont illustré un certain nombre de patterns communs aux portefeuilles étudiés.

Parmi les résultats significatifs, il est apparu que l'impact du sexe sur les taux d'incidence est principalement lié à la couverture des grossesses pathologiques.

Il est également apparu que l'effet de la franchise est très significatif sur le risque d'entrée en incapacité. Précisément, l'augmentation de 30 jours de la période de franchise s'est traduite par une division par 2 du risque. Ce constat tient à la fois au risque moral et au phénomène de troncature. De nombreuses études soulignent d'ailleurs l'existence d'une corrélation entre niveau de couverture et risque d'incapacité.

Déjà largement illustrés dans la littérature, des écarts significatifs de risque sont également apparus selon les départements de résidence. Les catégories socioprofessionnelles (CSP) ont également un impact important sur le risque d'incapacité. Pour les Travailleurs Non-Salariés (TNS), cette observation se décline au niveau des caisses professionnelles. Enfin, le niveau d'indemnisation se trouve positivement corrélé au risque d'entrer en incapacité.

Les résultats ont mis en exergue, sur certains portefeuilles, l'absence de corrélation entre l'instauration d'une surprime imposée à la suite de la sélection médicale et le risque. Ce constat doit amener l'assureur à s'interroger sur la sincérité des réponses apportées au questionnaire médical. Et peut-être à engager un travail de lutte contre la fraude.

ETUDE DU MAINTIEN EN INCAPACITE

Le maintien est la seconde composante du risque d'incapacité. Il s'agit d'une appréciation du temps écoulé entre le début et la fin de la période d'incapacité. La période d'incapacité peut prendre fin pour diverses raisons :

- La reprise du travail si les capacités sont retrouvées.
- L'entrée en invalidité si les capacités ne sont pas retrouvées par l'individu au bout d'une durée prévue aux termes du contrat. Un passage en invalidité est également possible en amont si l'état de l'assuré est considéré comme irréversible.
- La retraite ou le décès de l'individu.

La Sécurité Sociale prévoit l'entrée automatique en invalidité après 36 mois écoulés en incapacité. Cette durée est donc généralement la norme pour les contrats d'assurance prévoyant une garantie invalidité en plus de la garantie incapacité.

La mesure de la durée d'un épisode d'incapacité fait intervenir des modèles de survie dans lesquels l'évènement d'intérêt est la reprise du travail. Comme lors de la construction des lois d'entrée en incapacité, les données sont censurées à droite et tronquées à gauche. Les censures sont issues d'individus dont les prestations seraient arrêtées avant la reprise du travail (passage en invalidité, passage à la retraite, fin de la période d'étude, etc.). Les troncatures sont introduites par les franchises. Le choix de la méthode se porte naturellement sur l'estimateur de Kaplan-Meier, déjà présenté. Les fonctions de survie (i.e. de maintien en incapacité) sont généralement calculées par âge ce qui aboutit à une table à deux dimensions : âge / ancienneté.

Les taux bruts de sortie d'incapacité issus de ce calcul peuvent être lissés par la méthode de Whittaker-Henderson. Cette dernière est alors adaptée au cas à deux dimensions pour réaliser simultanément un lissage vertical et horizontal. En pratique, cette méthodologie nécessite un nombre important d'observations pour aboutir à des résultats fiables, particulièrement sur les anciennetés et âges extrêmes. Concernant les données étudiées, il a donc été choisi une méthode d'ajustement à une table externe de référence. Dans le cas des lois de maintien en incapacité, le choix se porte naturellement sur les tables du BCAC, et plus précisément sur la version de 2013 (non réglementaire)

Le modèle semi-paramétrique à référence externe de Brass permet de relier la loi d'expérience et la table de référence via une régression linéaire :

$$\text{logit}(q_{x,t}^{\text{ajusté}}) = \theta_1 + \theta_2 * \text{logit}(q_{x,t}^{\text{BCAC}}) + \varepsilon_{x,t},$$

$$\text{Et donc : } q_{x,t}^{\text{ajusté}} = \frac{\exp(\theta_1 + \theta_2 * \text{logit}(q_{x,t}^{\text{BCAC}}) + \varepsilon_{x,t})}{1 + \exp(\theta_1 + \theta_2 * \text{logit}(q_{x,t}^{\text{BCAC}}) + \varepsilon_{x,t})}.$$

Les taux de sortie sont reliés via leur transformation logistique (*logit*) afin de limiter les taux ajustés à l'intervalle [0,1] sur lequel se définissent les probabilités.

Il est ensuite d'usage de valider graphiquement les résultats en comparant les sorties observées aux sorties modélisées, d'une part par ancienneté et d'autre part par tranche d'âge. Les intervalles de confiance bilatéraux à 95% sont également calculés par approximation normale. La comparaison par ancienneté entre les sorties observées et les sorties modélisées est affichée dans la **Figure V**, ci-contre, pour le portefeuille d'assurance emprunteur. Seules les durées supérieures à la franchise de 90 jours sont affichées. L'ajustement semble capter correctement l'ensemble des anciennetés. Il apparaît des taux de sortie trop élevés sur le premier mois, traduisant un compromis trouvé par la régression entre les taux de sortie des quatre premiers mois. Sur les anciennetés plus élevées, les sorties modélisées et les sorties observées sont globalement superposées.

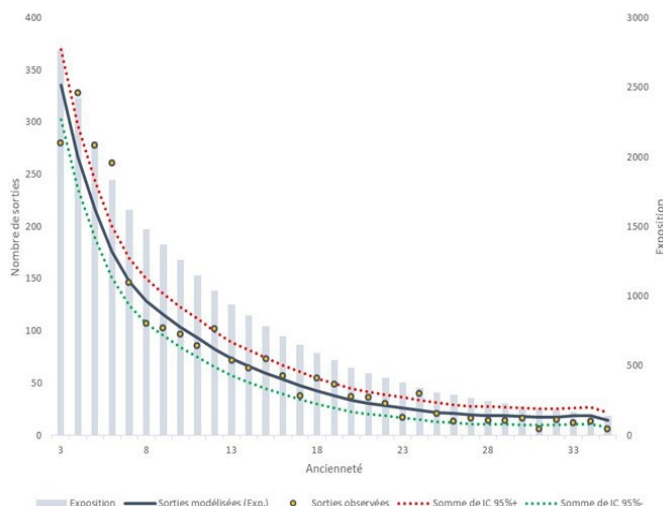


Figure V - Comparaison entre les sorties observées et les sorties modélisées, par ancienneté

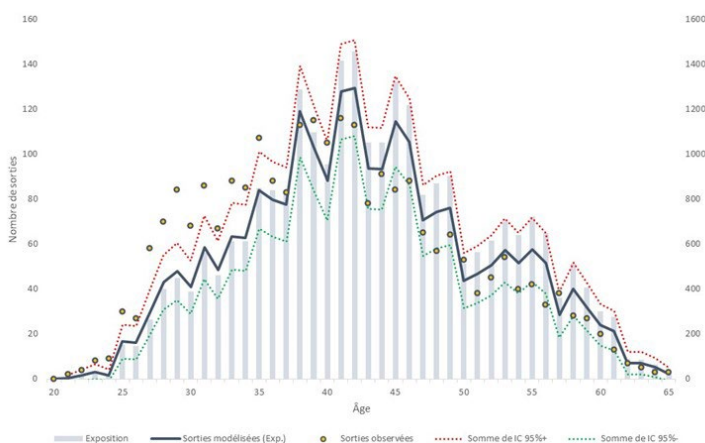


Figure VI - Comparaison entre les sorties observées et les sorties modélisées, par âge

La comparaison par âge, telle que présentée dans la **Figure VI** entre les sorties modélisées et les sorties observées fait apparaître une sous-estimation jusqu'à 35 ans, puis une surestimation après 40 ans. Sur cette deuxième plage, les sorties observées sont néanmoins encadrées par les intervalles de confiance des sorties modélisées. En imposant la même régression linéaire quel que soit l'âge, le modèle de Brass n'est pas en mesure de satisfaire plus précisément l'ensemble des points du modèle. Les pondérations ont néanmoins permis de trouver un bon compromis sur les intersections présentant les plus fortes expositions.

En tout état de cause, il convient de souligner que l'utilisation des taux ajustés permet ici d'atteindre un ratio SMR (sorties observées / sorties modélisées) global de 96% contre 132% en utilisant la table réglementaire du BCAC 2010, jusqu'à présent utilisée pour le calcul des provisions. Dans ces conditions, la

table d'expérience ajustée semble plus appropriée que la table réglementaire. Pour comparer ces deux tables, il est d'usage de calculer les espérances de durée résiduelle en incapacité :

$$e_{x,t} = \frac{1}{L_{x,t}} * \sum_{k=0}^{max-t} L_{x,k+t},$$

avec x l'âge de l'assuré à la survenance du sinistre, t l'ancienneté de l'état d'incapacité. Les espérances de durée résiduelle sont tracées dans la **Figure VII**, ci-contre, pour la table d'expérience. La forme est caractéristique : la baisse des taux de sortie se traduit dans un premier temps par un allongement des espérances de durée résiduelle. Puis le temps maximum restant à courir jusqu'au passage en invalidité étant de plus en plus faible, les durées résiduelles baissent jusqu'à l'ancienneté maximum de la table. Cette mesure peut être utilisée pour comparer les effets de provisionnement entre deux tables de maintien en incapacité. La comparaison est effectuée grâce au calcul des écarts relatifs entre les deux tables :

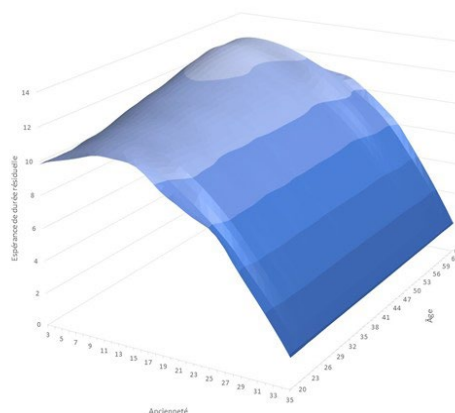


Figure VII - Espérances de durée résiduelle en

$$\frac{e_{x,t}^{BCAC_{2010}} - e_{x,t}^{XP}}{e_{x,t}^{XP}}.$$

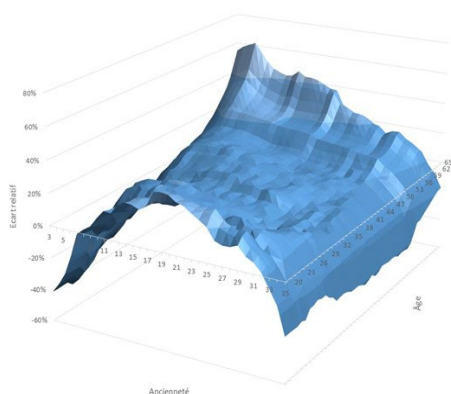


Figure VIII - Ecart relatif des barèmes de la table d'expérience et du BCAC

Les écarts relatifs entre durées de la table ajustée (**Figure VII**) et la table du BCAC 2010 sont présentés dans la **Figure VIII**. Ces derniers indiquent que l'utilisation de la table réglementaire conduit, dans cet exemple, à un excès de provisionnement de l'ordre de 10% à 15% sur les anciennetés et âges centraux. Les coefficients de la table ajustée sont cependant dégradés les premiers mois pour les moins de 35 ans, puis les trois derniers mois pour l'ensemble des âges. Cette dégradation sur les anciennetés élevées s'explique par les écarts entre la table réglementaire du BCAC 2010, par rapport à laquelle la comparaison est effectuée, et la table du BCAC 2013 ayant servi de référence externe au modèle. Sur cette dernière, les taux de sortie conditionnels sur les anciennetés élevées sont nettement

plus faibles. Globalement, l'écart relatif moyen pondéré entre les durées résiduelles de la table d'expérience ajustée et de la table du BCAC 2010 est d'environ 6% sur ce portefeuille.

En présence d'échantillons de petites tailles le modèle de Brass a permis de calibrer les taux de sortie en utilisant la table du BCAC 2013 comme référence externe. Cette méthode a apporté **une bonne stabilité au modèle, au prix d'un ajustement imparfait aux données d'origines. En particulier, la contrainte liée à l'utilisation d'un unique jeu de coefficients de régression a conduit à un ajustement unique sur tous les âges. Ces taux ajustés seront néanmoins plus adaptés à la sinistralité des portefeuilles que ceux issus de la table réglementaire du BCAC 2010 jusqu'à présent utilisée dans le calcul des provisions.**

SUMMARY

INTRODUCTION

A work stoppage leads to a loss of income in many situations. The loss is more significant for self-employed workers for whom the compensation mechanisms are less protective. A several weeks' interruption is therefore likely to have serious consequences on workers' total income. For this reason, private social protection actors (such as insurance companies) offer additional guarantees. These guarantees provide partial, or, in some cases, total compensation for the loss of income related to the work interruption.

Two components contribute to the risk of short-term disability leave. Firstly, the probability for an insured to experience a period of incapacity. It is expressed in the form of a frequency generally qualified as the rate of entry into incapacity or the incidence rate. The second component relates to the duration of the work interruption. It is generally expressed in the form of an expected residual duration in short-term disability or in the form of a probability of remaining in this state at each selected seniority step.

The purpose of this study is to estimate the probability laws that govern short-term disability entries and the study of exogenous variables, as well as the estimation of the durations of work stoppage on a provident portfolio. **This approach requires the use of specific statistical tools to the statistical modeling of duration phenomena, also referred to as survival analysis.**

DATA PRESENTATION

The study covers approximately 250,000 contracts (which represents about 1.3 million years of exposure). About half are mortgage insurance contracts. The other half are optional personal provident insurance contracts intended for employees and self-employed workers. In case of work stoppage, the guarantees offer the payment of daily allowances or the waiver of insurance premiums. They can be the main guarantee of the contract or an optional guarantee.

The population studied on the provident portfolio is mainly made up of working-class households. Originally composed mainly of male workers, this population has undergone a profound change with the decline of this type of employment in France in the 2000s to become more heterogeneous and mixed today. The population studied on the mortgage insurance portfolio is on the other hand mainly composed of upper-class workers because of the financial constraints linked to access to property.

Given this strong heterogeneity of policyholders, it would be necessary to construct segmented laws. But the volume of data being limited, such segmentation would result in high volatility of the estimated probabilities. To avoid this limit, Cox regressions are carried out to assess the effect of certain variables on the risk of work stoppage.

DETERMINATION OF CRUDE RATES OF ENTRY INTO INCAPACITY

The elaboration of the crude rates of entry in incapacity state consists, as a corollary, in estimating the survival function $S(t)$ of individuals able to carry out their professional activity. The choice of the survival function estimator is constrained by the impossibility for the insurer to fully observe all the survival durations given the incompleteness of the data (censors et truncations).

The Kaplan-Meier estimator is retained insofar as it is deemed to be the only consistent estimator of the survival function. The survival function estimator is given below with d_j and n_j , respectively the number of events of interest and the risk exposure on date j :

$$\hat{S}(t) = \prod_{T_j \leq t} \left(1 - \frac{d_j}{n_j}\right), \forall j \in [x, x+1[.$$

SMOOTHING OF INCAPACITY ENTRY RATES

As shown in **Chart I** opposite, unless you have a very large amount of data on any age, the crude rates from the previous calculation are erratic. Their pace is chaotic and presents jumps between adjacent age groups. This phenomenon is particularly marked at extreme ages, for which the population at risk is smaller. It is therefore difficult to read and interpret them in this form. Above all, these rates are perfectly adjusted to the training data. It is therefore unlikely that they will adjust correctly to the loss experience of future policyholders. **A smoothing work is therefore necessary to generalize the rates of entry into incapacity.**

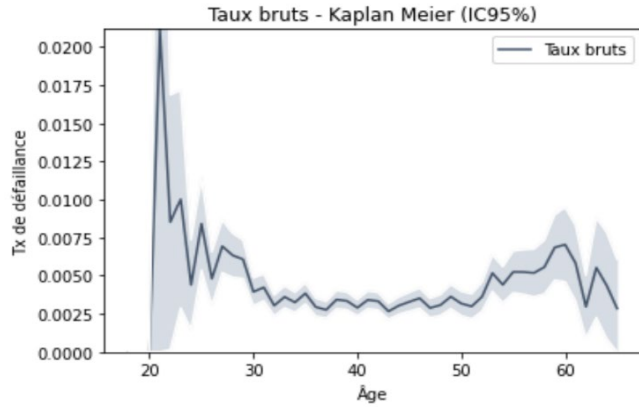


Chart I – Crude rates of entry in incapacity

The Whittaker-Henderson method is a smoothing method that seeks to minimize the sum M of a data fidelity measure F and a regularity measure S . A coefficient h acts as a normalization constant making it possible to calibrate the influence of each of the two measurements on the smoothing:

$$M = F + hS,$$

Fidelity is measured by the sum of the squared deviations weighted between the crude rates $\hat{q}_x$ and the smoothed rates g_x :

$$F = \sum_{x=x_{inf}}^{x=x_{sup}} w_x (g_x - \hat{q}_x)^2.$$

Regularity is measured by the square of the sum of the z-order forward differences of the smoothed rates:

$$S = \sum_{x=x_{inf}}^{x=x_{sup}-z} (\Delta^z g_x)^2 = \sum_{x=x_{inf}}^{x=x_{sup}-z} \left(\sum_{k=0}^z (-1)^{z-k} \binom{z}{k} g_{x+k} \right)^2.$$

In practice, it is necessary to have the matrix form of these expressions. The vector of smoothed rates is denoted g such that $g = (g_x)_{x_0 \leq x \leq x_0+n}$, the vector of crude rates is denoted $\hat{q}$ such that $\hat{q} = (\hat{q}_x)_{x_0 \leq x \leq x_0+n}$, the diagonal matrix of weights is denoted W such that $W = \text{diag}(w_x)_{x_0 \leq x \leq x_0+n}$.

The expression of the fidelity criterion F therefore becomes $F = {}^t(g - \hat{q})W(g - \hat{q})$.

A matrix K_z , composed of signed binomial coefficients of order z is introduced such that $\Delta^z g = K_z g$. The matrix expression of the regularity criterion S therefore becomes:

$$S = {}^t(K_z g)(K_z g),$$

So that $M = {}^t(g - \hat{q})W(g - \hat{q}) + h^t(K_z g)(K_z g)$ and whose minimum is given for the vector of smoothed rates:

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

If z is generally fixed on 2, there is a consensus on the absence of an ideal solution in the choice of h . **However, to avoid the arbitrary choice of the shape of the curve, a cross-validation approach is proposed.** This method consists of dividing the training set into K subsets of the same size, called folds. The folds are in turn used as a validation set while the $K - 1$ folds are used as a training set. This method allows all observations to be used for both training and validation. And each observation is used exactly once for validation. The advantage of using such a calibration strategy is to produce a more general law since it is tested on data that was not used to develop the smoothing. It must therefore make it possible to better predict the loss experience of future policyholders by avoiding the effects of overlearning.

The parameters tested are the degree of smoothing z , the normalization constant h and the use, or not, of the weight matrix. Thus, for each tested value of the parameters, a new incapacity entry law is calculated and then smoothed on each training set. The last fold is used to measure the performance of the model. In the case of smoothing, the performance of the model can be measured by a standardized mortality ratio (SMR ratio). However, an SMR ratio close to 100% calculated globally can mask compensations between ages. The choice therefore falls on the χ^2 homogeneity test statistic. This involves adding the relative squared deviations at each point on the curve:

$$X = \sum_{x_{min}}^{x_{max}} \frac{(d_x - \hat{d}_x)^2}{\hat{d}_x}.$$

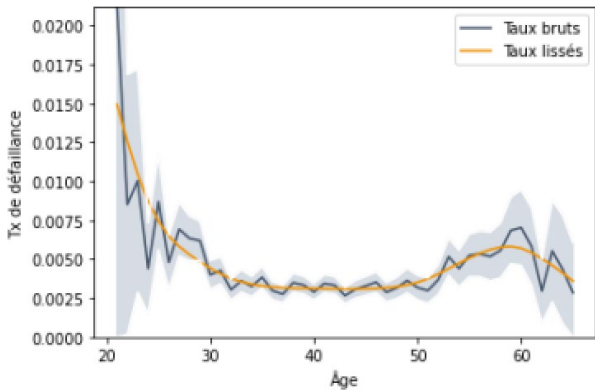


Chart II – Smoothed rates for $z=2$ et $h=23$

The number of degrees of freedom does not matter here, in the sense that it is not a question of determining a p -value but of determining the smoothed rates best suited to the validation samples, i.e., with the lowest test statistic. This method results in fixing $z = 2$, $h = 23$ and not using the weight matrix. The smoothed rates are shown in **Chart II** opposite (orange line). These have a much more regular appearance.

STUDY OF THE HETEROGENEITY OF THE PORTFOLIO

The probabilities of having a short-term disability, previously presented, are calibrated on the entire portfolio. They translate the average behavior of many heterogeneous individuals. To carry out pricing calculation, to refine provision calculations, or simply to get a better understanding in the underlying forces, it is necessary to refine these laws by evaluating the impact of exogenous variables on risk. More specifically, **the objective is to get risk increase or reduction coefficients for each of these variables. The statistical method is based on proportional hazard Cox regression.** The Cox model is based on the decomposition of default risk into two components.

- A basic risk denoted $\lambda_0(t)$, dependent on time t , and common to all individuals.
- An additional risk $\exp(\beta^T Z_i)$, independent on time but dependent on the characteristics of individuals. β corresponds to the vector of the p parameters and the z_i to value of the p covariates for the individuals i .

The default risk for individual i is thus given by $\lambda_i(t|Z_i, \beta) = \exp(\beta^T Z_i) \lambda_0(t)$. No assumption is made on the distribution of $\lambda_0(t)$ which corresponds to the hazard function, unknown, for $z = 0$. In this sense, the Cox model is said to be semi-parametric.

Under these conditions, the risk ratio between two individuals is independent of time:

$$\frac{\lambda_1(t|Z_1, \beta)}{\lambda_2(t|Z_2, \beta)} = \frac{\exp(\beta^T Z_1)}{\exp(\beta^T Z_2)} = \exp(\beta^T (Z_1 - Z_2)).$$

Since the Cox model is therefore based on a stable proportionality of risks over time, it is necessary to ensure that this constraint is respected. It is particularly useful to plot the standardized Schönefeld residuals. Concretely, this involves calculating, for each event, the difference between the value of the variable of the individual concerned and the weighted average of this variable for all the individuals at risk. If the proportional hazards assumption is hold, then the residuals should be homogeneously distributed over time. Therneau et Grambsch (1994) established that:

$$E[S_{kj}] + \hat{\beta}_j \approx \beta_j(t_k),$$

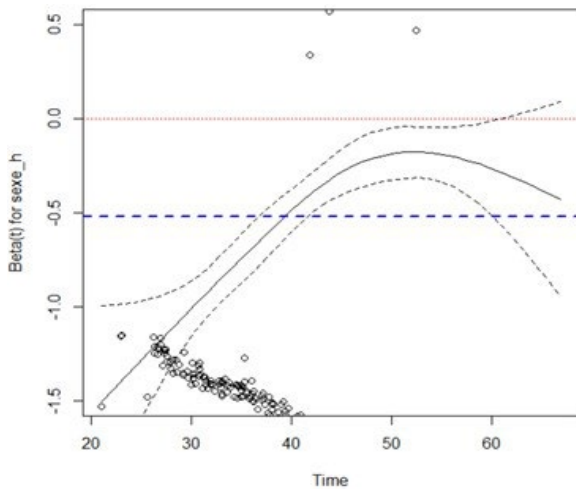


Chart III - Schönefeld residuals for the gender variable

Where S_{kj} is the standardized residual for the variable j at event k moment, and β_j is the time independent coefficient from Cox model. It is therefore possible to plot $S_{kj} + \hat{\beta}_j$ as a function of time (smoothed function) to reveal the form of $\beta(t)$ for each variable. **Chart III** opposite shows this representation for the sex variable. The blue line corresponds to the regression coefficient for a man relative to a woman. The Schönefeld residuals indicate that the effect of gender evolves over time. Therefore, the coefficient resulting from the regression is an average coefficient.

To manage the time dependence of the effect of certain variables, it is possible to partition the time by age groups. Each individual is duplicated over each time interval corresponding to the step selected. The duplication of policyholders over the different intervals has no impact on the law of survival since the population at risk is identical to the original model (exits by censoring). It is thus possible to create staircase functions to have a coefficient on each selected time interval (piecewise proportional hazard), as presented in **Chart IV**.

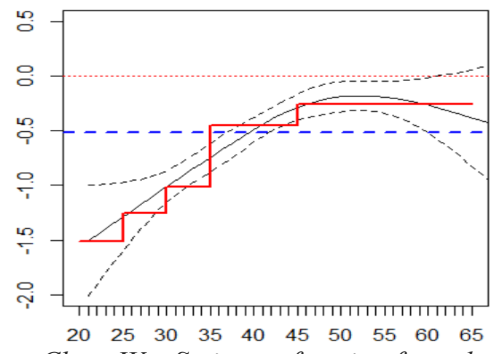


Chart IV – Staircase function from the partitioned model.

The coefficients from the regression have illustrated a certain number of patterns common to the portfolios studied.

Among the significant results, it appeared that the impact of gender on incidence rates is mainly related to the coverage of pathological pregnancies.

It also appeared that the deductible has significant effect on the risk of short-term disability. Specifically, a 30-day increase in the deductible period results in a reduction in risk by a factor of 2. This observation is due to both moral hazard and the phenomenon of truncation. Numerous studies underline a correlation between the level of coverage and the risk of incapacity.

Already widely illustrated in the literature, significant differences in risk have also appeared according to the place of residence. Socio-professional categories (CSP) also have a significant impact on the risk. For self-employed workers, this observation was applied to the professional funds. The level of compensation was found to be positively correlated with the risk of becoming incapacitated.

The results highlighted the absence of effect of an additional premium imposed following medical selection on certain portfolios. This observation should lead the insurer to question the sincerity of the answers given to the medical questionnaire. And maybe start some anti-fraud detection.

STUDY OF MAINTAINING IN INCAPACITY

The second component of incapacity risk is duration. This is the time elapsed between the entry and the end of the period of incapacity. The period of incapacity may end for several reasons:

- Resuming of work when the capacities are recovered.
- Entry into long-term disability if the capacities are not recovered by the individual at the end of a period provided in the contract. A change to long-term disability is also possible upstream if the capacities are considered as non-recoverable.
- Retirement or death of the individual.

Social Security automatically set entry into invalidity after 36 months elapsed in incapacity. This duration is therefore generally the norm for insurance contracts providing disability cover in addition to incapacity cover.

The duration measurement of a short-term disability episode intuitively involves survival models, in which the event of interest is resuming of work. As for the construction of the laws of entry, the data is affected by the intervention of censures on the right and truncations on the left. The censures come from individuals whose benefits would be stopped before returning to work (transition to invalidity, transition to retirement, end of the period of study etc.). Truncations are introduced by deductibles. The choice of method naturally falls on the Kaplan-Meier estimator, already presented. Survival functions (i.e., duration) are generally calculated by age, which results in a two-dimensional table: age / seniority.

The crude incapacity exit rates resulting from this calculation can be smoothed using the Whittaker-Henderson method. It is then adapted to the two-dimensional case to simultaneously achieve vertical and horizontal smoothing. Smoothing can be optimized by cross-validation. In practice, this methodology requires many observations to produce reliable results, particularly for extreme ages. In the case of this study, a method of adjustment with an external reference table was therefore preferred. In the case of incapacity duration laws, the choice naturally falls on the BCAC tables, and more specifically on the 2013 version.

The semi-parametric Brass model with external reference makes it possible to link the law of experience and the reference table via a linear regression:

$$\text{logit}(q_{x,t}^{adjust}) = \theta_1 + \theta_2 * \text{logit}(q_{x,t}^{BCAC}) + \varepsilon_{x,t},$$

$$\text{And so : } q_{x,t}^{adjust} = \frac{\exp(\theta_1 + \theta_2 * \text{logit}(q_{x,t}^{BCAC}) + \varepsilon_{x,t})}{1 + \exp(\theta_1 + \theta_2 * \text{logit}(q_{x,t}^{BCAC}) + \varepsilon_{x,t})}.$$

The exit rates are linked with their logistic transformation (*logit*) in order to limit the adjusted rates to the interval [0,1] on which the probabilities are defined.

It is customary to validate the results graphically by comparing the number of observed exits with the number of modeled exits. Firstly, by seniority and secondly by age group. Two-sided 95% confidence intervals are also calculated by normal approximation. The comparison by age of observed exits and modeled exits is displayed in **Chart V** opposite. The adjustment seems to correctly capture all the seniorities. Excessive exit rates appear in the first month, reflecting a compromise found by the regression between the exit rates for the first four months. On higher seniorities, the modeled exits and the observed exits are globally superimposed.

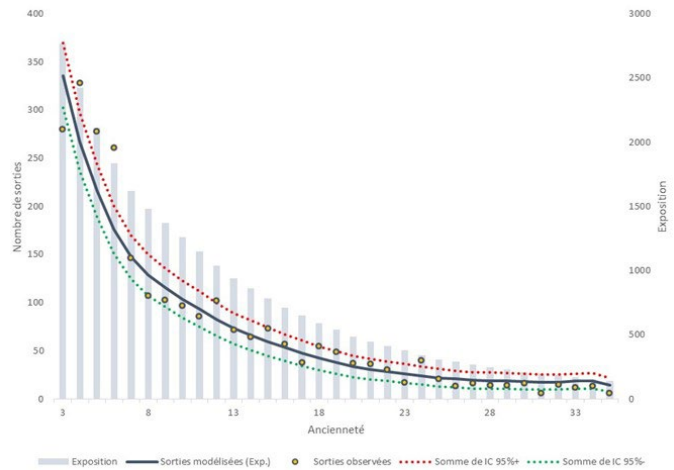


Chart V - Comparison between observed exits and modeled exits, by seniority

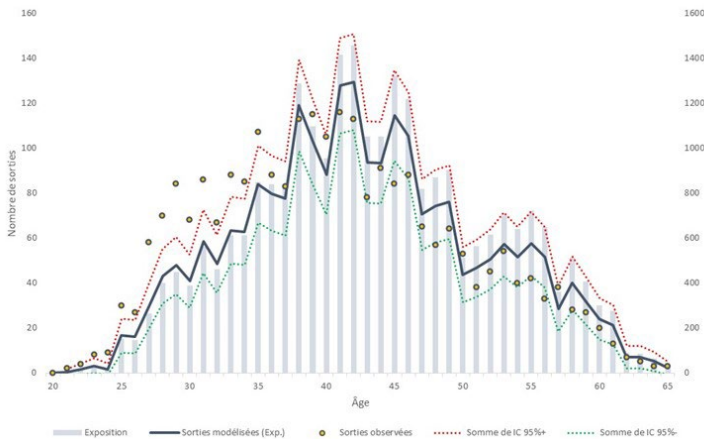


Chart VI - Comparison between observed exits and modeled exits, by age

The comparison by age, as presented in **Chart VI** between modeled exits and observed exits, shows an underestimation up to age 35, then an overestimation after age 40. On this second range, the observed exits are nevertheless framed by the confidence intervals of the modeled exits. By imposing a unique set of regression coefficients, the Brass model is not able to satisfy all the points precisely. The weightings nevertheless allows a good compromise on the intersections presenting the strongest exposures. However, it should be noted that the use of adjusted rates here makes it possible to achieve an overall SMR ratio of 96% compared to 132% using the BCAC

(2010) regulatory table. Under these conditions, the adjusted table seems more appropriate than the regulation table. To compare two tables, it is customary to calculate the residual duration expectations:

$$e_{x,t} = \frac{1}{L_{x,t}} * \sum_{k=0}^{max-t} L_{x,k+t},$$

with x the age of the insured at the occurrence of the claim, t the seniority. The residual duration expectations are plotted in **chart VII**, opposite, for the experience table. The form is characteristic: The fall in exit rates is initially reflected in a lengthening of residual duration expectations. Then the maximum time remaining until the transition to long-term disability being increasingly shorter, the residual durations decrease again. This measure can be used to compare the effects of provisioning between two incapacity duration tables. This approach can be carried out in relation to the regulatory table of the BCAC (2010).

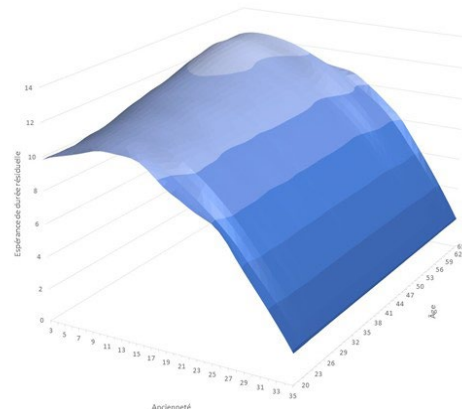


chart VII - Residual duration expectations

The comparison is performed by calculating the relative differences between the two tables:

$$\frac{e_{x,t}^{BCAC} - e_{x,t}^{XP}}{e_{x,t}^{XP}}.$$

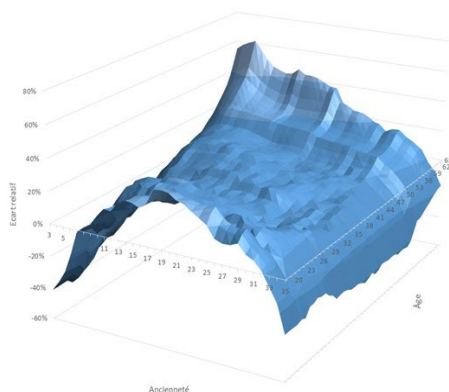


Chart VIII - Relative deviations

Chart VII durations are used to calculate deviations presented in **Chart VIII**. These indicate that the use of the BCAC regulatory table leads, in this example, to an excess of provisioning of the order of 10% at 15 % on seniority and central ages. The coefficients of the adjusted table are however degraded the first three months for the under 35s, then the last three months for all ages. This deterioration over long seniority is explained by the differences between the regulatory table of the BCAC (2010) against which the comparison is made, and the table of the BCAC 2013 which served as the external reference for the model. On the latter, the exit conditional rates on long seniority are much lower. Overall, the weighted average relative deviation between the residual durations of the

experience table and the BCAC table (2010) is approximately 6% for this portfolio.

In the presence of small sample sizes, smoothing crude rates using the Whittaker-Henderson method was not effective. **Under these conditions, the Brass model made it possible to calibrate exit rates using the BCAC 2013 table as an external reference. This method made it possible to bring better stability to the model, at the cost of an imperfect adjustment to the original data: using a single set of regression coefficients led to a single adjustment for all ages. These adjusted rates will nevertheless be more fitted to the loss experience of the portfolios than those taken from the BCAC 2010 regulatory table.**

REMERCIEMENTS

A *Françoise* de m'avoir transmis sa curiosité et son goût de l'effort, et à *André* pour son soutien et ses relectures.

A *Henri-Pierre RODRIGUES* d'avoir accepté de diriger ce mémoire et plus largement de m'avoir fait découvrir l'actuariat. Ses encouragements, ses conseils éclairés et ses relectures ont été d'une grande aide.

A *Frédéric PLANCHET* d'avoir accepté de partager son savoir au cours de multiples entretiens, et dont l'aide a été très précieuse.

A *Laurent DEGIOANNI* et *Franck SETTTON* pour leur confiance. Et plus généralement au Groupe Prévoir de m'avoir permis d'entreprendre cette aventure.

A l'équipe du CEA, enseignants et administratifs, pour leur temps et leur patience.

A tout ceux qui, de près ou de loin, ont contribué au bon déroulé de la formation puis à l'élaboration de ce mémoire.

TABLE DES MATIÈRES

1	Introduction	25
2	Contexte de l'étude et statistiques descriptives du portefeuille.....	27
2.1	Contexte de l'étude.....	27
2.2	Présentation et retraitement des données.....	29
2.2.1	Les contrats intégrés à l'étude	29
2.2.2	Consolidation, mise en forme et contrôle des données.....	32
2.3	Statistiques descriptives des données extraites.....	35
2.3.1	Les données de prévoyance classiques à destination des actifs.....	35
2.3.2	Les données emprunteur.....	42
3	Construction des tables d'entrée en incapacité.....	47
3.1	Estimation des taux bruts.....	47
3.1.1	Estimateur des moments de Hoem	48
3.1.2	Estimateur de Kaplan-Meier.....	50
3.1.3	Présentation des taux bruts (illustrée avec le portefeuille emprunteur).....	52
3.2	Lissage des taux bruts.....	53
3.2.1	Mesure de fidélité F.....	53
3.2.2	Mesure de régularité S.....	53
3.2.3	Forme matricielle et minimisation de M	54
3.2.4	Calibrage et résultats :	55
3.3	Régressions de Cox et effet de certaines caractéristiques sur le risque	58
3.3.1	Principe de la régression de Cox à hasard proportionnel.....	58
3.3.2	Mise en œuvre de régression de Cox sur un portefeuille emprunteur	59
3.3.3	Portefeuilles de contrats de prévoyance classiques à destination des actifs	69
3.3.4	Conclusion sur les lois d'entrée en incapacité.....	78
4	Etude du maintien.....	79
4.1	Définitions et réglementation	79
4.2	Elaboration des taux de maintien en incapacité pour le portefeuille de contrats de prévoyance classiques.....	80
4.2.1	Estimateur de Kaplan-Meier.....	80
4.2.2	Lissage de Whittaker-Henderson.....	83
4.2.3	Lissage par un ajustement à référence externe de Brass et prolongement de la loi.....	88
4.3	Elaboration des taux de maintien en incapacité pour le portefeuille emprunteur	93
5	Conclusion.....	97
6	Bibliographie	98
7	Annexes	99

1 INTRODUCTION

Des travaux de l'Institut de Recherche et de Documentation en Economie de la Santé (IRDES) soulignent qu'un arrêt de travail entraîne une perte de revenu dans de nombreuses situations. Une Interruption Temporaire de Travail (ITT) de plusieurs semaines est donc susceptible d'entraîner des conséquences plus ou moins lourdes en fonction du niveau de revenu des travailleurs. Pour cette raison, des acteurs privés de la protection sociale proposent des garanties complémentaires. Ces garanties prévoient une indemnisation pour compenser, au moins partiellement, la perte de revenus liée à l'arrêt de travail.

