Université Paris-Saclay
École doctorale de mathématiques Hadamard (EDMH, ED 574)

Établissement d'inscription : Université d'Evry-Val d'Essonne

Laboratoire d'accueil : Laboratoire de mathématiques et modélisation d'Évry, UMR 8071 CNRS-INRA

THÈSE DE DOCTORAT ÈS MATHÉMATIQUES

Spécialité : Mathématiques appliquées

Guillaume BIESSY

Modélisation semi-markovienne de la perte d'autonomie chez les personnes âgées : application à l'assurance dépendance

Date de soutenance : Lundi 28 novembre 2016

Après avis des rapporteurs : Michel DENUIT (Université catholique de Louvain)
Christian ROBERT (Université Claude Bernard Lyon 1)

Jury de soutenance :

Agathe GUILLOUX	(Université d'Evry Val d'Essonne) Examinatrice
Vincent LEPEZ	(SCOR Global Life SE) Co-encadrant de thèse
Olivier LOPEZ	(Université Pierre et Marie Curie) Président du jury
Catherine MATIAS	(CNRS) Directrice de thèse
Frédéric PLANCHET	(Université Claude Bernard Lyon 1) Examinateur
Christian ROBERT	(Université Claude Bernard Lyon 1) Rapporteur

NNT : 2016SACLE048

Il y a bien une critique des valeurs et des moyens de la science, mais l'art de trouver (quoiqu'on l'ait baptisé euristique), demeure aussi personnel que tous les autres arts.

Paul Valéry, *Entretiens*

Remerciements

Un grand nombre de personnes ont contribué peu ou prou au succès de cette thèse, et je tiens à saisir cette occasion unique pour les remercier.

En premier lieu, je voudrais remercier ma directrice de thèse Catherine Matias pour sa disponibilité tout au long de ces trois années de thèse. Nos échanges m'ont permis d'avancer sur de nombreux sujets et ont grandement contribué à améliorer la rigueur et la clarté du présent manuscrit. Aussi je lui pardonne volontiers toutes ses ratures et ses remarques au stylo rouge. Je voudrais également remercier Vincent Lepez, mon encadrant au sein de SCOR Global Life. Vincent a tout d'abord réussi à me transmettre son enthousiasme sur le sujet de la dépendance et a ensuite largement contribué à mon intégration au sein de l'entreprise. J'espère que lui me pardonnera les nombreuses déceptions que lui ont donné mes différentes tentatives dans le domaine de la pâtisserie.

Je tiens ensuite à remercier Michel Denuit et Christian Robert qui me font l'honneur de rapporter cette thèse ainsi qu'Olivier Lopez, Agathe Guilloux et Fréderic Planchet qui ont gracieusement accepté de faire partie du jury.

Je voudrais maintenant remercier tous mes collègues passés comme présents au sein de SCOR Global Life. Plus particulièrement, j'aimerais remercier les membres de l'équipe de Recherche et Développement sur l'Assurance Dépendance avec qui j'ai eu la chance de travailler : Ilan Cohen, Laure De Montesquieu, Charlotte Forien (merci également pour la recette de brownies) et Tiziana Torri. Grâce à eux, j'ai pu rester en contact avec les réalités opérationnelles du métier d'actuaire R&D et je l'espère, donner un côté pratique aux résultats obtenus lors de la thèse. Je tiens également à remercier les membres de l'équipe Longévité et plus particulièrement Julien Tomas avec qui j'ai pu avoir de fructueuses discussions sur les méthodes de vraisemblance locale, sans oublier Razvan (notamment pour son aide sur le modèle de Brass), Agne et Jonathan (merci pour la relecture). Un grand merci également à Daria Ossipova-Kachakhidze, responsable des deux équipes citées ci-avant pour l'intérêt qu'elle a bien voulu porter à mes travaux. Je tiens à remercier Yang Lu, ancien doctorant à la SCOR pour nos échanges et j'en profite pour souhaiter bonne chance aux nouveaux doctorants, Nesrine et Samuel, ainsi qu'à Laura (alternante à nulle autre pareille) pour tout ce qui les attend. Merci également aux médecins conseils de la SCOR, et en particulier à Gabriela, John et Pierre. Merci à Sophie pour son aide pour tous les aspects organisationnels.

Je remercie également les membres du LaMME pour leur accueil, et en particulier Michèle pour son aide dans l'organisation des déplacements. Je tiens à remercier les membres du projet Lolita pour les échanges dans le cadre du groupe de travail Longévité et des différents séminaires auxquels j'ai eu la chance de participer.

Je voudrais remercier mes amis : Dominique, Sylvie, Cyril, Cécile, Alexis, Clara, Nicolas, Élodie, Julie, Aude pour leur soutien tout au long de la thèse. Enfin, merci à mes parents à qui je serai à jamais reconnaissant pour leur amour, leur patience et leurs sacrifices.

Résumé

L'allongement de l'espérance de vie observé depuis le début du 20^{e}siècle dans les pays industrialisés pose un certain nombre de défis aux sociétés modernes. Parmi eux celui de la perte d'autonomie chez les personnes âgées, connue également sous le nom de dépendance. La dépendance des personnes âgées se définit comme un état d'incapacité à effectuer seul tout ou partie des Actes de la Vie Quotidienne (AVQ). La dépendance apparaît dans la grande majorité des cas sous l'effet d'une ou plusieurs pathologies chroniques liées au vieillissement. Les personnes concernées ont alors besoin de l'assistance d'une tierce personne, un proche ou un aidant professionnel ou même d'intégrer un Établissement d'Hébergement pour Personnes Âgées Dépendants (EHPAD) dans les cas les plus sévères. En France, une aide publique, l'Allocation Personnalisée pour l'Autonomie (APA), est destinée à couvrir les frais liés à la dépendance. Cependant, le montant des prestations accordées demeure faible devant les dépenses engendrées. Aussi, de nombreux assureurs ont développé des produits spécifiques destinés à compléter l'aide publique fournie par l'APA.

Afin de fixer les prix de ces produits et d'assurer le suivi du risque, les assureurs ont besoin de modéliser le processus de dépendance. Cette modélisation passe dans la majorité des cas par une représentation multi-états du processus dont les états sont l'autonomie, le décès ainsi qu'un ou plusieurs niveaux de dépendance. Pour prédire le risque il est alors nécessaire d'estimer les probabilités de transition entre ces états. Sous l'hypothèse de Markov, on considère que ces probabilités de transition dépendent uniquement de l'état actuel. En ce qui concerne l'étude du risque de dépendance, cette hypothèse peut paraître trop restrictive pour rendre compte de la complexité du phénomène étudié. Dans le cadre semi-markovien, plus général, les probabilités de transition peuvent également dépendre du temps passé dans l'état actuel. Au cours de cette thèse, nous nous attachons à montrer la nécessité d'une modélisation semi-markovienne du processus. Nous mettons ainsi en évidence l'impact du temps passé en dépendance sur les probabilités de décès. Nous montrons par ailleurs que la prise en compte de la diversité induite par les pathologies permet d'améliorer sensiblement l'adéquation du modèle proposé aux données étudiées. Plus encore, nous établissons que la forme particulière de la probabilité de décès en fonction du temps passé en dépendance peut être expliquée par le mélange des groupes de pathologies qui constituent la population des individus dépendants.

Le premier chapitre de la thèse propose une introduction du sujet et des principales méthodes utilisées pour sa quantification. Le deuxième chapitre est consacré à l'étude des probabilités de transitions pour les individus déjà dépendants sur la base de données publiques de l'APA. Dans le troisième chapitre, nous introduisons une démarche entièrement paramétrique pour l'estimation des probabilités de transition dans un modèle avec un seul niveau de dépendance sur la base de données de portefeuilles. Nous prenons notamment en compte le rôle du mélange des groupes de pathologies, quand bien même celles-ci ne sont pas observées. Enfin, le quatrième chapitre est consacré à l'étude des probabilités de transition associées à 4 groupes de pathologies : cancer, maladies neurologiques, démence et autres causes. Cette étude permet ainsi de valider les résultats empiriques établis au cours des chapitres précédents.

Abstract

Alongside the increase in life expectancy observed in developed countries since the beginning of the 20th century , numerous challenges arise for modern societies. Among them the loss of autonomy in elderly people, also known as Long-Term Care (LTC). Long-term care may be defined as a state of incapacity to perform autonomously part of the Activities of Daily Living (ADL). In most cases, long-term care is caused by one or several pathologies linked to aging. Disabled people therefore require help provided by a relative or professional caregiver or may even need to enter a nursing home. In France, a public aid called the *Allocation Personnalisée pour l'Autonomie* (APA), literally customized aid for autonomy, aims at covering the expenses caused by long-term care. Nevertheless, the amount of benefit is relatively small in regards of those expenses. Therefore, many insurers have designed products dedicated to complement the public aid.

In order to price those products and monitor the risk, insurers need to model the long-term care process. In most cases, one rely on multi-state modeling with states autonomy, death and one or several levels of LTC. To predict the risk one has to assess the transition probabilities between states. Under the Markov assumption, those probabilities are considered to only depend on the current state. As regards the study of LTC, this assumption may be seen as too restrictive to account for the complexity of the underlying risk. In a semi-Markov framework, those probabilities may also depend on the time spent in the current state. In this thesis, we emphasis the necessity of the semi-Markov modeling. We demonstrate the impact of time spent in LTC on death probabilities. Besides, we exhibit that taking into account the diversity induced by pathologies leads to sizable improvements in the fit of the model to experience data. Furthermore, we highlight that the peculiar shape taken by death probabilities as a function of time spent in LTC may be explained by the mixture of pathology groups among the disabled population.