Le Groupe Prévoir est un assureur de personnes fondé en 1910 à la veille de la première guerre mondiale, la même année que l'adoption de la loi sur les retraites ouvrières et paysannes. Sous l'impulsion de son fondateur André Angot, la Compagnie alors dénommée Compagnie Française d'Assurances Populaires, s'attache à intéresser les populations les plus modestes. Les premiers contrats émis garantissent le versement d'un capital en cas de décès. Naturellement, l'entreprise se développe dans les bassins industriels et miniers, c'est-à-dire principalement dans le nord et l'est de la France. Grâce à la volonté des descendants des familles fondatrices, le Groupe Prévoir, toujours détenu en grande partie par ces dernières, poursuit le même engagement. Par l'intermédiaire de sa filiale d'assurance, Prévoir Vie, la Compagnie intervient plus particulièrement auprès des foyers populaires (particuliers ou professionnels), par l'intermédiaire d'un réseau debout. Cette population étant particulièrement sensible à la baisse de revenus, la plupart des contrats de prévoyance de Prévoir (commercialisés ou en run-off) intègrent des garanties Incapacité Temporaire de Travail. Plus récemment, le Groupe a entamé un virage stratégique vers l'assurance de prêts, qui s'est notamment concrétisé par des acquisitions importantes dans l'univers du courtage. Cette trajectoire de développement renforce donc davantage le positionnement de Prévoir sur les garanties liées à l'arrêt de travail.

La couverture de ces populations s'inscrit dans un environnement de plus en plus concurrentiel qui incite les assureurs à diminuer leurs tarifs, notamment pour les contrats destinés aux Travailleurs Non-Salariés (TNS) ou pour les contrats d'assurance de prêts. Dans le cas de l'assurance de prêts, cette pression tarifaire conduit parfois à utiliser des matrices de déformation qui éloignent, à certains instants, le prix payé par l'assuré de la réalité de son risque. Dans ces conditions, il est primordial pour Prévoir Vie de connaître encore plus en profondeur le risque porté par les garanties incapacité, tant sur ses portefeuilles de prévoyance « classiques » que sur ceux d'assurance emprunteur. Cette nécessité s'applique aussi bien aux travaux de tarification qu'aux problématiques de provisionnement, particulièrement dans l'approche de Best Estimate. L'objectif principal de cette étude est donc de construire des lois d'expérience d'entrée et de maintien en incapacité.

Or, l'étude de ces garanties nécessite généralement des quantités importantes de données, dont ne dispose pas toujours un assureur comme Prévoir Vie. Ces travaux s'attachent donc à examiner le risque « incapacité » sous le prisme de la rareté des données sur 3 portefeuilles distincts : un portefeuille de prévoyance « classique » à destination des salariés, un portefeuille de prévoyance « classique » à destination des travailleurs non-salariés et un portefeuille d'assurance de prêts. Le choix des méthodes statistiques est contraint par la quantité limitée de données disponibles, quantité d'autant plus contrainte que l'analyse nécessite une segmentation plus fine. Pour cette raison également, les périodes d'observation sont particulièrement étendues. Des précautions sont donc prises afin de ne pas introduire de biais dans l'analyse.

Dans une première partie, quelques statistiques descriptives des données seront présentées. Le travail de retraitement des données sera également détaillé.

Dans une seconde partie, une estimation des taux bruts d'entrée en incapacité sera proposée. Ces taux bruts seront lissés afin de disposer de lois d'expérience. Puis une analyse de l'hétérogénéité du portefeuille sera effectuée. Pour cela, l'impact d'un certain nombre de variables sera évalué à l'aide du modèle de Cox à risques proportionnels.

Enfin, dans une dernière partie, des lois de maintien en incapacité seront calculées grâce au modèle à référence externe de Brass, puis comparées à la table réglementaire du BCAC.

2 CONTEXTE DE L'ETUDE ET STATISTIQUES DESCRIPTIVES DU PORTEFEUILLE

2.1 Contexte de l'étude

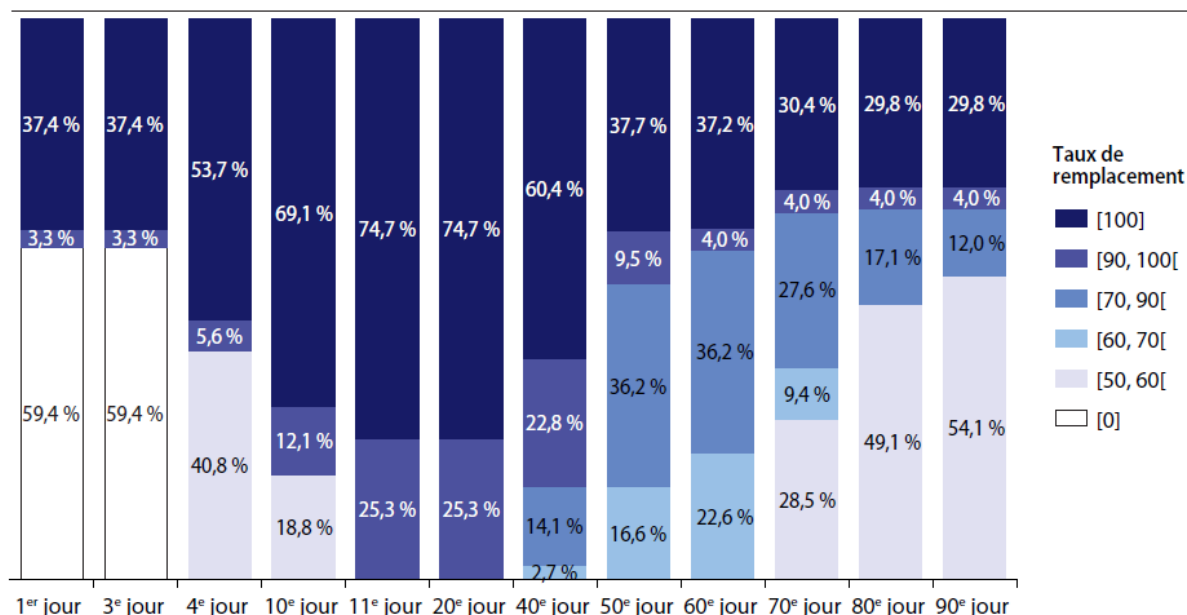
L'Incapacité Temporaire Totale (ITT) traduit l'état passager d'un travailleur, salarié ou non, dans l'incapacité d'exercer son activité professionnelle en raison d'une maladie ou d'un accident, dont l'origine peut être professionnelle ou non-professionnelle.

Le code de la sécurité sociale dispose que « **l'assurance maladie assure le versement d'indemnités journalières (IJ) à l'assuré qui se trouve dans l'incapacité physique constatée par le médecin traitant [...] de continuer ou de reprendre le travail** » (Article L321-1).

Le régime obligatoire de la sécurité sociale prévoit une franchise de 3 jours. La franchise est une période qui s'applique à partir de la survenance du sinistre et pendant laquelle l'assuré n'est pas indemnisé. Passée la franchise, l'indemnisation de l'assuré social diffère en fonction des régimes.

Les salariés se voient indemnisés à hauteur de 50% de leurs revenus de référence dans la limite de 1,8 fois le SMIC. Dans le cas des salariés disposant au minimum d'une année d'ancienneté dans l'entreprise, l'employeur est tenu de compléter les IJ du régime obligatoire de la sécurité sociale après l'application d'une franchise de 7 jours. Ce complément réhausse le taux de remplacement des revenus à 90% puis 66,6% pendant une période limitée qui dépend de l'ancienneté du salarié dans l'entreprise. Par ailleurs, les conventions collectives de branche professionnelle sont susceptibles de prévoir des dispositions plus favorables.

Des travaux menés en 2015 par l'IRDES sur la base Hygie (données 2005-2008) permettent d'apprécier la distribution du taux de remplacement selon le rang du jour d'arrêt et mettent en exergue une forte disparité entre les salariés. Ces taux de remplacement sont présentés dans la **Figure 1**.



Lecture : Les trois premiers jours d'arrêt ne sont pas pris en charge pour 59,4 % des salariés de l'échantillon. Le taux de remplacement au 20^e jour d'arrêt varie entre 90 et 99 % pour 25 % des salariés et est de 100 % pour les 75 % restants.

Source : Base Hygie 2005-2008 (Irdes), base des conventions collectives (Insee, Irdes).

Champ : Salariés relevant des conventions analysées, hors salariés non couverts par une convention collective.

Figure 1 : Taux de remplacement du revenu selon le rang du jour d'arrêt pour les salariés couverts par une convention collective (source : Document de travail de l'IRDES)

Les artisans-commerçants sont indemnisés à hauteur de 50% du revenu de référence dans la limite de 1 fois le PASS (Plafond Annuel de Sécurité Sociale : 41 136€ en 2021).

Enfin, les professionnels libéraux se voient indemnisés à hauteur de 50% de leur revenu de référence plafonné à 3 fois le PASS jusqu'au 90^{ème} jour. Passé ce délai, l'indemnisation est inégale selon les caisses professionnelles. Environ la moitié de caisses ne prévoit aucune indemnité en cas d'incapacité au-delà du 90^{ème} jour (notaires, pharmaciens etc.). L'autre moitié fait apparaître des taux de remplacement très variables et relativement faibles au regard des rémunérations généralement perçues par ces professions.

Les règles susmentionnées sont synthétisées dans la **Figure 2** ci-après.



Figure 2 : Synthèses des différents régimes d'indemnisation

Au regard de ces éléments, il apparaît qu'un arrêt prolongé de plusieurs semaines entraîne de fortes réductions de revenus pour la majorité des salariés. Cette perte de revenus, **particulièrement marquée pour les TNS**, est donc susceptible de bousculer l'équilibre économique d'un ménage. La tentation de la reprise précoce (avant consolidation de l'état de santé) de l'activité professionnelle peut alors engager des conséquences sanitaires néfastes.

Dans ce contexte, une couverture assurantielle privée peut être apportée par la souscription d'une garantie « incapacité » auprès d'une société d'assurance. Dans le cas des salariés, cette couverture vise à compléter les deux premiers niveaux du système d'indemnisation que sont le régime obligatoire de la Sécurité Sociale et la part employeur. Cette garantie peut être souscrite individuellement, comme dans le cas des contrats Prévoir, ou collectivement via l'employeur. Dans le cas des TNS, cette couverture est le seul complément au régime obligatoire.

Cette garantie prévoit une compensation de nature indemnitaire ou forfaitaire pour compenser au moins partiellement la perte de revenus. La prestation indemnitaire est calculée en fonction du niveau de revenus au moment de la survenance du sinistre, tandis que le niveau de la prestation forfaitaire est établi lors de la souscription de la garantie. Dans tous les cas, le principe de non-enrichissement de l'assuré doit s'appliquer et l'assuré transmet à l'assureur une preuve de revenus à la souscription et/ou au moment du sinistre.

Ces garanties peuvent être souscrites en tant que garanties obligatoires d'un contrat, ou comme des garanties optionnelles, et sont assorties d'une franchise. La franchise est habituellement choisie par l'assuré lors de la souscription et son niveau varie en fonction des populations cibles. Les durées de franchise choisies par les salariés sont généralement plus longues que celles choisies par les TNS qui ne bénéficient pas de couverture employeur. Des mécanismes de rechute et de prolongation de l'arrêt de travail sont néanmoins intégrés afin que la franchise ne soit pas appliquée plusieurs fois pour un même fait générateur. Il est d'usage de proposer une modulation de la franchise en cas d'accident, d'hospitalisation et pour certains types de maladies (Maladies Non Objectivables par exemple). Les notations sont alors les suivantes : franchise maladie / franchise hospitalisation / franchise accident.

Par ailleurs, les assureurs assortissent le plus souvent les garanties incapacité d'une période de carence (ou d'attente). Pendant cette période, généralement portée à 90 jours à compter de la souscription, les sinistres ne sont pas couverts sauf disposition contraire (par exemple en cas d'accident). Ce mécanisme a pour but de lutter contre l'antisélection.

Enfin, une durée maximale de couverture est prévue aux termes des contrats. Passée cette durée, l'assuré est considéré comme invalide et une garantie spécifique, fonction du taux d'invalidité, peut prendre le relais.

2.2 Présentation et retraitement des données

2.2.1 Les contrats intégrés à l'étude

Le Groupe Prévoir, assureur de personnes depuis 1910, intervient plus particulièrement auprès des foyers populaires (particuliers ou professionnels), par l'intermédiaire d'un réseau debout. La plupart des contrats de prévoyance commercialisés ou en run-off proposent des garanties d'Incapacité Temporaire de Travail. Ces garanties se déclinent sous forme d'indemnités journalières et d'exonération des cotisations. Trois types d'offres ont été retenues dans cette étude : les offres de prévoyance classiques à destination des salariés, les offres de prévoyance classiques à destination des TNS et une offre emprunteur distribuée par un courtier grossiste. Le portefeuille étudié est réparti de la façon suivante :

	Nombre d'assurés	Part du portefeuille
Salariés	85 909	34%
PREVOIRINCAP	10 201	4%
PREVOIR MULTISOLUTIONS	59 118	24%
SOLUTION PROTECTION ACTIF	16 590	7%
TNS	25 427	10%
PREVOIRPRO SOLUTION	1 750	1%
PREVOIRPRO ACTIF PREVOYANCE	3 754	1%
PREVOIRPRO ACTIF PREVOYANCE MADELIN	8 174	3%
SOLUTION PRO PREVOYANCE MADELIN	4 820	2%
SOLUTION PRO PREVOYANCE NON MADELIN	6 929	3%
Emprunteur	139 267	56%
Total	250 603	100%

Tableau 1 : Répartition du portefeuille de contrats étudié

2.2.1.1 Les offres de prévoyance classiques à destination des travailleurs salariés

- **PrévoirIncap : la garantie incapacité est la garantie principale du contrat.** Elle est donc obligatoire. Elle se décline sous forme d'indemnités journalières et d'une exonération du paiement des cotisations en période de versement des IJ.
 - Période de commercialisation : 02/2000 à 03/2007.
 - Age à la souscription : assuré contractant de 18 à 50 ans.
 - Option pour la franchise : 15/15/15, 40/40/40, 90/90/90.
 - Période de carence : 0 pour l'accident, 3 mois pour la maladie et 12 mois pour les maladies du dos et neuropsychologiques.
 - Autres garanties : un capital décès obligatoire de 15 000€.

- Prévoir MultiSolutions : il s'agit d'un produit multi-garanties de protection des actifs. La garantie principale couvre le décès de l'assuré. **La garantie indemnités journalières en cas d'ITT est optionnelle. Une exonération des cotisations en cas d'ITT est également proposée en option.**
 - Période de commercialisation : 04/2007 à 04/2016.
 - Age à la souscription : assuré contractant de 18 à 55 ans.
 - Option pour la franchise :
 - Indemnités journalières : 15/15/15, 40/40/40, 90/90/90.
 - Exonération des cotisations : 90/90/90.
 - Période de carence : 90 jours.
 - Autres garanties :
 - Obligatoires : capital décès, assistance.
 - Optionnelles : capital ITD, rente invalidité.

- Solution Protection actif : il s'agit d'une évolution du Prévoir MultiSolutions dont **la garantie incapacité est toujours optionnelle. L'exonération des cotisations est rendue obligatoire.**
 - Période de commercialisation : 05/2016 à 10/2021.
 - Age à la souscription : assuré contractant de 30 à 60 ans.
 - Option pour la franchise :
 - Indemnités journalières : 15/15/15, 40/40/40, 90/90/90, 180/180/180.
 - Exonération des cotisations : 90/90.
 - Période de carence : 90 jours.
 - Autres garanties :
 - Obligatoires : capital décès, assistance
 - Optionnelles : rente viagère en cas de décès, rente éducation, rente invalidité, capital invalidité, capital maladies redoutées, capital blessure.

Les taux de souscription aux garanties entrant dans le champ de l'étude sont présentés dans le **Tableau 2**. Ces éléments révèlent que **les indemnités journalières ne sont pas la principale motivation d'achat de ce type de produits pour les assurés** Prévoir Vie. Ainsi, lorsque que cette garantie est optionnelle, seul 1 assuré sur 5 y souscrit. Le taux de souscription de l'option exonération des cotisations est plus élevé (72% sur le Prévoir MultiSolutions) que celui à l'option IJ, ce qui s'explique vraisemblablement par le faible coût pour l'assuré d'une telle garantie.

	Taux de souscription	
	Exonération des primes en cas d'ITT	Indemnités journalières en cas d'ITT
Salariés	79%	29%
PREVOIRINCAP	100%	100%
PREVOIR MULTISOLUTIONS	72%	21%
SOLUTION PROTECTION ACTIF	100%	21%

Tableau 2 : Taux de souscription aux garanties incapacité pour les produits à destination des salariés

2.2.1.2 Les offres de prévoyance classiques à destination des TNS

- **PrévoirPro Solution** : la garantie principale couvre le décès. **La garantie incapacité est optionnelle** et peut être complétée d'une incapacité spéciale professions de santé. **Une exonération des cotisations est également proposée en option.**
 - Période de commercialisation : 06/2006 à 05/2010.
 - Age à la souscription : assuré contractant de 20 à 50 ans.
 - Option pour la franchise :
 - Indemnités journalières : 15/15/15, 30/30/30, 60/60/60.
 - Exonération des cotisations : 90/90/90.
 - Période de carence : 90 jours.
 - Autres garanties :
 - Obligatoires : rente viagère en cas de décès.
 - Optionnelles : rente éducation, indemnités journalières en cas d'invalidité.

- **PrévoirPro Actif Prévoyance Madelin** : évolution du Prévoir Pro Solution, proposant entre autres la **modulation de l'IJ optionnelle sur 3 périodes distinctes afin de s'adapter aux différents systèmes d'indemnisations**. La prestation peut être, au choix de l'assuré à la souscription, de nature indemnitaire ou forfaitaire. **L'exonération des primes en cas d'ITT est obligatoire** :
 - Période de commercialisation : 04/2014 à 05/2019.
 - Age à la souscription : 18 à 60 ans.
 - Option pour la franchise :
 - IJ forfaitaire : 15/15/15 ou, 30/30/30, 60/60/60, 90/90/90, 365/365/365.
 - IJ indemnitaire : 15/15/15, 30/30/30, 60/60/60, 90/90/90.
 - Exonération des cotisations : 90/90/90.
 - La franchise peut être réduite à 3 jours en cas d'accident ou d'hospitalisation, si souscription à l'option.
 - Période de carence : 90 jours.
 - Autres garanties :
 - Obligatoires : rente viagère
 - Optionnelles : rente éducation, rente invalidité.

- **PrévoirPro Actif Prévoyance** : complément non-Madelin du Prévoir Actif Prévoyance Madelin. Ce produit dispose notamment des options complémentaires suivantes :
 - ITT accident à la condition de ne pas souscrire à l'ITT toute cause de la version Madelin.
 - ITT spéciale Conjoint du collaborateur (IJ).
 - Frais professionnels (IJ).

- **Solution Pro Prévoyance** : évolution du PrévoirPro Actif Prévoyance. **L'exonération des primes en cas d'ITT est également obligatoire**. Le montant des IJ peut être modulé sur 3 périodes distinctes. Seules les IJ forfaitaires sont proposées. Les IJ Frais professionnels sont également proposés sur la version non-Madelin.
 - Période de commercialisation : 05/2019 à 12/2022.
 - Age à la souscription : 18 à 65 ans.
 - Option pour la franchise :
 - IJ Forfaitaires : 15/3/3, 30/3/3, 60/3/3, 90/3/3, 365/3/3.
 - IJ Frais professionnels : 15/15/15, 30/30/30, 60/60/60, 90/90/90.
 - Période de carence : 90 jours.
 - Autres garanties :
 - Obligatoires : rente viagère (version Madelin), Capital décès (version non-Madelin).

- Optionnelles : triplement accident (version non-Madelin), Invalidité, Invalidité Professionnelle (professions médicales), rente éducation.

Les taux de souscriptions aux garanties TNS entrant dans le champ de l'étude sont présentés dans le **Tableau 3**. Il apparaît que **les indemnités journalières, lorsqu'elles sont optionnelles, sont beaucoup plus fréquemment souscrites que dans le cas des salariés. Elles semblent donc être l'une des motivations principales d'achat de ce type de produits.**

	Taux de souscription					
	Exonération des primes	IJ toute cause	IJ Cause Accidentelle	IJ Frais Pro.	IJ Conjoint collaborateur	IJ Spéciales Professions de santé
TNS	64%	48%	4%	14%	0%	0%
PREVOIRPRO SOLUTION	93%	85%	0%	0%	0%	3%
PREVOIRPRO ACTIF PREVOYANCE MADELIN	100%	89%	0%	0%	0%	0%
PREVOIRPRO ACTIF PREVOYANCE	0%	0%	11%	29%	1%	0%
SOLUTION PRO PREVOYANCE MADELIN	97%	92%	0%	0%	0%	0%
SOLUTION PRO PREVOYANCE NON MADELIN	100%	34%	0%	23%	0%	0%

Tableau 3 : Taux de souscription des garanties incapacité pour les produits à destination des TNS

L'expression d'une telle différence de motivation entre les salariés et les TNS dans la souscription des garanties incapacité doit conduire à étudier distinctement les fréquences d'entrée en incapacité de ces deux populations, au moins dans un premier temps. Il s'agira alors de déterminer, le moment venu, si ce critère est discriminant dans les taux d'incidence.

2.2.1.3 L'offre emprunteur

L'offre emprunteur intégrée à l'étude est distribuée par un courtier grossiste et assurée par Prévoir Vie. Il s'agit donc d'un produit en marque blanche commercialisé entre 2011 et 2017.

Concernant les franchises, les banques disposent généralement d'un niveau standard. Ce dernier s'établit à 90 jours quelle que soit la cause de l'arrêt (90/90/90). Il est donc logique de constater une quasi-exclusivité de cette franchise dans le portefeuille (98%).

2.2.2 Consolidation, mise en forme et contrôle des données

Le retraitement des données permet de répondre à deux enjeux. D'une part, s'assurer de la cohérence de l'ensemble des informations, d'autre part, disposer des données sous une forme exploitable en vue de réaliser des calculs.

Les données sont issues de deux questionnaires différents. D'une part les contrats prévoyance classiques à destination des salariés et des TNS, commercialisés et gérés par Prévoir. D'autre part les contrats d'assurance emprunteur, commercialisés et gérés par le courtier grossiste. Afin de respecter les caractéristiques des données propres à ces deux environnements, deux travaux de retraitement distincts sont réalisés.

Les contrats de prévoyance classiques à destination des salariés et des TNS ne peuvent porter que sur une seule tête. Ils peuvent contenir plusieurs garanties incapacité avec des niveaux de franchise différents : des indemnités journalières et des exonérations de cotisations. Les travaux de préparation des données doivent ainsi aboutir à la création d'une table « contrats », disposant d'une ligne par couple « contrat / garantie » et d'une table « sinistres » disposant d'une ligne par couple « sinistre / garantie ».

Les contrats d'assurance emprunteur peuvent porter sur plusieurs têtes. Le risque étant calculé de façon individualisée, la table « contrats » comporte donc une ligne par couple « contrat / assuré ». En revanche une seule garantie incapacité est possible : l'exonération des mensualités du prêt.

2.2.2.1 Les données relatives aux contrats de prévoyance classiques à destination des salariés et des TNS

Chaque couple « contrat / garantie », doit faire l'objet d'une seule ligne dans la table éponyme. Les données contrats Prévoir Vie sont issues du progiciel GraphTalk AIA et sont extraites par l'équipe dédiée. La table fournie est composée d'une ligne par couple « contrat / garantie » pour chaque période de couverture.

Le comptage du nombre de contrats, toutes situations confondues, constitue le premier contrôle élémentaire. Il doit être cohérent avec les rapports usuellement communiqués dans l'entreprise : tableaux de bord issus du contrôle de gestion ou du marketing, lois de comportement construites par la Direction Technique etc.

Après avoir décodé le libellé des garanties, un second niveau de contrôle consiste à s'assurer que l'ensemble des garanties attendues sur chaque produit apparait strictement dans la table. Puis que les garanties obligatoires sont bien répertoriées pour chacune des polices. Par définition, le contrôle du nombre de garanties optionnelles est moins évident. Néanmoins, les taux de souscription des options indiquées dans le document des lois de comportement permettent de valider les ordres de grandeur attendus.

Enfin, un troisième niveau de contrôle, aléatoire, consiste à valider sur une dizaine de contrats par produit que les informations extraites sont bien identiques à celles accessibles depuis l'outil de gestion des contrats :

- Souscriptions aux garanties optionnelles.
- Niveau de la garantie choisie (franchise, IJ etc.).
- Période de couverture de la garantie.

A l'issue de ces trois niveaux de contrôles, synthétisés dans la **Figure 3**, l'extraction des données brutes, c'est-à-dire la « matière première » de l'étude, peut être considérée comme complète au niveau contrats.

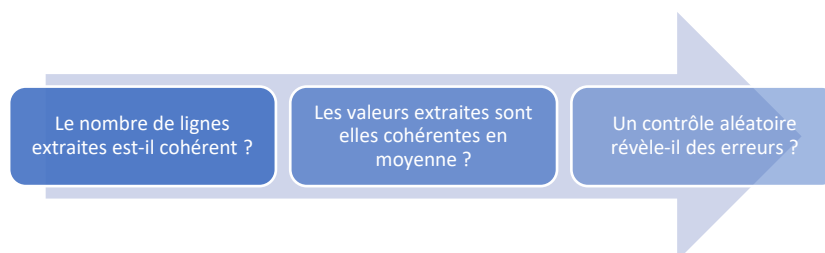


Figure 3 : Synthèse des 3 niveaux de contrôles effectués sur les données contrats

Afin de calculer des taux d'entrée en incapacité par assurés, et non par contrats, il est maintenant nécessaire de regrouper ces informations sur une seule entrée par couple assuré / période de couverture. Il convient de choisir une méthode de regroupement qui génère le moins de perte d'informations possible. Dès lors, deux options ont été envisagées.

La première option a consisté à construire un algorithme pour extraire des bases propres à une certaine segmentation. Par exemple propre aux garanties « Indemnités Journalières » sur les contrats « PrévoirIncap ». Dès lors trois écueils se sont présentés. D'abord, le temps de traitement de cet algorithme s'est révélé beaucoup trop long pour être utilisé à la volée (plusieurs heures nécessaires). Ensuite, l'utilisation de multiples bases de données a rendu les pistes d'audit quasiment impossibles à suivre. Enfin, la perte d'informations s'est révélée importante, réduisant le champ des possibles sur la création de variables de segmentation dans la suite des travaux.

La seconde option, qui a donc été retenue, consiste en la création d'une base de données intégrant des collections d'objets. Plus précisément, à chaque changement dans la couverture de l'assuré, une nouvelle ligne est générée. La piste d'audit est facilitée par le stockage des informations contrats sous forme de listes. Le langage Python permet d'effectuer de multiples opérations sur ces collections d'objets (identifier les doublons etc.). Il a par exemple été possible de stocker les informations suivantes :

- Les dates de naissance : le contrôle du caractère unique des dates de naissance pour chaque assuré pouvant ainsi être opéré facilement.
- Les garanties souscrites : cette liste a permis de calculer le nombre de garanties de chaque type (IJ, Exo).
- La liste des franchises souscrites selon la garantie permettant, par exemple, d'identifier les assurés disposant d'une réduction de franchise en cas d'accident (ce qui a fait l'objet d'une variable).

Le **Tableau 4** ci-après illustre le type de structure de la base de données générée pour un assuré fictif :

Assuré	Date de début de couverture	Date de fin de couverture	Contrats	Nomb re de garan ties IJ	Nombre de garanties Exo	Franchise garanties IJ	Franchise garanties Exo
1	01/01/2020	30/06/2020	[PI_001]	1	1	[30]	[30]
1	01/07/2020	31/12/2021	[PI_001, PMS_001]	1	2	[30]	[30,15]

Tableau 4 : Structure de la base de données

2.2.2.2 Les données relatives aux sinistres

Chaque sinistre doit pouvoir être rapproché d'un couple « contrat / garantie ». En effet, un même fait générateur est susceptible de déclencher plusieurs lignes de sinistres en fonction du nombre de garanties activées. Par exemple, le déclenchement des indemnités journalières dans un premier temps, puis de l'exonération des cotisations dans un second temps.

La table fournie par l'équipe informatique est composée d'une ligne par couple « sinistre / garantie / période de prestation ». Les périodes de prestation sont alignées sur les mois civils. Autrement dit, un sinistre couvert du 15/01/2021 au 15/02/2021 sera indemnisé sur deux périodes de prestation : 15/01/2021-31/01/2021 et 01/02/2021-15/02/2021. Un algorithme simple construit en Python permet de regrouper ces sinistres en une seule ligne combinant les périodes de prestation. Cet algorithme détecte également les phénomènes de rechute et prolongation.

Le contrôle de l'intégrité des données s'articule autour des dates des sinistres. Le premier contrôle consiste à s'assurer que chaque sinistre indemnisé s'étend bien sur une période couverte par une garantie.

Un second contrôle doit permettre de s'assurer que les franchises réellement constatées sont bien celles prévues aux termes du contrat. Ce contrôle est indispensable dans la mesure où le calcul des franchises par l'équipe de gestion des sinistres est manuel en cas d'annexe d'exclusion. Cette analyse permet d'établir qu'environ 2% des sinistres sont en erreur. Les erreurs sont fortement centrées autour de la franchise prévue. Pour les contrats dont le décalage est de maximum 3 jours, la date de survenance du sinistre est modifiée. Dans le cas contraire, l'assuré est retiré de l'analyse.

Un indicateur de durée d'indemnisation du sinistre est calculé afin d'identifier les valeurs aberrantes. En particulier, aucune indemnisation ne doit excéder la période maximale prévue par la garantie.

Le travail de retraitement des données n'est pas présenté pour le portefeuille emprunteur. Il s'est avéré plus simple compte tenu de la présence d'un seul produit et d'une base de données déjà partiellement agrégée.

2.3 Statistiques descriptives des données extraites

Les données contrats et sinistres sont maintenant considérées comme complètes et intègres. Elles sont en outre présentées sous une forme facilitant leur exploitation.

Les garanties portent sur des populations fortement hétérogènes dont certains facteurs sont observables pour l'assureur :

- Âge et sexe de l'individu.
- État de santé général de l'assuré à la souscription : mesurable grâce au questionnaire médical.
- Catégorie Socio-Professionnelle (CSP) : de nombreuses recherches mettent en exergue le lien entre l'absentéisme et les niveaux de rémunération ou les conditions de travail.
- Motivation d'achat du contrat : les conséquences d'une ITT sont différentes selon que l'assuré est salarié ou non. Dans le cas de l'assurance emprunteur, la souscription à la garantie incapacité est généralement une condition d'accès au crédit.
- Niveau de franchise : la littérature a montré qu'il existe un lien entre le niveau de couverture et la fréquence d'entrée en incapacité (aléa moral).
- Localisation géographique : il existe en France une forte inégalité des territoires face au taux d'entrée en incapacité. Dans son rapport de 2006 sur la Sécurité Sociale, la Cour des Comptes soulignait que « les fortes différences territoriales existantes qui varient toujours dans une proportion de 1 à 3 ne peuvent guère être expliquées par la structure socioprofessionnelle de la population active de ces départements » (Cour des comptes, 2006).

La forte hétérogénéité de la population assurée rend donc indispensable l'établissement de lois d'entrée segmentées et de coefficients de majoration / minoration par rapport à une population de référence. Dans ce contexte, il est d'usage d'effectuer une description statistique des données afin d'appréhender au mieux les caractéristiques de la population étudiée.

2.3.1 Les données de prévoyance classiques à destination des actifs

2.3.1.1 Caractéristiques des populations assurées

La première observation porte sur le sexe des assurés. Il existe, selon les produits, des déséquilibres entre les populations masculines et féminines. En particulier, les populations masculines sont plus fortement représentées sur les produits les plus anciens de l'étude : le Prévoir Pro Solution pour les TNS et le PrévoirIncap pour les salariés.

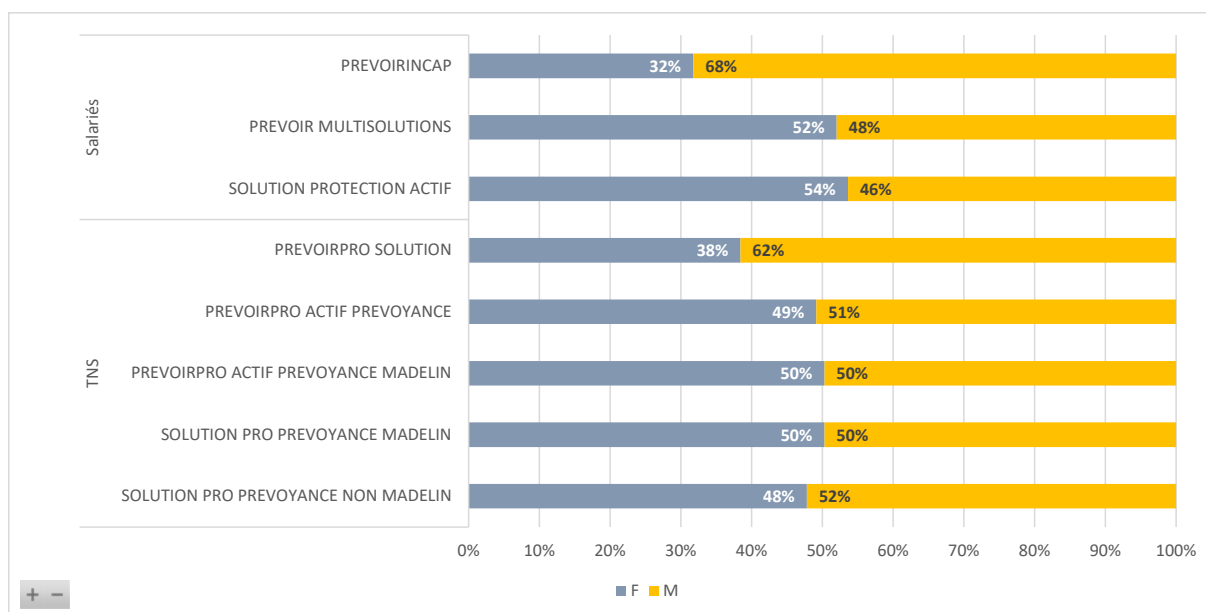


Tableau 5 : Répartition des sexes féminin et masculin dans le portefeuille étudié

La stratégie de développement commercial du Groupe s'appuyait historiquement sur les populations ouvrières et minières, largement masculines. Or les emplois industriels sont en déclin en France, particulièrement depuis les années 2000. **Ce bouleversement dans la structure des emplois a conduit Prévoir Vie à revoir son positionnement.** Ce dernier est dorénavant largement étendu aux populations féminines. Cela se traduit, par exemple, par la prise en charge de l'incapacité liée aux grossesses pathologiques sur les offres TNS les plus récentes. Les plaquettes commerciales des produits en attestent (**Figure 4** ci-après) :



Plaquette commerciale du produit PrévoirIncap (2000) : Un homme rentre du travail et est accueilli par deux enfants.

Plaquette commerciale du Prévoir MultiSolutions (2006) : Une femme est au centre de l'image.

Figure 4 : Plaquettes commerciales des produits PrévoirIncap et Prévoir MultiSolutions

Ces évolutions dans la structure des populations assurées doivent être prises en compte dans les futurs travaux de tarification. Si la réglementation interdit toute adaptation du tarif en fonction du sexe, la part des deux sexes dans la population de référence servant à tarifier devra être calibrée. Et le niveau retenu ne saurait donc être extrapolé des données trop anciennes du portefeuille.

La distribution des âges à la souscription est globalement homogène entre les deux sexes pour les salariés (**Figure 5**).

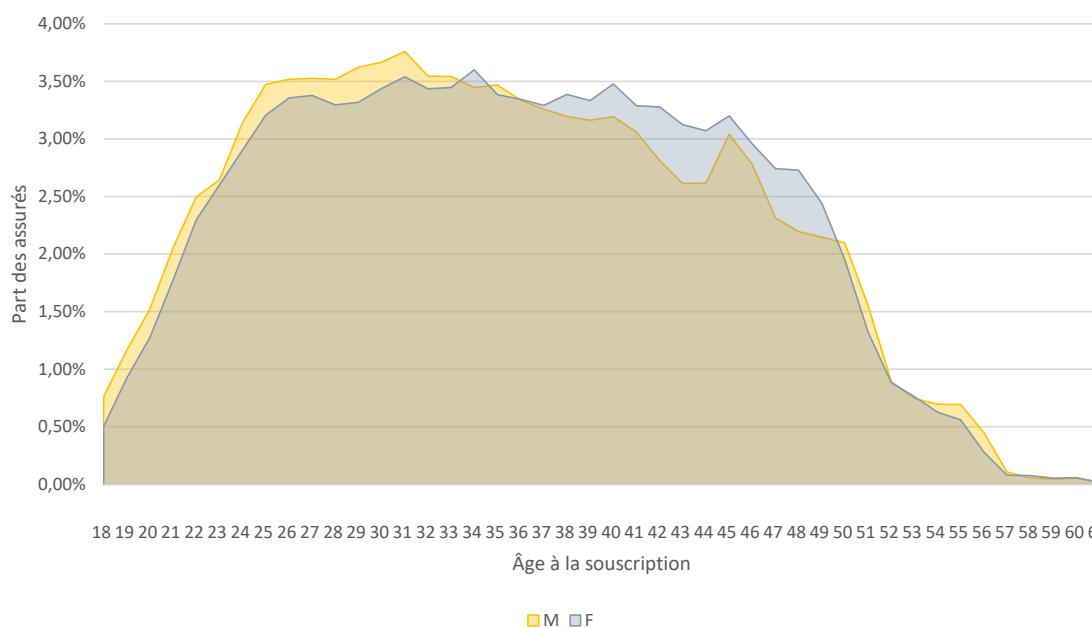


Figure 5 : Distribution des populations salariées féminines et masculines selon l'âge à la souscription

En ne conservant que les assurés ayant souscrit aux indemnités journalières, la répartition est cependant très différente. Si la population féminine souscrit légèrement plus tard en moyenne, sa distribution est nettement plus « aplatie » entre 24 et 32 ans (voir **Figure 6**). Le morcellement des carrières féminines, notamment dans les populations peu qualifiées et plus particulièrement à la suite d'une maternité, peut expliquer cette distribution. Ce phénomène est logiquement beaucoup moins visible sur le Solution Protection Actifs qui n'est ouvert à la souscription qu'à partir de 30 ans. Les indemnités journalières étant tarifées à l'âge atteint, il conviendra donc de tenir compte de la répartition des sexes selon l'âge.

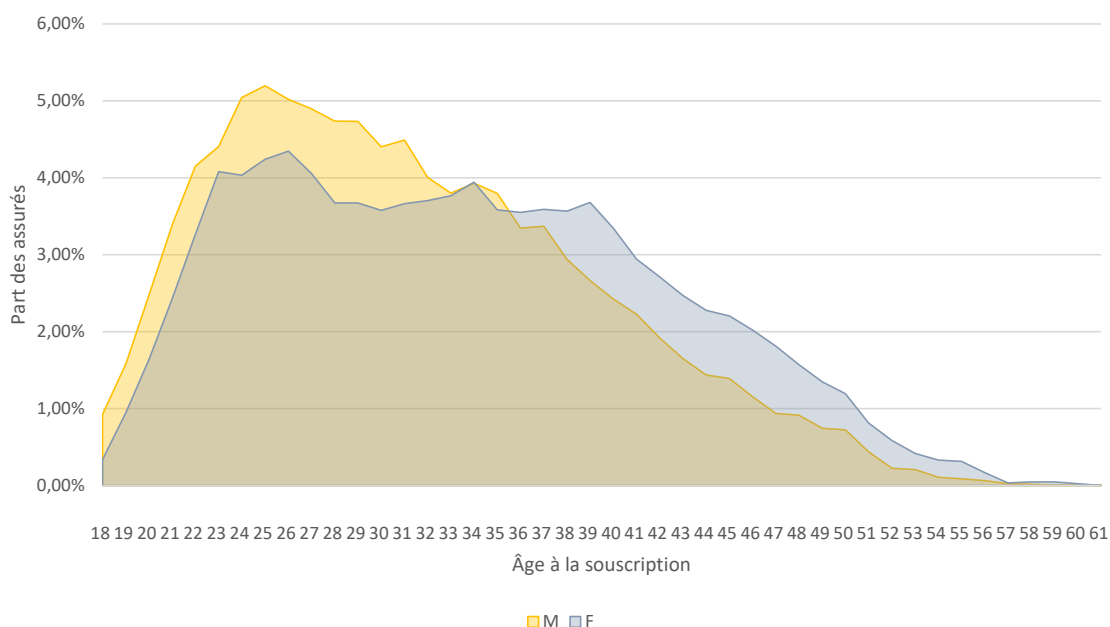


Figure 6 : Distribution des populations salariées féminines et masculines ayant souscrit aux indemnités journalières selon l'âge à la souscription

Ce phénomène ne se retrouve pas sur les TNS. Les distributions sont plus homogènes, bien que celle des femmes soit légèrement décalée à gauche par rapport à celle des hommes (**Figure 7**).

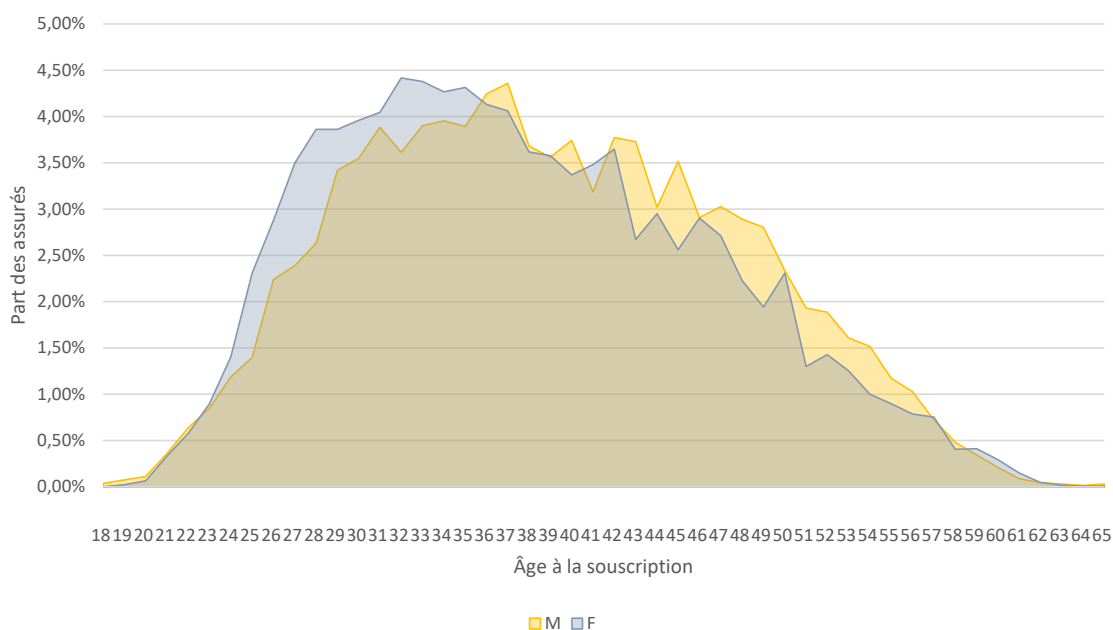


Figure 7 : Distribution des populations TNS féminines et masculines selon l'âge à la souscription

Concernant les écarts salariaux entre les femmes et les hommes, le revenu annuel déclaré à la souscription est plus faible chez les femmes que chez les hommes. Dans les deux cas, le revenu moyen est faible au regard du niveau national qui s'élevait en moyenne à 29 000€ pour un salarié du privé en 2019. Prévoir Vie étant historiquement positionné sur la protection des foyers populaires, il est logique de constater une telle distribution chez les salariés, présentée ci-dessous dans la **Figure 8**. En outre, le revenu moyen est nettement plus élevé sur les produits les plus récents de la gamme, confirmant les évolutions déjà mentionnées dans la stratégie commerciale.

Le revenu moyen est également plus élevé chez les TNS mais ne fait pas apparaître les mêmes écarts entre les femmes et les hommes.

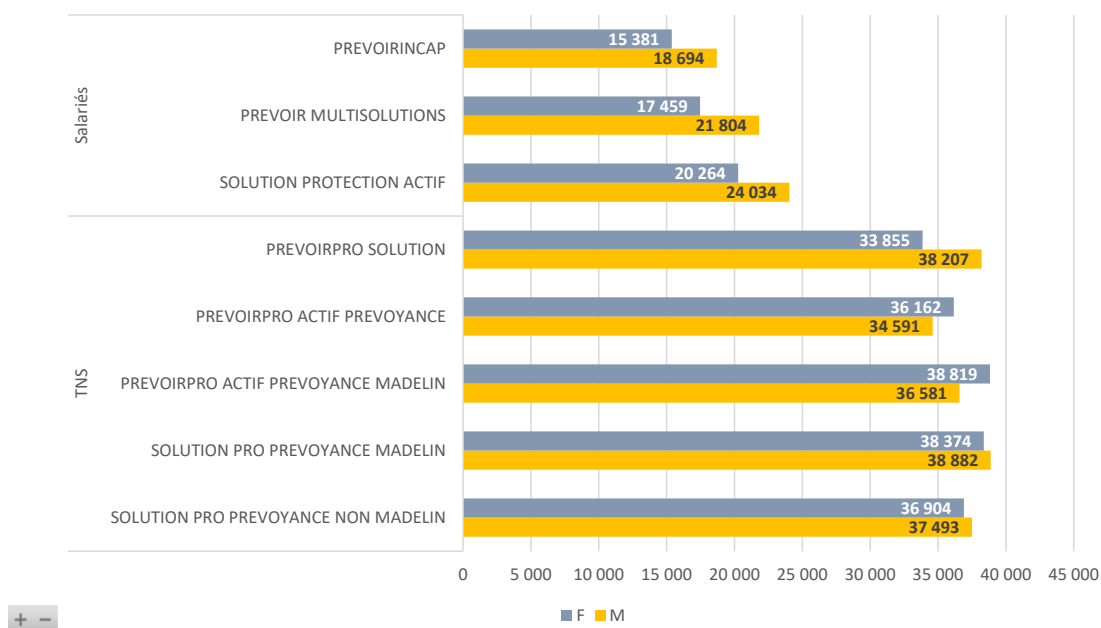


Figure 8 : Distribution du revenu des populations salariées et TNS

Les niveaux de franchise constatés sur les indemnités journalières diffèrent également entre les populations salariées et TNS. Le taux de remplacement des revenus pour les salariés est susceptible de baisser plus fortement à partir du 41^{ème} jour, pour atteindre son niveau le plus bas au 91^{ème} jour (**Figure 1**). Pour autant, les données observées sur le portefeuille Prévoir laissent apparaître **une prédominance de la franchise 15 jours pour cette population**. Ce phénomène peut s'expliquer par le caractère populaire des populations assurées. L'utilité marginale des revenus est traditionnellement présentée comme une fonction décroissante du niveau de vie. Plus précisément, les foyers populaires ne peuvent absorber une perte de revenus, même si cette dernière n'excède pas 10%. Il semble donc logique pour ces derniers de couvrir ce risque le plus tôt possible dans la chaîne d'indemnisation. Enfin, la part de la franchise 15 jours décroît sur les produits plus récents et devient minoritaire sur le Solution Protection Actifs, produit le plus récent de la gamme (**Figure 9**). Ce phénomène est cohérent avec l'augmentation du revenu moyen dans le portefeuille. Un autre élément susceptible d'exacerber cette observation est le caractère obligatoire de la garantie incapacité dans le PrévoirIncap, produit le plus ancien de l'étude. Il en résulte une motivation d'achat différente, et des assurés possiblement plus averses à ce risque. En outre, la prime est totalement dédiée à ce risque, réduisant potentiellement les arbitrages budgétaires à la souscription, et donc favorisant la franchise la plus onéreuse pour l'assuré.

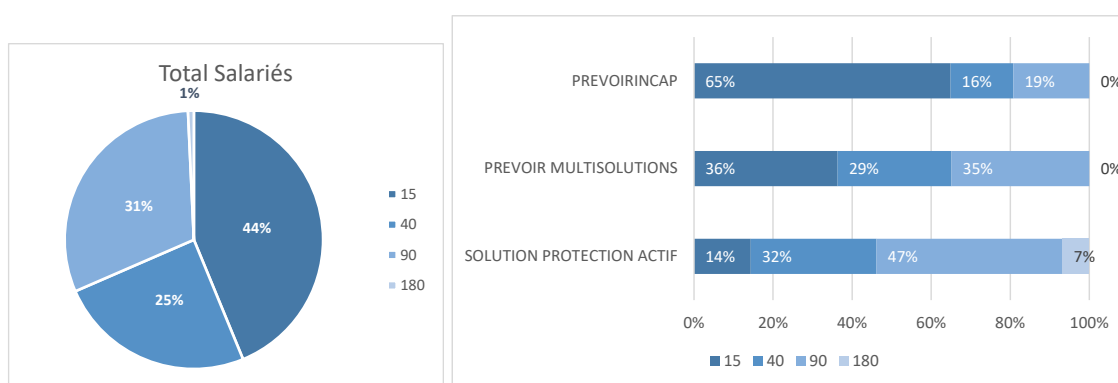


Figure 9 : Répartition des franchises souscrites pour les IJ toute cause

Les TNS souscrivent très majoritairement à la franchise 15 jours, quel que soit le produit (**Tableau 6**). Ces derniers ne bénéficiant pas du niveau employeur dans leur chaîne d'indemnisation, les indemnités journalières constituent une des principales motivations à l'achat pour ce type de produits. Il est donc logique de constater une telle répartition.

Franchise	Part
15	94%
30	2%
60	1%
90	1%
365	3%

Tableau 6 : Répartition des franchises souscrites pour les IJ Toute Cause

Afin d'illustrer et de mieux comprendre cette observation, il est utile d'extraire la répartition des TNS par rapport à leur régime obligatoire, duquel dépend les niveaux d'indemnisation (**Tableau 7**). Le niveau d'indemnisation indiqué est celui en vigueur à la souscription, c'est-à-dire avant la réforme de juillet 2021. Il apparaît que les femmes appartiennent majoritairement à CARPIMKO, la caisse de retraite et prévoyance des professions paramédicales. Cela s'explique par la forte représentation des infirmières libérales dans le portefeuille. Les hommes sont majoritairement affiliés au SSI qui regroupe notamment les artisans – commerçants.

Caisse	F	M	Total	Indemnisation avant réforme de juillet 2021
SSI	42%	83%	63%	50% du revenu de référence après 3 jours (max 1 PASS)
CARPIMKO	48%	8%	27%	≈55€ par jour à partir du 91 ^{ème} jour
CIPAV	4%	4%	4%	Aucune indemnisation
Autres	6%	5%	5%	na

Tableau 7 : Répartition des sections professionnelles parmi les TNS

Il convient de préciser que l'absence de la franchise 3 jours dans le cadre des IJ toute cause a vocation à lutter contre l'antisélection et les coûts de gestion liés à l'indemnisation de micro sinistres. Le prix d'une telle garantie serait en outre particulièrement élevé.

Enfin, l'ensemble des assurés étudiés sur les portefeuilles de prévoyance classique représente environ 500 000 années d'exposition (c'est-à-dire la somme des temps d'observation).

2.3.1.2 Caractéristiques des sinistres

Le nombre de sinistres observés dans cette étude est de 8 826. Sans opérer des calculs de taux d'entrée en incapacité, il est pourtant possible de fournir quelques éléments descriptifs sur les sinistres du portefeuille.

Un premier niveau de lecture consiste à observer la répartition des sinistres sur les garanties IJ toute cause en fonction des niveaux de franchises souscrits (**Figure 10**). Il apparaît ainsi que **les franchises les plus courtes sont les plus observées**, sur les salariés comme sur les TNS. La franchise 15 jours est ainsi systématiquement majoritaire dans les sinistres, y compris sur les produits où elle est la moins souscrite. La même analyse graphique n'a pas d'intérêt sur les TNS puisque la franchise 15 jours est souscrite à 94%.

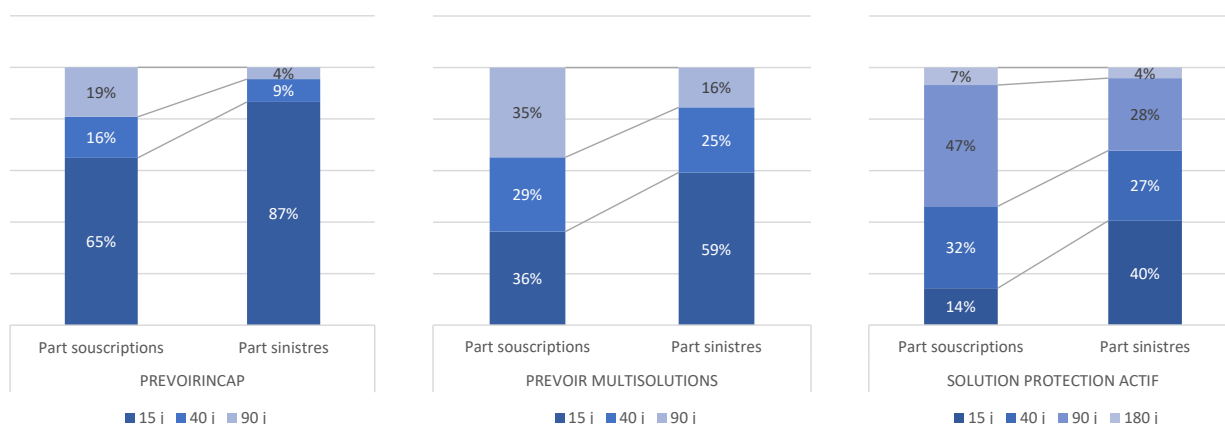


Figure 10 : Part des niveaux de franchise dans les souscriptions et les sinistres sur les garanties IJ toute cause pour les salariés

Trois phénomènes participent probablement à cette observation. Le plus important est lié au caractère décroissant de la fréquence des sinistres en fonction de leur durée. Autrement dit, les sinistres courts sont plus nombreux que les sinistres longs, comme l'illustre la distribution des durées des sinistres selon la table du BCAC (**Figure 11**), ci-après.

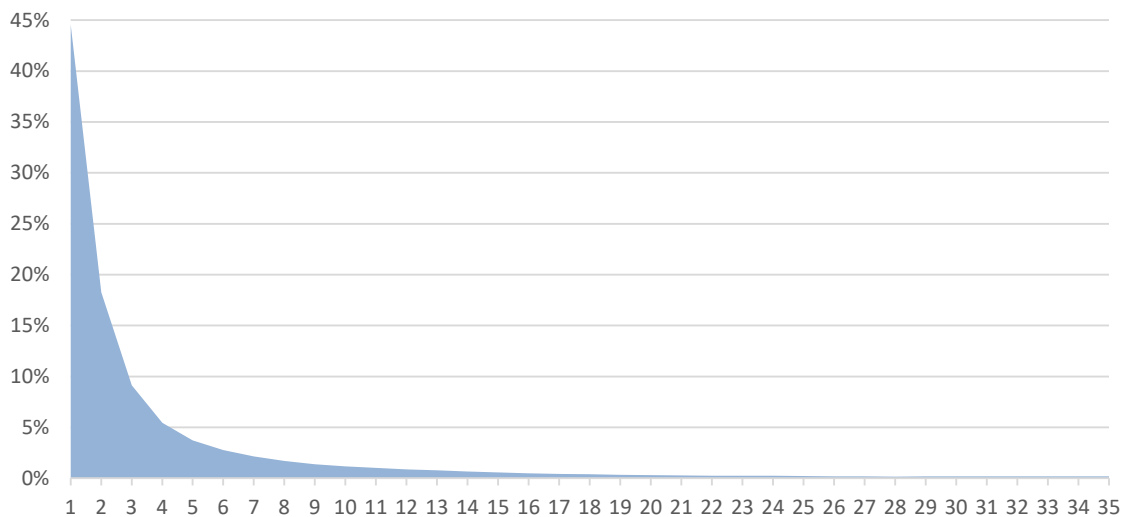


Figure 11 : Distribution des durées d'incapacité en mois, selon la table du maintien du BCAC (2013) pour un individu âgé de 40 ans

Ce phénomène se traduit par un effet dit de **troncature**. Plus précisément, **les assurés ne déclarent un état d'incapacité à l'assureur que lorsque ce dernier s'étend au-delà de la période franchise**. Plus la franchise est élevée, plus le nombre d'incidences inobservées par l'assureur est important, et donc plus le taux d'entrée en incapacité est faible. Cet élément, très structurant dans la construction des lois d'entrée en incapacité, sera détaillé dans le deuxième chapitre.

Le deuxième phénomène en jeu, plus difficile à démontrer, est comportemental. Un document de travail de l'IRDES publié en 2015 (IRDES, 2015) suggère que **le taux d'entrée en incapacité est fortement influencé par le niveau de couverture**. Autrement dit, la couverture par une garantie indemnité journalière dès le 15^{ème} jour pourrait inciter le recours à l'ITT. Ce phénomène peut être qualifié de **risque moral** (*moral hazard*).

Enfin le troisième phénomène en jeu est **l'antisélection**. Les franchises courte attirent intrinsèquement des profils plus risqués, se sachant plus susceptibles d'utiliser la garantie.

Un second niveau d'analyse consiste à étudier la distribution des sinistres en fonction de l'ancienneté du contrat. Il apparaît dans la **Figure 12** ci-après, que les sinistres sont fortement regroupés en début de vie du contrat. Il est logique de constater une telle distribution car le nombre de contrats est décroissant avec le temps, particulièrement dans le cas des contrats TNS qui chutent traditionnellement plus rapidement (environnement concurrentiel exacerbé et fragilité financière des TNS assurés par Prévoir Vie sur les anciens portefeuilles). Seul un rapprochement avec la population sous risque, dans le chapitre 2, permettra d'analyser finement cette variable. En revanche, un tel nombre de sinistres sur les 12 premiers mois de vie du contrat amène dès à présent à s'interroger sur la sincérité des réponses au questionnaire médical et sur l'antisélection en œuvre à la souscription.

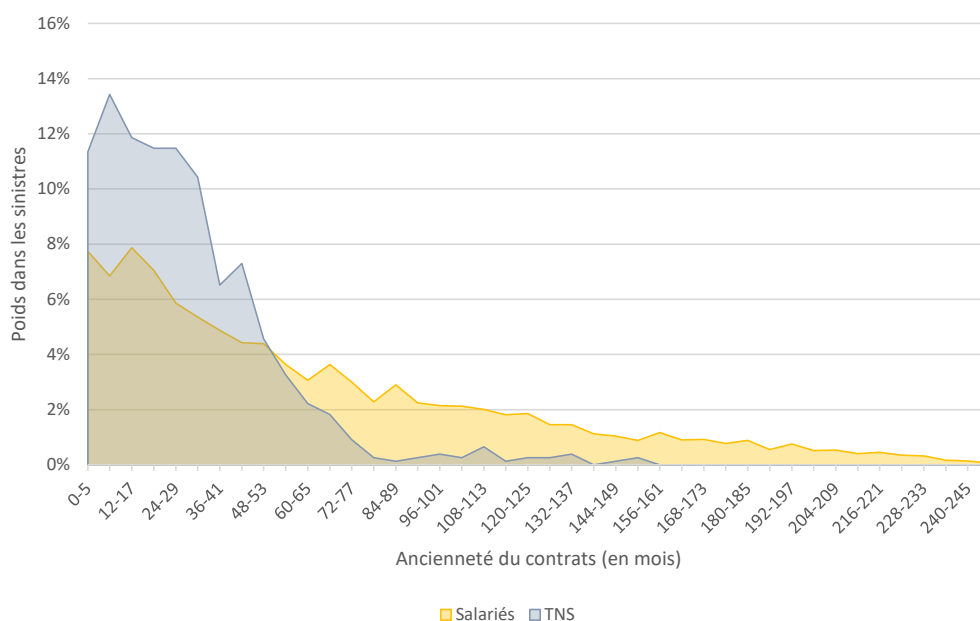


Figure 12 : Répartition des sinistres IJ toute cause en fonction de l'ancienneté des contrats

Enfin, les autres variables n'apportent à ce stade que peu d'éléments d'information sur la sinistralité. Elles seront en revanche étudiées beaucoup plus finement par la suite.

2.3.2 Les données emprunteur

Les contrats emprunteur sont commercialisés par un courtier grossiste. Les données transmises sont moins riches, mais permettent tout de même d'identifier un certain nombre de points remarquables. En particulier, les assurés présentent un profil très différent de celui observé sur les contrats de prévoyance Prévoir. La somme des temps d'exposition est d'environ 800 000 années.

2.3.2.1 Caractéristiques des populations assurées

Les contrats emprunteur sont majoritairement souscrits dans le cadre de l'acquisition d'un bien immobilier. Ils sont donc généralement souscrits par des populations disposant de revenus suffisants pour accéder à la propriété. Le revenu n'étant pas communiqué à la souscription du contrat d'assurance emprunteur, mais au moment de la souscription du prêt à la banque, cette information n'est pas disponible dans les rapports du courtier grossiste. Cependant, les professions sont indiquées, permettant d'extrapoler le niveau de richesse des individus. La **Figure 13**, ci-dessous, illustre la part majoritaire des cadres dans le portefeuille d'assurés.

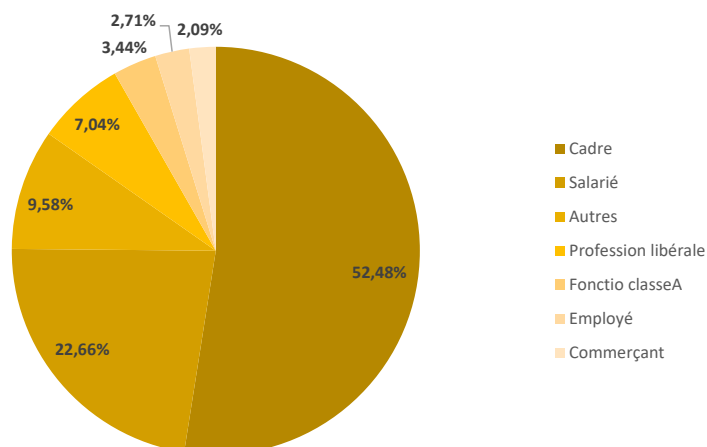


Figure 13 : Répartition des professions des assurés dans le portefeuille emprunteur

L'âge à la souscription révèle également les différences qui peuvent s'opérer entre les portefeuilles. Il apparaît en effet dans la **Figure 14** ci-dessous que les assurés des contrats emprunteur entrent en portefeuille plus tard que ceux des contrats prévoyance Prévoir. Ils souscrivent à un moment plus avancé de leur vie professionnelle. Cette observation tient probablement à deux phénomènes : d'une part, la nécessité de disposer d'un minimum de ressources financières pour accéder à la propriété, et d'autre part, la naissance d'un deuxième enfant juste avant 40 ans pouvant motiver un déménagement.

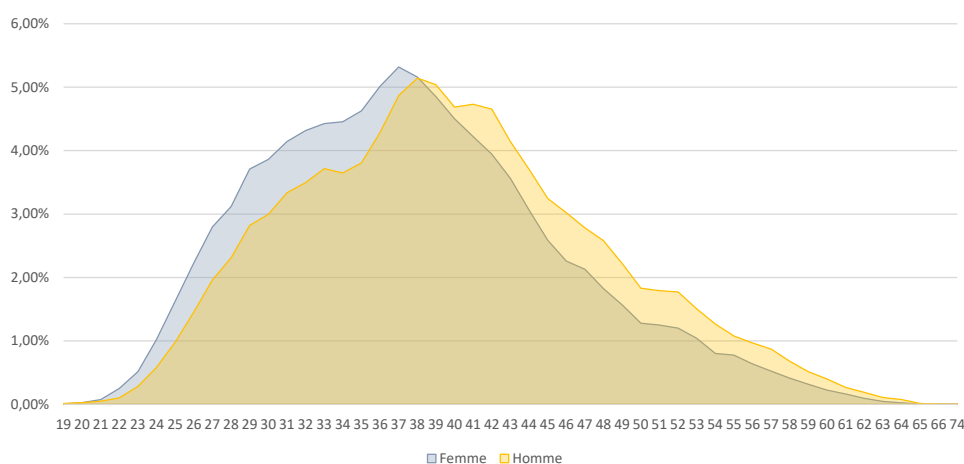


Figure 14 : Distribution des âges à la souscription

2.3.2.2 Caractéristiques des sinistres

Le nombre de sinistres observés sur ce portefeuille est de 2 972. Alors que la part des deux sexes dans la population assurée est quasi équivalente, la part des femmes dans les sinistres est globalement supérieure à celle des hommes (**Figure 15**). Il apparaît en outre clairement que les sinistres surviennent à des âges plus jeunes pour les femmes que pour les hommes. Cette observation s'explique sans doute en grande partie par les grossesses pathologiques.

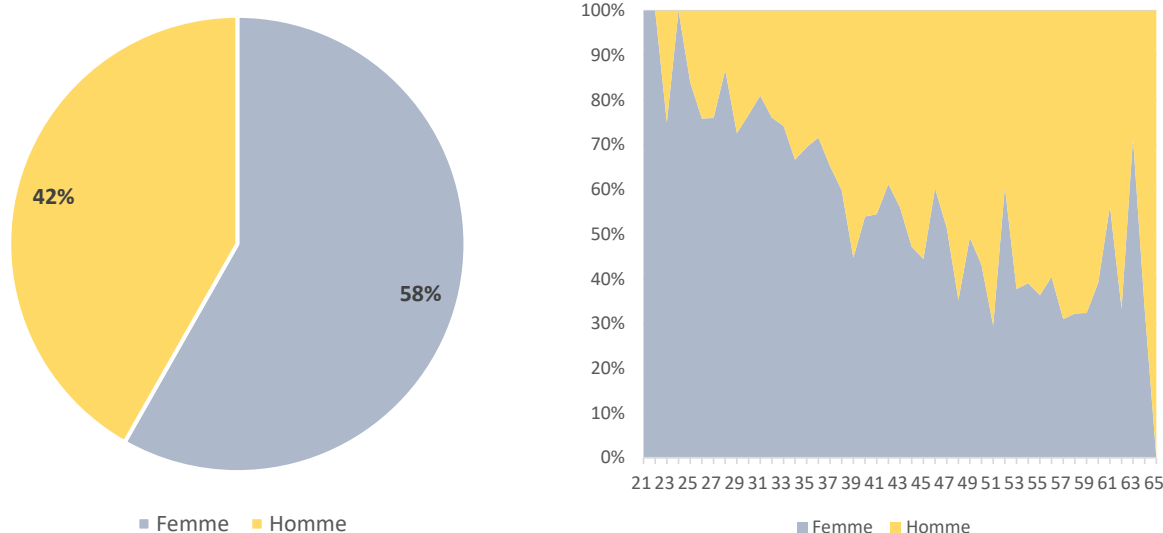


Figure 15 : Répartition des sinistres entre les femmes et les hommes

Il apparaît dans la **Figure 16** que la distribution des sinistres par ancienneté des femmes est très proche de celle des hommes. Parallèlement, ce graphique illustre également l'impact de la période de carence : les sinistres sont généralement moins nombreux en première année compte tenu du délai d'attente de 90 jours. En revanche, la baisse affichée à partir de l'année 4 ne traduit pas, à ce stade, une baisse de la sinistralité. Les effectifs sous risque sont en effet plus faibles sur les anciennetés les plus élevées.

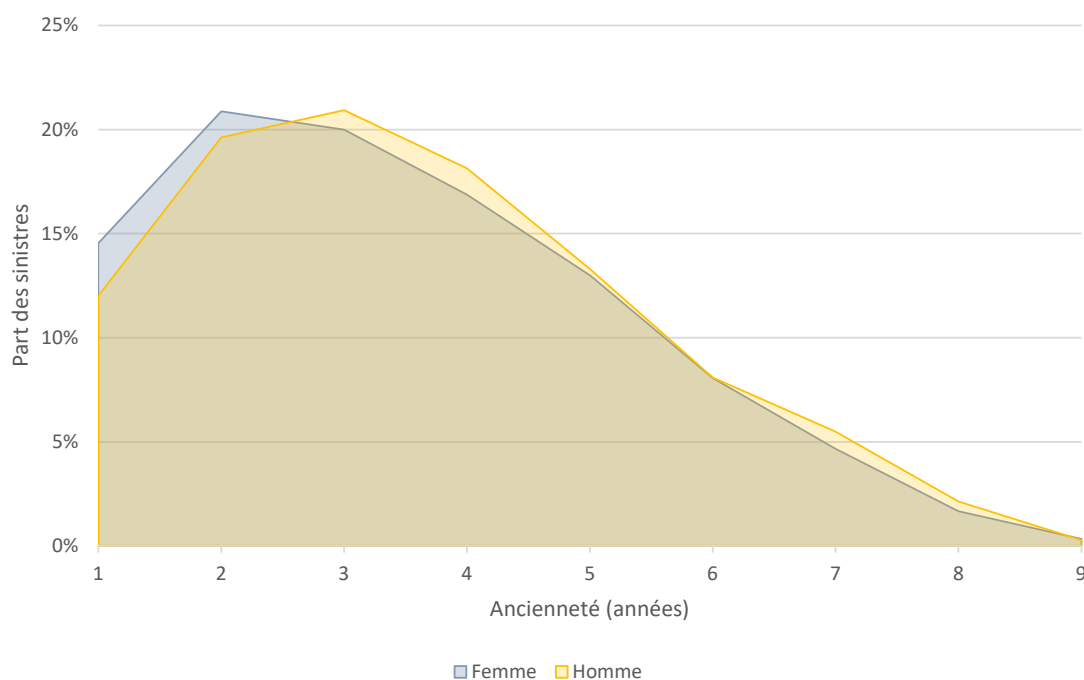


Figure 16 : Répartition des sinistres en fonction de l'ancienneté du contrat

Ces éléments permettent d'apprécier les portefeuilles étudiés avec un peu plus de précisions. Ils traduisent la **forte hétérogénéité des assurés dans ces portefeuilles**. Ils montrent également qu'au sein des portefeuilles de contrats de prévoyance classiques distribués par Prévoir, de profondes mutations s'opèrent, en lien avec l'histoire de la Compagnie. En ce sens, ces travaux permettent de forger une intuition sur le caractère non-

universel des lois qui seront définies dans la suite de cette étude. Dans ces conditions, **il conviendra de mesurer l'impact d'un certain nombre de variables sur les taux d'entrée en incapacité.**

La partie 3 a donc pour objet la description des techniques statistiques mises en œuvre dans le but de produire des lois d'entrée en incapacité. Il sera également question d'évaluer l'impact des différentes variables. La détermination de coefficients de majoration ou de minoration du risque pour chacune d'elle permettra alors de calibrer les lois en fonction des travaux à réaliser.

3 CONSTRUCTION DES TABLES D'ENTREE EN INCAPACITE

La construction de tables d'entrée en incapacité relève du domaine de l'analyse de survie. Les démarches s'inscrivant dans cette discipline ont vocation à étudier les durées de vie. Il peut s'agir de la durée de vie d'un individu, du temps qui s'écoule avant la défaillance d'une machine, ou de façon générale du temps qui s'écoule jusqu'à l'apparition d'un événement précis, manifestant un changement d'état. Cet événement est qualifié d'évènement d'intérêt, ou génériquement de décès.

La fonction de survie $S(t)$ mesure la probabilité que la durée de vie, assimilée à la réalisation de la variable aléatoire T , excède une durée écoulée t . Elle mesure donc la proportion d'individus encore en vie à chaque point t . Il s'agit par définition d'une fonction décroissante qui est notée sous la forme suivante :

$$S(t) = \mathbb{P}(T > t).$$

La fonction de hasard $h(t)$ mesure la probabilité pour un individu de décéder entre t et $t + dt$ sachant qu'il était vivant juste avant t . Cette mesure est parfois qualifiée de force de mortalité ou de taux de hasard instantané.

$$h(t) = \lim_{dt \rightarrow 0} \frac{\mathbb{P}(t \leq T < t + dt \mid T \geq t)}{dt}.$$

En pratique, la fonction de hasard permet d'identifier, en tout instant, l'intensité du risque de changement d'état. Elle n'est donc pas nécessairement croissante mais elle est en revanche positive ou nulle.

L'entrée en incapacité traduit un changement d'état : passage de l'état « capable » à l'état « incapable ». L'étude de cet évènement s'inscrit donc naturellement dans l'analyse de survie. En pratique, la fonction de survie des individus en état de travailler n'est pas connue. L'objectif est donc ici de développer des outils statistiques permettant de **proposer une estimation, à partir des observations, de la loi de probabilité régissant, en fonction de leur âge, le passage des individus en incapacité.**

La construction des tables d'entrée en incapacité peut être présentée en deux étapes. D'une part, l'élaboration des taux brut d'entrée en incapacité. D'autre part, le lissage des taux bruts pour obtenir une courbe régulière.

3.1 Estimation des taux bruts

Idéalement, l'assureur souhaiterait obtenir une distribution continue de la probabilité d'entrée en incapacité en fonction de l'âge. L'actuaire pourrait également calculer la probabilité d'entrée une seule fois en incapacité, puis deux fois etc. pour chaque tranche d'âge. Mais de tels travaux statistiques nécessiteraient une quantité très importante de données pour obtenir des résultats significatifs. Pour cette raison, il est proposé de calculer, à chaque âge entier, la probabilité d'entrer au moins une fois en incapacité, c'est-à-dire de l'évènement complémentaire de « ne pas entrer en incapacité ».

Une autre problématique pour l'assureur tient à l'impossibilité d'observer intégralement l'ensemble des durées de survie compte tenu de l'incomplétude des données. Deux phénomènes en particulier doivent être considérés :

- D'une part, une observation est dite **censurée** lorsque l'assureur dispose d'un intervalle de temps pour toute information sur la date de survenance de l'évènement, généralement en dehors de la période suivie. En particulier, un individu est censuré lorsque ce dernier quitte l'étude sans que l'évènement d'intérêt soit survenu (censure à droite). L'absence de survenance de l'évènement sur l'intervalle de temps étudié constitue donc une information. En revanche il n'est pas possible de déterminer à quelle date surviendra l'évènement.

- D'autre part, une observation est dite **tronquée** lorsque l'assureur n'est pas en mesure d'observer de durées supérieures ou inférieures à certains seuils. Particulièrement, les troncatures sont introduites par les entrées en portefeuille entre deux âges entiers. Les franchises introduisent également des troncatures dans la mesure où seuls les sinistres dont le maintien en incapacité est supérieur au niveau de la franchise sont connus. Pour cette raison, il est courant d'établir des lois d'entrée en incapacité pour chaque niveau de franchise.

Intuitivement, le calcul de la population sous risque sur une tranche d'âge donnée ne peut donc se résumer au simple dénombrement des assurés en portefeuille. Une approche naïve pourrait consister à intégrer et sortir les individus de l'étude uniquement sur des âges entiers. Mais la problématique de censure concernant la quasi-totalité des individus, la perte d'information serait trop importante. Dans ces conditions, **il est préférable d'utiliser un estimateur statistique des taux d'entrée qui permette de capter ce phénomène. Les estimateurs de Hoem et de Kaplan-Meier sont adaptés.**

3.1.1 Estimateur des moments de Hoem

Jan Hoem (1939 – 2017) est un statisticien suédois. Il dirigea notamment l'Institut Max Planck pour la recherche démographique (Allemagne) et a particulièrement œuvré dans le domaine des sciences actuarielles. Dans un article publié en 1977, il proposa une généralisation de l'estimateur binomial. Il s'agit d'un estimateur paramétrique des taux d'entrée dans le sens où le nombre d'événements d'intérêt est supposé suivre une loi binomiale. Cet estimateur est connu sous le nom de « estimateur des moments de Hoem ».

En pratique, l'estimateur s'articule autour de la décomposition du temps en intervalles $[x ; x + 1[$. Au sein de ces intervalles, qui peuvent par exemple correspondre à des âges entiers, le temps d'exposition au risque de chacun des n_x individus i peut être formalisé par un intervalle $[s_i - r_i]$.

Plus précisément, en supposant que l'individu i entre à l'âge $x + r$ et sorte à l'âge $x + s$, son temps d'observation est égal à $s - r$. En absence de troncature, de censure ou d'événement d'intérêt, l'individu est présent sur toute la tranche d'âge, r et s sont nuls et son temps d'observation est égal à $x + 1 - x = 1$. Dans le cas contraire, le temps d'observation, illustré ci-dessous, est compris dans l'intervalle $]0,1[$.



La probabilité pour un individu i d'entrer en incapacité sur la période d'observation, sachant qu'il n'y était pas au début de la période d'observation (c'est à dire la période commençant en $x + r$ et s'étalant sur une durée de $s - r$) se note de la façon suivante :

$${}_{s_i-r_i}q_{x+r_i}. \quad (1)$$

Il est fait l'hypothèse que la variable aléatoire X_i , qui vaut 1 en cas d'entrée en incapacité de l'individu i entre $x + r$ et $x + s$ et vaut 0 sinon, suit une loi de Bernoulli :

$$X_i \sim \beta({}_{s_i-r_i}q_{x+r_i}).$$

Il est également fait l'hypothèse que dans l'intervalle $[x_i ; x_i + 1[$ la probabilité pour un individu d'âge x_i d'entrer en incapacité entre l'âge x_i et $x_i + t$ est une fonction linéaire du temps t :

$${}_tq_x = tq_x \forall t \in [0 ; 1[. \quad (2)$$

L'expression (1) peut donc être approchée par $(s_i - r_i) * q_x$ ce qui implique que $\mathbb{E}[X_i] = (s_i - r_i) * q_x$.