The first chapter of this thesis provides an introduction of the long-term care risk and different tools to quantify it. The second chapter focuses on death probabilities among disabled, using the APA database. In the third chapter, we introduce a fully parametric approach to estimate transition probabilities in a model with a single state of LTC, relying on data from an insurance portfolio. Lastly, the fourth chapter study transition probabilities for 4 distinct groups of pathologies: cancer, neurological diseases, dementia and other causes. This validates the empirical results obtained in the previous chapters.

Table des matières

Chapitre 1

Introduction

1.1 Contexte

1.1.1 Le risque dépendance

L'allongement de l'espérance de vie observé depuis le début du 20^{e}siècle dans les pays industrialisés pose un certain nombre de défis aux sociétés modernes. Parmi eux celui de la perte d'autonomie chez les personnes âgées, connue également sous le nom de dépendance. La dépendance des personnes âgées se définit comme un état d'incapacité à effectuer seul tout ou partie des Actes de la Vie Quotidienne (AVQ). La définition de ces AVQ et leur nombre varie selon les pays. En France, les plus courants sont : se nourrir, se laver, se déplacer, s'habiller. La dépendance apparaît dans la grande majorité des cas sous l'effet d'une ou plusieurs pathologies chroniques liées au vieillissement. Les personnes concernées ont alors besoin de l'assistance d'une tierce personne, un proche ou un aidant professionnel ou même d'intégrer un Établissement Hospitalier pour Personnes Âgées Dépendants (EHPAD) dans les cas les plus sévères. En France, une aide publique, l'Allocation Personnalisée pour l'Autonomie (APA), est destinée à couvrir les frais liés à la dépendance. Cependant, le montant des prestations accordées demeure faible devant les dépenses engendrées surtout pour les personnes intégrant un EHPAD dont le coût peut dépasser 4 000 € par mois. Devant ce risque financier auquel une part croissante de la population se retrouve ainsi exposée, de nombreux assureurs ont développé des produits spécifiques destinés à compléter l'aide publique fournie par l'APA.

1.1.2 L'offre dépendance en France

En France, l'offre d'assurance dépendance se caractérise par un certain nombre de spécificités. Les produits d'assurance individuels vendus prévoient en général le versement de cotisations par l'assuré tant qu'il est autonome, le montant de ces cotisations étant fixé au cours du temps et déterminé lors de la souscription du produit. En contrepartie l'assuré dépendant reçoit une rente de l'entrée en dépendance jusqu'au décès. Cette approche est dite forfaitaire et la rente versée à l'assuré peut ainsi être utilisée à la discrétion de celui-ci. Cette approche est à opposer aux produits indemnitaires développés par exemple aux États-Unis pour lesquels l'assureur prend en charge les frais liés à la dépendance des assurés se qualifiant à concurrence d'un certain montant défini lors de la souscription. Autre différence notable, la dépendance est définie par un grand nombre d'assureurs français comme un état d'incapacité *consolidé* et *permanent*. Cette notion d'irréversibilité est absente des définitions proposées dans d'autres grand marchés de l'assurance dépendance dans le monde comme les États-Unis, Israël ou Singapour. En contrepartie, le versement de la rente ne s'arrête qu'au décès de l'assuré, à la différence des trois exemples cités ci-dessus. Au niveau de la définition du niveau de dépendance requis pour bénéficier des

prestations, celui-ci se base principalement sur 4 AVQ mentionnés ci-avant : se nourrir, se laver, se déplacer, s'habiller. L'état de dépendance totale, qui donne droit au montant maximum de prestation, requiert en général l'incapacité d'effectuer 3 de ces 4 AVQ. Les produits les plus récents proposent aujourd'hui un montant plus faible pour la dépendance partielle définie comme la perte de 2 AVQ sur 4, de l'ordre de 50 à 60 % de la rente en dépendance totale. Notons également que la grille AGGIR (pour Autonomie Gérontologie Groupes Iso Ressources), une définition alternative de la dépendance utilisés par les pouvoirs publics pour le versement de l'APA, est également utilisée par certains assureurs, seule ou en complément de la définition avec des AVQ, auquel cas l'assuré doit se qualifier selon les deux définitions pour bénéficier des prestations. On trouvera une présentation plus détaillée de la grille AGGIR dans le Chapitre 2. Certains produits proposent également en option le versement d'un capital lors de l'entrée dans un état de dépendance partielle. Celui-ci est destiné à couvrir les frais d'équipement du logement de l'assuré pour faire face à la situation de dépendance.

1.1.3 L'estimation du risque, un réel défi

Afin de fixer le prix des produits et d'assurer un suivi efficace de l'évolution du risque, les actuaires doivent être capables de modéliser de manière assez fine le processus de dépendance. Cette tâche est rendue très ardue par plusieurs facteurs. Tout d'abord, il s'agit d'un processus très complexe, dont les causes, à savoir les pathologies liées au vieillissement, sont très variées. Parmi ces causes, les plus fréquentes sont la maladie d'Alzheimer, le cancer, les maladies cardiovasculaires et les maladies neurologiques. Ensuite, le volume et la qualité des données disponibles sont en général très limités. En effet, il faut noter que les premiers produits dépendance ne sont apparus en France que vers la fin des années 1980. Or la dépendance est un phénomène qui concerne majoritairement les personnes très âgées. Nos travaux indiquent ainsi que l'âge médian d'entrée en dépendance se situe entre 85 et 90 ans. Les produits d'assurance sont souscrits en moyenne à l'âge de 65 ans, ce qui signifie qu'il faut donc attendre plus de 25 ans avant d'obtenir un retour d'expérience satisfaisant sur le risque assuré. Par ailleurs, à l'inverse du risque de mortalité, la détermination de l'état de dépendance d'un assuré n'est pas une chose aisée. Les définitions utilisées par les assureurs ont beaucoup évolué depuis les premiers contrats, de même que leur politique de sélection médicale à l'entrée du contrat et leur politique de gestion des sinistres. A cela s'ajoute la propre évolution du risque dans le temps sous l'effet des progrès de la médecine et de facteurs socioéconomiques. De ce fait, il est particulièrement difficile et risqué d'agréger différentes sources afin d'augmenter le volume des données disponibles.

Ces effets permettent de comprendre pourquoi la littérature sur le sujet de la dépendance est peu fournie en applications. Les modèles multi-états ont été identifiés très tôt comme un outil adapté pour étudier ce risque. le lecteur interessé pourra ainsi se référer à Haberman and Pitacco (1998) ou Denuit and Robert (2007). Des solutions alternatives existent comme par exemple l'utilisation de modèles à risques proportionnels (introduits dans Cox, 1972) utilisés par exemple par Czado and Rudolph (2002) pour étudier l'effet du sexe et du niveau de dépendance sur le risque à partir de données de portefeuilles allemandes. Néanmoins, sur les données dont nous disposons, l'hypothèse de proportionnalité des risques n'est pas vérifiée, ce qui, compte-tenu de la complexité du phénomène considéré, n'a rien d'étonnant. Dans cette thèse, nous considérons un cadre multi-états que nous allons maintenant introduire.

1.2 Modélisation multi-états

Considérons un processus en temps continu $(Z_u)_{u\geqslant 0}$ càdlàg (continu à droite, limite à gauche) qui prend ses valeurs dans un ensemble d'états $E = \{e_1, e_2, \ldots, e_p\}$ et notons $(\mathcal{F}_u)_{u\geqslant 0}$ la filtration engendrée par ce processus.

Dans le contexte de la dépendance, ce processus représente l'état de santé de l'assuré. Ce dernier peut être autonome, décédé, ou dans un état de dépendance plus ou moins sévère. La variable d'indexation du processus u correspond quant à elle à l'âge de l'assuré et c'est ainsi que nous la désignerons dorénavant. Pour une introduction plus complète de la théorie markovienne et ses applications en assurance, le lecteur intéressé pourra se référer à Janssen and Manca (2007).

1.2.1 Notion de processus markovien

Le processus $(Z_u)_{u\geqslant 0}$ est dit markovien s'il vérifie la relation suivante dite propriété de Markov pour tout $0 \leqslant x \leqslant y$ et tout $e \in E$

$$\mathbb{P}(Z_y = e|\mathcal{F}_x) = \mathbb{P}(Z_y = e|Z_x).$$

En termes moins formels, pour un processus markovien l'état futur ne dépend pas du passé si l'état actuel est connu. Aussi, les processus markoviens sont dit sans mémoire. Pour $i, j \in E$ et $0 \leqslant x \leqslant y$, notons

$$p_{i,j}(x, y) = \mathbb{P}(Z_y = j|Z_x = i)$$

les probabilités de transition du processus. Dans le cas où $p_{i,j}(x, y)$ ne dépend de (x, y) qu'à travers la quantité $y - x$, le processus markovien est dit homogène. En assurance vie, les risques étudiés dépendent généralement de l'âge des assurés et il est rare que cette hypothèse soit vérifiée. Néanmoins, l'homogénéité par morceaux du processus constitue dans la plupart des cas une approximation raisonnable, à la clé de certaines méthodes d'estimation dont l'une est présentée dans la suite de ce chapitre.