On pose alors $Z_i = \frac{X_i}{s_i - r_i}$, si bien que $\mathbb{E}[Z_i] = \frac{\mathbb{E}[X_i]}{s_i - r_i} = q_x$.

Les variables Z_i étant supposées mutuellement indépendantes et de même loi, alors la loi des grands nombres indique que pour un nombre suffisamment élevé d'observations, la moyenne empirique $\bar{Z}$ converge en probabilité vers la moyenne théorique $\mathbb{E}[Z]$. Il est dès lors possible d'en déduire l'estimateur $\hat{q}_x$:

$$\hat{q}_x = \frac{\sum_{i=1}^{n_x} X_i}{\sum_{i=1}^{n_x} (s_i - r_i)} = \frac{d_x}{\sum_{i=1}^{n_x} (s_i - r_i)}.$$

La qualité d'un estimateur est généralement mesurée par l'erreur quadratique, qui peut être définie comme l'espérance de l'écart quadratique entre la vraie valeur du paramètre et l'estimateur :

$$EQ(\hat{q}_x) = E_q(\hat{q}_x - q_x)^2 = V_q(\hat{q}_x) + E_q(\hat{q}_x) - q_x.$$

Autrement dit, la qualité de l'estimateur peut être évaluée par la somme de sa variance et de son biais, c'est-à-dire de l'écart entre son espérance et la vraie valeur du paramètre. Le biais correspond à l'erreur structurelle. La variance mesure l'erreur statistique, la dispersion par rapport à l'espérance. Il est possible de démontrer que l'estimateur de Hoem est un estimateur sans biais, c'est-à-dire que son espérance est égale à la vraie valeur du paramètre.

$$E[\hat{q}_x] = E\left[\frac{\sum_{i=1}^{n_x} X_i}{\sum_{i=1}^{n_x} s_i - r_i}\right] = \frac{E[\sum_{i=1}^{n_x} X_i]}{\sum_{i=1}^{n_x} s_i - r_i} = \frac{\sum_{i=1}^{n_x} (s_i - r_i) * q_x}{\sum_{i=1}^{n_x} (s_i - r_i)} = q_x.$$

Dès lors, l'erreur quadratique se confond avec sa variance, qui peut être exprimée ainsi :

$$V(\hat{q}_x) = V\left(\frac{Y_x}{\sum s_i - r_i}\right) = \frac{V(Y_x)}{(\sum s_i - r_i)^2} = \frac{\sum (s_i - r_i) q_x (1 - q_x)}{(\sum s_i - r_i)^2} = \frac{p_x q_x}{\sum s_i - r_i}.$$

Il apparaît donc que la variance $V(\hat{q}_x)$ tend vers 0 lorsque la taille n de l'échantillon est suffisamment grande. Dès lors, l'estimateur $\hat{q}_x$ a la propriété de converger en probabilité vers q_x . C'est à dire qu'il prend des valeurs proches de q_x avec une probabilité d'autant plus proche de 1 que la taille n de l'échantillon sera grande.

L'erreur quadratique est donc asymptotiquement nulle, et la forme de la variance indique que cet estimateur converge à la vitesse $\frac{1}{n}$.

En faisant intervenir le théorème central limite, il est possible de construire des intervalles de confiance au niveau $1 - \alpha$, avec $u_{\alpha/2}$ le fractile d'ordre $1 - \frac{\alpha}{2}$ d'une loi normale centrée réduite et $N_x = \sum s_i - r_i$ la population sous risque :

$$IC(\alpha) = \left[\hat{q}_x \pm u_{\alpha/2} \sqrt{\frac{\hat{q}_x(1 - \hat{q}_x)}{N_x}} \right] \forall N_x \geq 30, N_x q \geq 5 \text{ et } N_x(1 - q) \geq 5.$$

L'approximation d'une loi binomiale $\beta(n, q)$ par une loi normale $N(nq, nqp)$ est généralement considérée comme valable pour $n \geq 30, nq \geq 5$ et $n(1 - q) \geq 5$ (Lecoutre, 2019).

3.1.2 Estimateur de Kaplan-Meier

L'estimateur de Kaplan-Meier est issu de la collaboration du mathématicien Edward L. Kaplan (1920-2006) et du biostatisticien américain Paul Meier (1924-2011). Dans un article publié en 1958 dans la revue *Journal of the American Statistical Association* et intitulé « Nonparametric estimation from incomplete observations » (Kaplan & Meier, 1958), ils proposent un estimateur non-paramétrique sous le nom de « estimateur produit-limite ». Cet estimateur est non-paramétrique dans le sens où aucune hypothèse n'est faite sur la distribution des taux d'entrée. Il s'agit d'un estimateur de la fonction de survie.

Le principe énoncé est que la probabilité de survivre jusqu'à $t > s$ est équivalente à la probabilité de survivre jusqu'à s , puis, sachant qu'on a survécu jusqu'à s , de survivre jusqu'à t . Plus formellement :

$$S(t) = P(T > s)P(T > t|T > s) = S(s)P(T > t|T > s).$$

Cette expression fait apparaître un produit de termes, à l'origine du nom donné à cet estimateur. Le temps est fractionné en autant de dates u_j qu'il y a d'instantanés pendant lesquels se produisent des événements, qu'il s'agisse d'entrées, de sorties ou pertes (censures). En notant p_j la probabilité de survivre sur l'intervalle $]u_{j-1}, u_j]$ sachant que l'individu était vivant en u_{j-1} et $q_j = 1 - p_j$, l'expression de la fonction de survie $S(t)$ devient alors :

$$S(t) = \prod_{j=1}^n P(T > u_j | T > u_{j-1}) = \prod_{j=1}^n p_j = \prod_{j=1}^n (1 - q_j).$$

S'agissant d'une fréquence, un estimateur naturel est donné par $\hat{q}_j = \frac{d_j}{n_j}$ avec d_j le nombre d'événements d'intérêts (ici des entrées en incapacité) survenus sur $]u_{j-1}, u_j]$ et n_j la population sous risque dans cet intervalle. Dès lors, l'expression de l'estimateur de la fonction de survie devient :

$$\hat{S}(t) = \prod_{T_j \leq t} \left(1 - \frac{d_j}{n_j}\right).$$

L'estimation du taux brut sur la tranche d'âge $[x, x + 1[$ est ainsi défini par :

$$\hat{q}_x = 1 - \hat{S}_x(t) = 1 - \prod_{T_j \leq t} \left(1 - \frac{d_j}{n_j}\right), \forall j \in [x, x + 1[.$$

En l'absence d'ex æquo, c'est-à-dire en considérant qu'à chaque instant il ne peut y avoir qu'un seul et unique événement (censure ou événement d'intérêt), $n_j = n - i + 1$ et $d_i = 1$ ou 0.

Plus généralement, la présence d'ex æquo impose, par convention, de considérer que les observations non censurées précèdent systématiquement les censures. Ainsi $n_j = n_{j-1} - d_{j-1} - \lambda_{j-1} + \theta_{j-1}$ avec λ_j et θ_j respectivement les nombres de pertes et d'entrées entre $]u_{j-1}, u_j]$.

En pratique, l'estimateur de Kaplan-Meier nécessite de construire une table sur laquelle figure autant de dates que d'événements. Pour chacune de ces dates, ordonnées de façon décroissante, le nombre d'événements survenus est consigné : entrées, décès, pertes. Si bien qu'il est possible de calculer la population sous risque à chacun des points : $n_j = n_{j-1} - d_{j-1} - \lambda_{j-1} + \theta_j$. Enfin, l'estimateur $\hat{q}_j = \frac{d_j}{n_j}$ est ajouté à la table. Le **Tableau 8** ci-après, est donné à titre d'exemple.

Intervalles		Décès	Pertes	Entrées	At risk	$\hat{q}_j$
u_{j-1}, u_j	j	δ_j	λ_j	θ_j	n_j	δ_j/n_j
0, 1	1	3	0	0	100	3/100
1, 3	2	2	1	5	97	2/97
3, 4	3	0	3	1	99	0/99
4, 9	4	4	2	1	97	4/97
9, 10	5	3	4	0	92	3/92

Tableau 8 : Illustration table des évènements

Enfin, dans les situations pour lesquelles les dates exactes des événements ne seraient pas connues, ou simplement dans une démarche d'allègement des calculs, une discrétisation du temps peut être opérée. Les événements peuvent par exemple être regroupés sur des pas mensuels en comptabilisant les entrées en incapacité mois par mois. Il s'agit alors d'une approche paramétrique, puisqu'une hypothèse sur la distribution binomiale des entrées en incapacité doit être retenue. L'estimateur du paramètre de la loi est donné par la résolution de la log vraisemblance et est de même forme que celui énoncé dans la version non paramétrique (Planchet, Modèles de durée : Statistique des modèles non paramétriques, 2021).

La variance de la fonction de survie peut être approchée par l'estimateur de de Greenwood :

$$\hat{V}(\hat{S}(t)) = \hat{S}(t)^2 \sqrt{\sum_{T_{(i)} \leq t} \frac{d_i}{n_i(n_i - d_i)}}.$$

Enfin, dans leur ouvrage *Modélisation statistique des phénomènes de durée*, Planchet et Thérond précisent que **l'estimateur de Kaplan-Meier est convergent et cohérent mais est toutefois positivement biaisé et surestime donc en moyenne la vraie valeur de la fonction de survie. Il est l'unique estimateur cohérent de la fonction de la survie. Il est en outre asymptotiquement gaussien.** Ces propriétés font de l'estimateur de Kaplan-Meier l'estimateur le plus utilisé pour l'estimation d'une fonction de survie (Planchet & Thérond, Modélisation statistique des phénomènes de durée, 2011).

Cette dernière propriété permet de calculer les intervalles de confiance asymptotiques ponctuels dont les bornes sont données par :

$$IC(\alpha) = \hat{S}_i \left(1 \pm u_{\frac{\alpha}{2}} \sqrt{\sum_{T_{(i)} \leq t} \frac{d_i}{n_i(n_i - d_i)}} \right).$$

avec $u_{\alpha/2}$ le fractile d'ordre $1 - \frac{\alpha}{2}$ d'une loi normale centrée réduite.

3.1.3 Présentation des taux bruts (illustrée avec le portefeuille emprunteur)

Les taux bruts issus des estimateurs de Kaplan Meier et de Hoem sont très proches, comme l'illustre la représentation graphique ci-après sur le portefeuille emprunteur (**Figure 17**). L'ensemble des calculs suivants seront détaillés pour le portefeuille d'assurance emprunteur. Les résultats seront directement communiqués pour les portefeuilles de prévoyance classiques.

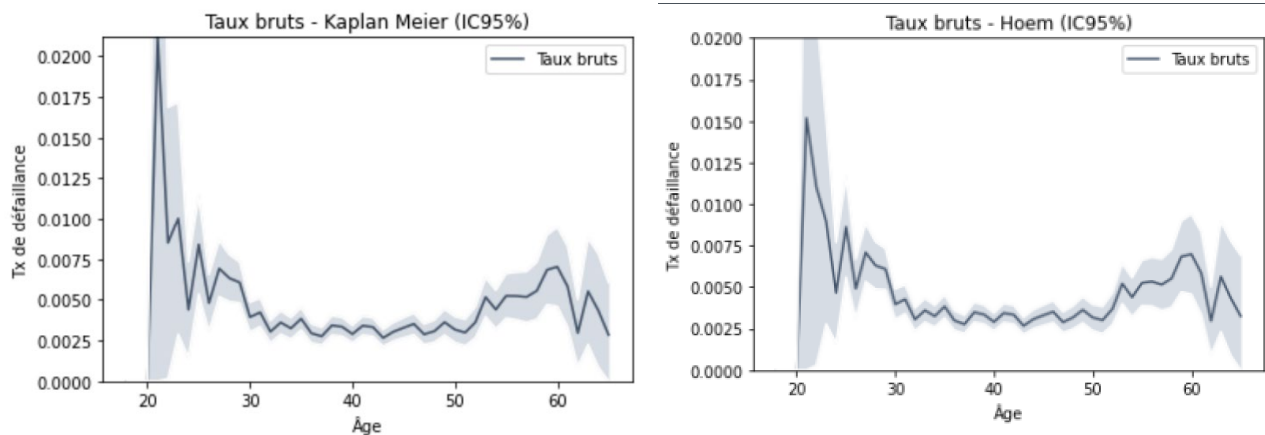


Figure 17 : Taux bruts d'incidence à partir des estimateurs de Kaplan Meier et de Hoem

L'estimateur de Kaplan-Meier étant le plus couramment utilisé, et étant particulièrement bien documenté dans la littérature, ce dernier est retenu pour la suite des travaux présentés.

Il apparaît que le tracé des taux de sortie prend une forme convexe. C'est-à-dire qu'il décroît jusqu'à environ 35 ans, puis croît jusqu'à environ 60 ans. Il décroît ensuite de nouveau jusqu'à 65 ans. La décomposition de ces taux bruts par sous-populations permet de mieux comprendre ce taux moyen. En particulier, les taux élevés chez les plus jeunes semblent hérités de la population féminine, comme l'illustre la représentation graphique des taux bruts selon le sexe (**Figure 18**) :

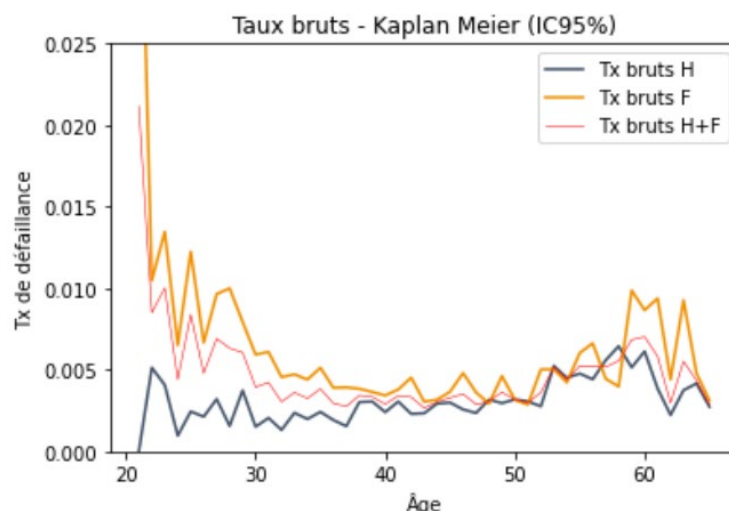


Figure 18 : Taux bruts par sexe

Les taux bruts de la population féminine sont nettement supérieurs chez les jeunes. Avec le vieillissement, les taux bruts des deux sexes se rejoignent pour quasiment se confondre vers 45 ans. Il est très probable que ce phénomène illustre l'impact des grossesses pathologiques sur le taux de risque des femmes.

La segmentation en populations plus étroites se traduirait à ce stade par une volatilité trop importante des taux bruts. Les effets de certaines caractéristiques seront étudiés plus précisément par la suite.

Enfin il convient de ne conserver que les classes d'âge pour lesquelles les données sont suffisantes. En pratique, s'agissant de l'estimateur de Hoem, il est d'usage d'appliquer les critères d'approximation d'une loi binomiale par une loi normale, c'est-à-dire $n_x \geq 30, n_x * \hat{q}_x \geq 5, n(1 - \hat{q}_x) \geq 5$ (Lecoutre, 2019). Ces critères apparaissent généralement dans la littérature sous l'appellation « critères de Cochran ». Dans le cas de l'estimateur de Kaplan-Meier, il convient plutôt de fixer une valeur limite sur l'amplitude des intervalles de confiance. L'application de telles conditions amène ici à ne conserver que les âges compris entre 21 et 65 ans. Les âges compris entre 18 et 20 sont donc rejetés.

3.2 Lissage des taux bruts

A moins de disposer d'une quantité très importante de données à tout âge, les taux issus des calculs précédents sont erratiques. C'est-à-dire que leur allure est chaotique et présente des sauts entre les tranches d'âge adjacentes. Ce phénomène est particulièrement marqué sur les âges extrêmes, pour lesquels la population sous risque est moins importante et donc la volatilité plus importante. Il est ainsi difficile de les lire et de les interpréter sous cette forme. Dans la mesure où ces taux sont parfaitement ajustés aux données d'apprentissage, il est probable qu'ils introduisent un biais élevé par rapport à la sinistralité des futurs assurés. Dans ces conditions, **un travail de lissage des taux bruts est essentiel et n'est pas dicté par les seules considérations esthétiques.**

La méthode de Whittaker-Henderson est une méthode de lissage. Elle cherche à minimiser une mesure M , qui peut être entendue comme une fonction de coût. Cette mesure se décompose comme la somme d'un critère F de la fidélité aux taux bruts et d'un critère S de régularité (smoothness). Un coefficient h permet de contrôler l'influence de chacun de deux critères sur la mesure M . F et S n'ayant pas la même unité, h peut également être interprété comme une constante de normalisation.

$$M = F + hS.$$

3.2.1 Mesure de fidélité F

La fidélité est mesurée par la somme des écarts quadratiques entre les taux lissés g_x et les taux bruts $\hat{q}_x$. Afin de limiter l'impact des points aberrants, un poids, noté w_x peut-être ajouté à chaque âge :

$$F = \sum_{x=x_{inf}}^{x=x_{sup}} w_x (g_x - \hat{q}_x)^2.$$

Les poids retenus correspondent généralement à l'effectif à l'âge x rapporté à l'effectif moyen sur l'ensemble des âges.

$$w_x = \frac{n_x}{\frac{\sum_{x=x_{inf}}^{x=x_{sup}} n_x}{x_{sup} - x_{inf} + 1}}.$$

3.2.2 Mesure de régularité S

Afin de mesurer la régularité S de la courbe, Whitaker-Henderson s'appuie sur le calcul de distance entre chacun des points adjacents sur la courbe. Formellement, il s'agit de la différence avant par pas de 1, qui peut être définie comme : $\Delta f(x) = f(x + 1) - f(x)$, et qui peut être interprétée comme une version discrétisée de la dérivée.

Plus précisément, afin de calibrer l'intensité du lissage, des différences avant d'ordres supérieurs peuvent être calculées. $\Delta^z f(x)$ correspond à la différence avant d'ordre z de $f(x)$, c'est-à-dire à la composée de $\Delta f(x)$, z fois, qui est donc définie par $\Delta^z f(x) = (\Delta \circ \Delta \circ \dots \circ \Delta)(f(x))$ et qui peut être interprétée comme la dérivée $z^{\text{ème}}$ discrétisée de $f(x)$.

Il est aisé de développer $\Delta^z f(x)$ pour $z = 3$ et $z = 4$ par exemple :

$$\Delta^3 f(x) = f(x+3) - 3f(x+2) + 3f(x+1) - f(x),$$

$$\Delta^4 f(x) = f(x+4) - 4f(x+3) + 6f(x+2) - 4f(x+1) + f(x).$$

Les premiers termes de chaque monôme correspondant, au signe près, aux coefficients binomiaux d'ordre z , ces expressions peuvent être généralisées en utilisant la formule de Pascal :

$$\Delta^z f(x) = \sum_{k=0}^z (-1)^{z-k} \binom{z}{k} f(x+k).$$

S'agissant d'un calcul de distance, les différences avant d'ordre z sont élevées au carré. En remplaçant $f(x)$ par les taux lissés g_x , l'expression de S , comme la somme des carrés des différences avant d'ordre z des taux lissés, est la suivante :

$$S = \sum_{x=x_{\inf}}^{x=x_{\sup}-z} (\Delta^z g_x)^2 = \sum_{x=x_{\inf}}^{x=x_{\sup}-z} \left(\sum_{k=0}^z (-1)^{z-k} \binom{z}{k} g_{x+k} \right)^2.$$

3.2.3 Forme matricielle et minimisation de M

En pratique, il convient de disposer de la forme matricielle de ces expressions. Le vecteur des taux lissés est noté g tel que $g = (g_x)_{x_0 \leq x \leq x_0+n}$, le vecteur des taux bruts est noté $\hat{q}$ tel que $\hat{q} = (\hat{q}_x)_{x_0 \leq x \leq x_0+n}$, la matrice diagonale des poids est notée W telle que $W = \text{diag}(w_x)_{x_0 \leq x \leq x_0+n}$.

L'expression du critère de fidélité F devient donc $F = {}^t(g - \hat{q})W(g - \hat{q})$.

Une matrice K_z , composée des coefficients binomiaux signés d'ordre z est introduite telle que $\Delta^z g = K_z g$. Le vecteur g des taux lissés étant de dimension n , la matrice K_z est impérativement composée de n colonnes. L'expression de S faisant intervenir g_{x+z} , afin que toutes les valeurs soient correctement définies, le nombre de lignes de K_z doit donc être de $(n - z)$. Si bien que la matrice $K_z g$ est de dimension $(n - z) \times n$. A titre d'illustration pour $n = 5$ et $z = 3$:

$$q = \begin{bmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \\ q_5 \end{bmatrix} \text{ et } K = \begin{bmatrix} -1 & 3 & -3 & 1 & 0 \\ 0 & -1 & 3 & -3 & 1 \end{bmatrix}.$$

Si bien que $K_z g = \begin{bmatrix} -1q_1 & 3q_2 & -3q_3 & 1q_4 & 0 \\ 0 & -1q_2 & 3q_3 & -3q_4 & 1q_5 \end{bmatrix}$, conformément à l'expression de $\Delta^3 f(x)$ développée ci-avant.

Plus généralement, l'expression matricielle du critère de régularité S devient donc :

$$S = {}^t(K_z g)(K_z g),$$

Si bien que $M = {}^t(g - \hat{q})W(g - \hat{q}) + {}^t(K_z g)(K_z g) = {}^t g W g - 2 {}^t g W \hat{q} + {}^t \hat{q} W \hat{q} + {}^t K_z K_z {}^t g g$.

Le problème d'optimisation peut être résolu en annulant la dérivée partielle par rapport au vecteur des taux lissés :

$$\frac{\partial M}{\partial g} = 2Wg - 2W\hat{q} + 2h^t K_z K_z g = 2(W + h^t K_z K_z)g - 2W\hat{q}.$$

S'agissant d'une matrice diagonale constituée de valeurs strictement positives, W est inversible. Avec une valeur de h suffisamment faible, la matrice $C = W + h^t K_z K_z$ se situe au voisinage de W et est donc également inversible. Dès lors, $\frac{\partial M}{\partial g} = 0$ pour $g = (W + h^t K_z K_z)^{-1} W \hat{q}$.

3.2.4 Calibrage et résultats :

La mise en place de l'équation matricielle ne relève pas du défi sur R ou sur Python mais le calibrage des constantes est en revanche plus délicat. Si z est généralement fixé à 2, il existe un consensus dans la littérature sur l'absence de solution idéale dans le choix de h . Les éléments suivants cherchent à explorer deux protocoles de calibrage de la constante de normalisation.

3.2.4.1 Calibrage par la statistique du Chi-deux

Dans un article publié en 1981 et intitulé *Use of the Chi-square statistic to set Whittaker-Henderson smoothing coefficients* (utilisation de la statistique du Chi-deux pour calibrer les coefficients de lissage de Whittaker-Henderson), Lee Giesecke propose un protocole de calibrage faisant intervenir une version adaptée de la statistique du Chi-deux (Giesecke, 1981).

Dans sa version simplifiée, Giesecke démontre que, pour u'' le vecteur des taux lissés et u le vecteur des taux bruts, la statistique de test $\sum \frac{w_i(u_i'' - u_i)^2}{u_i(1 - u_i)}$ suit une loi du Chi-deux dont le nombre de degrés de liberté est égal au rang de la matrice $B = (I - (W + hK'K)^{-1}W)$.

L'auteur précise que « (...) les statisticiens utilisent généralement la partie supérieure de la queue de distribution du Chi-deux pour évaluer si une statistique de test est supérieure à un niveau attendu (à une certaine probabilité α). Mais qu'il s'agit ici de s'intéresser aux deux queues de distribution, dès lors que l'objectif est également de mesurer si les taux lissés sont plus proches des taux bruts qu'attendus pour une valeur particulière du degré de lissage z (...) ».

En pratique, il s'agit donc de réinitialiser itérativement le coefficient de lissage h jusqu'à ce que la probabilité de recevoir une valeur plus élevée (ou plus faible) du Chi-deux soit suffisamment proche de 50%.

Enfin, Giesecke précise « (...) qu'il est rare que le lissage au cinquantième centile du Chi-deux soit aussi performant que le lissage avec des coefficients définis instinctivement. » et que « (...) le 50^{ème} centile procure a minima une bonne solution initiale. » En particulier, si les taux bruts sont déjà relativement lisses, les itérations du Chi-deux se traduiront par un lissage trop intense. *A contrario*, dans le cas de taux bruts trop irréguliers, ce protocole se traduira par un lissage insuffisant.

Dans le cas des données incapacité du portefeuille emprunteur, le cinquantième centile du Chi-deux correspond à $h = 7$ pour $z = 2$. La représentation graphique de ce lissage est donnée dans la **Figure 19** ci-après :

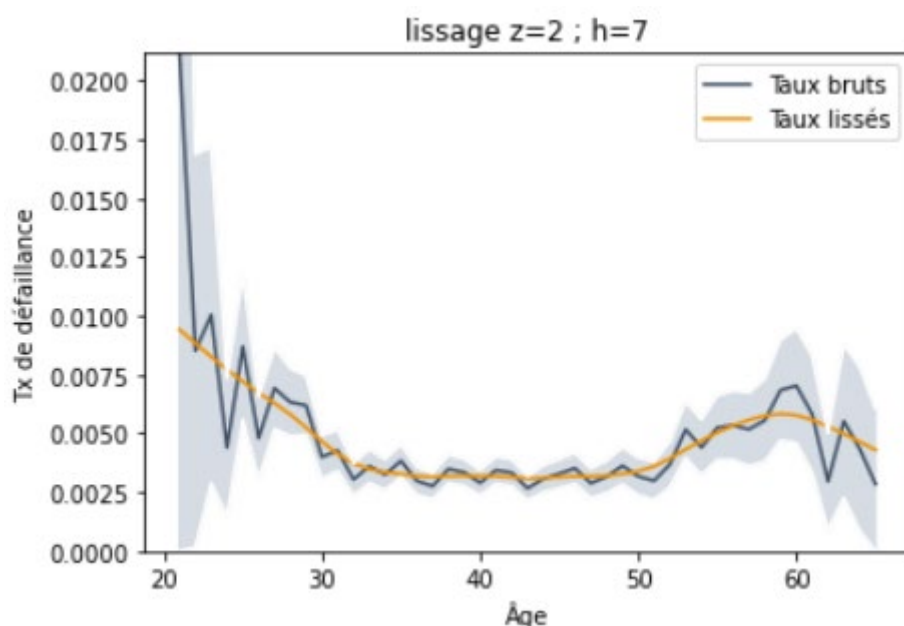


Figure 19 : Lissage pour $z=2$ et $h=7$ et l'utilisation de la matrice des poids

3.2.4.2 Calibrage par validation croisée K-fold

La validation croisée, ou cross-validation est une méthode couramment utilisée pour le calibrage des hyper paramètres dans les modèles d'apprentissage supervisé. Cette méthode consiste à découper le jeu d'entraînement en K sous-ensembles de même taille, appelés plis. Les plis sont à tour de rôle utilisés comme ensemble de validation tandis que les $K - 1$ plis sont utilisés comme ensemble d'entraînement (scikit-learn.org, s.d.). Cette méthode permet à toutes les observations d'être utilisées à la fois pour l'entraînement et la validation. Chaque observation est utilisée exactement une seule fois pour la validation. L'avantage de l'utilisation d'une telle stratégie de calibrage est de produire une loi plus généraliste puisqu'elle est testée sur des données n'ayant pas servi à l'élaboration de son lissage. Elle doit donc permettre de mieux prédire la sinistralité des futurs assurés en évitant les effets de surapprentissage.

Il est donc attendu que le lissage permette de prédire le mieux possible des points non utilisés dans l'apprentissage de la loi. La **Figure 20** ci-dessous illustre la décomposition du jeu de données dans le cas où $K = 4$.

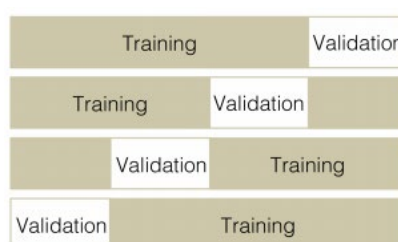


Figure 20 : Illustration de la validation croisée (source : scikit-learn.org)

Les paramètres testés sont le degré de lissage z , la constante de normalisation h et l'utilisation, ou non, de la matrice des poids. Ainsi, pour chaque valeur testée des paramètres, une nouvelle loi d'entrée en incapacité est calculée puis lissée sur chaque jeu d'entraînement. Le dernier pli est utilisé pour mesurer la performance du modèle.

Dans le cas du lissage, la performance du modèle peut être mesurée par un ratio de mortalité standardisé (ratio SMR). Cependant, un ratio SMR proche de 100% calculé globalement peut masquer des compensations entre les âges. Le choix se porte donc sur la statistique de test d'homogénéité du Chi-deux. Il s'agit d'additionner les écarts quadratiques relatifs en chacun des points de la courbe :

$$X = \sum_{x_{min}}^{x_{max}} \frac{(d_x - \hat{d}_x)^2}{\hat{d}_x}.$$

Le nombre de degrés de liberté n'a pas d'importance ici, dans le sens où il ne s'agit pas de déterminer une p-valeur mais de déterminer les taux lissés les plus adaptés aux échantillons de validation, c'est à dire présentant la statistique de test la plus faible.

Après avoir testé toutes les valeurs de h comprises entre 1 et 100 et les valeurs de z comprises entre 1 et 3, la valeur la plus faible de la mesure de performance est ainsi obtenue pour $z = 2$ et $h = 23$, sans utilisation de la matrice des poids. Les taux lissés sont présentés dans la **Figure 21** ci-après.

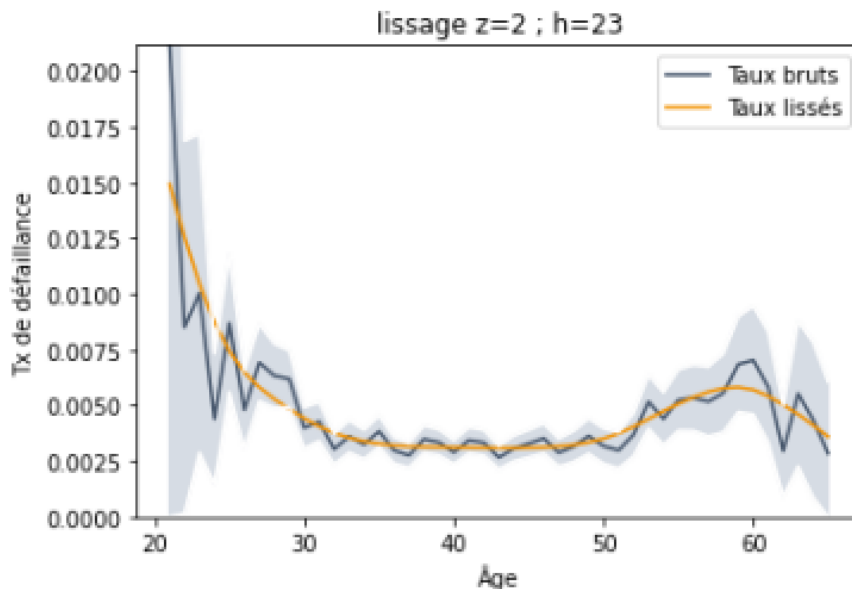


Figure 21 : Lissage pour $z=2$ et $h=23$

Dans la pratique, les mesures de performance calculées sont très proches, qu'il s'agisse du ratio SMR ou des écarts quadratiques relatifs. **Les résultats par validation croisée sont conservés dans la mesure où leur construction doit, en principe, garantir une meilleure généralisation de la loi** et donc sa meilleure adéquation à la sinistralité future.

Les taux lissés permettent de mieux apprécier la différence dans le cadencement du risque, précédemment exposé, entre les femmes et les hommes. Le graphique ci-après illustre ce propos.

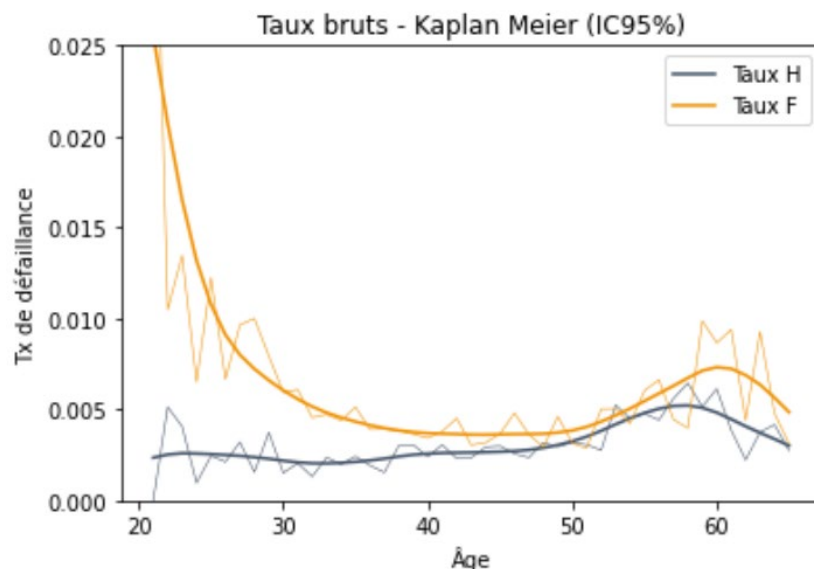


Figure 22 : Taux lissés par sexe

3.3 Régressions de Cox et effet de certaines caractéristiques sur le risque

A ce stade, les estimateurs des taux d'incidence calculés par Kaplan-Meier sont calibrés sur l'intégralité du portefeuille. Ils traduisent le comportement moyen de nombreux individus hétérogènes. Afin de réaliser des travaux de tarification, d'affiner les calculs de provisions, ou simplement de mieux comprendre les forces sous-jacentes, il est indispensable d'affiner ces lois en évaluant l'impact de variables exogènes sur le risque. A titre d'exemple, tarifier au même prix un cadre sédentaire et un ouvrier conduirait à être peu compétitif pour le cadre. *A contrario* le tarif serait particulièrement bon marché pour l'ouvrier ce qui exposerait l'assureur à des comportements anti-sélectifs. Plus précisément, **l'objectif est de disposer de coefficients de majoration ou de minoration du risque pour chacune de ces variables**. La méthode statistique retenue est la régression de Cox, dite à hasard proportionnel.

3.3.1 Principe de la régression de Cox à hasard proportionnel

David Cox est un statisticien britannique dont la plupart des travaux datent de la seconde moitié du XX^{ème} siècle. Il propose en 1972 une méthode permettant d'évaluer l'effet de covariables sur le risque instantané de défaillance (i.e. sur la fonction de hasard).

Le modèle de Cox repose sur la décomposition du risque de défaillance en deux composantes :

- Un risque de base noté $\lambda_0(t)$, dépendant du temps t et commun à tous les individus (i.e. indépendant des caractéristiques des individus).
- Un risque additionnel $\exp(\beta^T Z_i)$ ne dépendant pas du temps mais des caractéristiques des individus. β correspond au vecteur des p paramètres et z_i à la valeur des p covariables pour l'individu i .

La fonction de hasard pour l'individu i est ainsi donnée par $\lambda_i(t|Z_i, \beta) = \exp(\beta^T Z_i) \lambda_0(t)$. Aucune hypothèse n'est faite sur la distribution de $\lambda_0(t)$ qui correspond à la fonction de hasard, inconnue, pour $z = 0$. En ce sens, le modèle de Cox est dit semi-paramétrique.

La non-dépendance au temps du vecteur p -dimensionnel des covariables est une hypothèse forte ayant donné son nom aux modèles à hasard proportionnel. C'est-à-dire que **l'influence des variables sur la fonction de hasard est considérée comme stable dans le temps** (donc sur tous les âges dans le cas du risque d'incapacité). Dans ces conditions, le rapport des fonctions de hasard (*risk ratio*) entre deux individus, $\frac{\lambda_1(t|Z_1, \beta)}{\lambda_2(t|Z_2, \beta)} = \frac{\exp(\beta^T Z_1)}{\exp(\beta^T Z_2)} = \exp(\beta^T (Z_1 - Z_2))$, est indépendant du temps.

Pour un événement d'intérêt donné en t_j , la probabilité que la défaillance provienne de l'individu i est donnée par la vraisemblance partielle :

$$L_i(\beta) = \mathbb{P}(\text{évèn. de } i \text{ en } t_j \mid \text{évèn. en } t_j) = \frac{\mathbb{P}(\text{évèn. de } i \text{ en } t_j)}{\mathbb{P}(\text{évèn. en } t_j)}.$$

Dans le cas du modèle de Cox, la contribution de l'individu i à la vraisemblance est donc donnée par :

$$L_i(\beta) = \frac{\exp(\beta^T Z_i)}{\sum_{j \in \mathcal{R}_{t_i}} \exp(\beta^T Z_j)},$$

Avec $\mathcal{R}_{t_i}$ l'ensemble des assurés sous risque en t_i .

La vraisemblance partielle totale de Cox peut donc être obtenue :

$$L(\beta) = \prod_{i=1}^D L_i(\beta) = \prod_{i=1}^D \frac{\exp(\beta^T Z_i)}{\sum_{j \in \mathcal{R}_{t_i}} \exp(\beta^T Z_j)} = \prod_{i=1}^n \left[\frac{\exp(\beta^T Z_i)}{\sum_{j \in \mathcal{R}_{t_i}} \exp(\beta^T Z_j)} \right]^{d_i}.$$

Avec D le nombre d'entrées en incapacité et d_i la variable indicatrice d'entrée en incapacité : dans le contexte continu, et donc dans l'hypothèse d'absence de concomitance des événements ou des censures, il est possible d'ordonner les durées observées de sorte qu'à chaque date correspond une censure ou un événement d'intérêt. Si bien que $d_i = 1$ en cas d'entrée en incapacité et 0 en cas de censure. La fonction de log-vraisemblance qui en découle est alors la suivante :

$$\text{Log}(L(\beta)) = \sum_{i=1}^n d_i (\beta^T Z_i - \ln \sum \exp(\beta^T Z_j)).$$

Le point maximisant la fonction de log-vraisemblance est également celui qui annule $\frac{\partial \text{Log}(L(\beta))}{\partial \beta}$, dérivée partielle par rapport à β . Cette équation est couramment résolue à l'aide de l'algorithme de Newton-Raphson, qui, par itération, permet de déterminer le point d'intersection entre une fonction et la droite des abscisses.

Enfin, le temps étant généralement discrétisé, il est courant d'observer des ex æquo sur les dates d'événements. Il est dès lors nécessaire d'adapter l'estimation de la vraisemblance. En pratique, le logiciel R prévoit d'utiliser par défaut la méthode dite d'Efron qui calcule un risque moyen pour tous les événements survenus simultanément.