Les processus markoviens ont été utilisés pour les études des risques biométriques dont participe la dépendance. Ainsi la tarification des premiers produits d'assurance lancés sur le marché français reposait sur ces modèles comme en témoigne SCOR (1995). Ce rapport fournit un aperçu assez fidèle des bonnes pratiques de l'époque, qui n'ont encore aujourd'hui pas beaucoup changé chez certains assureurs. Il est vrai que ces modèles offrent un cadre théorique très simple pour l'estimation des probabilités de transition entre états. Dans le milieu académique, les modèles markoviens ont fait l'objet de publications récentes comme les travaux de Pitacco (2015). Cependant, l'hypothèse de Markov est très vite mise en défaut lorsque l'on s'intéresse aux probabilités de transition de l'état de dépendance vers le décès. En effet, on observe que la probabilité de décès des individus dépendants varie très fortement en fonction du temps passé dans l'état de dépendance. Elle est maximale juste après l'entrée dans cet état et décroit ensuite très rapidement au cours des mois qui suivent. Au cours de cette thèse, nous nous attacherons à expliquer ce phénomène, selon nous lié à l'évolution de la part des pathologies dans la population des dépendants au cours du temps. Cet effet n'est généralement pas capté dans le cadre markovien, pour lequel l'impact du temps passé sur le risque est le même, qu'il s'agisse de temps passé dans un état d'autonomie ou de dépendance. Il est néanmoins possible d'intégrer le temps passé en dépendance comme une covariable à travers un modèle à risques proportionnels comme le fait Helms et al. (2005). Ces auteurs se proposent ainsi d'introduire un coefficient pour chaque année successive passée dans l'état de dépendance. Néanmoins, cette approche conduit à introduire un nombre élevé de paramètres, ce qui, compte-tenu du nombre limité de données disponible, conduit à une grande volatilité des coefficients calculés. Par ailleurs, elle ne permet pas de prendre en compte les effets croisés de l'âge et du temps passé en dépendance. La prise en compte du temps passé en dépendance dans les modèles se fait de manière naturelle dans le cadre semi-markovien que nous allons maintenant introduire.

1.2.2 Notion de processus semi-markovien

Reprenons les notations précédentes et introduisons des processus $(T_n)_{n\in\mathbb{N}}$ et $(X_n)_{n\in\mathbb{N}}$ définis comme suit :

$$T_n = \min\{u > T_{n-1} | Z_u \neq Z_{T_{n-1}}\}$$

et

$$X_n = Z_{T_n}$$

pour $n \in \mathbb{N}^*$ en posant $T_0 = 0$. Par commodité, introduisons également le processus de comptage $(N_u)_{u\geqslant 0}$ qui correspond au nombre de transitions effectuées par le processus à la date u :

$$N(u) = \text{Card}\{k \in \mathbb{N}^* | T_k \leqslant u\}.$$

La suite X_n correspond aux états successifs occupés par le processus et T_n est la suite des instants de sauts du processus. Par définition, le processus ne change pas d'état entre deux instants de sauts. La connaissance du processus Z_x est donc équivalente à celle des processus X_n et T_n.

En utilisant ces notations, le processus $(Z_u)_{u\geqslant 0}$ est dit semi-markovien si $(X_n, T_n)_{n\in\mathbb{N}}$ est markovien. Pour un processus semi-markovien, l'état futur ne dépend pas du passé si l'état actuel et le temps depuis la dernière transition sont connus. On pourra trouver une présentation plus détaillée des processus semi-markoviens dans Cinlar (1969). Dans le cas semi-markovien, on peut également définir des probabilités de transition

$$p_{i,j}(x, y, z) = \mathbb{P}(Z_z = j | Z_y = i, T_{N(y)} = x)$$

ainsi que des probabilités associées au prochain changement d'état

$$\overline{p}_{i,j}(x, y, z) = \mathbb{P}(X_{N(y)+1} = j, T_{N(y)+1} \leqslant z | Z_y = i, T_{N(y)} = x).$$

La donnée de l'une ou l'autre de ces probabilités permet de définir l'évolution du modèle. Si les probabilités de transition ne dépendent de x, y et z qu'à travers les quantités $y - x$ et $z - y$ le processus est dit semi-markovien homogène. Notons que les probabilités de transitions peuvent faire intervenir plusieurs changements d'état, ce qui les rend en général plus difficiles à estimer ou à interpréter que les probabilités associées au prochain changement d'état.

1.2.3 Noyau semi-markovien

On définit le noyau semi-markovien $Q_{i,j}(x, t)$ par

$$Q_{i,j}(x, t) = \mathbb{P}(X_{N(x)+1} = j, T_{N(x)+1} - T_{N(x)} \leqslant t | X_{N(x)} = i, T_{N(x)} = x)$$

pour $(i, j) \in E$, $0 \leqslant x$ et $0 \leqslant t$.

Le noyau semi-markovien s'interprète comme la probabilité, vue à l'instant d'arrivée au temps x dans l'état i, que la prochaine transition ait lieu vers l'état j et avant qu'une durée t ne se soit écoulée. La donnée de la loi initiale et du noyau semi-markovien permet de décrire entièrement le processus. De plus le noyau semi-markovien admet une décomposition de la forme

$$Q_{i,j}(x, t) = s_{i,j}(x) F_{i,j}(x, t)$$

où

$$s_{i,j}(x) = \mathbb{P}(X_{N(x)+1} = j | X_{N(x)} = i)$$

est appelée probabilité de saut et

$$F_{i,j}(x, t) = \mathbb{P}(T_{N(x)+1} - T_{N(x)} \leqslant t | X_{N(x)} = i, X_{N(x)+1} = j)$$

est la fonction de répartition du temps de séjour connaissant le prochain état. On notera que les probabilités de saut peuvent être vues comme les probabilités de transitions du processus X_n qui est un processus de Markov en temps discret, appelé également chaîne de Markov. On pourra se référer à Limnios and Oprisan (2012) pour une étude plus approfondie du noyau semi-markovien.

1.2.4 Intensité de transition

On définit les intensités de transition du modèle par

$$\mu_{i,j}(x,t) = \lim_{h\to 0}\frac{1}{h}\mathbb{P}\left(Z_{x+t+h} = j | Z_{x+t} = i, T_{N(x+t)} = x\right).$$

On a ainsi

$$\mathbb{P}\left(Z_{x+t+h} = j | Z_{x+t} = i, T_{N(x)} = x\right) = \mu_{i,j}(x,t)h + o(h).$$

Lorsque $h \to 0$, cette probabilité de transition est ainsi équivalente à $\mu_{i,j}(x,t)h$ ce qui lui vaut le nom de probabilité instantanée de transition. Notons toutefois que l'intensité de transition n'est pas à proprement parler une probabilité et peut en effet prendre des valeurs supérieures à 1.

Le noyau semi-markovien peut par ailleurs s'exprimer en fonction des intensités de transition du modèle

$$Q_{i,j}(x,t) = \int_{u=0}^{t} \mu_{i,j}(x,u) \exp\left(-\int_{v=0}^{u}\sum_{k\neq i}\mu_{i,k}(x,v)\mathrm{d}v\right)\mathrm{d}u.$$

Ainsi l'intensité de transition permet au même titre que celui-ci de définir l'évolution du processus.

1.2.5 Approches de modélisation possibles

Les processus semi-markoviens sont depuis longtemps utilisés dans les domaines de l'épidémiologie pour étudier différents processus médicaux complexes comme l'infection par le VIH (Mathieu, 2006) ou la réaction à une greffe rénale (Foucher et al., 2007). Cependant, les applications à l'étude du risque de dépendance sont peu courantes, et ce malgré de grandes similitudes avec les processus mentionnés ci-dessus. A partir des quantités introduites ci-dessus, on peut dégager plusieurs grandes approches de modélisation. Helms et al. (2005) propose une estimation directe des probabilités de transition dans un cadre markovien mais qui pourrait être étendu au cadre semi-markovien. Dans la pratique, les actuaires procédant à l'étude du risque s'intéressent en général aux probabilités associées au prochain changement d'état. Celles-ci sont souvent estimées grâce à des méthodes non paramétriques inspirées de l'étude de la mortalité. On pourra trouver une présentation très concrète de telles méthodes dans Hardy et al. (2011). Notons cependant que les formules utilisées pour la tarification des produits font apparaître les probabilités de transition et non les probabilités associées au prochain changement d'état, ce qui n'est pas sans provoquer une certaine confusion. Il est également possible d'adopter une approche basée sur la modélisation paramétrique du noyau semi-markovien, directement inspirée du domaine de l'épidémiologie et peu usitée dans le cadre de la dépendance. Elle a néanmoins été choisie par Lepez et al. (2013) pour la définition d'un modèle à plusieurs niveaux de dépendance basé sur les données de l'APA. Cet auteur fournit ainsi une forme paramétrique pour chacune des deux composantes du noyau, à savoir les probabilités de sauts et les lois de durées conditionnelles associées. Nous reprendrons et étendrons cette approche dans le Chapitre 2 de la présente thèse. Enfin, il est possible d'estimer directement les intensités de transition du modèle. Notons que dans cette dernière approche n'y a plus de confusion possible entre probabilités de transition et probabilités associées au prochain changement d'état. En effet, dans une

modélisation en temps continu, la probabilité d'observer deux transitions au cours du même pas de temps est négligeable. L'emploi de modèles en temps continu reste cependant très rare dans le domaine de l'actuariat où l'utilisation de tables est prépondérante. Dans le Chapitre 3, nous employons une approche reposant sur une modélisation paramétrique des intensités de transition d'un modèle à 3 états. Dans le Chapitre 4, nous adoptons une approche non-paramétrique par maximum de vraisemblance local toujours dans le but d'estimer les intensités de transition du même modèle. Nous allons maintenant sortir du cadre général de cette section pour introduire sans plus tarder le modèle *illness-death* qui fera l'objet de beaucoup d'attention dans ces deux chapitres.