3.3.2 Mise en œuvre de régression de Cox sur un portefeuille emprunteur

Le nombre de variables disponibles pour introduire une segmentation est important. Or, il existe un risque de surapprentissage dans un modèle trop segmenté. Ce risque est d'autant plus élevé que la quantité de données disponibles est faible. Au regard de la quantité de données disponibles pour cette étude, le choix est donc fait de sélectionner un nombre restreint de variables explicatives. Il s'agit dans un premier temps d'identifier les variables significatives.

3.3.2.1 Portefeuille emprunteur

3.3.2.1.1 Définition des variables

Dans un premier temps, 14 variables sont introduites dans le modèle. Ces dernières sont présentées dans le **Tableau 9** ci-après.

Code variable	Type de variable	Définition
sexe_h	Binaire	Sexe de l'assuré(e). Codée 1 pour les hommes.
tabac	Binaire	Consommation de tabac. Codée 1 pour oui.
CSP	Ordinale	Profil de risque de l'activité professionnelle. De CSP1 à CSP4.
franchise	Ordinale	Franchise minimum souscrite. Codée 0 pour 90 jours.
anc	Ordinale	Ancienneté du contrat
geo	Nominale	Départements de résidence, regroupés par quartiles de risque
quotite	Binaire	Quotité moyenne. Codée 0 pour 100%, 1 sinon.
majo_tarif	Binaire	Majoration moyenne du tarif. Codée 0 pour 0%, 1 sinon.
option_confort	Binaire	Taux de souscription à l'option confort (dos). Codée 1 pour oui.
option_confort_plus	Binaire	Taux de souscription à l'option confort+ (dos + psy). Codée 1 pour oui.
option_IPP	Binaire	Taux de souscription à l'option IPP. Codée 1 pour oui.
inval_med	Binaire	Invalidité spéciale professions de santé. Codée 1 pour oui.
rabais_couple	Binaire	Rabais offert en cas de souscription en couple. Codée 1 pour oui.
capital	Binaire	Capital cumulé assuré à la souscription. Codée 1 si > médiane (220K€)

Tableau 9 : Définition des variables introduites dans le modèle emprunteurs

La plupart des variables sont binaires, c'est-à-dire qu'elles sont codées 1 ou 0. Dans ce cas, la modalité 0 sera considérée comme le risque de base du modèle. Par exemple sur la variable « tabac », la modalité 0 correspond à l'absence de consommation de tabac. Le risque de base du modèle de Cox sera donc calibré sur un individu ne consommant pas de tabac. Et le coefficient attribué à cette variable correspondra au surrisque

ou sous-risque correspondant à la consommation de tabac. Trois variables sont catégorielles. Le risque de base est alors calculé sur la modalité codée numériquement dont la valeur est la plus faible. Puis un coefficient est calculé pour chacune des modalités suivantes.

Par ailleurs, **le modèle de Cox n'est pas en mesure de gérer des variables dont la valeur évolue avec le temps** : la valeur de la variable est fixée à la date de la sortie. Afin de s'affranchir de cette limite, il est néanmoins possible de modifier la structure de la base de données. Chacun individu est dupliqué et censuré sur chaque intervalle d'ancienneté retenu. Ainsi, chaque individu est dorénavant observé et intégré à la régression sur chaque pas d'ancienneté. Cette démarche est sans impact sur la loi de survie puisqu'en tout instant la population sous risque est identique au modèle d'origine. La puissance de calcul étant limitée, le pas d'ancienneté doit être adapté. En l'espèce, un pas mensuel générerait des centaines de millions de lignes, impossibles à traiter avec des outils conventionnels. En pratique, il est apparu une augmentation du risque dès la deuxième année pour atteindre un maximum entre l'année 3 et l'année 4. Deux modalités ont donc été retenues pour cette variable :

0	=< 12 ou > 48 mois
1	>12 et =< 48 mois

Enfin il convient à ce stade de noter que l'importance de ces variables ne doit pas être résumée à la valeur des coefficients de Cox mais également à la distribution de ces dernières dans le portefeuille. A titre d'exemple, la franchise 30 jours n'est souscrite que par environ 600 individus, ce qui limite donc l'intérêt d'une telle segmentation. Néanmoins, l'objectif étant ici de s'intéresser à l'hétérogénéité du risque, cette dernière est conservée pour la suite de l'étude. Les effectifs par modalités sont portés dans le **Tableau 10** ci-après. La variable ancienneté n'est pas indiquée puisque les contrats peuvent porter sur plusieurs anciennetés.

	0	1
sexe_h	47%	53%
tabac	95%	5%
quotite	61%	39%
majo_tarif	94%	6%
option_confort	89%	11%
option_confort_plus	60%	40%
option_IPP	94%	6%
inval_med	99%	1%
rabais_couple	20%	80%
capital	50%	50%

	Groupe 0	Groupe 1	Groupe 2	Groupe 3
geo	47%	25%	16%	12%

	CSP1	CSP2	CSP3	CSP4
CSP	60%	30%	9%	0%

	90	180	60	30
franchise	98%	1%	0%	0%

Tableau 10 : Distribution des modalités par variable

3.3.2.1.2 Résultats de la régression de Cox

Un premier modèle M_0 est calibré. Les résultats sont donnés dans le **Tableau 11** ci-après. Le seuil α retenu pour la significativité des coefficients est de 5%.

	coef	exp(coef)	se(coef)	z	Pr(> z)	
sexe_h	-0.51654	0.59658	0.03838	-13.458	< 2e-16	***
tabac	0.24123	1.27281	0.07818	3.085	0.002033	**
CSP2	0.56848	1.76558	0.04285	13.267	< 2e-16	***
CSP3	0.74757	2.11185	0.05868	12.741	< 2e-16	***
CSP4	1.55553	4.73760	0.17238	9.024	< 2e-16	***
quotite	-0.32228	0.72450	0.04490	-7.178	7.06e-13	***
majo_tarif	0.49963	1.64811	0.06484	7.705	1.31e-14	***
option_confort	0.36841	1.44543	0.06127	6.013	1.82e-09	***
option_confort_plus	0.51829	1.67916	0.04220	12.282	< 2e-16	***
inval_med	0.46771	1.59634	0.12291	3.805	0.000142	***
capital	-0.07402	0.92866	0.04112	-1.800	0.071871	.
anciennete_13_48	0.28974	1.33608	0.03723	7.782	7.14e-15	***
franchise_30	1.53544	4.64339	0.12180	12.607	< 2e-16	***
franchise_60	0.83825	2.31232	0.26841	3.123	0.001790	**
franchise_180	-1.05476	0.34827	0.27817	-3.792	0.000150	***
rabais_couple	-0.01647	0.98366	0.04820	-0.342	0.732502	
option_IPP	-0.02125	0.97898	0.04433	-0.479	0.631710	
geo_groupe_1	0.60868	1.83800	0.05338	11.403	< 2e-16	***
geo_groupe_2	0.80772	2.24278	0.05605	14.410	< 2e-16	***
geo_groupe_3	1.13315	3.10544	0.05594	20.258	< 2e-16	***

Tableau 11 : Sortie R de la régression de Cox sur le modèle M_0

Il ressort de ces premiers éléments que 3 coefficients ne sont pas significatifs au seuil de 5% : le montant du capital assuré à la souscription (capital), la présence d'une réduction en cas de souscription en couple (rabais_couple) et la souscription à l'option IPP (option_IPP). Le montant du capital assuré à la souscription est régulièrement identifié dans la littérature comme une variable exogène. De ce point de vue, il est donc surprenant qu'elle ne soit pas significative dans le cas de ce portefeuille. L'effet de cette variable devient cependant significatif si la zone géographique est retirée de la régression, ce qui pourrait suggérer une certaine proximité entre les deux.

3.3.2.1.3 Hypothèse de non-redondance

Plus généralement la non-significativité de certaines variables peut s'expliquer par des redondances entre ces dernières. Autrement dit, une variable peut apparaître non significative car « gênée » par une autre variable qui peut elle-même être non-significative. Il est donc nécessaire de valider l'hypothèse de non-redondance des variables, c'est à dire de rechercher des colinéarités entre elles. Si les variables sont très corrélées, cela se traduit par une variance élevée dans l'estimation des paramètres. Les corrélations empiriques entre chaque couple de variables explicatives sont calculées. A cette fin, il est possible d'effectuer des régressions simples entre chaque variable et d'exploiter les R^2 afin de calculer les facteurs d'inflation de variance (VIF) :

$$V_j = \frac{1}{(1 - R_j^2)}$$

Où R_j^2 est le coefficient de détermination de la régression de la variable X_j sur les autres variables. Par définition, plus R_j^2 est proche de 1, plus X_j est linéairement proche des autres variables. Et donc plus V_j est grand. En pratique, le VIF est calculé pour chaque variable explicative. Si plusieurs variables ont un VIF supérieur à 5, la variable dont le VIF est le plus élevé est supprimée. Puis la procédure est renouvelée jusqu'à ce que toutes les variables aient un VIF inférieur à 5. Les résultats sur le jeu de données sont donnés ci-après dans le **Tableau 12** :

	VIF
sexe_h	1,0306
tabac	1,0139
CSP2	1,2947
CSP3	1,2417
CSP4	1,0268
quotite	1,1561
majo_tarif	1,0114
option_confort	1,1850
option_confort_plus	1,2749
inval_med	1,0978
capital	1,1473
anciennete_13_48	1,0053
franchise_30	1,0770
franchise_60	1,0035
franchise_180	1,0007
rabais_couple	1,1082
option_IPP	1,1034
geo_groupe_1	1,6642
geo_groupe_2	1,6327
geo_groupe_3	1,7626

Tableau 12 : Calcul des facteurs d'inflation de la variance sur les covariables du modèle

Il apparait que les facteurs d'inflation de variance sont inférieurs à 5 pour l'ensemble des variables. **L'hypothèse de non-redondance des variables peut donc être validée et les trois variables non-significatives sont retirées une à une.** La régression est alors recalculée (modèle M_1). Les résultats de cette dernière sont présentés dans la sortie R ci-après.

	coef	exp(coef)	se(coef)	z	Pr(> z)	
sexe_h	-0.51919	0.59500	0.03832	-13.551	< 2e-16	***
tabac	0.24549	1.27825	0.07813	3.142	0.001677	**
CSP2	0.57983	1.78573	0.04236	13.688	< 2e-16	***
CSP3	0.76159	2.14167	0.05803	13.123	< 2e-16	***
CSP4	1.57150	4.81388	0.17217	9.128	< 2e-16	***
quotite	-0.34028	0.71157	0.04262	-7.983	1.43e-15	***
majo_tarif	0.49266	1.63667	0.06465	7.621	2.52e-14	***
option_confort	0.36171	1.43578	0.06083	5.946	2.75e-09	***
option_confort_plus	0.50958	1.66459	0.04066	12.533	< 2e-16	***
inval_med	0.44537	1.56107	0.12200	3.651	0.000262	***
anciennete_13_48	0.28906	1.33518	0.03722	7.766	8.08e-15	***
franchise_30	1.53373	4.63545	0.12148	12.626	< 2e-16	***
franchise_60	0.84108	2.31886	0.26826	3.135	0.001717	**
franchise_180	-1.05578	0.34792	0.27815	-3.796	0.000147	***
geo_groupe_1	0.61984	1.85863	0.05291	11.715	< 2e-16	***
geo_groupe_2	0.82209	2.27526	0.05537	14.846	< 2e-16	***
geo_groupe_3	1.15128	3.16223	0.05494	20.953	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

Tableau 13 : Sortie R de la régression de Cox sur le modèle M_1

3.3.2.1.4 Validation de l'hypothèse des risques proportionnels

Les coefficients présentés ci-avant dans le **Tableau 13** ne diffèrent pas selon l'âge de l'assuré. En effet, **le modèle de Cox repose sur une hypothèse forte : la proportionnalité des risques est constante dans le temps (hypothèse de hasard proportionnel)**. Le rapport entre les risques de deux individus à l'âge α est supposé être le même à l'âge β (cf. 3.3.1 page 58). Il est donc nécessaire de s'assurer que cette contrainte est satisfaite. Deux méthodes peuvent être utilisées pour diagnostiquer l'absence de proportionnalité constante des

risques dans le temps. La première méthode est l’affichage des fonctions de survie. Sur une échelle log-log complémentaire, les deux fonctions de survie testées doivent apparaître approximativement translatées. La seconde méthode est le recours aux résidus de Schönefeld standardisés.

Concrètement, il s’agit de calculer, pour chaque événement d’intérêt k , la différence entre la valeur de la variable j de l’individu concerné et la moyenne pondérée de cette variable pour l’ensemble des individus à risque. **Si l’hypothèse de hasard proportionnel est vérifiée, les résidus S_{kj} doivent être distribués de façon homogène au cours du temps.** Un test de corrélation entre les résidus et le temps peut être mené en utilisant le Chi-deux. Si la statistique de test est inférieure à 5%, l’hypothèse de hasard proportionnels est rejetée (au seuil de 5%). Les résultats du test, indiqués dans le **Tableau 14** ci-après, montrent que l’hypothèse de hasard proportionnel n’est pas vérifiée pour les 5 variables indiquées par une flèche bleue.

	chisq	df	p	
sexe_h	59.3248	1	1.3e-14	←
tabac	2.6239	1	0.10526	
CSP2	7.0970	1	0.00772	←
CSP3	0.1688	1	0.68119	
CSP4	11.9878	1	0.00054	←
quotite	2.1742	1	0.14034	
majo_tarif	0.1812	1	0.67037	
option_confort	2.9275	1	0.08708	
option_confort_plus	0.0092	1	0.92360	
inval_med	0.0446	1	0.83279	
anciennete_13_48	2.9447	1	0.08616	
franchise_30	4.6173	1	0.03165	←
franchise_60	0.9439	1	0.33128	
franchise_180	0.0224	1	0.88092	
geo_groupe_1	3.8435	1	0.04994	
geo_groupe_2	1.4187	1	0.23361	
geo_groupe_3	20.5090	1	5.9e-06	←
GLOBAL	110.1450	17	1.1e-15	

Tableau 14 : Résultat du test de corrélation des résidus de Schönefeld standardisés pour l’ensemble des variables

Les CSP4 et la franchise 30 jours sont très minoritaires dans le portefeuille. Pour ces deux variables il est donc fait le choix de ne pas tenir compte du résultat du test, ce qui constitue une approximation mais qui permettra de simplifier le modèle.

En complément du test de corrélation, une analyse graphique doit être menée. Cette analyse repose sur les travaux de Therneau et Grambsch (1994) qui ont montré le lien suivant :

$$E[S_{kj}] + \hat{\beta}_j \approx \beta_j(t_k), \text{ où}$$

- S_{kj} est le résidu standardisé pour la variable j à la $k^{\text{ème}}$ date d’évènement,
- $\hat{\beta}_j$ est l’estimation obtenue à partir du modèle de Cox (indépendant du temps),
- $\beta_j(t_k)$ est un coefficient dépendant du temps pour la variable j au moment à la $k^{\text{ème}}$ date d’évènement

Il est dès lors possible d’utiliser la représentation graphique de $S_{kj} + \hat{\beta}_j$ en fonction du temps (via une fonction lissée et ses intervalles de confiance) pour **révéler la forme fonctionnelle de $\beta(t)$ pour chacune des variables. En pratique, s’il est possible de tracer une droite horizontale dans l’intervalle de confiance, l’hypothèse de hasard proportionnel est confortée pour la variable concernée.** La représentation des résidus de Schönefeld standardisés pour la variable correspondant à la souscription des options Dos et Psy (option_confort_plus) est donnée ci-après à titre d’illustration (**Figure 23**). Elle confirme les résultats du test de corrélation qui avaient validé l’hypothèse de hasard proportionnel pour cette variable.

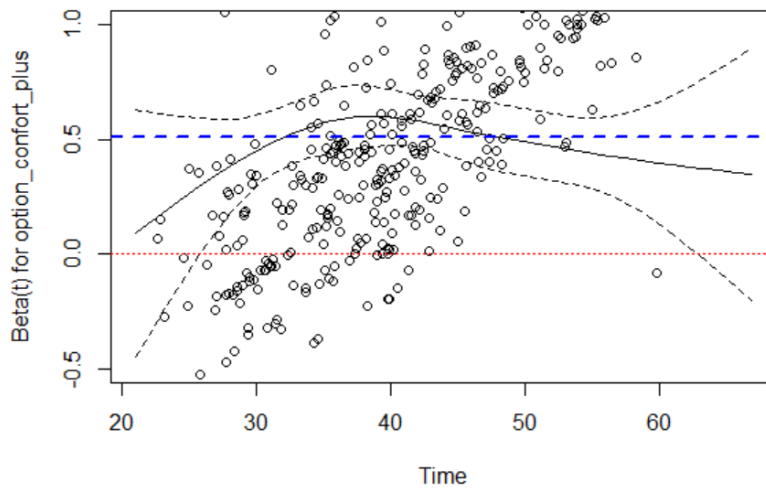


Figure 23 : Représentation des résidus de Schönefeld standardisés pour la variable *option_confort_plus* (*option dos et psy*)

Afin de faciliter la lecture, le coefficient $\hat{\beta}_j$ issu du modèle de Cox est représenté par une ligne horizontale (**bleu**). Une ligne **rouge**, correspondant à l'absence d'effet de la variable sur le taux d'incidence est également tracée. Les points (cercles) correspondent aux $S_{kj} + \hat{\beta}_j$.

Dans le cas de deux variables (quotité et CSP2), l'analyse graphique s'est révélée être en contradiction avec les résultats du test de corrélation. Concernant la variable quotité, l'analyse graphique révèle que l'hypothèse de hasard proportionnel ne devrait probablement pas être retenue contrairement à ce qu'indiquait le test de corrélation. Concernant la variable CSP2, il est possible de tracer une droite horizontale dans l'intervalle de confiance de la fonction lissée de $S_{kj} + \hat{\beta}_j$, tandis que le test de corrélation indique que l'hypothèse de hasard proportionnel n'est pas vérifiée. Par prudence, la proportionnalité du risque est donc considérée comme non-homogène dans le temps pour ces deux variables. Ces exemples soulignent la **complémentarité du test de Chi-deux et de l'analyse graphique**.

Les variables dont l'effet dépend de l'âge (i.e. celles dont l'hypothèse de hasard proportionnel n'est pas vérifiée) ne peuvent pas être évaluées en l'état par la régression de Cox au risque d'obtenir un coefficient moyen qui ne renseignera pas sur le risque réel sur tous les âges. Afin de gérer la dépendance au temps de l'effet de ces variables, il convient de partitionner le modèle sur des tranches d'âges à l'image de ce qui a été fait pour l'ancienneté des contrats. Il est ainsi possible de **créer des fonctions en escalier afin de disposer d'un coefficient sur chaque intervalle de temps sélectionné (hasard proportionnel par morceaux)**. Le nombre d'individus sous risque reste inchangé dans temps. Pour cette raison, cette opération est neutre pour les autres variables. Il aurait été possible de partitionner le modèle sur chaque âge. Mais plus les intervalles sont courts, plus le risque de surajustement est élevé. Les représentations graphiques de ces partitionnements sont présentées dans la **Figure 24** ci-après. Les abscisses correspondent au temps (âge), les ordonnées à la valeur du coefficient. Le tracé bleu correspond aux coefficients du modèle linéaire, tandis que le tracé rouge correspond aux coefficients du modèle partitionné.

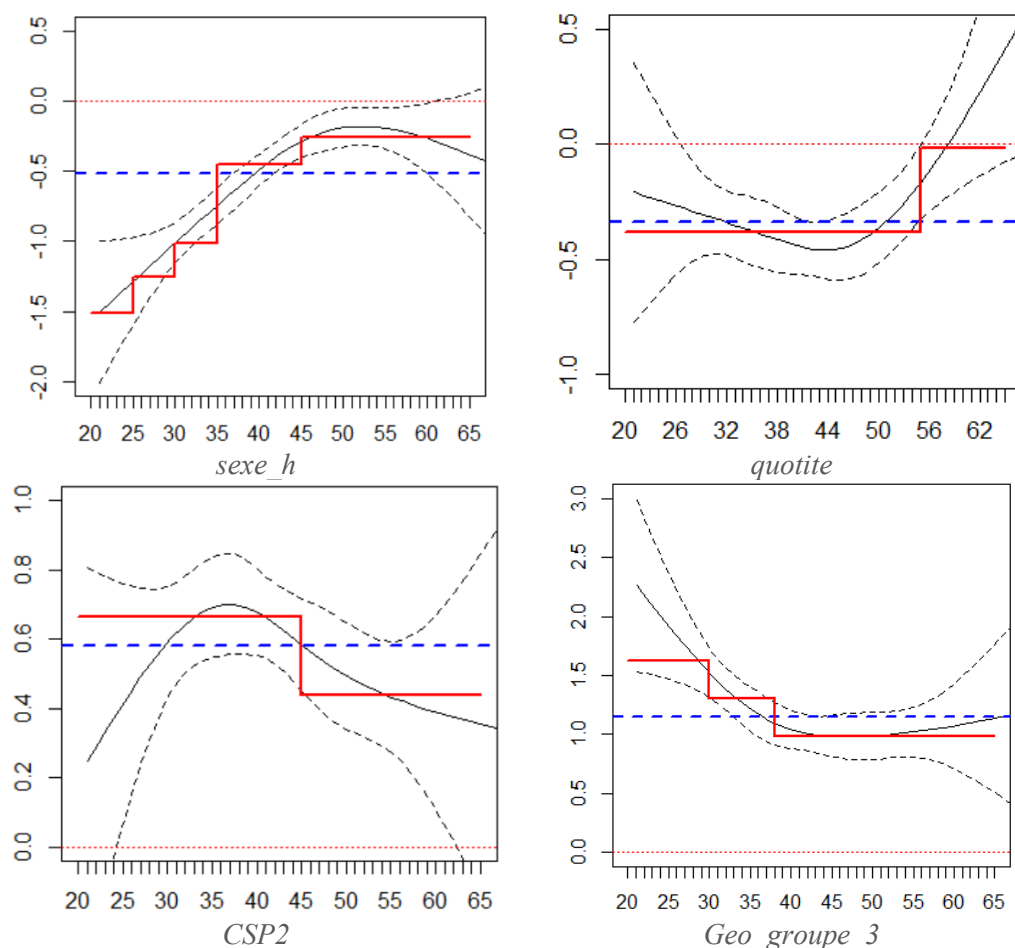


Figure 24 : tracés des coefficients du modèle de Cox

3.3.2.1.5 Interprétation des résultats

Après s'être assuré de la significativité des coefficients et du respect (ou du contournement) de l'hypothèse de hasard proportionnel, les coefficients définitifs sont présentés séparément ci-après. La sortie R est présentée en **Annexe 1**.

3.3.2.1.5.1 Impact du sexe

Sur ce portefeuille emprunteur, **les hommes ont un risque d'entrée en incapacité inférieur à celui des femmes (sexe_h)**. La lecture du tableau s'effectue de la façon suivante : pour un homme âgé de 36 à 45 ans, la probabilité d'entrer en incapacité est inférieure d'un tiers à celle d'une femme du même âge (64%). Comme l'illustre le **Tableau 15**, l'effet de cette variable diminue avec l'âge pour se stabiliser après 45 ans (le coefficient se rapproche de 100%, se stabilisant à 78%). Ce résultat, cohérent avec l'analyse graphique effectuée sur les taux bruts et les taux lissés, suggère un fort impact des grossesses pathologiques sur les taux d'incidence féminins avant 45 ans. Ce phénomène n'est pas négligeable puisqu'environ 13% des états d'incapacité observés s'expliquent par des grossesses pathologiques.

Variable	Tranche d'âge	exp(coef)	Pr(> z)	
Sexe				
Femme (base)	[18 ; 65]	100%	na	na
Homme	[18 ; 25]	23%	1,714%	*
Homme	[25 ; 30]	29%	0,000%	***
Homme	[30 ; 35]	36%	0,000%	***
Homme	[35 ; 45]	64%	0,000%	***
Homme	[45 ; 65]	78%	0,005%	***

Tableau 15 : Coefficients variable sexe_h

3.3.2.1.5.2 Impact du niveau de franchise

Le risque d'incapacité diminue avec l'augmentation de la période de franchise. A titre d'illustration

Variable	Tranche d'âge	exp(coef)	Pr(> z)	
Franchise				
90 jours (base)	[18 ; 65]	100%	na	na
30 jours	[18 ; 65]	472%	0,000%	***
60 jours	[18 ; 65]	233%	0,160%	**
180 jours	[18 ; 65]	35%	0,015%	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tableau 16 : Coefficients variables franchises

(Tableau 16), le passage d'une franchise de 30 jours à une franchise de 90 jours permet ainsi de diviser le risque par 5. Le phénomène de troncature, déjà développé précédemment peut expliquer ce constat : seules les incapacités dont la durée excède la franchise sont connues de l'assureur. Aussi, plus la période de franchise est longue, moins le nombre de

sinistres déclarés à l'assureur est important. Ce phénomène peut être mesuré par les taux de maintien en incapacité issus de la table du BCAC (2010). D'après cette dernière, les nombres d'individus toujours présents en incapacité aux seuils des différentes franchises, et rapportés à la franchise de référence (90 jours) sont les suivants à 45 ans :

Franchises	Coefficients BCAC	Coefficients Cox
30	200%	464%
60	129%	231%
90	100%	100%
180	62%	35%

Tableau 17 : Maintien du BCAC comparé aux coefficients de Cox des niveaux de franchise

La lecture du tableau est la suivante : d'après la table du BCAC, parmi les individus toujours en incapacité au 30^{ème} jour, seule la moitié est toujours en incapacité au 90^{ème} jour (100% / 200%). C'est-à-dire qu'un assureur aurait théoriquement observé un taux d'entrée en incapacité deux fois plus faible sur la franchise de 90 jours que sur la franchise de 30 jours. L'effet de la franchise sur les taux d'incidence, mesuré par la régression de Cox, est donc nettement supérieur au seul biais de troncature extrapolé des coefficients du BCAC : sur 464 individus toujours en incapacité au 30^{ème} jour, seul 22% sont toujours en incapacité au 90^{ème} jour (100% / 464%).

Deux causes peuvent expliquer cet écart :

- Une différence de comportement (antisélection et risque moral) entre les assurés ayant souscrit une franchise réduite (30 jours ou 60 jours) et les assurés ayant souscrit la franchise de référence (90 jours). Dans un document de travail, l'IRDES rappelle que « *Les résultats empiriques montrent que le niveau d'indemnisation est positivement lié au taux d'absence (Frick et Malo, 2005 ; Bonato et Lusinyan, 2004 ; Chaupain-Guillot et Guillot, 2009), au nombre de jours d'absence pour maladie (Osterkamp et Röhn, 2007 ; Malo, 2005) ou à la durée d'arrêt maladie (Johansson et Palme, 2005, Ben Halima et Regaert, 2012, Galizzi et Boden, 2003, Spierdijk et al., 2009)* » (IRDES, 2015).
- Les durées d'incapacité pourraient être structurellement plus faibles sur le portefeuille étudié que dans les données utilisées pour la construction de la table du BCAC. Ce cas de figure pourrait s'expliquer en partie par les effets de la sélection médicale et par la population du portefeuille, plutôt composée de cadres et pourra être validé par l'étude du maintien.

3.3.2.1.5.3 Impact de la Catégorie Socio-Professionnelle

Il apparaît que **la Catégorie Socio-Professionnelle (CSP) est un facteur déterminant dans le risque de survenance d'un épisode d'incapacité**. Ce résultat était attendu puisque les CSP correspondent aux classes de risque identifiées dans le tarif. Ces segmentations correspondent généralement aux types de tâches effectuées dans l'activité professionnelle au moment de la souscription et sont, entre autres, évaluées sur les critères de la sédentarité et de la dangerosité de l'environnement de travail. La CSP1 est la moins risquée et la CSP4 la plus risquée. Ce portefeuille est majoritairement constitué de CSP1 et de CSP2. Ces derniers ont un risque accru de +91% jusqu'à 45 ans, puis de +60% par rapport aux CSP1 (**Tableau 18**). La baisse du risque relatif des CSP2 après 45 ans traduit vraisemblablement des changements dans les carrières professionnelles liés à l'âge et à l'expérience (avancements, reconversions etc.).

Variable	Tranche d'âge	exp(coef)	Pr(> z)	
CSP				
<i>CSP1 (base)</i>	<i>[18 ; 65]</i>	<i>100%</i>	<i>na</i>	<i>na</i>
CSP2	[18 ; 45]	191%	0,000%	***
CSP2	]45 ; 65]	160%	0,000%	***
CSP3	[18 ; 65]	223%	0,000%	***
CSP4	[18 ; 65]	500%	0,000%	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tableau 18 : Coefficients variables CSP

3.3.2.1.5.4 Impact du département de résidence

Le département de résidence est également un facteur significatif du risque d'incapacité. Les départements ont été découpés en 4 zones qui correspondent aux différents quartiles de risque (la liste des départements n'est pas communiquée pour des raisons de confidentialité). Comme l'illustre le **Tableau 19** ci-contre, les assurés du 4^{ème} quartile (groupe 3) ont par exemple 5 fois plus de risque de survenance d'un épisode d'incapacité entre 18 et 30 ans que les assurés du groupe de référence correspondant au 1^{er} quartile (groupe 0). Dans un dossier publié en septembre 2022, la DRESS établit également un lien entre le département de résidence et l'accès aux soins : « *dans les zones de faible densité médicale, le risque de renoncement aux soins est accru pour les personnes les plus défavorisées* » (Direction de la recherche, des études, de l'évaluation et des statistiques (DREES), 2022).

Variable	Tranche d'âge	exp(coef)	Pr(> z)	
Département				
Groupe 0 (base)	[18 ; 65]	100%	na	na
Groupe 1	[18 ; 65]	188%	0,000%	***
Groupe 2	[18 ; 65]	231%	0,000%	***
Groupe 3	[18 ; 30]	503%	0,000%	***
Groupe 3	]30 ; 38]	362%	0,000%	***
Groupe 3	]38 ; 65]	267%	0,000%	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tableau 19 : Coefficients variable départements

3.3.2.1.5.5 Impact des options Confort (dos) et Confort+ (dos et psy)

Les options Confort et Confort+ étendent le champ des causes acceptées pour l'activation de la garantie incapacité. Il est donc cohérent que **leur souscription conduise à une majoration du risque d'incapacité (Tableau 20)**. Il est vraisemblable que ces options se caractérisent par une forte antisélection.

Variable	Tranche d'âge	exp(coef)	Pr(> z)	
Options				
Aucune (base)	[18 ; 65]	100%	na	na
Confort (Dos)	[18 ; 65]	142%	0,000%	***
Confort+ (Dos+Psy)	[18 ; 65]	165%	0,000%	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tableau 20 : Coefficient souscription aux options

3.3.2.1.5.6 Lien entre les majorations de tarifs liées à l'analyse du dossier médical et le risque

Dans un certain nombre de dossiers, la sélection médicale peut conduire à l'ajout d'une surprime individuelle au tarif. La présence d'une surprime traduit donc un état de santé jugé davantage risqué. Comme l'illustre le **Tableau 21** ci-après, **la sélection médicale semble efficace** puisque les individus concernés par une surprime ont bien un risque accru (+61%). Il pourrait être intéressant de s'intéresser, dans de futurs travaux, au lien entre le taux de surprime et l'accroissement du risque d'entrée en incapacité (même si la surprime intègre également la probabilité de maintien).

Variable	Tranche d'âge	exp(coef)	Pr(> z)	
Majoration tarif - Surprime				
Aucune (base)	[18 ; 65]	100%	na	na
Surprime	[18 ; 65]	161%	0,000%	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tableau 21 : Coefficient majoration tarif - Surprime

3.3.2.1.5.7 Impact de l'exercice d'une profession médicale révélée par la souscription à l'invalidité spéciale profession médicale

La souscription à l'option invalidité spéciale traduit l'exercice **d'une profession médicale, qui est en effet un facteur aggravant du risque de survenance d'un épisode d'incapacité (Tableau 22)**. Plusieurs hypothèses peuvent être émises. D'une part les blessures sont très incapacitantes pour des professions qui nécessitent une certaine dextérité (chirurgiens, dentistes, kinésithérapeutes etc.). D'autre part l'accès aux certificats d'incapacité est facilité pour ces professions. Enfin la présence de cette option révèle probablement des comportements anti-sélectifs.

Variable	Tranche d'âge	exp(coef)	Pr(> z)	
Majoration tarif - Prof. Médicales				
Aucune (base)	[18 ; 65]	100%	na	na
Professions médicales	[18 ; 65]	156%	0.027%	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tableau 22 : Coefficient majoration tarif - Professions médicales

3.3.2.1.5.8 Impact de la quotité

Plus la quotité est élevée, plus l'assuré est indemnisé en cas d'incapacité. Une corrélation positive ayant été établie entre le niveau d'indemnisation et le risque de survenance d'un épisode d'incapacité (IRDES, 2015), il est cohérent de constater que **les assurés bénéficiant d'une quotité inférieure à 100% ont un risque de survenance minoré (Tableau 23 ci-après)**. Toutefois ce coefficient n'est plus significatif après 55 ans (il a donc été remplacé par 100% dans le tableau ci-dessous). Enfin, il pourrait être utile, dans un second temps de calculer un coefficient relatif au niveau de quotité.

Variable	Tranche d'âge	exp(coef)	Pr(> z)	
Quotite				
100% (base)	[18 ; 65]	100%	na	na
< 100%	[18 ; 55]	69%	0,000%	***
< 100%	[56 ; 65]	100%	na	na

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tableau 23 : Coefficients variable quotité

3.3.2.1.5.9 Impact de la consommation de tabac

La consommation de tabac se traduit par un risque accru de survenance d'un état d'incapacité. Le risque est supérieur d'environ 30% par rapport à un individu non-fumeur (Tableau 24 ci-contre). Ce taux ne doit pas être comparé aux publications médicales : d'une part il est calculé sur un échantillon non-représentatif de la population d'un pays comme la France. D'autre part, **compte tenu des surprimes élevées appliquées aux fumeurs, seule une partie d'entre eux déclare effectivement la consommation de tabac** (seulement 5% de fumeurs déclarés dans le portefeuille).

Variable	Tranche d'âge	exp(coef)	Pr(> z)	
Tabac				
Non-fumeur (base)	[18 ; 65]	100%	na	na
Fumeur	[18 ; 55]	129%	0,105%	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tableau 24 : Coefficient tabac

3.3.2.1.5.10 Impact de l'ancienneté

Les contrats dont l'ancienneté est comprise entre 13 et 48 mois ont un risque accru de survenance d'un épisode d'incapacité par rapport aux autres anciennetés. Le moindre risque avant le 13^{ème} mois s'explique principalement par la période de carence de 3 mois. Pour les contrats les plus anciens, il est probable qu'un certain nombre d'états d'incapacité ne soit pas déclaré par oubli. Un échange avec les équipes en charge de la gestion des contrats permettra d'approfondir cette question.

Variable	Tranche d'âge	exp(coef)	Pr(> z)	
Ancienneté				
=< 12 et > 48 (base)	[18 ; 65]	100%	na	na
> 12 et =< 48	[18 ; 55]	133%	0,000%	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tableau 25 : Impact de l'ancienneté

3.3.3 Portefeuilles de contrats de prévoyance classiques à destination des actifs

Les méthodes utilisées sur les portefeuilles de prévoyance classiques sont les mêmes que celles mentionnées supra sur le portefeuille d'assurance emprunteur. Elles ne sont donc pas développées à nouveau.

Les contrats de prévoyance classiques (salariés et TNS) sont plus hétérogènes que ceux du portefeuille emprunteur. Le comportement des travailleurs salariés et des travailleurs non-salariés diffère. Ces derniers présentent des taux d'incidence globalement plus faibles. Il est donc essentiel d'analyser ces deux populations séparément afin d'éviter certains biais analytiques. A titre d'illustration, la **Figure 25** ci-après, présente les taux d'entrée pour l'ensemble assurés disposant d'une garantie Indemnité Journalière (IJ) avec une franchise de 15 jours. Cette figure suggère que l'incidence est plus faible dans les 36 premiers mois suivant la souscription.

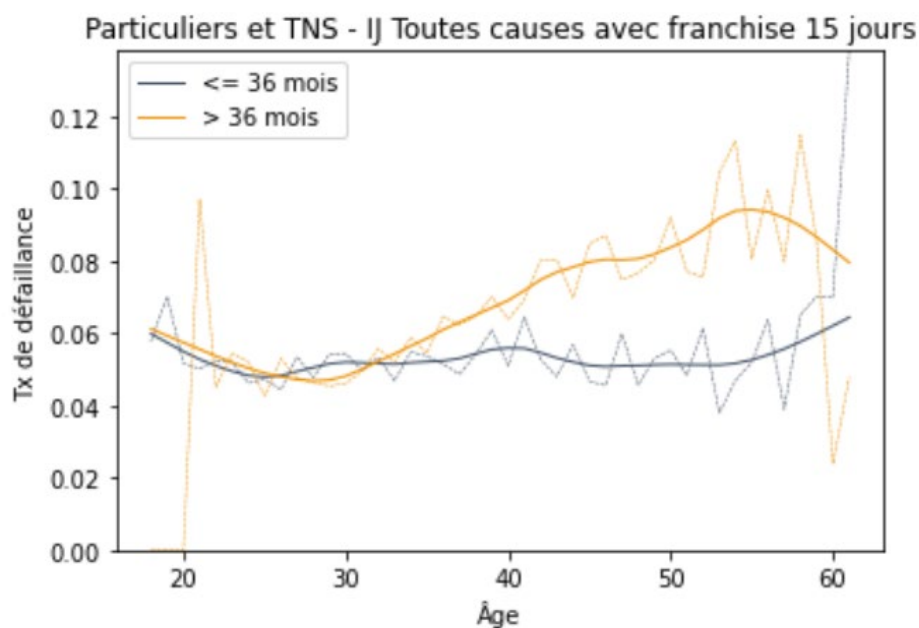


Figure 25 : Taux d'incidence par ancienneté (franchise 15 jours)

Pourtant, l'écart entre les deux courbes traduit également la baisse du poids des TNS dans le portefeuille : ces derniers présentent des taux d'entrée en incapacité plus faibles mais également des taux de chute plus importants. La **Figure 26** ci-après distingue les deux populations et présente une situation plus ambiguë. Pour les TNS, les taux avant et après 36 mois se confondent globalement. S'agissant des particuliers, les taux d'entrée sont effectivement légèrement inférieurs pour les contrats les plus récents, à partir de 50 ans environ.