1.3 Cas particulier du modèle *illness-death*

Considérons de nouveau un processus en temps continu $(Z_u)_{u\geqslant 0}$ *càdlàg* mais supposons cette fois que le processus prend ses valeurs dans un ensemble à 3 états $E = \{A, I, D\}$, où A correspond à l'autonomie, I (comme *illness*) correspond à la dépendance, défini à l'aide d'un niveau unique et D au décès. Ce modèle dont on pourra trouver une introduction complète dans Pitacco (2014) est connu sous le nom de modèle *illness-death*. Il est utilisé dans le domaine de l'assurance pour la modélisation de nombreux risques, à savoir l'incapacité, la dépendance, l'hospitalisation, et le chômage. Le modèle *illness-death* est l'un des modèles multi-états les plus simples utilisables pour l'étude du risque de dépendance. Il s'avère néanmoins d'un grand intérêt pratique. Ce modèle est avant tout utilisé pour la tarification des produits proposant un seul niveau de garantie. Dans le cas des produits proposant plusieurs niveaux de garantie, correspondant à plusieurs niveaux de dépendance, chaque garantie peut être évaluée séparément grâce à un modèle *illness-death*. Cette approche présente comme principale limite de ne pas tenir compte du lien entre les différentes garanties. Dans une telle modélisation, la probabilité de tomber en dépendance lourde sera la même pour un individu autonome et un individu en état de dépendance partielle, ce qui n'est pas conforme à la réalité. Une représentation du modèle et des transitions possibles dans le cas le plus général est donné par la Figure 1.1. Il n'y a pas de transitions possibles à partir de l'état de décès aussi cet état est dit absorbant. Dans le cadre du risque de dépendance, il est possible de considérer que la transition de la dépendance vers l'autonomie a une probabilité nulle, ce que nous nous attachons maintenant à justifier.

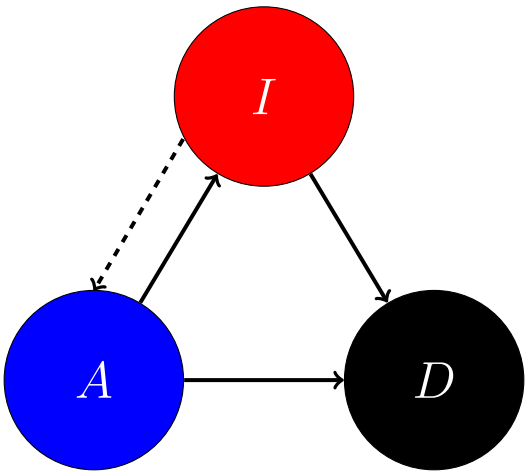


FIGURE 1.1 – Le modèle *illness-death* et ses transitions possibles.

1.3.1 Sur l'hypothèse de non-retour vers l'état d'autonomie

L'hypothèse de non retour vers l'autonomie peut être défendue par plusieurs arguments de natures différentes. En premier lieu, notons que dans la définition adoptée par une très large

majorité des assureurs français, la dépendance est un état de perte d'autonomie consolidé et irréversible. Il n'est bien sûr pas toujours évident d'apporter un diagnostic fiable, surtout dans le cas des maladies neurologiques ou de la démence. Cependant, les retours vers l'autonomie restent exceptionnels, surtout pour ce qui concerne les états de dépendance les plus sévères. A cette définition, il convient d'opposer celle utilisée par les assureurs sur d'autres marchés comme aux États-Unis pour laquelle l'irréversibilité n'est plus nécessaire à l'attribution des prestations. Ainsi, une personne qui aurait temporairement besoin d'un aidant par exemple en raison d'une immobilisation prolongée suite à une chute se qualifierait comme dépendante aux États-Unis mais pas en France. Du fait de la différence entre les deux définitions, le nombre de retours dans le modèle français sera donc sensiblement plus faible que celui observé sur d'autres marchés. Autre différence majeure entre les deux marchés, les prestations en France sont forfaitaires : un montant de rente est versé à l'assuré dépendant tous les mois et l'utilisation qui en est faite est à la discrétion de celui-ci. Dans le modèle américain, les prestations sont indemnitaires, et l'assureur s'engage à rembourser les frais liés à la dépendance, à concurrence d'un certain montant défini dans le contrat. Principale conséquence de ce choix, pour un assuré retrouvant son autonomie aux États-Unis, le versement des prestations cesserait de lui-même. En France, pour faire cesser ce versement, l'assureur devrait être capable de prouver le retour à l'autonomie de l'assuré. Cela nécessiterait de contrôler de manière périodique le niveau de dépendance de l'assuré, ce qui aurait un coût prohibitif pour l'assureur. Sans contrôle, l'assureur français se doit alors de verser des prestations à l'assuré dépendant jusqu'à son décès. Dans cette situation, la probabilité de retour n'intervient pas dans la tarification du produit, aussi n'est-il pas nécessaire de la prendre en compte pour obtenir le prix du produit. Par ailleurs, notons que la prise en compte de cette transition accroît sensiblement la complexité du modèle, en le rendant cyclique. En vertu de tous ces éléments, la prise en compte du retour vers l'autonomie dans le modèle, facultative du point de vue de la tarification, ne devrait être envisagée que si elle permet d'aboutir à une meilleure compréhension du risque. Comme nous l'avons souligné, les retours vers l'autonomie sont peu fréquents, sous-déclarés et la prise en compte de ces retours complexifie beaucoup le modèle. Aussi, de notre point de vue, ces éléments justifient de considérer une probabilité de retour nulle dans le modèle, ce que nous ferons dorénavant.

1.3.2 Intensités de transition du modèle

Sous l'hypothèse de non-retour, il reste trois transitions possibles dans le modèle : l'entrée en dépendance, le décès des autonomes et celui des dépendants. Dans le cadre le plus général, les intensités de transitions dépendent de tout l'historique du processus. Lors de la section précédente, nous avons montré que cet historique pouvait être représenté par la suite des états visités d'une part et la suite des instants de transition d'autre part. Dans le cas du modèle à 3 états et pour les individus toujours vivants, cet historique se limite donc à un maximum de 2 états visités dont le dernier est l'état actuel du processus, à l'âge d'entrée dans ces états et à l'âge actuel. Dans le cadre de l'assurance individuelle qui nous intéresse ici, les assurés sont soumis à un processus de sélection qui permet de s'assurer que ceux-ci sont autonomes à la souscription. L'âge d'entrée dans l'état d'autonomie est donc l'âge à la souscription du produit. Il est légitime de s'interroger sur le rôle éventuel joué par cet âge à la souscription sur le risque. En première analyse, celui-ci a un impact puisque que le rôle de la sélection médicale est de vérifier que l'assuré ne présente pas de problèmes pouvant favoriser l'apparition de la dépendance. La probabilité d'un individu sélectionné de tomber en dépendance aux cours des années suivant la souscription sera donc *a priori* plus faible que pour un individu autonome tiré au sort dans la population française. Cet effet s'atténue cependant avec le temps écoulé depuis l'instant de souscription. On peut alors considérer que passé une dizaine d'années, l'âge de souscription n'affecte plus le risque. Deux individus ayant souscrit l'un à 60 ans et l'autre à 70 ans, âgés tous les deux aujourd'hui de 80 ans auront ainsi le même risque. La vérification de

cette hypothèse est cependant très délicate à cause d'autres facteurs pouvant jouer sur le risque, aussi nous serons ici contraints de l'admettre. Sous ses conditions, les probabilités de transition des autonomes ne dépendent que de l'âge actuel, et la probabilité de décès des dépendants dépend de l'âge d'entrée en dépendance et de l'âge actuel uniquement. Comme le temps passé en dépendance s'obtient comme la différence entre l'âge actuel et l'âge d'entrée en dépendance, ce modèle est alors nécessairement, dans le cadre le plus général, semi-markovien non homogène.

Nous noterons x le temps passé dans l'état d'autonomie depuis la naissance de l'individu et t le temps passé dans l'état dépendance. De cette manière, x représente à la fois l'âge actuel de l'individu autonome et l'âge d'entrée en dépendance pour l'individu dépendant. De cette manière, on peut introduire les intensités de transition du modèle comme des fonctions de x et de t. On notera λ, μ_a et μ_i les intensités de transition correspondant à l'entrée en dépendance, le décès des autonomes et le décès des dépendants respectivement

$$\begin{aligned}
\mu_a(x) =& \lim_{h\to 0} \frac{1}{h}\mathbb{P}\left(Z_{x+h} = D | Z_x = A\right), \\
\lambda(x) =& \lim_{h\to 0} \frac{1}{h}\mathbb{P}\left(Z_{x+h} = I | Z_x = A\right), \\
\mu_i(x,t) =& \lim_{h\to 0} \frac{1}{h}\mathbb{P}\left(Z_{x+t+h} = D | Z_{x^-} = A, Z_x = I, Z_{x+t} = I\right).
\end{aligned}$$

Une représentation schématique du processus et ses intensités de transition est donnée par la Figure 1.2. Nous allons maintenant présenter une méthodologie pour l'estimation des intensités de transition du modèle et leur utilisation à des fins de tarification.

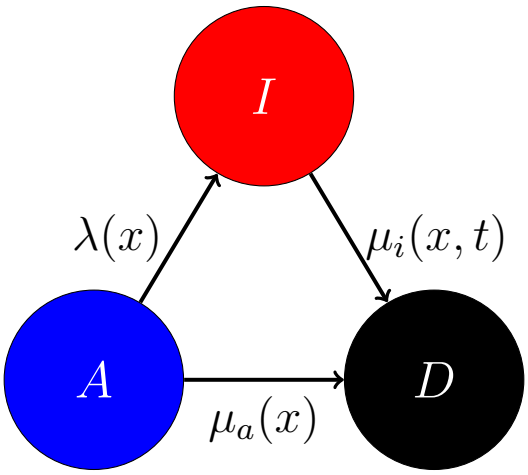


FIGURE 1.2 – Le modèle *illness-death* sans retour possible, adapté au cas de l'assurance dépendance.