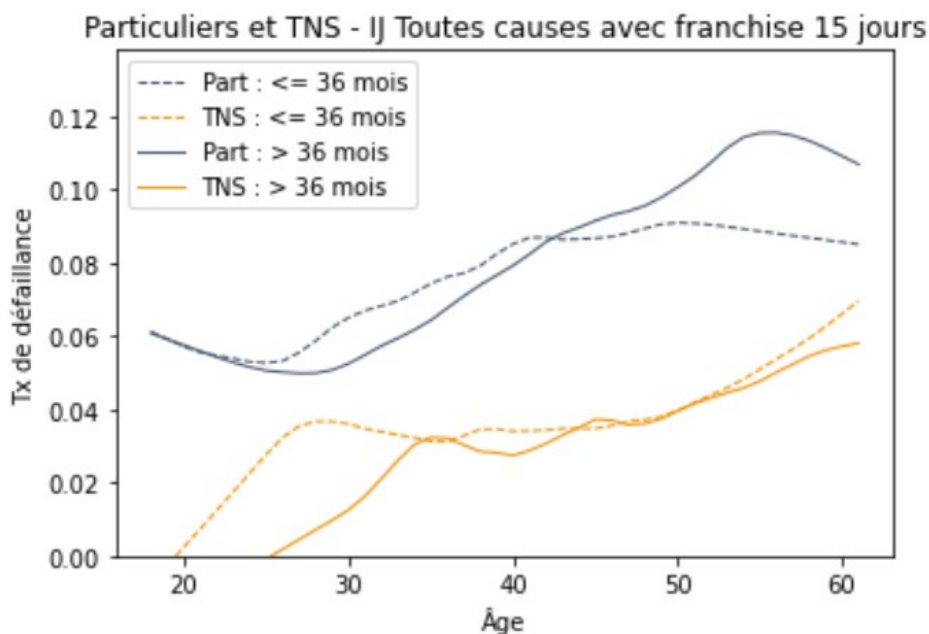


Figure 26 : Taux d'incidence par ancienneté Salariés et TNS (franchise 15 jours)

Par ailleurs, les garanties étudiées sont également hétérogènes. Certains individus disposent de garanties prévoyant le versement d'indemnités journalières (au regard des montants indemnisés, ces garanties se rapprochent du portefeuille emprunteur), tandis que d'autres ne bénéficient que de l'exonération de leurs cotisations d'assurance prévoyance. Enfin, certains contrats à destination des TNS ne prévoient une indemnisation qu'en cas d'accident. Pour ces raisons, les particuliers et les TNS feront l'objet de lois propres. Les garanties prévoyant l'exonération des cotisations feront également l'objet d'une loi distincte. Les interactions entre les deux garanties seront cependant étudiées.

3.3.3.1 Les contrats à destination des salariés

Pour un même niveau de franchise, il existe une nette différence d'entrée en incapacité entre les assurés disposant d'au moins une garantie prévoyant le versement d'indemnités journalières et ceux ne pouvant bénéficier que d'une exonération de leurs cotisations prévoyance. Deux effets se combinent probablement :

- D'une part l'antisélection et le risque moral liés au niveau d'indemnisation plus élevé pour les assurés couverts par des indemnités journalières.
- D'autre part, les garanties exonération sont généralement obligatoires et leur niveau de couverture est faible : quelques euros dans la majorité des cas. Il est donc probable que **certaines victimes, uniquement couvertes par la garantie exonération, ne déclarent pas leur sinistre à l'assureur** (par oubli ou pour éviter les démarches administratives).

Ces différences d'incidence sont illustrées dans la **Figure 27** ci-dessous :

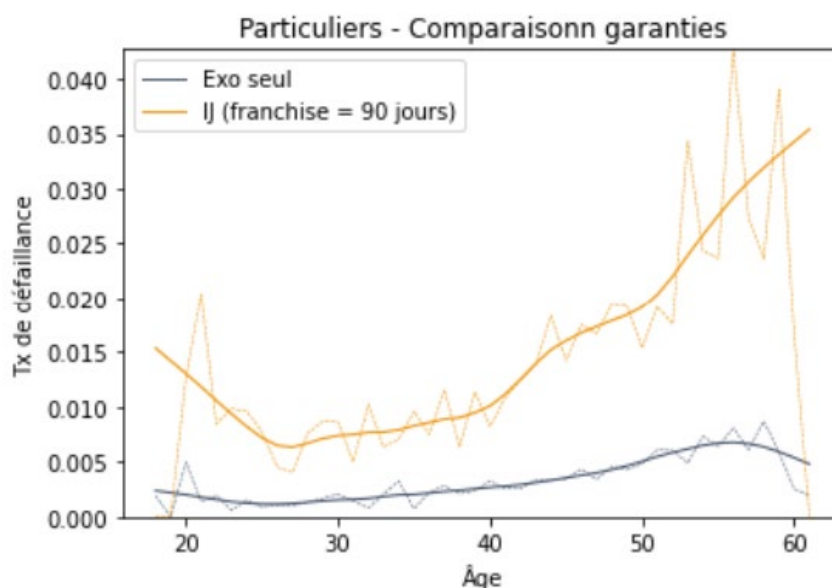


Figure 27 : Taux d'incidence par garantie

Dans ces conditions, il est préférable de traiter séparément les garanties IJ et exonération des cotisations. En particulier, le choix est fait de se concentrer sur les garanties IJ dans cette étude.

La mise en œuvre des garanties IJ toutes causes résulte de facteurs de risques très hétérogènes. En particulier, la **Figure 28** ci-après illustre les différences de taux d'incidence entre les niveaux de franchise. **Comme observé sur le portefeuille emprunteur, les taux d'entrée augmentent lorsque la franchise diminue.** Le rapport entre les risques semble approximativement identique sur tous les âges. La franchise 180 jours est trop peu souscrite pour être affichée correctement dans cette figure mais sera tout de même étudiée par la suite.

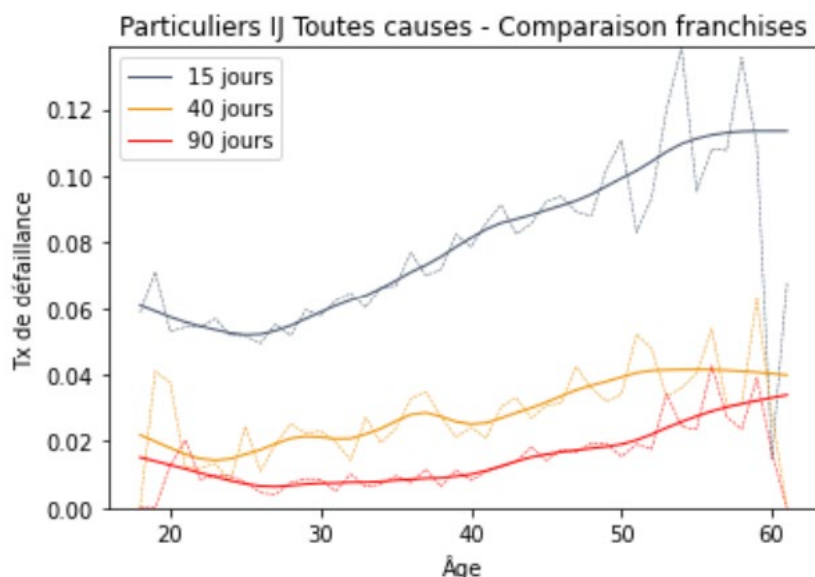


Figure 28 : Taux d'incidence par franchise

En l'absence de couverture des grossesses pathologiques, la forme des courbes de taux de défaillance est quasi identique entre les hommes et les femmes. Les taux des hommes se situent en revanche à un niveau plus élevé avant 55 ans. La **Figure 29** ci-après illustre cette situation pour les individus disposant d'une garantie IJ. L'étude de l'hétérogénéité du portefeuille par la régression de Cox permettra d'apporter des éléments explicatifs.

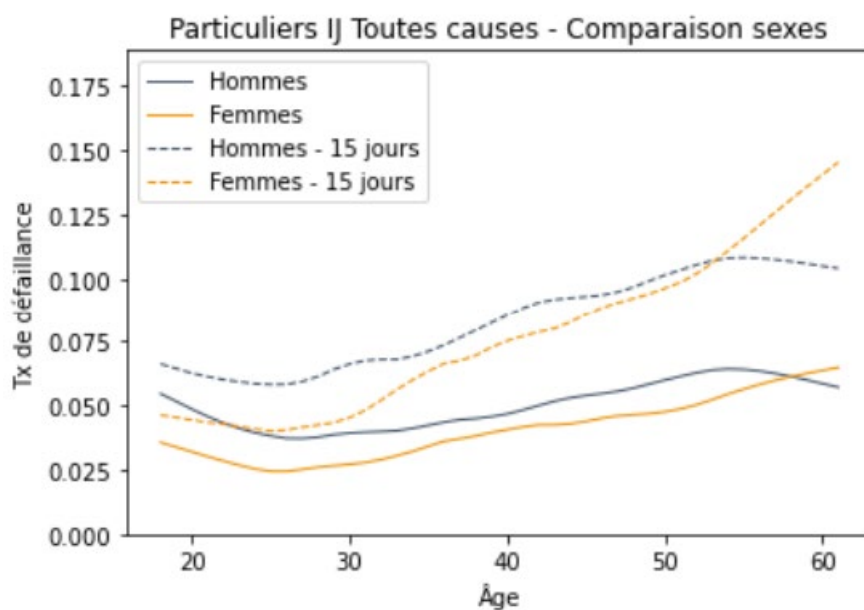


Figure 29 : Taux lissés par sexe

Une régression de Cox est réalisée pour cette population disposant d'une garantie IJ toutes causes afin d'alimenter cette l'analyse. Les variables finalement retenues sont présentées dans le **Tableau 26** ci-après. Dans la mesure où les assurés peuvent avoir souscrit plusieurs produits, ajouté des garanties ou modifié la valeur des IJ, la répartition des modalités est présentée pour les couples « assuré / période de garantie ».

Variable	Description	Modalités			
		]0 ; 10]	]10 ; 15]	]15 ; inf[	
detail_IJ_actuelles	Montant de l'indemnité journalière souscrite	25%	31%	43%	
		15	40	90	180
detail_franchises_IJ	Franchise la plus courte des garanties IJ actives sur la période.	44%	25%	31%	1%
		0	1		
flag_produit_PI	Codée 1 si l'individu est bénéficiaire d'un contrat PrévoirIncap sur la période, 0 sinon	65%	35%		
		0	1		
flag_CSP	Codée 0 pour les employés, commerçants, professions intermédiaires et les non renseignés. Codée 1 sinon (principalement ouvriers).	60%	40%		
		0	1	2	
detail_code_postal	Numéro de département	71%	17%	12%	
		0	1		
surprime	Présence d'une surprime sur les garanties incapacité	88%	12%		
		0	1		
anc	Ancienneté du contrat (>12 & =< 24 mois)	(non concerné)			

Tableau 26 : Présentation des variables pour les contrats à destination des particuliers

Les résultats sont présentés ci-après. Leur établissement a requis les mêmes outils et la même méthodologie que ceux décrits pour le portefeuille emprunteur. Ils ne sont donc pas exposés à nouveau. La sortie R est communiquée en **Annexe 2**. En outre, afin de maximiser la quantité de données disponibles pour l'étude, un temps d'observation de près de 20 ans a été retenu. Il a donc été nécessaire de s'assurer au préalable que, toutes choses étant égales par ailleurs, les probabilités d'entrer en incapacité étaient homogènes sur toute la période d'observation. Cette propriété a été vérifiée en intégrant à la régression les périodes d'expositions, regroupées sur des plages de 5 ans.

3.3.3.1.1 Impact de la variable sexe

L'effet du sexe n'est pas significatif au seuil de 5%. Ce résultat suggère que **la différence de taux observée entre les sexes (Figure 29 page 72) provient d'autres variables** elles-mêmes introduites dans la régression de Cox (par exemple la CSP). Il s'agissait vraisemblablement d'une variable confondante. Les contrats étudiés ne comportent pas de couverture en cas de grossesse pathologique, ce qui explique cette observation (*a contrario*, les contrats du portefeuille emprunteur couvraient les grossesses pathologiques. Passés les âges habituels de grossesse, les taux des deux sexes se confondaient.). Cette variable a donc été retirée du modèle.

3.3.3.1.2 Impact du niveau de franchise

Comme décelé également par l'analyse graphique, **l'effet de la franchise est très significatif (Tableau 27)**. L'écart de risque entre la variable 15 jours (risque base) et la variable 90 jours se réduit très légèrement avec le vieillissement. Ces résultats sont cohérents avec l'analyse effectuée sur le portefeuille emprunteur, dont les causes ont déjà été illustrées (3.3.2.1.5.2 page 66).

Variable	tranche d'âge	exp(coef)	Pr(> z)	
Franchise				
15 jours (base)	[18 ; 65]	100%	na	na
40 jours	[18 ; 45]	41%	0,000%	***
180 jours	[18 ; 65]	26%	0,005%	***
90 jours	[18 ; 45]	15%	0,000%	***
90 jours	[45 ; 51]	21%	0,000%	***
90 jours	[51 ; 55]	25%	0,000%	***
90 jours	[55 ; 60]	33%	0,000%	***

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tableau 27 : Impact du niveau de franchise

3.3.3.1.3 Impact du montant des indemnités journalières

Plus le niveau d'IJ souscrit est important, plus les taux d'incidence sont élevés. Il s'agit probablement d'une manifestation des comportements anti-sélectifs et du risque moral déjà illustrés et en lien avec le niveau d'indemnisation (IRDES, 2015). Comme l'illustre le **Tableau 28**, pour les niveaux élevés (modalité 2 : >15€), cette observation se vérifie principalement pour les âges les plus jeunes et n'est plus significative au-delà de 45 ans. L'effectif sous risque est alors peu conséquent. Plus généralement, en entrant le montant d'IJ souscrit comme une variable continue dans le modèle de Cox, il a été possible de déduire que le risque augmente en moyenne d'environ +1,1% pour chaque euro supplémentaire (non-utilisé *in fine*).

Variable	tranche d'âge	exp(coef)	Pr(> z)	
Montant des IJ (€)				
=< 10€ (base)	[18 ; 65]	100%	na	na
> 10€ et =< 15€	[18 ; 65]	109%	1%	*
> 15€	[18 ; 35]	137%	0,000%	***
> 15€	[35 ; 45]	125%	0,000%	***

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tableau 28 : Impact du montant des IJ

3.3.3.1.4 Impact de la Catégorie Socio-Professionnelle (CSP)

Les ouvriers présentent des taux d'incidence plus élevés que les autres professions représentées dans le portefeuille. D'une part leurs tâches nécessitent une bonne condition physique, probablement incompatible avec un certain nombre de pathologies. D'autre part, l'exercice de ces professions a un impact défavorable sur l'état de santé des ouvriers. La DRESS précise que « les conditions de travail sont [...] liées à la santé, selon des causalités complexes qui jouent dans les deux sens (impact des expositions professionnelles sur la santé, mais aussi de l'état de santé sur les trajectoires professionnelles) ». En particulier, les enquêtes menées sur les conditions de travail révèlent que les ouvriers, qualifiés ou non, sont bien plus exposés aux contraintes physiques dans l'exécution de leur travail que les autres catégories professionnelles (Direction de la recherche, des études, de l'évaluation et des statistiques (DREES), 2022). Le **Tableau 29** illustre ce phénomène : les professions ont été réunies en deux groupes correspondants aux niveaux de risque observés. Le groupe 1 est principalement constitué d'ouvriers.

Variable	tranche d'âge	exp(coef)	Pr(> z)	
CSP				
Groupe 0 (base)	[18 ; 65]	100%	na	na
Groupe 1	[18 ; 65]	118%	0.000%	***

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tableau 29 : Impact de la CSP

3.3.3.1.5 Lien entre les majorations de tarifs liées à l'analyse du dossier médical et le risque

L'utilité de la sélection médicale est démontrée puisque les individus soumis à une surprime présentent un risque significativement supérieur. Comme l'illustre le **Tableau 30** l'impact de cette variable est davantage marqué chez les jeunes.

Variable	tranche d'âge	exp(coef)	Pr(> z)	
Surprime				
Non (base)	[18 ; 65]	100%	na	na
Oui	[18 ; 28]	152%	0,000%	***
Oui	]28 ; 65]	126%	0,000%	***

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tableau 30 : Impact de la sélection médicale

3.3.3.1.6 Impact de la zone géographique

Comme régulièrement évoquée dans la littérature et déjà démontrée sur le portefeuille emprunteur, **la zone géographique de résidence est un facteur de risque d'entrée en incapacité (Tableau 31)**. Deux types de zones ont été retenues dans cette étude : une présentant des taux plus faibles (groupe 2), en vert dans la cartographie ci-après (**Figure 30**) et une autre présentant des taux plus élevés (groupe 1) en rouge. Enfin la zone de référence est présentée en jaune. La taille des ronds correspond au nombre d'événements observés. Il est en particulier remarquable que les départements au sein du groupe 1 (rouge) sont plutôt distribués en dessous d'une ligne Marseille-Brest. Les départements du groupe 2 (vert) sont plutôt distribués au-dessus.

Variable	tranche d'âge	exp(coef)	Pr(> z)	
Département de résidence				
Groupe 0 (base) 	[18 ; 65]	100%	na	na
Groupe 1 	[18 ; 65]	120%	0,000%	***
Groupe 2 	[18 ; 65]	68%	0,000%	***
signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tableau 31 : Impact de la zone géographique

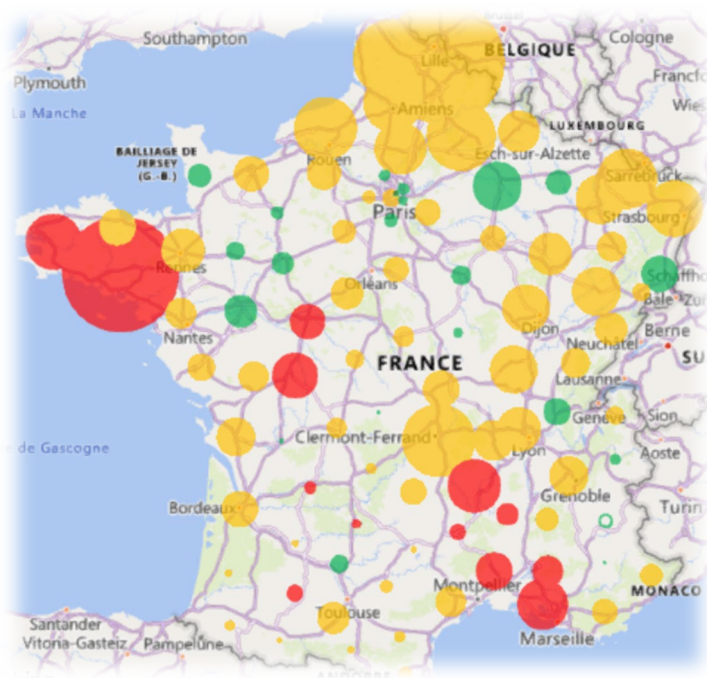


Figure 30 : Carte des groupes de départements

3.3.3.1.7 Impact de l'ancienneté

Une **sinistralité légèrement plus importante est constatée entre le 13^{ème} et le 24^{ème} mois suivant la souscription des contrats (Tableau 32 ci-après)**. La sinistralité de première année est en particulier limitée par la carence. La remontée du risque en deuxième année traduit peut-être la volonté pour les assurés de « rentabiliser » les primes versées. Ce comportement pourrait être comparable à ce qui peut être observé sur les contrats de complémentaire santé individuelle à adhésion facultative sur les premiers mois de souscription. Il est en revanche assez différent de celui observé sur l'assurance emprunteur, pour lequel les motivations d'achat ne sont pas les mêmes. En particulier, l'assurance emprunteur dispose d'un caractère quasi obligatoire dans le sens où il s'agit d'un prérequis pour accéder à l'emprunt.

Variable	tranche d'âge	exp(coef)	Pr(> z)	
Ancienneté				
=< 12 ou > 24 (base)	[18 ; 65]	100%	na	na
> 12 et =< 24	[18 ; 65]	118%	0%	***

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tableau 32 : Impact de l'ancienneté

3.3.3.1.8 Impact de la motivation d'achat

Par ailleurs, les assurés disposant d'un contrat PrévoirIncap ont été identifiés par une variable (flag_produit_PI). En effet les IJ en cas d'incapacité sont optionnelles sur le Prévoir MultiSolution et le Solution Protection Actif. Sur le PrévoirIncap, elles constituent en revanche la garantie principale et donc la principale motivation d'achat. **La majoration du risque (Tableau 33) pour les assurés disposant de ce type de contrat suggère une forte antisélection.**

Variable	tranche d'âge	exp(coef)	Pr(> z)	
Bénéficiaire d'une Prévoir Incap				
Non (base)	[18 ; 65]	100%	na	na
Oui	[18 ; 65]	153%	0.000%	***

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tableau 33 : Majoration du risque des bénéficiaires du PrévoirIncap

3.3.3.2 Les contrats à destination des TNS

Les résultats de la régression sont exposés directement ci-après. La sortie R est présentée en **Annexe 3**. A nouveau, l'homogénéité du risque sur toute la période d'observation a été préalablement vérifiée (toutes choses étant égales par ailleurs).

3.3.3.2.1 Impact du sexe

Contrairement aux salariés, **le sexe des assurés a un impact significatif sur les taux d'incidence des TNS**. Toutefois, cet impact n'est pas homogène selon les produits souscrits (**Tableau 34**). En effet, **le produit le plus récent du portefeuille étudié (Solution Pro Prévoyance (SPP)) prévoit la prise en charge des grossesses pathologiques** (sur les contrats les plus récents, une carence d'un an a finalement été ajoutée). Si bien que parmi les bénéficiaires de ce produit, les femmes ont un risque majoré d'environ 47% avant 45 ans (non significatif après 45 ans). Tandis que sur les autres produits, leur risque est minoré d'environ 27% sur la même tranche d'âge. **Il est probable que la minoration du risque hors grossesses pathologiques provienne de la population d'artisans, majoritairement masculine, et qui partage la même caisse que les commerçants** (donc non capté par la variable caisses professionnelles).

Variable	tranche d'âge	exp(coef)	Pr(> z)	
Sexe				
Homme (base)	[18 ; 65]	100%	na	na
Femme sans SPP	[18 ; 45]	73%	0,038%	***
Femme avec SPP	[18 ; 45]	147%	0,082%	***
Femme sans SPP	]45 ; 65]	81%	4,020%	*
Femme avec SPP	]45 ; 65]	92%	65,009%	

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tableau 34 : Impact du sexe

3.3.3.2.2 Impact du niveau de franchise

L'impact du niveau de franchise sur le risque de survenance d'un épisode d'incapacité est également démontré pour les TNS (Tableau 35). En revanche, les coefficients des franchises de 60 jours et de 90 jours sont apparus non-significatifs, probablement compte tenu du très faible volume de données disponibles sur ces modalités. Il pourra être envisagé d'utiliser les majorations observées sur les salariés.

Variable	tranche d'âge	exp(coef)	Pr(> z)	
Franchise				
Franchise 15 jours (base)	[18 ; 65]	100%	na	na
Franchise 30 jours	[18 ; 65]	63%	0,002%	***
Franchise 365 jours	[18 ; 65]	18%	0,013%	***
signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tableau 35 : Impact du niveau de franchise

3.3.3.2.3 Lien entre les majorations de tarifs liées à l'analyse du dossier médical et le risque

Il n'a pas été possible de démontrer le lien entre les majorations de tarifs liées à l'analyse du dossier médical et le risque d'entrée en incapacité. L'absence d'impact significatif de la variable surprime constitue un point remarquable et doit amener à s'interroger sur l'intégrité des données, la sincérité des réponses au questionnaire de santé ou sur le protocole de sélection. Il convient de préciser qu'un dispositif dérogatoire prévoyant la suppression de la carence et un allègement des formalités médicales est prévu en cas de reprise à la concurrence :

- PrévoirPro Solution (PPS) : pas de dispositif dérogatoire.
- PrévoirPro Actif Prévoyance (PPAP) : suppression du délai de carence en cas de reprise à la concurrence (sous certaines conditions).
- Solution Pro Prévoyance (SPP) : suppression du délai de carence et allègement des formalités médicales (sous certaines conditions).

3.3.3.2.4 Impact du montant des indemnités journalières (IJ)

Comme pour les particuliers, **le montant des IJ souscrites impacte fortement les taux d'entrée en incapacité (Tableau 36).** Chaque euro supplémentaire augmente le risque d'environ +0,6% en moyenne. Cette observation illustre à nouveau les effets d'antisélection et de risque moral.

Variable	tranche d'âge	exp(coef)	Pr(> z)	
Montant des IJ (€)				
<i>IJ ≤ 40 (base)</i>	<i>[18 ; 65]</i>	<i>100%</i>	<i>na</i>	<i>na</i>
IJ > 40 et ≤120	[18 ; 65]	145%	0,000%	***
IJ > 120 et ≤200	[18 ; 40]	310%	0,000%	***
	]40 ; 65]	190%	0,000%	***
IJ > 200	[18 ; 65]	338%	0,000%	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tableau 36 : Impact du montant des IJ

3.3.3.2.5 Impact de la zone géographique

L'impact des départements de résidence est également significatif, en conservant le même zonage que pour les particuliers. Les taux sont par ailleurs globalement comparables comme l'illustre le **Tableau 37** ci-après.

Variable	tranche d'âge	exp(coef)	Pr(> z)	
Département de résidence				
Groupe 0 (base)	[18 ; 65]	100%	na	na
Groupe 1	[18 ; 65]	118%	1,501%	*
Groupe 2	[18 ; 65]	81%	1,304%	*

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tableau 37 : Impact de la zone géographique

3.3.3.2.6 Impact des caisses professionnelles

L'impact des caisses professionnelles est également significatif. Néanmoins, compte tenu du regroupement d'un certain nombre de caisses professionnelles sous le régime du SSI, il devient délicat d'utiliser cette variable pour la segmentation du risque.

Variable	tranche d'âge	exp(coef)	Pr(> z)	
Caisse professionnelles				
Groupe 0 (base)	[18 ; 65]	100%	na	na
Groupe 1	[18 ; 65]	79%	0,222%	**

signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Enfin, la présence d'une réduction de franchise en cas d'accident ou d'hospitalisation a été testée mais son impact n'est pas significatif.

3.3.4 Conclusion sur les lois d'entrée en incapacité

Dans la mesure où les garanties d'indemnisation des périodes d'incapacité relèvent de l'analyse de survie, les lois d'entrée en incapacité ont été produites grâce à l'estimateur de Kaplan-Meier. Afin de les généraliser, ces dernières ont ensuite fait l'objet d'un lissage de Whittaker-Henderson, calibré par validation croisée.

Dans un second temps, l'impact d'un certain nombre de variables a été mesuré par une régression de Cox à hasard proportionnel. L'hypothèse de hasard proportionnel a été vérifiée par un test de corrélation et par une analyse graphique, qui se sont révélés complémentaires.

Pour les variables dont l'effet n'est pas stable sur tous les âges, un partitionnement du modèle a été nécessaire. L'ensemble de ces étapes est récapitulé dans un logigramme communiqué en annexe 4. Les coefficients attribués à ces variables seront à appliquer à la loi de base (non communiquée pour des raisons de confidentialité), c'est à dire calibrée sur les modalités du risque de base de la régression. En fonction des travaux, un nombre restreint de variables pourra être sélectionné, et la loi correspondant au risque de base recalculée. En particulier, un arbre de survie pourra être utile pour affiner le choix des variables.

4 ETUDE DU MAINTIEN

4.1 Définitions et réglementation

La deuxième composante du risque d'incapacité est le maintien. Il s'agit du **temps écoulé entre l'entrée et la fin de la période d'incapacité**. La période d'incapacité peut prendre fin pour diverses raisons :

- la reprise du travail si les capacités sont retrouvées.
- l'entrée en invalidité si les capacités ne sont pas retrouvées par l'individu au bout d'une durée prévue aux termes du contrat. Un passage en invalidité est également possible en amont si les capacités sont considérées comme non-recouvrables.
- la retraite ou le décès de l'individu.

La Sécurité Sociale prévoit l'entrée automatique en invalidité après 36 mois écoulés en incapacité. Cette durée est donc généralement la norme pour les contrats d'assurance prévoyant une garantie invalidité en plus de la garantie incapacité.

Une table réglementaire est prévue pour la composante maintien du calcul des provisions de sinistres en incapacité. Ces provisions, d'abord encadrées par le Code Assurances à l'article A. 331-22 (abrogé), sont dorénavant encadrées par l'article 143-12 du règlement de l'Autorité des Normes Comptables (ANC) qui dispose :

« (...) Le calcul des provisions techniques de prestations d'incapacité de travail et d'invalidité est effectué à partir des [...] lois de maintien en incapacité de travail et en invalidité indiquées aux articles 600-2, 600-3 et 600-4 du présent règlement. Toutefois, il est possible pour une entreprise d'assurances d'utiliser une loi de maintien établie par ses soins et certifiée par un actuaire indépendant de cette entreprise agréé à cet effet par l'une des associations d'actuaire reconnues par l'Autorité de contrôle prudentiel et de résolution (...) »

La table réglementaire du maintien en incapacité est celle du BCAC (Bureau Commun d'Assurances Collectives). Il s'agit d'une table à double entrée (âge de survenance d'une part et durée du sinistre d'autre part) prolongée en 2010 indiquant, pour chaque durée, les effectifs restant en incapacité sur une cohorte initiale de 10 000 personnes. Une version actualisée est parue en 2013 mais ne dispose pas du caractère réglementaire. Le BCAC propose également une table de passage en invalidité et une table de maintien en invalidité. Ces deux derniers éléments ne sont pas traités dans la suite de ce mémoire.

La mesure de la durée d'un épisode d'incapacité fait intervenir des modèles de survie, dans lesquels l'évènement d'intérêt est la sortie d'incapacité par la reprise du travail. Comme pour la construction des lois d'entrée en incapacité, les données comportent des censures à droite et des troncatures à gauche. Les censures sont issues d'individus dont les prestations sont arrêtées avant la reprise du travail (passage en invalidité, passage à la retraite, fin de la période d'étude, décès de l'assuré). Les troncatures sont introduites par les franchises. **En effet, seuls les sinistres dont la durée est supérieure à la franchise peuvent être observés.**

Dans le cadre de cette étude, le choix de la méthode se porte naturellement sur l'estimateur de Kaplan-Meier, déjà présenté dans le chapitre 2. Cet estimateur est en effet retenu dans la majorité des publications et des documentations disponibles. Il est de ce fait particulièrement bien documenté.

A titre d'illustration, les fonctions de survie de la table réglementaire du BCAC (2010) sont données dans la **Figure 31** ci-après.

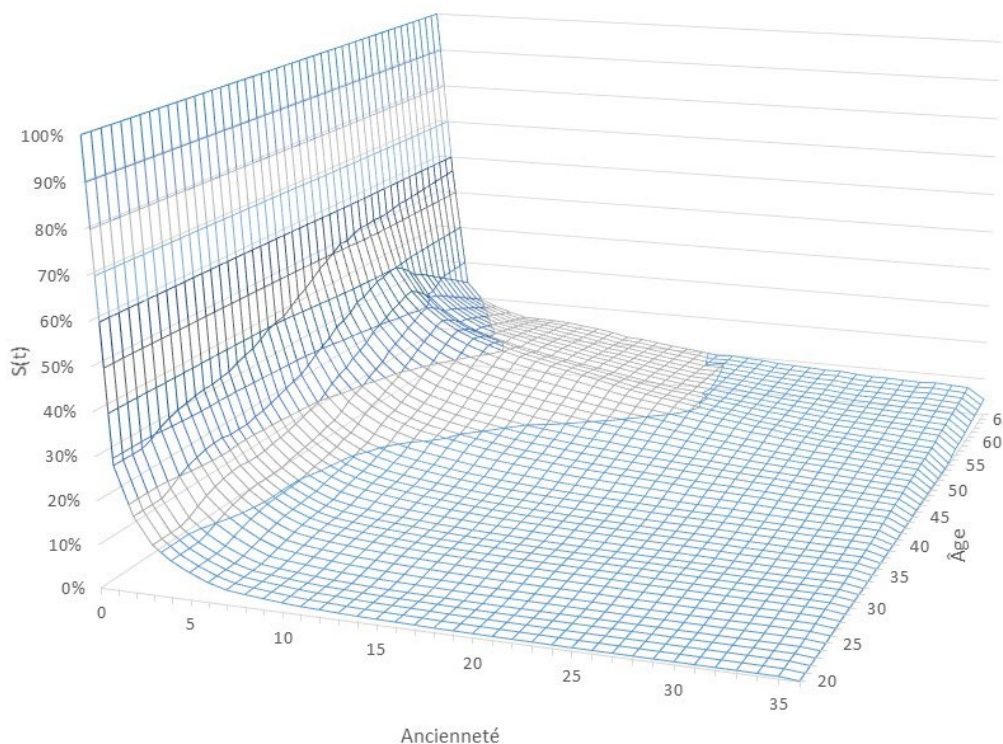


Figure 31 : Fonctions de survie du BCAC (2010)

La **Figure 31** illustre la forme convexe et décroissante des fonctions de survie en incapacité. Le nombre de sorties mensuelles décroît très fortement les premiers mois, puis converge à un niveau très faible. En particulier, sur la table réglementaire du BCAC, entre 40% et 70% (selon l'âge au moment de la survenance) des individus ont retrouvé leurs capacités à l'issue du premier mois de maintien en incapacité. Les arrêts courts sont donc les plus nombreux et les arrêts les plus longs sont plus rares. Enfin, les durées de maintien en incapacité sont également plus longues sur les âges avancés comme l'illustre la crête centrée autour de 54 ans. Passé cet âge, les durées se raccourcissent à nouveau, à l'approche de l'âge de départ à la retraite.

4.2 Elaboration des taux de maintien en incapacité pour le portefeuille de contrats de prévoyance classiques

4.2.1 Estimateur de Kaplan-Meier

La construction des lois de maintien en incapacité repose sur le même jeu de données que celui utilisé pour la construction des lois d'entrée en incapacité. Des données initiales, ne sont conservés que les sinistres, c'est-à-dire uniquement les lignes pour lesquelles l'évènement « entrée en incapacité » est observé. Parmi ces sinistres, il convient de déterminer ceux dont l'ensemble de l'épisode d'incapacité est observable, c'est dire ne faisant pas l'objet d'une censure à droite. Pour ces derniers, l'évènement d'intérêt « retour à l'activité professionnelle », est observé. Les autres sinistres, c'est-à-dire ceux dont la fin de période d'incapacité est inobservable, sont censurés. Sont notamment considérées comme des censures les périodes d'incapacité non-suivies d'une nouvelle période de couverture pour l'assuré.

Les taux bruts de sortie d'incapacité sont donnés dans la **Figure 32** ci-après.

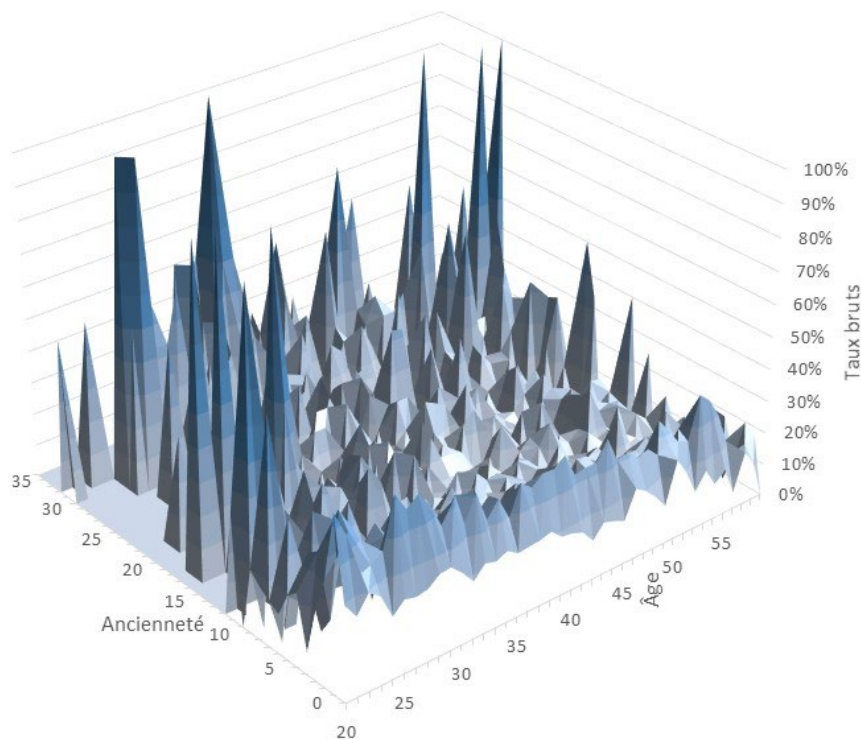


Figure 32 : Taux bruts de sortie par âge et ancienneté

Il apparaît que les taux bruts sont erratiques, particulièrement aux âges et aux anciennetés extrêmes. **Il est en effet nécessaire de disposer d'un grand nombre de sinistres pour être en mesure d'observer suffisamment d'événements sur chacune des anciennetés et chacun des âges.** Afin d'illustrer ce phénomène, l'amplitude des intervalles de confiance des fonctions de survie peut être mesurée.

4.2.1.1 Calcul des intervalles de confiance

La détermination des intervalles de confiance fait intervenir la variance. Comme présentée lors du calcul des taux d'entrée en incapacité, une approximation de la variance est donnée par l'estimateur de Greenwood :

$$\hat{v}(\hat{S}(t)) = \hat{S}(t)^2 \sqrt{\sum_{T(i) \leq t} \frac{d_i}{n_i(n_i - d_i)}}.$$

Dès lors, l'intervalle de confiance bilatéral se calcule de la façon suivante :

$$IC(\alpha) = \hat{S}_{(t)} \left[1 \pm u_{1-\frac{\alpha}{2}} \sqrt{\sum_{i=1}^t \frac{d_i}{n_i(n_i - d_i)}} \right].$$

L'amplitude de l'intervalle de confiance correspond à l'écart entre la borne supérieure et la borne inférieure. La **Figure 33** ci-dessous indique l'amplitude des intervalles de confiance pour chacune des fonctions de survie.