1.4 Estimation du risque

1.4.1 Structure des données

Avant de présenter différentes méthodes pouvant être utilisées pour l'estimation du risque, nous allons au préalable introduire la structure des données nécessaire à l'usage de ces méthodes. Dans la pratique, les données récoltées au sein des différentes compagnies d'assurance sont issues d'un ou plusieurs systèmes de gestion. La forme sous laquelle sont stockées ces données et leur qualité varie d'un portefeuille à l'autre. Par soucis de concision, nous ne détaillerons pas les différentes étapes permettant de résoudre les problèmes rencontrés sur les données comme la présence de valeurs manquantes, les valeurs erronées et la présence de doublon dans la base. Nous nous bornerons à noter qu'il n'existe pas de méthode universelle pour réaliser cette étape qui nécessite à la fois du bon sens et de la patience. Notons toutefois que certaines données

ne permettent pas de reconstituer un historique complet des trajectoires des assurés. C'est le cas par exemple lorsque la cause de sortie de l'état d'autonomie n'est pas précisée où lorsque les données ne fournissent pas d'historique des données mais des photographies du portefeuille à plusieurs dates données. Cette dernière situation porte le nom de censure par intervalles et des méthodologies spécifiques devront être utilisées. Ces situations sont abordées dans Satten and Sternberg (1999) pour un cadre multi-états acyclique et par Joly et al. (2002) dans le cas particulier du modèle *illness-death*.

Dans la plupart des cas, les assureurs disposent de deux bases, éventuellement fusionnées. La base des cotisants, notée C, regroupe les informations disponibles sur tous les assurés, jusqu'à leur entrée en dépendance, leur décès ou à l'arrêt du versement des cotisations. Cette base permet d'estimer les intensités de transition pour l'entrée en dépendance et le décès des autonomes. Elle contient *a minima* les informations suivantes

- La date de naissance notée DN,
- La date de souscription notée DS,
- La date de fin de versement des cotisations notée DF,
- La cause de fin de versement des cotisations notée CF,
- Le sexe noté S.

Nous fixons par ailleurs une période d'observation (τ_d, τ_f) sur laquelle sera réalisée l'étude des données. Soit n_C le nombre d'individus présents dans la base. A partir de ces quantités, on construit pour chaque $p \in \{1, \dots, n_C\}$ les variables suivantes utilisées directement dans l'estimation des lois

- L'âge de début d'observation x_p défini par $x_p = \max(\mathrm{DS}_p, \tau_d) - \mathrm{DN}_p$,
- L'âge de fin d'observation y_p défini par $y_p = \min(\mathrm{DF}_p, \tau_f) - \mathrm{DN}_p$,
- La cause de fin d'observation c_p valant 1 pour les décès, 2 pour les entrées en dépendances et 0 pour toutes les autres causes ;
- Le sexe de l'individu g_p (g pour gender) valant 1 pour les hommes et 2 pour les femmes.

On peut alors écrire symboliquement $C = (x_p, y_p, c_p, g_p)_{1 \leqslant p \leqslant n_C}$. Ces informations seront celles utilisées pour l'estimation des intensités d'entrée en dépendance et de décès des autonomes.

La base des rentiers, notée R, rassemble les informations sur les assurés dépendants recevant des prestations sous forme de rente jusqu'à leur décès. Elle contient *a minima* les informations suivantes

- La date de naissance notée notée DN,
- La date d'entrée en dépendance notée DE,
- La date de fin de versement de la rente notée DF,
- La cause de fin de versement de la rente notée CF,
- Le sexe noté S.

Nous fixons par ailleurs une période d'observation (τ_d, τ_f) sur laquelle sera réalisée l'étude des données. Notons que l'on peut ici choisir une période différente de celle considérée pour la base des cotisants. Il est nécessaire de disposer d'un volume suffisant d'informations pour l'estimation du modèle. En même temps, si la période est trop longue, l'existence d'une tendance d'évolution

du risque dans le temps risque de se faire sentir et les données obtenues au début et à la fin de la période ne correspondront pas exactement au même niveau de risque observé. Le choix de la période d'observation doit permettre de faire un compromis entre ces deux nécessités. Pour la base des cotisants, on dispose en général d'un volume de données conséquent et le choix d'une période de 5 ou 10 ans est en général bien adapté. Pour la base des rentiers, comme le volume est plus faible, il faudra en général considérer une période supérieure à 10 ans ou prendre en compte tout l'historique. Soit n_R le nombre d'individus présents dans la base. A partir de ces quantités, on construit pour chaque $p \in \{1, \dots, n_R\}$ les variables suivantes utilisées directement dans l'estimation des lois

- L'âge de début d'observation x_p défini par $x_p = \mathrm{DE}_p - \mathrm{DN}_p$,
- L'âge de fin d'observation y_p défini par $y_p = \min(\mathrm{DF}_p, \tau_f) - \mathrm{DN}_p$,
- La cause de fin d'observation c_p valant 1 pour les décès et 0 pour toutes les autres causes ;
- Le sexe de l'individu g_p (g pour gender) valant 1 pour les hommes et 2 pour les femmes.

On peut alors écrire symboliquement $R = (x_p, y_p, c_p, g_p)_{1 \leqslant p \leqslant n_R}$. Ces informations seront celles utilisées pour l'estimation de la loi de décès des dépendants.

Il est possible d'inclure d'autres variables explicatives du risque dans l'une ou l'autre des bases, de même qu'il est possible de distinguer d'autres causes de sortie. Dans le Chapitre 4, nous prendrons en compte le groupe de pathologie comme variable explicative de l'intensité de décès des dépendants. Ici, nous avons uniquement introduit le sexe car il s'agit d'une information très fréquente dont l'impact sur le risque est avéré. S'il est envisageable de tenir compte de cet impact à l'aide d'un modèle à risque proportionnels tel qu'introduit par Cox (1972), l'hypothèse de proportionnalité n'est quand à elle pas vérifiée en pratique. En effet, le sexe a non seulement un impact sur chacune des lois mais conditionne également la distribution des pathologies sous-jacentes. On observe ainsi plus de cancer chez les hommes et plus de cas de démence chez les femmes. L'impact du sexe sur le risque est donc beaucoup plus complexe qu'un simple coefficient multiplicatif. Dorénavant nous omettrons la variable sexe pour ne pas alourdir les notations, étant entendu que nous estimerons de manière indépendante les intensités de transition pour les hommes et pour les femmes. Nous allons maintenant présenter trois méthodes très différentes pour l'estimation des intensités de transition du modèle. Rappelons qu'il est bien sûr possible d'adopter d'autres approches, aussi bien au niveau de la méthode que de la quantité estimée.

1.4.2 Estimation non-paramétrique discrétisée

Reprenons le formalisme introduit dans la section précédente en considérant une population contenant n individus telle que le séjour de chaque individu $p \in \{1, \dots, n\}$ dans un état donné soit défini par un triplet (x_p, y_p, c_p) où x_p correspond à l'âge d'entrée dans l'état, y_p l'âge de sortie de l'état et c_p la cause de sortie associée : 1 pour le décès, 2 pour l'entrée en dépendance et 0 pour les autres causes de sortie.

Cas univarié

Considérons une suite d'instants de transition $0 \leqslant t_1 < t_2 < \cdots < t_m$ et faisons l'hypothèse que pour tout $k \in \{1, \dots, (m-1)\}$, l'intensité de transition est constante entre les instants t_k et t_{k+1}. Nous introduisons l'exposition centrale qui correspond au temps global passé par l'ensemble de la population dans l'état considéré entre les instants a et b

$$E(a,b) = \sum_{p=1}^{n} \int_{x_p}^{y_p} \mathbf{1}_{[a,b]}(u)\mathrm{d}u = \sum_{p=1}^{n} \max[0, \min(y_p, b) - \max(x_p, a)].$$

Nous notons également $N^i(a,b)$ le nombre d'évènements de type i survenus entre les instants a et b s'écrit quant à lui

$$N^i(a,b) = \sum_{p=1}^{n} \mathbf{1}_{\{c_p=i\}} \mathbf{1}_{[a,b[}(y_p).$$

Un estimateur de l'intensité de transition vers l'état i entre les instants t_k et t_{k+1} est alors donné par

$$\hat{\mu}_k^i = \frac{N^i(t_k, t_{k+1})}{E(t_k, t_{k+1})}.$$

Par ailleurs, cet estimateur est asymptotiquement gaussien d'écart-type

$$\hat{\sigma}_k^i = \frac{\sqrt{N^i(t_k, t_{k+1})}}{E(t_k, t_{k+1})}.$$

Cet estimateur peut-être utilisé afin d'estimer les intensités de transition de l'autonomie vers la dépendance ou le décès. Dans ce cas on choisira les instants t_k comme des âges entiers consécutifs et l'intensité de transition sera ainsi approchée par une fonction en escalier, constante entre deux âges entiers consécutifs.

Cas bivarié

Dans le cas où l'intensité de transition dépend de deux variables, comme pour la mortalité des dépendants dans le modèle *illness-death*, nous considérons deux suites d'instants de transition $0 \leqslant s_1 < s_2 < \cdots < s_l)$ et $0 \leqslant t_1 < t_2 < \cdots < t_m$ et faisons l'hypothèse que pour tout $j \in \{1, \ldots, (l-1)\}$ et tout $k \in \{1, \ldots, (m-1)\}$, l'intensité de transition est constante sur des produits d'intervalles de la forme $[s_j, s_{j+1}] \times [t_k, t_{k+1}]$. Dans ce cadre l'exposition centrale associée s'écrit

$$E(a,b,c,d) = \sum_{p=1}^{n} \mathbf{1}_{[a,b]}(x_p) \int_{x_p}^{y_p} \mathbf{1}_{[c,d]}(u - x_p) \mathrm{d}u = \sum_{p=1}^{n} \mathbf{1}_{[a,b]}(x_p) \max[0, \min(y_p - x_p, d) - c]$$

et le nombre d'évènements associés

$$N(a,b,c,d) = \sum_{p=1}^{n} \mathbf{1}_{\{c_p=1\}} \mathbf{1}_{[a,b[}(x_p) \mathbf{1}_{[c,d[}(y_p - x_p)$$

L'estimateur de l'intensité de décès pour un âge d'entrée dans l'état compris entre s_j et s_{j+1} et un temps passé dans l'état compris entre t_k et t_{k+1} s'écrit

$$\hat{\mu}_{j,k} = \frac{N(s_j, s_{j+1}, t_k, t_{k+1})}{E(s_j, s_{j+1}, t_k, t_{k+1})}$$

et son écart-type

$$\hat{\sigma}_{j,k} = \frac{\sqrt{N(s_j, s_{j+1}, t_k, t_{k+1})}}{E(s_j, s_{j+1}, t_k, t_{k+1})}.$$

Dans la pratique, les temps s_j correspondent aux âges entiers consécutifs et $[t_k, t_{k+1}]$ sont des intervalles consécutifs d'un mois ou un an avec $t_0 = 0$.

Discussion

Cet estimateur présente de nombreux avantages dont le premier est sa grande facilité de mise en œuvre. En effet les formules proposées peuvent être implémentées directement dans un tableur comme **Microsoft Excel** contrairement aux deux autres estimateurs que nous allons introduire. Par ailleurs, il s'interprète simplement grâce aux quantités E et N_i.

1.4.3 Estimation paramétrique par maximum de vraisemblance

Dans le modèle en temps continu introduit, il est possible d'exprimer la vraisemblance associée à chaque trajectoire en fonction des intensités de transition du modèle. La méthode du maximum de vraisemblance est alors un choix naturel afin de procéder à l'estimation de ces intensités. Pour un individu p dans la base des cotisants, la log-vraisemblance associée à la trajectoire en autonomie s'écrit

$$\begin{aligned} l_p(\mu_a, \lambda) =& \delta^1_{c_p} \log(\mu_a(y_p)) + \delta^2_{c_p} \log(\lambda(y_p)) - \int_{x_p}^{y_p} [\mu_a(u) + \lambda(u)] \, \mathrm{d}u \\ =& l_p(\mu_a) + l_p(\lambda) \end{aligned}$$

et pour un individu p dans la base des rentiers, la log-vraisemblance associée à la trajectoire en dépendance s'écrit

$$l_p(\mu_i) = \delta^1_{c_p} \log(\mu_i(x_p, y_p - x_p)) - \int_{x_p}^{y_p} \mu_i(x_p, u - x_p) \mathrm{d}u$$

où pour $k, l \in \mathbb{N}$, $\delta^l_k = \begin{cases} 1 \text{ si } k = l, \\ 0 \text{ sinon} \end{cases}$ est la fonction symbole de Kronecker. Ces expressions sont parfaitement adaptées à une estimation paramétrique de l'intensité de transition. Dans l'idéal, on choisira des formes dont l'intégrale peut être calculée explicitement. À défaut, l'utilisation d'une méthode numérique permettra un calcul approché de cette intégrale.

Exemple dans le cas univarié : loi de Gompertz

Si l'on suppose que l'intensité de décès des autonomes suit une loi de Gompertz introduite pour la première fois par Gompertz (1825)

$$\mu_a(x) = e^{a_a x + b_a},$$

l'estimateur du maximum de vraisemblance nous conduit alors à

$$\overline{(a_a, b_a)} = \underset{(a,b)}{\arg\max} \sum_{p \in C} \left[\delta^2_{c_p} (a y_p + b) - \frac{1}{a} (e^{a y_p + b} - e^{a x_p + b}) \right].$$

Exemple dans le cas univarié : loi logistique

Si l'on suppose maintenant que l'intensité d'incidence en dépendance suit une loi logistique à 4 paramètres introduite pour la première fois par Perks (1932)

$$\lambda(x) = \frac{e^{a_\lambda x + b_\lambda}}{1 + e^{a_\lambda x + c_\lambda}} + d_\lambda,$$

l'estimateur du maximum de vraisemblance nous conduit alors à

$$\overline{(a_\lambda, b_\lambda, c_\lambda, d_\lambda)} = \underset{(a,b,c,d)}{\arg\max} \sum_{p \in C} \left[\delta^2_{c_p} \log \left(\frac{e^{a y_p + b}}{1 + e^{a y_p + c}} + d \right) - \frac{e^{b-c}}{a} \log \left(\frac{1 + e^{a y_p + c}}{1 + e^{a x_p + c}} \right) - d(y_p - x_p) \right].$$

Exemple dans le cas bivarié : mélange de lois

Si l'on suppose que l'intensité de décès des dépendants suit un mélange de lois d'intensités respectives $\Delta_{i,1}(x)$ et $\Delta_{i,2}(x)$ et si l'on note $\theta_i(x)$ la probabilité d'être dans le second groupe (correspondant à l'intensité de décès $\Delta_{i,2}(x)$) lors de l'entrée en dépendance à l'âge x, on peut montrer que l'intensité de décès des dépendants s'écrit

$$\mu_i(x,t) = \Delta_{i,1}(x) + \frac{\theta_i(x)\left[\Delta_{i,2}(x) - \Delta_{i,1}(x)\right]}{\theta_i(x) + \left[1-\theta_i(x)\right]\exp\left\{\left[\Delta_{i,2}(x) - \Delta_{i,1}(x)\right]t\right\}}.$$

Si $\Delta_{i,1}(x)$, $\Delta_{i,2}(x)$ et $\theta_i(x)$ sont des formes paramétriques, alors l'estimateur du maximum de vraisemblance nous donne

$$\begin{aligned}\overline{(\Delta_{i,1},\Delta_{i,2},\theta_i)} =& \underset{(\Delta_1,\Delta_2,\Theta)}{\arg\max} \sum_{p\in R}\Big[\delta^1_{c_p}\log\left(\Delta_1(x_p) + \frac{\theta(x_p)\left[\Delta_2(x_p)-\Delta_1(x_p)\right]}{\theta(x_p)+\left[1-\theta(x_p)\right]\exp(\left[\Delta_2(x_p)-\Delta_1(x_p)\right]\left[y_p-x_p\right])}\right) \\ &- \Delta_2(x_p)\left[y_p-x_p\right] + \log\Big\{\theta(x_p)+\left[1-\theta(x_p)\right]\exp(\left[\Delta_2(x_p)-\Delta_1(x_p)\right]\left[y_p-x_p\right])\Big\}\Big]\end{aligned}$$

où $\overline{(\Delta_{i,1},\Delta_{i,2},\theta_i)}$ dénote abusivement l'estimateur du maximum de vraisemblance des paramètres intervenant dans les quantités $\Delta_{i,1}(x)$, $\Delta_{i,2}(x)$ et $\theta_i(x)$.

Discussion

La résolution de ce type d'équation nécessite d'avoir recours à un algorithme d'optimisation numérique. Sous le langage **R** (R Core Team, 2016), on pourra utiliser la fonction **optim**. Pour avoir plus de chance de trouver l'optimum global de la fonction, on privilégiera un algorithme permettant plusieurs initialisations aléatoires des valeurs initiales des paramètres, comme la fonction **gosolnp** du package **Rsolnp**. Une des limites de cet estimateur, caractéristique des méthodes paramétriques en général, est le risque de mauvaise spécification du modèle proposé. Pour mitiger ce risque, on peut tester différents modèles et les comparer à l'aide d'un critère tel le *Bayesian Information Criterion* (BIC) dont on trouvera une bonne présentation des propriétés dans Lebarbier and Mary-Huard (2006). Indépendamment, il est toujours souhaitable d'utiliser un estimateur non-paramétrique à des fins de comparaison pour détecter les éventuelles limites du modèle choisi. L'approche paramétrique présente néanmoins plusieurs avantages. Elle utilise toutes les données disponibles sans qu'il soit nécessaire de les segmenter. Elle produit également des courbes déjà lissées et que l'on peut naturellement extrapoler aux âges pour lesquels les données ne sont pas disponibles. Enfin cette approche permettent d'obtenir des résultats même quand le nombre d'observation est très limité. On trouvera des applications de l'approche paramétrique dans les Chapitres 2 et 3.

1.4.4 Estimation non-paramétrique par maximum de vraisemblance local

Ce troisième estimateur est le résultat d'une approche que l'on peut qualifier de localement paramétrique. On fait l'hypothèse qu'en chaque point de la surface que l'on cherche à estimer, la dite surface peut être approchée par un polynôme. On estime les coefficients de ce polynôme en utilisant les observations proches, avec un estimateur du maximum de vraisemblance modifié pour que chaque observation soit pondérée en fonction de la distance au point d'estimation. Le lecteur pourra par ailleurs se référer à Loader (1999) pour une introduction complète et didactique des méthodes de régression et de vraisemblance locales, qui peuvent être appliquées à une grande variété de sujets.