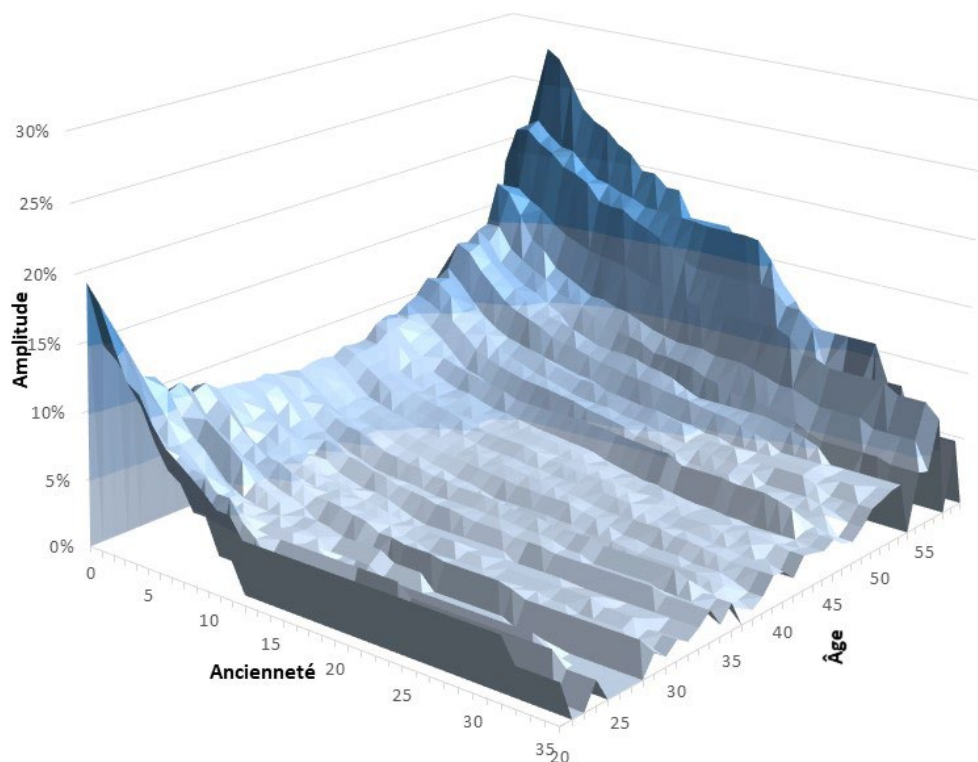


Figure 33 : Amplitude des intervalles de confiance

Il apparaît que **l'amplitude des intervalles de confiance est très élevée sur les âges extrêmes**, particulièrement après 55 ans et avant 25 ans. Ce constat tient à la trop faible quantité de données sur ces âges. Plus généralement, les intervalles de confiance paraissent, en apparence, se resserrer illogiquement au fur et à mesure que l'ancienneté augmente, c'est-à-dire au fur et à mesure que la quantité d'information diminue. L'expression de la formule permettant de déterminer les intervalles de confiance indique cependant que ces derniers accumulent bien « l'incertitude » depuis l'entrée en incapacité, mais que **la valeur de l'intervalle est relative à la valeur de la fonction de survie estimée**. Il est donc plus judicieux d'observer des **amplitudes relatives**, c'est-à-dire normalisées par la valeur de la fonction de survie en chaque point. La **Figure 34** ci-après indique la valeur de ces amplitudes relatives.

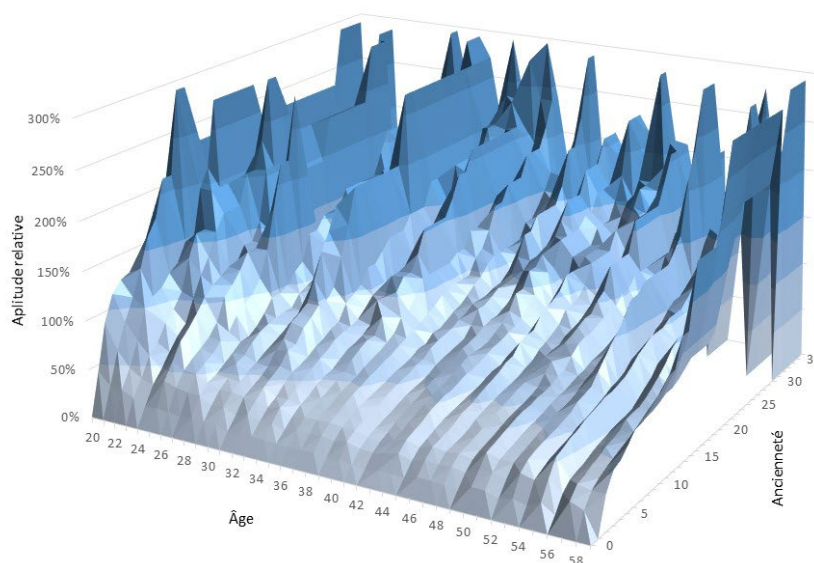


Figure 34 : Amplitude relative des intervalles de confiance des fonctions de survie

L'incertitude augmente donc bien avec l'ancienneté. Il apparaît également **qu'au-delà des âges extrêmes, les anciennetés élevées sont peu exploitables compte tenu du manque d'information disponible**. Seule une partie des couples « âges / anciennetés » peut effectivement être conservée. Afin de déterminer les points exploitables il peut être utile limiter arbitrairement l'amplitude relative des intervalles de confiance de la fonction de survie. Les points affichant une amplitude relative inférieure à 50% pour l'intervalle de confiance sont indiqués par une teinte foncée dans la **Figure 35** ci-après.

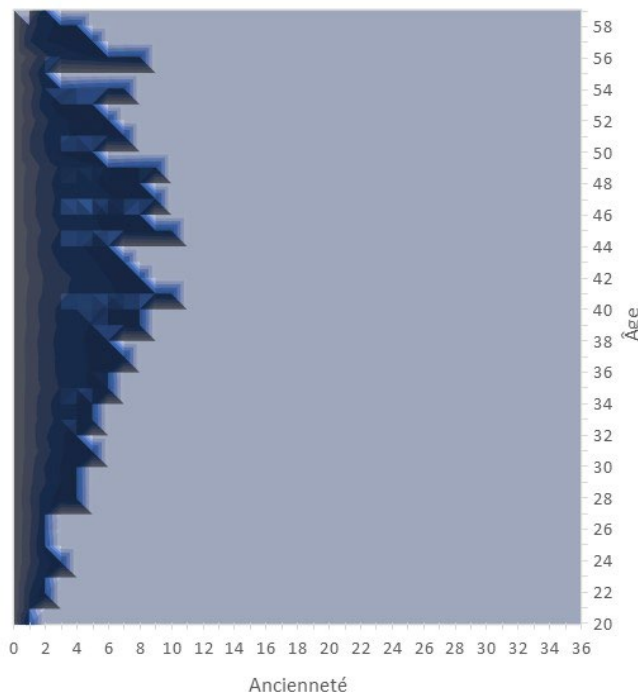


Figure 35 : Points satisfaisant les conditions (teinte foncée)

Ainsi, **quel que soit l'âge, l'information est insuffisante au-delà de 10 mois d'ancienneté**. Il serait néanmoins possible, en principe, de réaliser par la suite une extrapolation via la table du BCAC afin de disposer d'une loi complète.

En outre, les fonctions de survie sont estimées à partir de taux bruts et présentent un aspect irrégulier. Il est donc nécessaire de réaliser un lissage, à l'image des travaux déjà réalisés sur les taux d'entrée en incapacité. Deux méthodes sont présentées par la suite. D'une part un lissage de Whittaker-Henderson, adapté au cas à deux dimensions (utile avant 10 mois d'ancienneté, mais inopérant après). D'autre part un ajustement via une table de référence externe, qui permettra également de prolonger la loi d'expérience.

4.2.2 Lissage de Whittaker-Henderson

4.2.2.1 Adaptation de la méthode aux cas en dimension 2

Les lois de maintien étant à deux dimensions (durée et âge), il est nécessaire d'effectuer un double lissage. En effet, il convient **d'obtenir une régularité entre les pas d'ancienneté, mais également entre les âges**. Il est possible d'adapter la méthode de Whittaker-Henderson pour combiner les critères de régularité et de fidélité de chacune des dimensions. Les adaptations de la méthode, présentées dans le chapitre 2, sont formulées ci-après.

Le critère de fidélité s'adapte naturellement afin de capter tous les points.

	1 dimension	2 dimensions
Fidélité	$F = \sum_{i=1}^p w_i (g_i - \hat{q}_i)^2$	$F = \sum_{i=1}^p \sum_{j=1}^q w_{ij} (g_{ij} - \hat{q}_{ij})^2$

Le critère de régularité à deux dimensions peut être défini comme la double somme d'un critère de régularité vertical et d'un critère de régularité horizontal. Chacun de ces deux critères étant la somme de mesures de régularité à une dimension, déjà présentée dans le chapitre 2. Le degré de régularité z et le paramètre de normalisation h peuvent être définis pour chacune des dimensions afin de calibrer différemment le lissage vertical et le lissage horizontal. Si bien que la mesure de régularité devient la suivante :

	1 dimension	2 dimensions
Régularité	$S = h \sum_{i=1}^{p-z} (\Delta^z g_i)^2$	$\left[S_v = \sum_{j=1}^q \sum_{i=1}^{p-z_v} (\Delta_v^{z_v} g_{ij})^2 \right] \text{ et } \left[S_h = \sum_{i=1}^p \sum_{j=1}^{q-z_h} (\Delta_v^{z_h} g_{ij})^2 \right]$

La mesure de lissage devient $M = F + h_v S_v + h_h S_h$. La résolution de ce problème s'obtient en revenant dans un cadre unidimensionnel (Planchet, Modèles de durée : Méthodes de lissage et d'ajustement, 2021) :

- Le vecteur des taux bruts, noté u , de dimension $p * q$ est construit en intégrant ligne par ligne la matrice $\hat{q}$ des taux bruts, de dimension $p \times q$. Plus formellement, $u_{q(i-1)+j} = \hat{q}_{ij}$.
- La matrice des poids W^* est construite en intégrant ligne à ligne la matrice des poids W_{ij} sur sa diagonale. Plus formellement, $W_{q(i-1)+j, q(i-1)+j}^* = W_{ij}$ est donc une matrice carrée de dimension $p * q$.
- Le vecteur u des taux bruts étant de dimension $p * q$, la matrice des coefficients binomiaux $K_{z_v}^v$ est impérativement composée de $p * q$ colonnes. L'expression de S faisant intervenir q fois g_{p+z} , afin que toutes les valeurs soient correctement définies, le nombre de lignes de $K_{z_v}^v$ doit donc être de $q(p - z)$. Si bien que la matrice $K_{z_v}^v$ est de dimension $q(p - z) \times (p * q)$. En suivant ce raisonnement, la matrice $K_{z_h}^h$ est de dimension $p(q - z) \times (p * q)$. L'implémentation des coefficients binomiaux dans ces matrices est particulièrement délicate (notamment pour le lissage vertical) et a nécessité la construction d'un algorithme pour généraliser leur construction.

4.2.2.2 Présentation des résultats

Le lissage de Whittaker-Henderson en dimension 2 peut être calibré grâce à l'utilisation des techniques présentées lors du lissage à une dimension des taux d'entrées. Les méthodes du Chi-deux et de la validation croisée ont présenté des résultats très proches. Les paramètres retenus pour le lissage sont les suivants : $z_v = z_h = 2$; $h_v = 25$; $h_h = 5\,000$ (Afin de pouvoir illustrer cette méthode, les points pour lesquels les données étaient jugées insuffisantes ont été réintroduits).

A ce stade, les taux de sorties lissés présentés dans la **Figure 36** ci-après, n'arborent pas la forme classique et attendue sur les premiers mois. La hausse des taux le premier mois traduit l'effet des troncatures, c'est-à-dire des sinistres dont la durée totale est inférieure à la franchise et qui sont donc inobservés pour l'assureur. **L'absence de sinistres de moins de 15 jours dans les données étudiées a pour conséquence une surestimation du maintien sur les premiers mois (dans mesure où il manque les sinistres les plus courts).** *A contrario*, la table du BCAC est construite sur une franchise 0 jours, c'est-à-dire qu'elle capte l'ensemble de sinistres, y compris ceux de moins de 15 jours.

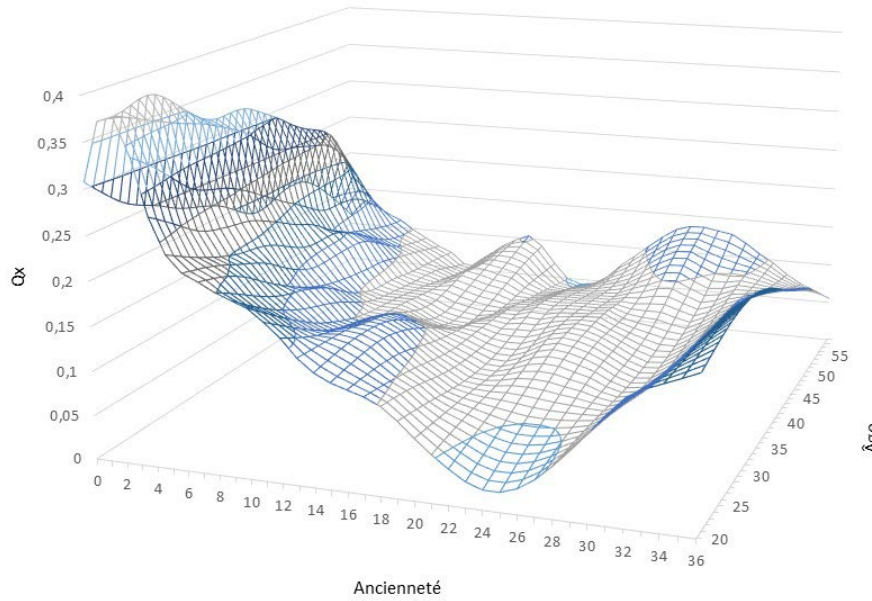


Figure 36 : Taux de sortie d'incapacité lissés sur les portefeuilles de prévoyance

Si cette loi d'expérience est en principe utilisable en l'état pour calculer des barèmes de provisionnement sur les portefeuilles de prévoyance classiques Prévoir Vie, il est nécessaire de procéder à une transformation des taux du premier mois afin d'harmoniser la loi avec les standards de marché et de pouvoir la comparer à la table réglementaire.

4.2.2.3 Calcul de la loi d'expérience en franchise 0 jours.

Afin de corriger la loi d'expérience, il convient, à chaque âge, de déterminer la part des sinistres dont la durée est inférieure à 15 jours. Aucune base nationale accessible aux assureurs n'est disponible. Il est donc proposé de procéder à une interpolation linéaire à partir de la table du BCAC. Compte tenu de la convexité supposée de la courbe sur l'intervalle $[0,1]$, il est probable que cette méthode surestime légèrement les taux de maintien à 15 jours. Cette démarche est néanmoins prudente. Le nombre d'individus toujours en incapacité à 15 jours sur la table du BCAC est donc donné par :

$$l_{(0,5)}^{BCAC} = \left(\frac{l_1^{BCAC} - l_0^{BCAC}}{2} \right) + l_0^{BCAC}.$$

Et la part des sinistres dont la durée est inférieure à 15 jours est donnée par : $\Delta = \frac{l_0^{BCAC} - l_{0,5}^{BCAC}}{l_0^{BCAC}}$.

Il est dès lors possible de recalculer la fonction de survie de la loi d'expérience en majorant la cohorte initiale des sinistres dont la durée est inférieure à 15 jours :

$$l_0^{\exp 0j} = \frac{l_0^{\exp 15j}}{1 - \Delta} \text{ et donc } S_{(t)}^{\exp 0j} = \frac{l_t^{\exp 15j}}{l_0^{\exp 0j}}, \forall t \in [1,35].$$

Comme l'illustre la **Figure 37** ci-après, la forme de cuvette des taux de sortie est désormais reconnaissable. Les taux décroissent fortement les premiers mois pour atteindre le fond de la cuvette aux environs du mois 18. Puis les taux de sortie croissent à nouveau à l'approche du mois 36 qui marque le passage en invalidité.

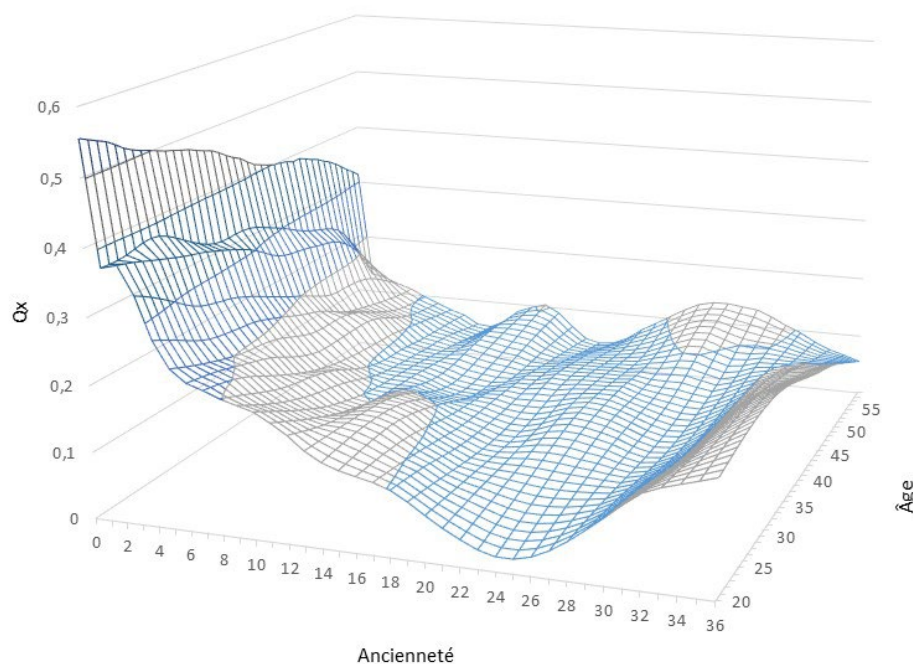


Figure 37 : Taux de sortie d'incapacité lissés sur les portefeuilles de prévoyance classiques ramenés à une franchise 0 jours

Ces taux de sortie permettent de tracer les fonctions de survie lissées qui sont présentées dans la **Figure 38** ci-dessous :

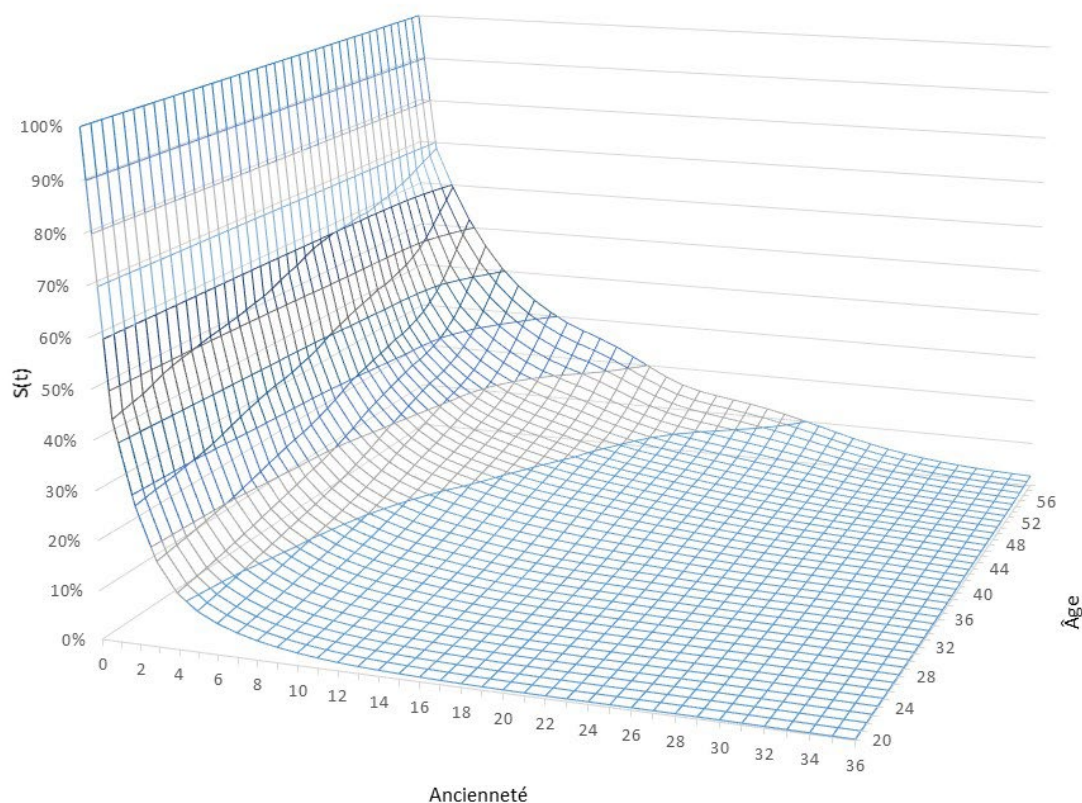


Figure 38 : Fonctions de survie en franchise 0 jours

Le tracé des fonctions de survie issues de la loi d'expérience fait apparaître une forme assez proche du tracé des fonctions de survie issues de la table du BCAC. Il est en particulier possible d'observer le début de

la crête caractéristique sur les âges élevés. Cette dernière n'apparaît cependant pas complètement puisque la loi d'expérience s'arrête à 59 ans faute de données disponibles après cet âge.

Afin de conforter et de mieux appréhender les lois d'expérience de maintien en incapacité, il convient de les comparer à des tables de référence (table réglementaire, loi d'expérience plus ancienne, etc.). Afin de faciliter la comparaison, il est généralement d'usage de calculer les espérances de durée résiduelle de maintien en incapacité.

4.2.2.4 Espérances de durée résiduelle et comparaison avec la table réglementaire

Les lois de maintien en incapacité sont principalement utilisées pour le calcul des barèmes de provisionnement. Ce barème est l'expression des espérances de durée résiduelle actualisées. Ce calcul rend beaucoup plus aisée la comparaison entre deux lois de maintien. Les espérances de durée (ou vie) résiduelle sont calculées de la façon suivante :

$$e_{x,t} = \frac{1}{L_{x,t}} * \sum_{k=0}^{max-t} L_{x,k+t},$$

avec x l'âge de l'assuré à la survenance du sinistre, t l'ancienneté de l'état d'incapacité. Les espérances de durée résiduelle sont tracées dans la **Figure 39** ci-dessous pour la table réglementaire et la table d'expérience :

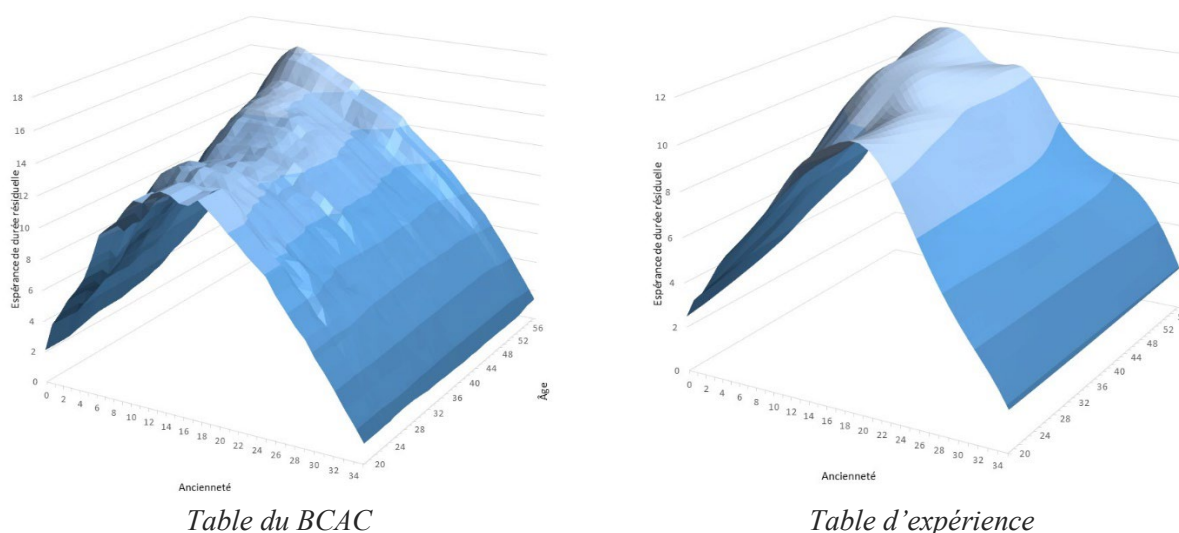


Figure 39 : Espérances de durée résiduelle

Il apparaît une relative proximité entre les deux tracés, dont la forme est caractéristique : **la baisse des taux de sortie se traduit dans un premier temps par un allongement des espérances de durée résiduelle. Puis le temps maximum restant à courir jusqu'au passage en invalidité étant de plus en plus faible, les durées résiduelles baissent jusqu'à l'ancienneté maximum de la table.** La lecture de cette figure semble indiquer que la table du BCAC conduit à un léger sur-provisionnement. Cette impression générale est confirmée par le calcul des écarts relatifs entre les deux tables :

$$\frac{e_{x,t}^{BCAC} - e_{x,t}^{XP}}{e_{x,t}^{XP}}.$$

Les écarts relatifs sont tracés dans la **Figure 40** ci-après.

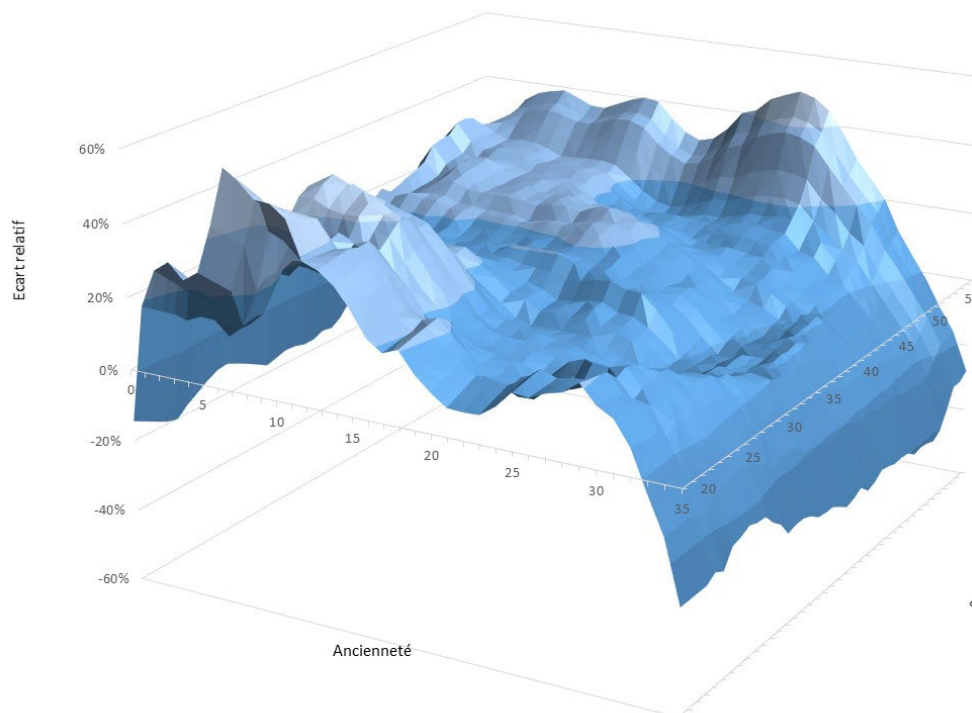


Figure 40 : Ecart relatif entre les espérances de durée résiduelle d'expérience et du BCAC

Il apparaît en effet que **la table du BCAC (2010) conduit globalement à un sur-provisionnement (à taux 0%)**. Ce dernier est d'environ 20% à 30% sur les âges et les anciennetés centrales. En revanche, au moment de l'entrée en incapacité, le provisionnement du BCAC (2010) est plus faible que celui de la loi d'expérience. Enfin, la table du BCAC tend à minimiser le provisionnement sur les derniers mois d'ancienneté. **Compte tenu des réserves émises sur la quantité de données disponibles, il ne serait de toute façon pas prudent d'utiliser la loi d'expérience ainsi établie.** En outre, cette dernière ne couvre pas tous les âges nécessaires. Afin de disposer d'une loi complète et plus fiable, il est possible d'utiliser une méthode statistique moins consommatrice d'informations. Cette seconde approche est proposée grâce à l'utilisation d'un modèle de Brass à référence externe.

4.2.3 Lissage par un ajustement à référence externe de Brass et prolongement de la loi

Dans le cas où les données sont insuffisantes, notamment pour certains points de la table, il est préférable d'ajuster la loi d'expérience grâce à une table externe de référence. Dans le cas des lois de maintien en incapacité, le choix se porte naturellement sur les tables du BCAC. Le modèle semi-paramétrique à référence externe de Brass permet de relier la loi d'expérience et la table de référence via une régression linéaire :

$$\text{logit}(q_{x,t}^{\text{ajusté}}) = \theta_1 + \theta_2 * \text{logit}(q_{x,t}^{\text{BCAC}}) + \varepsilon_{x,t}.$$

$$\text{Et donc } q_{x,t}^{\text{ajusté}} = \frac{\exp(\theta_1 + \theta_2 * \text{logit}(q_{x,t}^{\text{BCAC}}) + \varepsilon_{x,t})}{1 + \exp(\theta_1 + \theta_2 * \text{logit}(q_{x,t}^{\text{BCAC}}) + \varepsilon_{x,t})}.$$

Les taux de sortie sont reliés via leur transformation logistique (*logit*). En effet, dans le cadre d'une régression linéaire, la droite passant dans le nuage de points ne peut pas être bornée et il n'est donc pas possible de la limiter à l'intervalle $[0,1]$ sur lequel se définissent les probabilités p . Le *logit*, défini comme $\text{logit}(p) = \ln\left(\frac{p}{1-p}\right)$, est en revanche bien défini sur $\mathbb{R}$ et permettant donc d'effectuer des régressions linéaires. Le paramètre θ_1 peut être interprété comme l'ordonnée à l'origine (intercept, ou niveau). Il ne dépend ni de l'âge, ni de l'ancienneté du sinistre. Le paramètre θ_2 correspond à la pente de la droite. Il agit comme un

coefficient multiplicatif à appliquer sur la table de référence. Son apport est donc relatif. Enfin le bruit est noté $\varepsilon_{x,t}$. Il est considéré i.i.d.

En pratique, il est nécessaire de générer trois vecteurs de même taille. Le vecteur X correspondant aux transformations logistiques des taux de sortie de la table de référence, le vecteur Y correspondant aux transformations logistiques des taux bruts de la loi d'expérience empirique, et le vecteur W correspondant aux poids à appliquer pour la résolution du problème d'optimisation des moindres carrés.

Finalement, les coefficients issus de la régression linéaire sont les suivants : $\theta_1 = 0,17$; $\theta_2 = 1,04$. L'ajustement des taux bruts a été réalisé sur la table du BCAC 2013, qui est la version la plus récente mais non réglementaire. L'utilisation de cette table a abouti à des résultats plus satisfaisants que la version réglementaire de 2010. Plus précisément, il est d'usage de valider graphiquement les résultats en comparant les sorties observées aux sorties modélisées, d'une part par ancienneté et d'autre part par tranche d'âge. Les intervalles de confiances bilatéraux à 95% sont également calculés par approximation normale. La comparaison des sorties observées et des sorties modélisées est affichée dans la **Figure 41** ci-dessous.

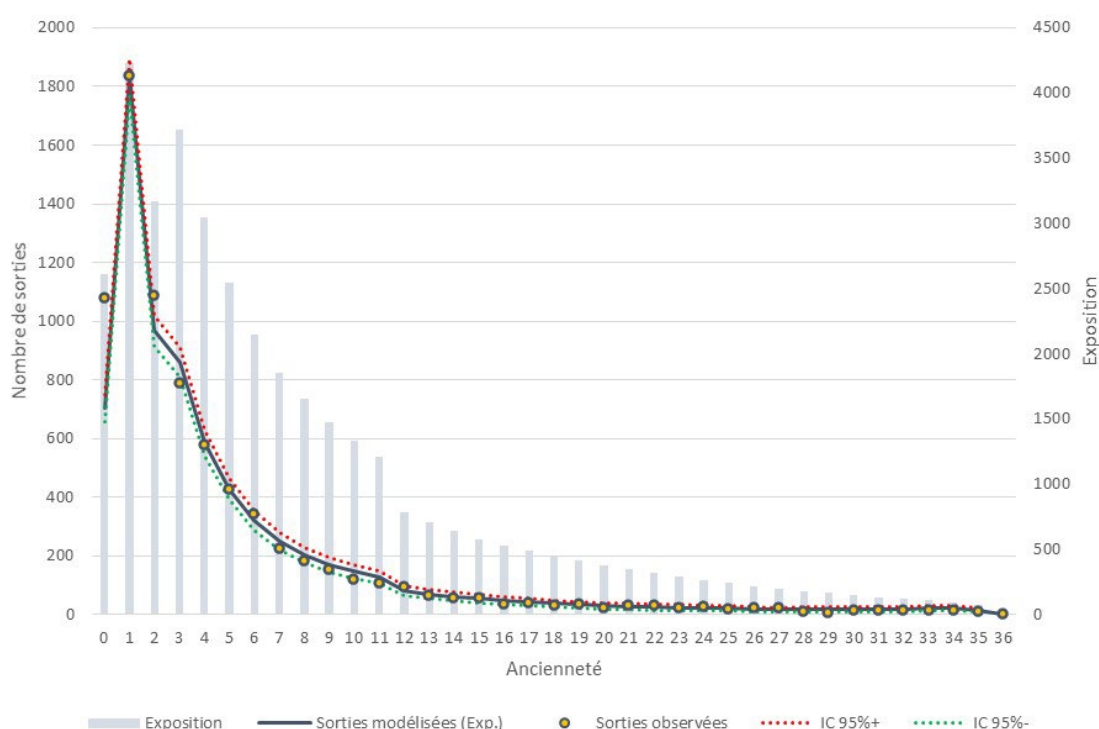


Figure 41 : Comparaison entre les sorties observées et les sorties modélisées (par ancienneté)

Le premier constat tient à la forme particulière du nombre de sorties. Il est généralement attendu que le nombre de sorties soit plus élevé le premier mois d'ancienneté, ce qui n'est pas le cas sur les données Prévoir. Cette particularité est une traduction des effets de franchise. D'une part, compte tenu de la franchise minimale à 15 jours, l'exposition maximale pour un contrat est de 0,5 sur le premier mois d'ancienneté. D'autre part, un nombre important de troncatures entre dans l'exposition à partir du deuxième mois d'ancienneté (franchises 30 et 40 jours).

Par ailleurs, la loi d'expérience ainsi générée semble sous-estimer le nombre de sorties sur le premier mois d'ancienneté. Cette sous-estimation est vraisemblablement un effet collatéral du calcul permettant l'approximation des forces de mortalité ($\mu_x \approx -\ln(1 - q_x)$) via les quotients de mortalité (q_x). Cette approximation repose sur l'hypothèse de forces de mortalités constantes par morceau entre les âges et les années entières. Or, cette hypothèse n'est probablement pas vérifiée compte tenu de l'effet de la franchise minimale à 15 jours et de la forme convexe des taux des sorties. Cependant **cette sous-estimation confère à la loi d'expérience un caractère plus prudent.**

Sur les anciennetés plus élevées, il apparaît que l'ajustement semble globalement bien équilibré, si bien que l'intervalle de confiance des sorties modélisées encadre les sorties observées sur la quasi-totalité des points de données. La **Figure 42** ci-après, focalisée sur les anciennetés de plus de 9 mois, permet en effet d'apprécier l'ajustement sur les anciennetés les plus élevées.

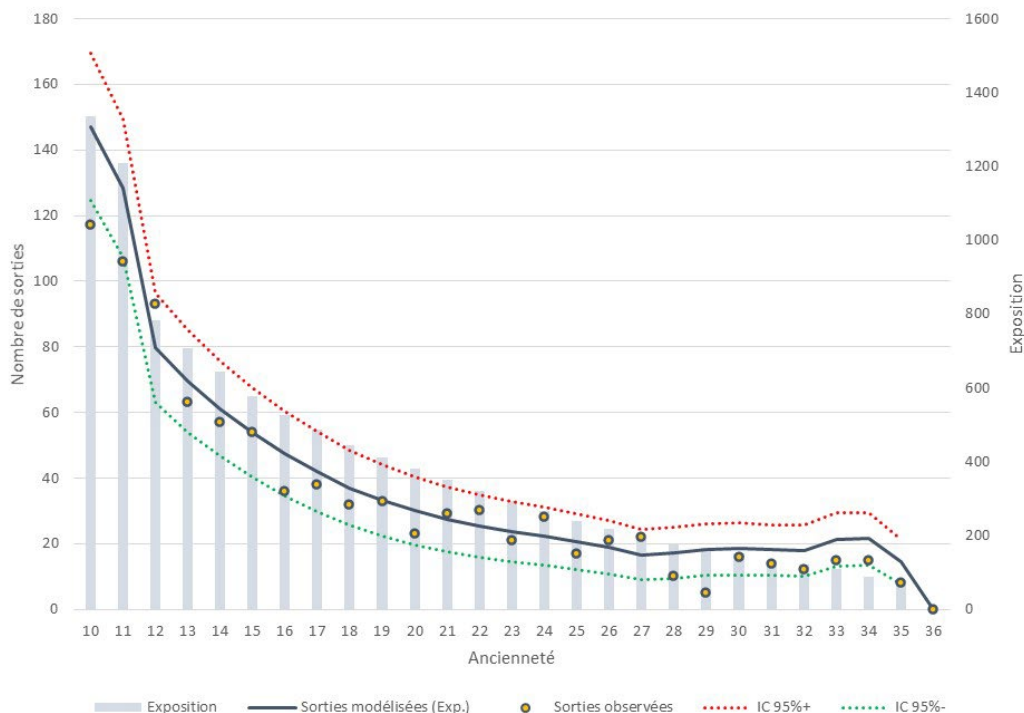


Figure 42 : Comparaison entre les sorties observées et les sorties modélisées par ancienneté (détail)

L'autre dimension de la validation de la régression consiste à comparer les sorties observées et modélisées en fonction des âges, comme présenté dans la **Figure 43** ci-dessous :

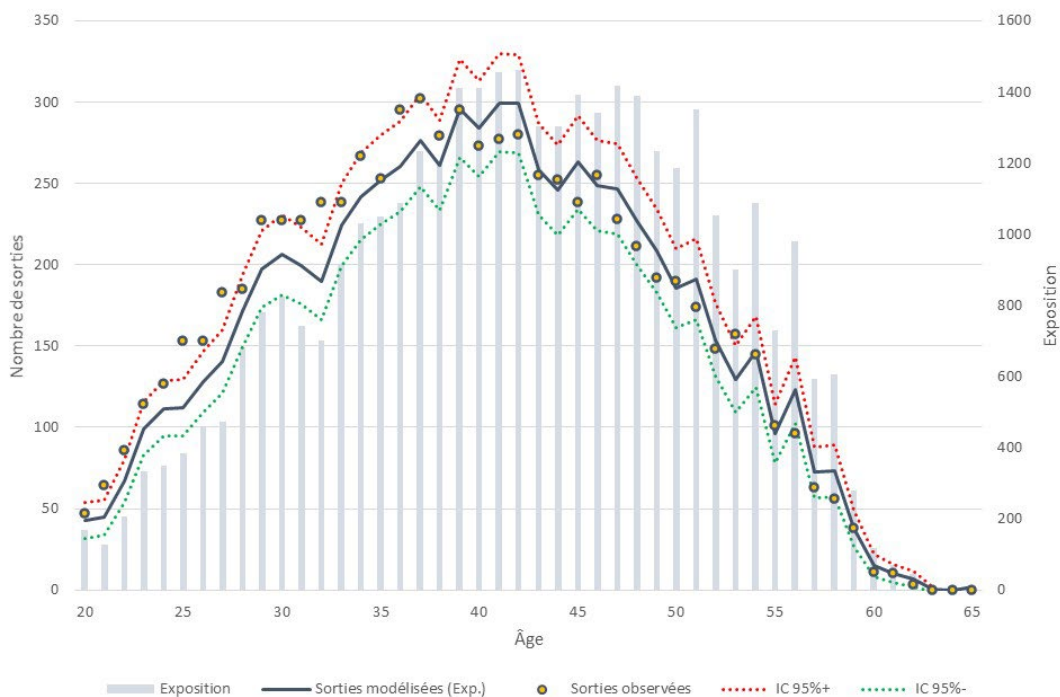


Figure 43 : Comparaison entre les sorties observées et les sorties modélisées par âge

L'intervalle de confiance encadre globalement les sorties observées. Il apparaît cependant une légère sous-estimation des sorties pour les assurés les plus jeunes au moment de la survenance du sinistre (entre 20 et 37 ans). En revanche, l'ajustement semble particulièrement bien adapté à partir de 38 ans comme en atteste la superposition des sorties observées et des sorties modélisées. La qualité de cet ajustement tient à l'exposition plus importante sur les âges élevés, et donc aux pondérations utilisées dans la régression. **Il s'agit de l'illustration de la limite de cette méthode qui « réside dans le fait qu'il peut être difficile d'admettre une relation linéaire unique pour l'ensemble des âges »** (Aubin & Rolland, 2010). Il serait en principe possible d'aménager la méthode de Brass en autorisant des changements de pente entre deux groupes d'âges autour d'un âge pivot (Blanpain & Chardon, 2011). Cet aménagement pourra faire l'objet de futurs travaux.