1.4.5 Cas univarié

On reprend les notations précédentes et on note $\mu(x)$ l'intensité que l'on cherche à estimer, où x est l'âge atteint par l'individu. On utilise la fonction de lien logarithme et on suppose que $\log \mu(u)$ peut être approché localement par un polynôme dont les coefficients sont $a = (a_0, a_1, \ldots, a_d)^T$ de telle manière que

$$\begin{aligned}\log \mu(u) =& a_0 + a_1(u-x) + \cdots + a_d(u-x)^d + \mathrm{o}((u-x)^d) \\ =& \langle a, A(u-x)\rangle + \mathrm{o}((u-x)^d)\end{aligned}$$

où $A(u) = (1, u, \ldots, u^d)^T$.

Nous introduisons la fonction de log-vraisemblance local au point $x \in \mathbb{R}$

$$\mathcal{L}_x(a) = \sum_{i=1}^{n}(1-c_i)W\left(\frac{y_i - x}{h(x)}\right)\langle a, A(y_i - x)\rangle - \int_{x_{min}}^{y_{max}} N(u)W\left(\frac{u-x}{h(x)}\right)e^{\langle a, A(u-x)\rangle}\mathrm{d}u$$

où

$$N(u) = \sum_{i=1}^{n}\mathbb{I}\{x_i < u \leqslant y_i\}$$

est le nombre d'individus sous risque à l'âge u, W est une fonction noyau (c'est à dire une fonction de $\mathbb{R}$ dans $\mathbb{R}$ d'intégrale 1), h est la fonction bande-passante (*i.e.* une fonction positive), $x_{min} = \min_{i\in\{1,\ldots,n\}} x_i$ et $y_{max} = \max_{i\in\{1,\ldots,n\}} y_i$. Un estimateur $\hat{a} = (\hat{a}_0, \ldots, \hat{a}_d)^T$ de a peut alors être obtenu en maximisant la log-vraisemblance locale

$$(\hat{a}_0, \ldots, \hat{a}_d)^T = \operatorname*{argmax}_a \mathcal{L}_x(a).$$

On obtient alors très simplement un estimateur $\hat{\mu}(x)$ de $\mu(x)$ en posant

$$\hat{\mu}(x) = \exp(\hat{a}_0).$$

1.4.6 Cas bivarié

Dans le cas où l'intensité de transition toujours notée μ dépend de deux variables, l'âge d'entrée dans l'état x et le temps passé dans l'état t, reprenons les notations précédentes en notant pour chaque individu $t_i = y_i - x_i$ le temps passé dans cet état.

Pour plus de clarté, on prend l'exemple du cas log-quadratique qui correspond à une fonction de lien logarithme et le choix d'un polynôme de degré $d = 2$. Pour tout couple (x, t) et pour (u, v) proche de (x, t), $\log \mu(u, v)$ peut être localement approché par un polynôme de degré 2 dont les coefficients sont notés $a = (a_0, a_1, \ldots, a_5)^T$ tels que

$$\begin{aligned}\log \mu(u,v) =& a_0 + a_1(u-x) + a_2(v-t) + a_3(u-x)^2 \\ &+ a_4(u-x)(v-t) + a_5(v-t)^2 + \mathrm{o}((u-x)^2 + (v-t)^2) \\ =& \langle a, A(u-x, v-t)\rangle + \mathrm{o}((u-x)^2 + (v-t)^2)\end{aligned}$$

où $A(u, v) = (1, u, v, u^2, uv, v^2)^T$.

Nous introduisons la fonction de log-vraisemblance

$$\begin{aligned}\mathcal{L}_{x,t}(a) =& \sum_{i=1}^{n}(1-c_i)\, W\left(\frac{\rho(x_i - x, t_i - t)}{h(x,t)}\right)\langle a, A(x_i - x, t_i - t)\rangle \\ &- \sum_{i=1}^{n}\int_0^{t_i} W\left(\frac{\rho(x_i - x, u - t)}{h(x,t)}\right)e^{\langle a, A(x_i - x, u-t)\rangle}\mathrm{d}u\end{aligned}$$

où ρ est une distance, W une fonction noyau et h est une fonction bande passante.

Une estimation $\hat{a} = (\hat{a}_0, \ldots, \hat{a}_5)^T$ de a peut alors être obtenue en maximisant la log-vraisemblance locale

$$(\hat{a}_0, \ldots, \hat{a}_5)^T = \underset{a}{\operatorname{argmax}}\ \mathcal{L}_{t,x}(a).$$

On en déduit une estimation $\hat{\mu}(x,t)$ de $\mu(x,t)$ en posant

$$\hat{\mu}(x,t) = \exp(\hat{a}_0)$$

On note deux différences majeures avec le cas univarié. Tout d'abord, comme h est une fonction bivariée, il est nécessaire de définir la distance entre deux points (x,t) et (u,v). Un choix naturel est la distance euclidienne mais comme le rôle des deux composantes n'est ici pas symétrique, il peut être intéressant de leur attribuer des poids différents. On remarque également que dans le cas bivarié, les trajectoires ne peuvent plus être regroupées en utilisant le nombre d'individus sous risque N. En effet, l'intensité de transition est ici conditionnée par l'âge d'entrée en dépendance qui est différent pour chaque individu.

Discussion

Cette méthode combine les avantages des méthodes paramétriques et discrétisées présentées ci-avant. Elle nécessité néanmoins la spécification de plusieurs éléments mentionnés ci-dessus, à savoir le degré d du polynôme considéré, la fonction W noyau choisie et enfin la bande-passante h utilisée. La complexité de cette méthode réside ainsi dans le choix de ces 3 éléments qui seront en général définis à l'aide d'un ou plusieurs paramètres dont le choix va déterminer la quantité de lissage appliquée. Ce choix ne sera pas discuté ici mais on trouvera un exemple d'application dans le Chapitre 4. Si le lissage est trop important, on court le risque de supprimer certaines caractéristiques des données. S'il est trop faible, cela risque au contraire de se traduire par du sur-apprentissage sur les données. Par ailleurs, nous conclurons cette section en remarquant que cette méthode peut être appliquée à des données discrétisées, à savoir l'exposition centrale et le nombre d'évènements définis pour l'estimateur discrétisé. Nous invitons lecteur intéressé à se référer à Tomas and Planchet (2013) pour une application en assurance dépendance basée sur cette approche. L'approche par maximum de vraisemblance locale sera utilisée tout au long du Chapitre 4.

1.4.7 Tarification et calcul des provisions

Description du produit d'assurance considéré

Considérons un produit d'assurance prévoyant le versement d'une rente d'un montant annuel R dès l'entrée en dépendance et jusqu'au décès de l'assuré, et un capital K versé lors de l'entrée en dépendance. En contrepartie, supposons que l'assuré s'engage à verser jusqu'à son entrée en dépendance ou son décès un montant de prime fixé lors de la souscription et constant dans le temps. Notons τ le taux d'actualisation en temps continu utilisé. Le principe de la tarification est de déterminer le montant de prime tel que les engagements de l'assureur et ceux de l'assuré soient égaux en valeur actuelle nette lors de la souscription. La majorité des produits proposés aujourd'hui prévoient des versements mensuels des cotisations de l'assuré, en début de mois et des versements mensuels des prestations de l'assureur en début ou fin de mois. Par commodité, nous utiliserons un modèle en temps continu dans lequel ces capitaux seront versés en quantités infinitésimales et de manière continue. Ce modèle ne correspond à aucune réalité mais du point de vue de la tarification, les écarts avec un produit prévoyant un versement mensuel sont négligeables. Le modèle en temps continu est équivalent, aux effets d'actualisations près, à un modèle *pro rata temporis* ou l'assuré (resp. l'assureur) verserait en fin de mois un montant de cotisations (resp. de prestations) correspondant à la fraction du mois passé dans l'état d'autonomie (resp. de dépendance).

Probabilités de survie

Reprenons les notations précédentes et posons $0 \leqslant x \leqslant y$ et $0 \leqslant t \leqslant s$. Nous introduisons les fonctions de survie dans l'état d'autonomie (resp. de dépendance)

$$\begin{aligned} A(x,y) =& \mathbb{P}(Z_y = A | Z_x = A) \\ I_x(t,s) =& \mathbb{P}(Z_{x+s} = I | Z_{x^-} = A, Z_x = I, Z_{x+t} = I). \end{aligned}$$

Il est possible de montrer que

$$\begin{aligned} A(x,y) =& \exp\left(-\int_x^y [\mu_a(u) + \lambda(u)] \, \mathrm{d}u \right), \\ I_x(t,s) =& \exp\left(-\int_t^s \mu_i(x,u) \mathrm{d}u \right). \end{aligned}$$

Les formules de tarification considérées feront intervenir les fonctions de survie actualisées définies comme suit

$$\begin{aligned} \overline{A}(x,y) =& e^{-\tau(y-x)} A(x,y) = \exp\left(-\int_x^y [\mu_a(u) + \lambda(u) + \tau] \, \mathrm{d}u \right), \\ \overline{I}_x(t,s) =& e^{-\tau(s-t)} I_x(t,s) = \exp\left(-\int_t^s [\mu_i(x,u) + \tau] \, \mathrm{d}u \right). \end{aligned}$$

On se souviendra qu'une intensité de transition est homogène à un taux d'intérêt, il est donc naturel que ces deux quantités soient additionnées dans la formule précédente.