L'ajustement étant conservé en l'état pour la suite des travaux, il est possible de tracer les durées résiduelles dans la **Figure 44** ci-après.

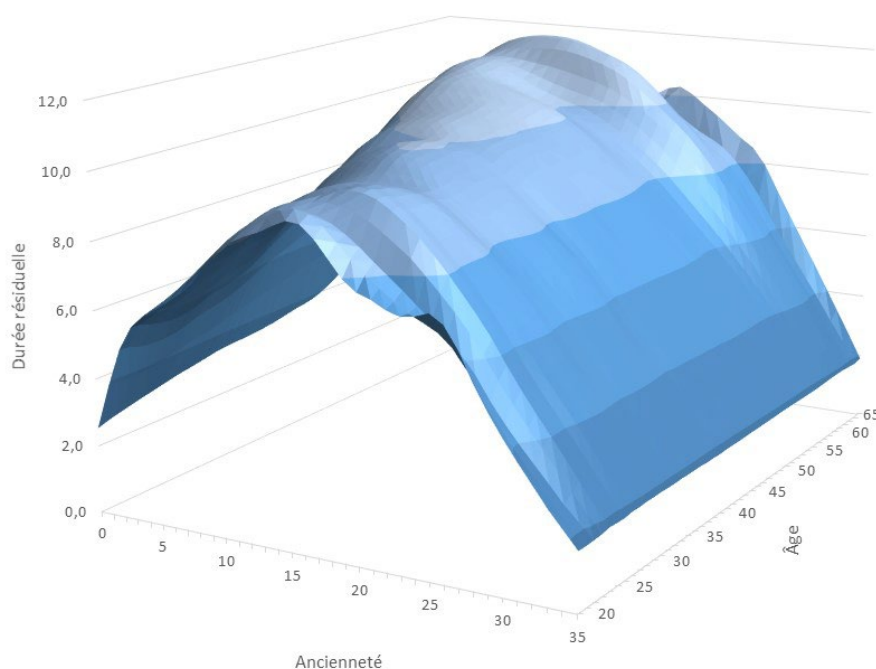


Figure 44 : Espérances de durée résiduelle de la loi d'expérience ajustée

La table de référence du BCAC 2013 étant lissée, la surface des durées résiduelles de la loi d'expérience l'est également. L'utilisation d'une régression linéaire impose en effet la même forme à la loi d'expérience qu'à la table de référence utilisée. Enfin les écarts relatifs avec la table réglementaire du BCAC (2010) sont tracés dans la **Figure 45** ci-après.

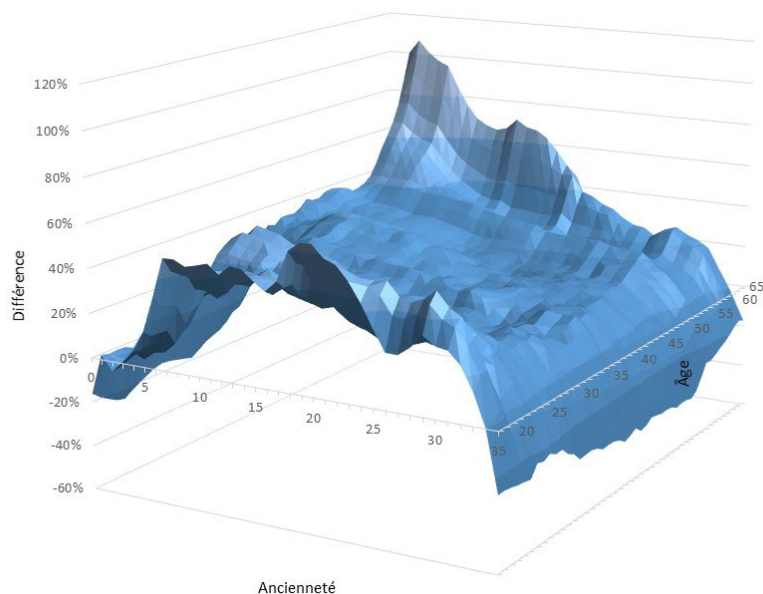


Figure 45 : Comparaison avec les espérances de durée résiduelle du BCAC (2010)

Il apparaît en premier lieu que l'utilisation de la table réglementaire conduit à un provisionnement plus important d'environ 20% à 35% sur les âges et anciennetés centraux. **L'écart relatif moyen pondéré par les expositions est de 25% entre les espérances de durée résiduelle de la table d'expérience et de la table du BCAC (2010).** Les espérances de durée résiduelle sont globalement plus basses après 34 mois passé en incapacité. Cette observation tient aux taux de sortie élevés de la table réglementaire du BCAC (2010) sur le mois 35 et non reconduits sur la version de 2013 ayant servi de référence au modèle de Brass.

Enfin, il est intéressant, à titre informatif, de comparer les espérances de durée résiduelle issues du lissage de Whittaker-Henderson et du lissage par ajustement de Brass dans la **Figure 46** ci-après ($\frac{e_{x,t}^{WH} - e_{x,t}^{ajustée}}{e_{x,t}^{ajustée}}$).

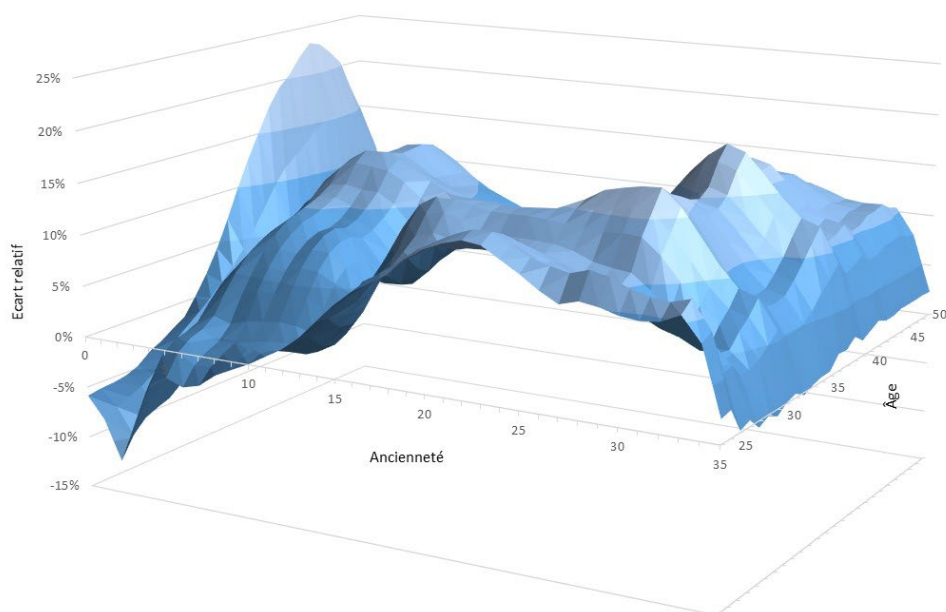


Figure 46 : Comparaison avec les espérances de durée résiduelle du lissage du WH

Sur les âges centraux, les écarts se situent globalement entre 5% et 15%, ce qui signifie que l'utilisation de la table ajustée par Brass conduit en principe à une économie de provisionnement. Elle est donc, en principe, moins prudente. Pour autant, à la suite des réserves émises sur la validité de la table lissée via Whittaker-Henderson (faible quantité de données), cette observation ne remet pas en cause le choix de la méthode : **compte tenu de la meilleure stabilité du modèle de Brass en présence de petits échantillons, ce dernier est finalement retenu.**

4.3 Elaboration des taux de maintien en incapacité pour le portefeuille emprunteur

Du fait de la surreprésentation de la franchise 90 jours (98% des assurés) et de la faiblesse des taux d'entrée en incapacité, comparativement aux portefeuilles de contrats de prévoyance classiques distribués par Prévoir, le nombre de sinistres observés sur le portefeuille emprunteur est limité. Dans ces conditions, l'élaboration d'une loi lissée par Whittaker-Henderson, même partielle n'est pas envisageable. **Le modèle de Brass est donc directement utilisé.**

Afin de valider la régression, les sorties modélisées et les sorties observées sont comparées. La **Figure 47** ci-après présente les sorties en fonction des anciennetés. Seules les durées supérieures à la franchise de 90 jours sont affichées.

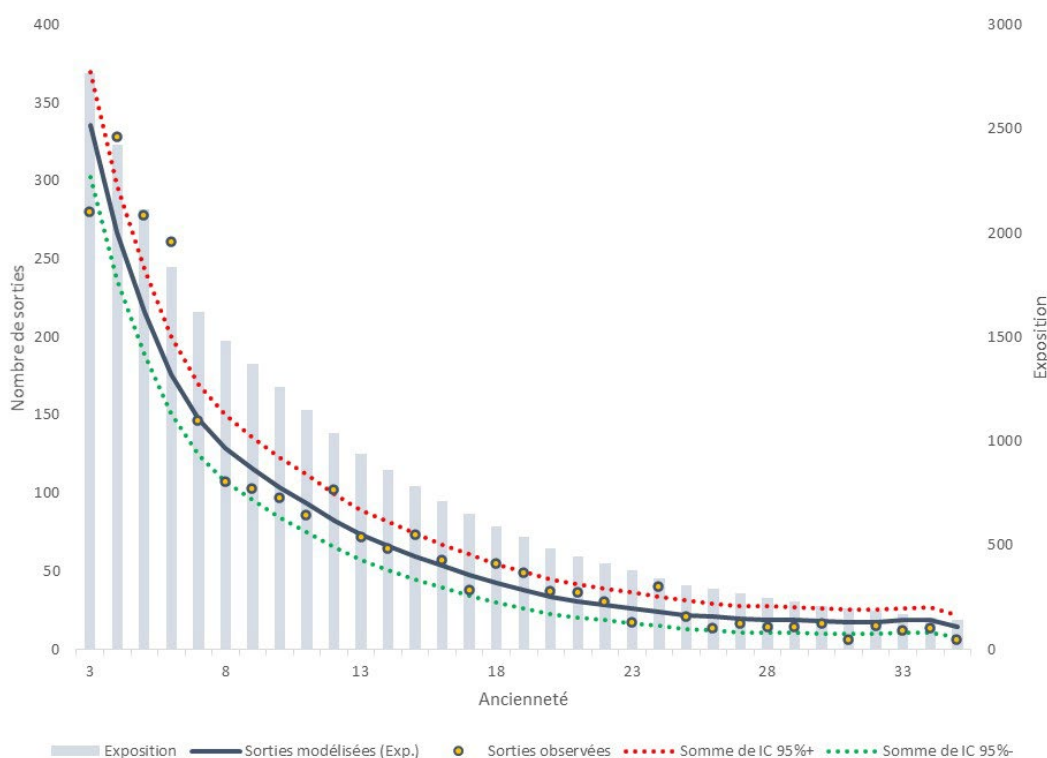


Figure 47 : Comparaison entre les sorties observées et les sorties modélisées par ancienneté

Il apparaît que les taux de sortie du premier mois suivant la franchise (mois 3 dans le graphique) sont surestimés par rapport aux données. Puis les taux de sortie sont sous-estimés sur les trois mois suivants, ce qui traduit un compromis trouvé par la régression entre les taux de sortie des quatre premiers mois. Sur cette période, les sorties modélisées se situent en dehors des intervalles de confiance. Sur les anciennetés plus élevées, les sorties modélisées et les sorties observées sont globalement superposées et se situent en général entre les bornes de l'intervalle de confiance.

Les mêmes éléments, par âge, sont appréciés dans la **Figure 48** ci-après.

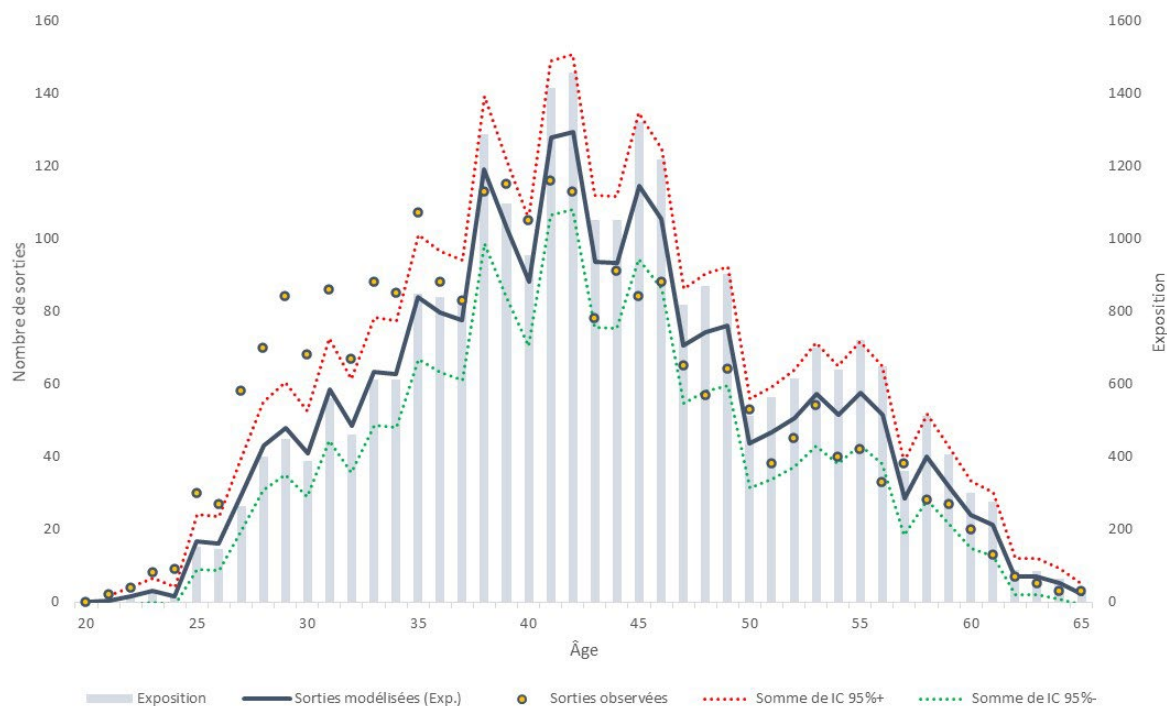


Figure 48 : Comparaison entre les sorties observées et les sorties modélisées par âge

La comparaison par âge entre les sorties modélisées et les sorties observées fait apparaître une sous-estimation des sorties avant 35 ans. Sur ces âges, les sorties observées se situent généralement au-delà de la borne supérieure de l'intervalle de confiance. Puis, passé 35 ans les sorties observées sont encadrées par les intervalles de confiance des sorties modélisées, traduisant un meilleur ajustement.

En imposant un unique jeu de coefficients de régression, l'ajustement de Brass n'est pas en mesure de satisfaire plus précisément l'ensemble des points du modèle. Les pondérations ont néanmoins permis un bon ajustement sur les intersections présentant les plus fortes expositions. En tout état de cause, il convient de souligner que l'utilisation des taux ajustés par Brass permet d'atteindre un ratio SMR global (sorties observées / sorties modélisées) de 104% contre 82% en utilisant la table réglementaire du BCAC (2010). Dans ces conditions, **la table ajustée semble plus appropriée que la table réglementaire** jusqu'à présent utilisée pour les calculs de provisionnement et les études.

Les espérances de durée résiduelle calculées sur les taux ajustés sont donc présentées dans la **Figure 49** ci-après.

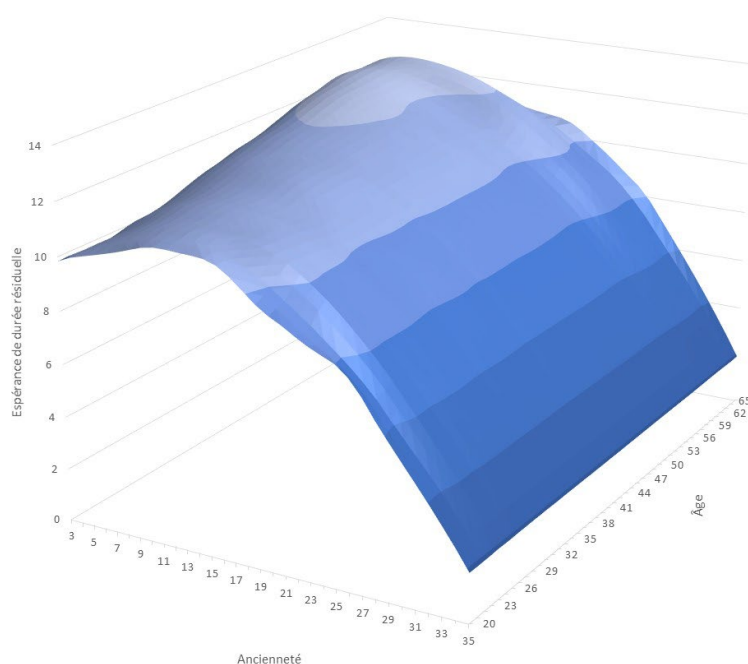


Figure 49 : Espérances de durée résiduelle

L'allure du tracé des espérances de durée résiduelle est très proche de celui du BCAC (2013) ayant servi de référence. Les espérances de durée résiduelle issues des taux ajustés et de la table réglementaire du BCAC (2010) sont comparées dans la **Figure 50** ci-après.

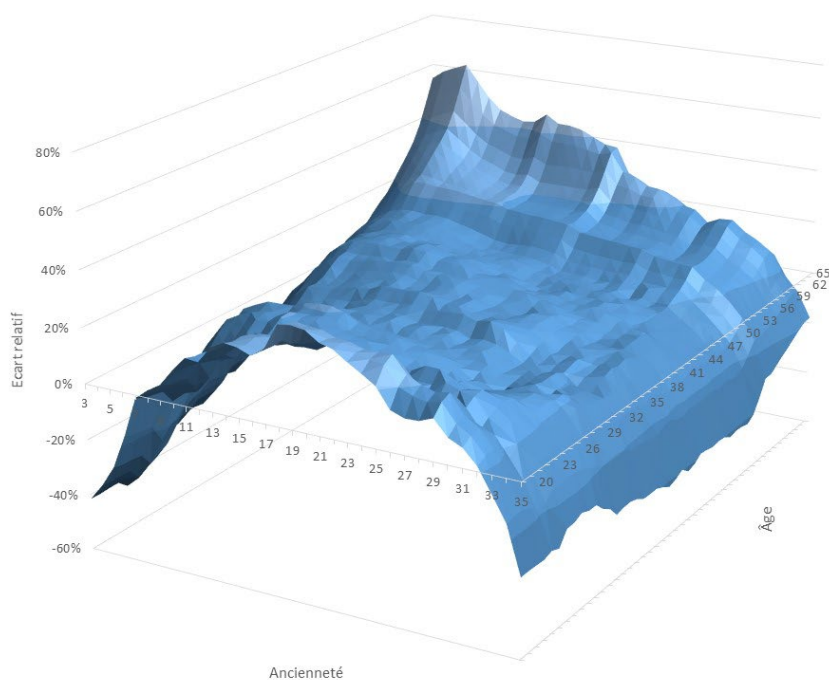


Figure 50 : Comparaison entre les espérances de durée résiduelle

Il apparaît globalement, que l'utilisation de la table réglementaire du BCAC (2010) se traduit par un excès de provisionnement (à taux 0%) de l'ordre de 10% à 15% sur les âges et anciennetés centraux. Les espérances de durée résiduelle de la table ajustée sont cependant plus élevées les trois premiers mois pour les moins de 35 ans, puis les trois derniers mois pour l'ensemble des âges. Cette dégradation sur les anciennetés élevées

s'explique par les écarts entre la table réglementaire du BCAC (2010) par rapport à laquelle la comparaison est effectuée, et la table du BCAC (2013) ayant servi de référence externe au modèle. Sur cette dernière, les taux de sorties conditionnels sur les anciennetés élevées sont nettement plus faibles. Finalement, **l'écart relatif moyen pondéré entre les espérances de durée résiduelle de la table d'expérience et de la table du BCAC (2010) est d'environ 6%.**

En présence d'échantillons de petites tailles, le lissage des taux brut *via* l'utilisation de la méthode de Whittaker-Henderson n'est pas efficace. Le modèle de Brass a permis de calibrer les taux de sortie en utilisant la table du BCAC comme référence externe. **Cette méthode a permis d'apporter une meilleure stabilité au modèle, au prix d'un ajustement imparfait aux données d'origines.** En particulier, la contrainte liée à l'utilisation d'un unique jeu de coefficients de régression a conduit à un ajustement unique sur tous les âges. Ces taux ajustés seront néanmoins plus adaptés à la modélisation de la sinistralité des portefeuilles que ceux issus de la table réglementaire du BCAC (2010). Ces étapes sont récapitulées dans un logigramme communiqué en annexe 4.

5 CONCLUSION

La France est considérée comme un pays offrant un niveau de protection sociale élevé. Pour autant, les incapacités temporaires totales de travail se traduisent généralement par une perte de revenu pouvant atteindre jusqu'à 50%. Elle peut donc perturber l'équilibre économique d'un ménage. Pour cette raison, les assureurs de personnes proposent des garanties prévoyant l'indemnisation partielle ou totale de ces périodes d'incapacité. Dans le cas de l'assurance emprunteur, cette garantie est généralement rendue obligatoire par les créanciers.

Les travaux menés dans ce mémoire ont permis d'approfondir la connaissance des risques portés sur ces garanties dans les portefeuilles de Prévoir Vie. Ils ont notamment mis en exergue l'hétérogénéité des risques, d'une part entre les garanties, mais également entre les populations assurées. En effet, le profil des assurés en portefeuille a significativement évolué ces vingt dernières années compte tenu de l'orientation de la stratégie commerciale. Dans ce contexte, la mise à jour des lois d'entrée et de maintien devenait indispensable. Elle permettra d'affiner le calcul des provisions, qui peuvent représenter des niveaux importants sur ce type de garanties et auxquelles sont donc très attentifs les Dirigeants. La mise à jour des lois d'expérience permettra également d'alimenter les travaux de tarification, particulièrement complexes sur le marché très dynamique des garanties liées à l'arrêt de travail.

L'étude de l'incapacité de travail repose sur des modèles de durée qui nécessitent de disposer de grandes quantités de données pour offrir des résultats fiables. Ce besoin est particulièrement marqué pour l'estimation des durées de maintien. L'utilisation de techniques statistiques spécifiques permet néanmoins de se soustraire, au moins partiellement, à cette problématique.

Concernant l'estimation des taux d'entrée en incapacité, la régression de Cox a permis de tirer profit de l'hétérogénéité du portefeuille. Malgré la relative rareté des données, l'impact de nombreuses variables a pu être évalué en vue d'associer à chaque loi construite, un jeu de coefficients permettant de la déformer pour mieux prendre en compte l'hétérogénéité des risques. Dans le cas de l'estimation des durées de maintien en incapacité, les données étaient insuffisantes pour construire des lois s'étendant sur tous les âges et toutes les anciennetés. Le modèle de Brass a néanmoins permis d'utiliser la table du BCAC (2013) comme référence externe pour les étendre et les lisser. L'ensemble de ces étapes est récapitulé dans un logigramme communiqué en annexe 4. Ces éléments vont à présent permettre d'alimenter des réflexions sur l'amélioration de la construction des tarifs, le calcul des provisions, ou plus généralement, sur les équilibres techniques des portefeuilles.

Mais, ces travaux s'inscrivent dans un environnement qui évolue. En particulier sur l'emprunteur, la loi « Lemoine » a été adoptée pendant la rédaction de ce mémoire. Elle entérine le droit à la résiliation à tout moment et ouvre un droit à l'oubli plus protecteur de l'assuré. Elle met également fin, sous certaines conditions, au questionnaire obligatoire de santé. De futurs travaux seront nécessaires pour en évaluer l'impact sur les risques portés par les compagnies.

Enfin, l'étude du portefeuille a parfois soulevé des interrogations, notamment sur la sincérité des réponses au questionnaire médical chez les travailleurs non-salariés. Sur ce point, la poursuite des travaux semble appropriée. En particulier, la recherche d'outils de détection et de lutte contre la fraude pourra faire l'objet de futurs développements. La construction d'un modèle d'apprentissage supervisé pourrait se traduire par une meilleure maîtrise de la sinistralité sur ces portefeuilles. L'augmentation du niveau de marge, ou la baisse des prix dans un contexte hyper concurrentiel, pourrait alors être un atout pour la Compagnie.

6 BIBLIOGRAPHIE

- Ali Ben Halima, M., Debrand, T., & Regaert, C. (2012, Juin). Arrêts maladie : comment expliquer les disparités départementales ? *Institut de recherche et documentation en économie de la santé (IRDES)*.
- Art. 143-2 ANC. (2015, 11 26).
- Aubin, I., & Rolland, A. (2010). *Lignes directrices de la construction des lois de maintien en incapacité et en invalidité*. Institut des Actuaire.
- Blanpain, N., & Chardon, O. (2011). *Document de travail : Les inégalité sociales face à la mort*. INSEE.
- Cour des comptes. (2006). *La sécurité sociale*.
- Direction de la recherche, des études, de l'évaluation et des statistiques (DREES). (2022). *L'état de santé de la population en France*.
- Giesecke, L. (1981). Use of the chi-square statistic to set Whittaker-Henderson smoothing coefficients.
- IRDES. (2015). *Quel est l'impact du système d'indemnisation maladie sur la durée des arrêts de travail*. Document de travail.
- Kamega, A., Planchet, F., & Wolfrum, R. (2014). Présentation et comparaison des nouvelles tables BCAC.
- Kaplan, E. L., & Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. (A. S. Association, Éd.) *Journal of the American Statistical Association*, 53(282), 457-481.
- Lecoutre, J.-P. (2019). *Statistiques et probabilités* (éd. 7e). Dunod.
- Planchet, F. (2021). Modèles de durée : Méthodes de lissage et d'ajustement. *Support de cours 2021-2022, Version 4.5*. ISFA.
- Planchet, F. (2021). Modèles de durée : Statistique des modèles non paramétriques. *Supports de cours 2021-2022, Version 4.6*. ISFA.
- Planchet, F. (2022). Modèles de durée : Introduction. *Support de cours 2022-2023, Version 4.7*. ISFA.
- Planchet, F., & Thérond, P. (2011). *Modélisation statistique des phénomènes de durée*. Economica.
- scikit-learn.org*. (s.d.). Récupéré sur Scikit-learn: https://scikit-learn.org/stable/modules/cross_validation.html
- Simon, Q. (2018). Modèles semi-paramétriques de survie en temps continu sous R. *Série des documents de travail « Méthodologie Statistique » de la Direction de la Méthodologie et de la Coordination Statistique et Internationale*.

7 ANNEXES

ANNEXE 1 : SORTIE R DE LA REGRESSION DE COX SUR LE PORTEFEUILLE EMPRUNTEUR

```
coxph(formula = Surv(tstart, tstop, tevent) ~ sexe_h:strata(tgroup_sexe) +
      tabac + CSP2:strata(tgroup_CSP2) + CSP3 + CSP4 + quotite:strata(tgroup_quotite) +
      majo_tarif + option_confort + option_confort_plus + inval_med +
      anciennete_13_48 + franchise_30 + franchise_60 + franchise_180 +
      geo_groupe_1 + geo_groupe_2 + geo_groupe_3:strata(tgroup_geo3),
      data = data_april_m1)

n= 515969, number of events= 2935
```

	coef	exp(coef)	se(coef)	z	Pr(> z)
tabac	0.256107	1.291891	0.078141	3.277	0.001047 **
CSP3	0.803619	2.233609	0.058302	13.784	< 2e-16 ***
CSP4	1.609298	4.999301	0.172238	9.343	< 2e-16 ***
majo_tarif	0.478927	1.614342	0.064718	7.400	1.36e-13 ***
option_confort	0.353172	1.423576	0.060859	5.803	6.51e-09 ***
option_confort_plus	0.500899	1.650204	0.040699	12.307	< 2e-16 ***
inval_med	0.444549	1.559787	0.122131	3.640	0.000273 ***
anciennete_13_48	0.287670	1.333317	0.037217	7.730	1.08e-14 ***
franchise_30	1.551815	4.720030	0.121524	12.770	< 2e-16 ***
franchise_60	0.846459	2.331377	0.268290	3.155	0.001605 **
franchise_180	-1.052538	0.349051	0.278160	-3.784	0.000154 ***
geo_groupe_1	0.633341	1.883894	0.052972	11.956	< 2e-16 ***
geo_groupe_2	0.835914	2.306922	0.055446	15.076	< 2e-16 ***
sexe_h:strata(tgroup_sexe)tgroup_sexe=1	-1.478444	0.227992	0.620215	-2.384	0.017137 *
sexe_h:strata(tgroup_sexe)tgroup_sexe=2	-1.249992	0.286507	0.152336	-8.205	2.30e-16 ***
sexe_h:strata(tgroup_sexe)tgroup_sexe=3	-1.011756	0.363580	0.112604	-8.985	< 2e-16 ***
sexe_h:strata(tgroup_sexe)tgroup_sexe=4	-0.448056	0.638869	0.060462	-7.411	1.26e-13 ***
sexe_h:strata(tgroup_sexe)tgroup_sexe=5	-0.253063	0.776419	0.062638	-4.040	5.34e-05 ***
CSP2:strata(tgroup_CSP2)tgroup_CSP2=1	0.646038	1.907966	0.051921	12.443	< 2e-16 ***
CSP2:strata(tgroup_CSP2)tgroup_CSP2=2	0.468154	1.597043	0.066097	7.083	1.41e-12 ***
quotite:strata(tgroup_quotite)tgroup_quotite=1	-0.374032	0.687955	0.045718	-8.181	2.81e-16 ***
quotite:strata(tgroup_quotite)tgroup_quotite=2	0.008327	1.008362	0.118119	0.070	0.943797
geo_groupe_3:strata(tgroup_geo3)tgroup_geo3=1	1.616060	5.033222	0.120585	13.402	< 2e-16 ***
geo_groupe_3:strata(tgroup_geo3)tgroup_geo3=2	1.286143	3.618803	0.087302	14.732	< 2e-16 ***
geo_groupe_3:strata(tgroup_geo3)tgroup_geo3=3	0.982351	2.670729	0.066309	14.815	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Concordance= 0.716 (se = 0.005)
Likelihood ratio test= 1916 on 25 df, p=<2e-16
Wald test = 1882 on 25 df, p=<2e-16
Score (logrank) test = 2102 on 25 df, p=<2e-16

ANNEXE 2 : SORTIE R DE LA REGRESSION DE COX SUR LE PORTEFEUILLE DE PREVOYANCE SALARIES

```
coxph(formula = Surv(tstart, tstop, tevent) ~ niveau_fran_IJ_TC_40 +  
niveau_fran_IJ_TC_90:strata(tgroup_fr90) + niveau_fran_IJ_TC_180 +  
niveau_eur_IJ_1 + niveau_eur_IJ_2:strata(eur_IJ_2) + anc +  
flag_CSP + surprime:strata(tgroup_surprime) + code_postal_group_1 +  
code_postal_group_2 + flag_produit_PI + (flag_exo == 0),  
data = data_strata)
```

n= 144165, number of events= 6382

	coef	exp(coef)	se(coef)	z	Pr(> z)	
niveau_fran_IJ_TC_40	-0.89544	0.40843	0.03738	-23.954	< 2e-16	***
niveau_fran_IJ_TC_180	-1.36987	0.25414	0.33496	-4.090	4.32e-05	***
niveau_eur_IJ_1	0.08302	1.08657	0.03346	2.481	0.01308	*
anc	0.11723	1.12438	0.02955	3.967	7.27e-05	***
flag_CSP	0.16731	1.18212	0.02567	6.518	7.12e-11	***
code_postal_group_1	0.18539	1.20369	0.02966	6.251	4.08e-10	***
code_postal_group_2	-0.39119	0.67625	0.05618	-6.963	3.33e-12	***
flag_produit_PI	0.42790	1.53403	0.03068	13.946	< 2e-16	***
flag_exo == 0TRUE	0.17823	1.19509	0.06213	2.869	0.00412	**
niveau_fran_IJ_TC_90:strata(tgroup_fr90)tgroup_fr90=1	-1.86815	0.15441	0.06025	-31.005	< 2e-16	***
niveau_fran_IJ_TC_90:strata(tgroup_fr90)tgroup_fr90=2	-1.54919	0.21242	0.09474	-16.352	< 2e-16	***
niveau_fran_IJ_TC_90:strata(tgroup_fr90)tgroup_fr90=3	-1.38016	0.25154	0.13538	-10.195	< 2e-16	***
niveau_fran_IJ_TC_90:strata(tgroup_fr90)tgroup_fr90=4	-1.10555	0.33103	0.16545	-6.682	2.35e-11	***
niveau_fran_IJ_TC_90:strata(tgroup_fr90)tgroup_fr90=5	-0.85062	0.42715	0.67943	-1.252	0.21059	
niveau_eur_IJ_2:strata(eur_IJ_2)eur_IJ_2=1	0.31786	1.37418	0.04697	6.768	1.31e-11	***
niveau_eur_IJ_2:strata(eur_IJ_2)eur_IJ_2=2	0.22279	1.24956	0.04627	4.815	1.48e-06	***
niveau_eur_IJ_2:strata(eur_IJ_2)eur_IJ_2=3	0.01163	1.01170	0.05419	0.215	0.83007	
surprime:strata(tgroup_surprime)tgroup_surprime=1	0.41796	1.51887	0.08256	5.063	4.14e-07	***
surprime:strata(tgroup_surprime)tgroup_surprime=2	0.23269	1.26198	0.03439	6.767	1.32e-11	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Concordance= 0.72 (se = 0.003)

Likelihood ratio test= 3845 on 19 df, p=<2e-16

Wald test = 3013 on 19 df, p=<2e-16

Score (logrank) test = 3675 on 19 df, p=<2e-16

ANNEXE 3 : SORTIE R DE LA REGRESSION DE COX SUR LE PORTEFEUILLE DE PREVOYANCE DES TNS

```
coxph(formula = Surv(tstart, tstop, tevent) ~ detail_sexe:strata(flag_produit_SPP_SPPM):(strata(sexe_SPP)) +
  niveau_fran_IJ_TC_TNS30 + niveau_fran_IJ_TC_TNS365 + niveau_eur_IJ_TNS1 +
  niveau_eur_IJ_TNS2:strata(eur_IJ_2) + niveau_eur_IJ_TNS3 +
  code_postal_group_1 + code_postal_group_2 + caisse_activite,
  data = data_strata)

n= 58283, number of events= 1426
```

	coef	exp(coef)	se(coef)	z	Pr(> z)
niveau_fran_IJ_TC_TNS30	-0.46347	0.62910	0.10961	-4.228	2.36e-05 ***
niveau_fran_IJ_TC_TNS365	-1.71813	0.17940	0.44871	-3.829	0.000129 ***
niveau_eur_IJ_TNS1	0.36951	1.44703	0.05783	6.389	1.66e-10 ***
niveau_eur_IJ_TNS3	1.21666	3.37590	0.21420	5.680	1.35e-08 ***
code_postal_group_1	0.16469	1.17903	0.06771	2.432	0.015009 *
code_postal_group_2	-0.20902	0.81138	0.08419	-2.483	0.013037 *
caisse_activite	-0.23113	0.79363	0.07555	-3.059	0.002218 **
niveau_eur_IJ_TNS2:strata(eur_IJ_2)eur_IJ_2=1	1.13180	3.10124	0.12528	9.034	< 2e-16 ***
niveau_eur_IJ_TNS2:strata(eur_IJ_2)eur_IJ_2=2	0.64442	1.90487	0.13909	4.633	3.60e-06 ***
detail_sexe:strata(flag_produit_SPP_SPPM)flag_produit_SPP_SPPM=0:strata(sexe_SPP)sexe_SPP=1	-0.31353	0.73087	0.08830	-3.551	0.000384 ***
detail_sexe:strata(flag_produit_SPP_SPPM)flag_produit_SPP_SPPM=1:strata(sexe_SPP)sexe_SPP=1	0.38261	1.46611	0.11434	3.346	0.000819 ***
detail_sexe:strata(flag_produit_SPP_SPPM)flag_produit_SPP_SPPM=0:strata(sexe_SPP)sexe_SPP=2	-0.20708	0.81296	0.10093	-2.052	0.040204 *
detail_sexe:strata(flag_produit_SPP_SPPM)flag_produit_SPP_SPPM=1:strata(sexe_SPP)sexe_SPP=2	-0.08214	0.92114	0.18107	-0.454	0.650091

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Concordance= 0.614 (se = 0.008)
Likelihood ratio test= 232.2 on 13 df, p=<2e-16
Wald test = 232.1 on 13 df, p=<2e-16
Score (logrank) test = 251.3 on 13 df, p=<2e-16

Logigramme des travaux