Formules de tarification

Soit $x_s, t \geqslant 0$ et $x, x_i \geqslant x_s$. Définissons

- La valeur actuelle probable de l'engagement d'un assuré autonome à l'âge x pour 1 € de prime annuelle s'écrit
$$P(x) = \int_{u=x}^{\omega} \overline{A}(x,u) \mathrm{d}u,$$
- La provision pour sinistres à payer (PSAP) : il s'agit de la valeur actuelle probable des engagements de l'assureur pour un assuré entré à l'âge x en dépendance, ayant passé une durée t dans cet état
$$\mathrm{PSAP}(x,t) = K \mathbf{1}\{t = 0\} + R \int_{u=t}^{\omega - x} \overline{I}_x(t,u) \mathrm{d}u,$$
- La valeur actuelle probable de l'engagement de l'assureur pour un assuré autonome d'âge à la souscription x_s, d'âge atteint x
$$\Pi(x) = \int_{u=x}^{\omega} \lambda(u) \overline{A}(x,u) \mathrm{PSAP}(u,0) \mathrm{d}u,$$

- Le montant de prime pure p^* assurant l'équilibre des engagements de l'assureur et de l'assuré en valeur actuelle probable pour un assuré souscrivant à l'âge x_s

$$p^*(x) = \frac{\Pi(x)}{P(x)},$$

- La provision pour risque croissant (PRC) : il s'agit de la valeur actuelle probable des engagements de l'assureur pour un assuré autonome ayant souscrit à l'âge x_s d'âge actuel x

$$\mathrm{PRC}(x_s, x) = \Pi(x)\left[p^*(x) - p^*(x_s)\right].$$

Ces formules de tarification font intervenir directement les intensités de transition. Le calcul de ces quantités nécessite l'emploi de méthodes numériques afin d'évaluer les différentes intégrales.

1.5 Apprentissage statistique et dépendance

A l'heure où nous écrivons ces lignes, les méthodes à base d'arbres de décision (introduits dans Breiman et al., 1984), de Support Vector Machine (SVM) (dont on trouvera une introduction dans Smola and Schölkopf, 2004) ou de réseaux de neurones (présentés par exemple dans Smith, 1993), regroupées sous le nom de méthodes d'apprentissage statistique connaissent un intérêt très fort dans beaucoup d'industries ainsi qu'au sein de la profession actuarielle. Cet engouement est notamment porté par les récents succès de ces méthodes dans les domaines de l'intelligence artificielle et de la reconnaissance de l'image ou de la voix, par les volumes énormes de données échangés sur les réseaux sociaux et sur internet en général, et enfin par la capacité de stockage et la puissance de calcul toujours plus importantes dont l'on dispose aujourd'hui. Au cours de cette thèse, nous nous sommes intéressés à plusieurs de ces méthodes, et nous conclurons ce chapitre d'introduction par un résumé de notre expérience sur l'emploi de ces méthodes. Parmi leurs avantages, on retiendra leur grande flexibilité par rapport aux méthodes classiques d'analyse de survie pour lesquelles le rôle des variables explicatives est contraint par le modèle. Ces méthodes sont ainsi d'un grand intérêt lorsque le nombre de variables explicatives est grand, le rôle de certaines variables est obscur ou lorsque les interactions entre plusieurs variables sont complexes. Par ailleurs, l'emploi de ces méthodes nécessite de disposer d'un volume de données très conséquent pour l'estimation du modèle. Dans le cadre de l'estimation du risque de dépendance, ces conditions ne sont que partiellement remplies. Les variables explicatives se limitent dans la majorité des cas à l'âge et au sexe de l'assuré. Le volume de donnée disponible est également assez limité, surtout en ce qui concerne l'état de dépendance. Néanmoins, dans le cas où l'on dispose d'informations concernant les pathologies, celles-ci se qualifient comme des variables explicatives dont l'effet sur le risque et l'interaction avec les autres variables est méconnu et a priori assez complexe. Dans ce contexte, l'utilisation de méthodes d'apprentissage statistique paraît justifié.

L'un des obstacles majeurs à la mise en place de ces méthodes concerne la présence de censure dans les données. A notre connaissance, aucune des méthodes évoquées plus haut ne permet de gérer telle quelle la censure. Il faut donc en amont transformer les données avant de pouvoir utiliser ces techniques. Parmi les solutions possibles, la méthode *Inverse Probability of Censoring Weights* (IPCW) est la plus utilisée dans la littérature. Elle consiste dans un premier temps à estimer la fonction de survie associée à la variable de censure grâce à l'estimateur de Kaplan-Meier. A chaque observation non-censurée, on associe ensuite un poids défini comme l'inverse de cette probabilité de survie pour la durée observée. On peut alors travailler uniquement avec les observations non censurées. Sous les bonnes hypothèses, cette méthode possède plusieurs propriétés remarquables. L'espérance de vie calculée à partir des observations non-censurées

avec les poids calculés est égale à l'espérance associée à la loi de durée originale. L'idée derrière cette méthode est par ailleurs très naturelle. Il s'agit de prédire pour chaque trajectoire non-censurée, combien de trajectoires censurées ne sont pas observées à cause de la censure. Les poids introduits vont ainsi permettre de contrebalancer l'effet de sous-estimation lié à la censure en sur-pondérant les observations les plus longues peu représentées dans l'échantillon à cause de la censure. On pourra trouver une application véritablement innovante de cette méthode à des problématiques d'assurance dans Lopez et al. (2015).

Cependant, un certain nombre d'hypothèses nécessaires à la mise en œuvre de cette méthodologie ne sont pas vérifiées dans le cas de l'assurance dépendance. Tout d'abord, la fonction de survie de la variable de censure ne doit pas dépendre de la valeur des covariables. Prenons l'exemple simplifié d'un portefeuille de dépendance dont tous les assurés souscrivent au même âge, disons 65 ans pour fixer les idées, et la même année, par exemple 1990. En 2015, date d'extraction de la base de donnée et donc date de la censure, qui est de type administrative, tous les assurés auront ainsi $65 + 2015 - 1990 = 90$ ans. Si l'on s'intéresse à la durée de vie en dépendance et que l'on considère l'âge d'entrée en dépendance comme variable explicative, alors la variable de censure est corrélée à 100 % avec l'âge d'entrée en dépendance. Un assuré devenu dépendant à l'âge de 87 ans sera ainsi censuré au bout d'une durée de $90 - 87 = 3$ ans. Pour ce type de portefeuille, très peu de sinistres seront censurés aux âges jeunes d'entrée en dépendance et beaucoup aux âges élevés. En effet, un assuré entré en dépendance à 60 ans doit survivre 20 ans pour être censuré contre 3 ans pour un assuré devenu dépendant à 87 ans. L'application de la méthodologie IPCW va alors non seulement conduire à garder uniquement les trajectoires non-censurées, ce qui veut dire que les âges jeunes seront sur-représentés, mais également à sur-pondérer ces trajectoires. Aussi dans cet exemple, la méthodologie ne corrige pas l'effet de la censure. Au contraire, elle l'aggrave. Bien sûr, dans la réalité, l'âge à la souscription des assurés est variable (néanmoins la majorité des souscriptions se font entre 60 et 70 ans) de même que la date de souscription (cependant les souscriptions sont souvent concentrées sur une période de 5 à 10 ans et le portefeuille est ensuite fermé aux nouveaux entrants). La corrélation entre âge à la souscription et variable de censure est donc bien inférieure à 100 % mais elle demeure néanmoins très importante. Parmi les autres contraintes de la méthode IPCW, la valeur maximale atteinte par la variable de censure doit être inférieure à la valeur maximale prise par la durée de survie. En effet, si la plus longue trajectoire observée est censurée, sa contribution ne sera que partiellement prise en compte par la méthode. Supposons par exemple que la plus grande trajectoire non censurée dure 20 ans. Dans ce cas, que la plus longue trajectoire censurée dure 20, 25 ou 30 ans n'impacte pas le résultat de l'estimation. L'information après 20 ans n'est pas prise en compte par la méthode. Cela créée ainsi un biais qui amène à la sous-estimation des durées de vie. Dans la pratique, cette hypothèse n'est pas vérifiée dans le cas de la dépendance. Autre point important, la détermination des poids, rappelons-le, se fait grâce à l'inverse de l'estimateur de Kaplan-Meier. Lorsque la fonction de survie se rapproche de 0, les poids deviennent donc à la fois très importants et très volatiles. Ainsi, si l'on retire la plus longue trajectoire observée, ou que l'on change la période d'observation de telle manière que la plus longue observation censurée ne le soit plus, cela peut avoir un gros impact sur les résultats. Compte-tenu de tous ces éléments, la méthodologie IPCW nous paraît ainsi adaptée aux problématiques d'assurance vie pour lesquelles l'on dispose de durées bornées. C'est le cas par exemple de l'assurance incapacité en France pour laquelle les prestations ont une durée limitée à 3 ans et la période de couverture se termine à l'âge légal de départ à la retraite. Dans le cas de l'assurance dépendance en France, les prestations et la période de couverture sont toutes deux illimitées dans le temps.

Pour étendre la méthodologie en prenant en compte les covariables, on peut songer à segmenter les données selon les valeurs de covariables, par exemple en regroupant les sinistres par tranches d'âge d'entrée en dépendance. Cependant, en reprenant l'exemple précédent et en construisant des tranches d'âges de 10 ans, il n'y aurait aucune trajectoire observée d'une durée

supérieure à 10 ans dans la tranche d'âge 80 à 90 ans. La méthodologie ne permettrait pas par exemple de calculer une espérance de vie en dépendance pour cette tranche car le poids de la queue de distribution ne pourrait pas être déterminé. Une approche plus fine consiste à utiliser des estimateurs à noyaux pour tenir compte des covariables dans l'estimation de la fonction de survie associée à la censure. On pourra se référer à Lopez (2011) et Lopez et al. (2013) qui fournissent des éléments de réponse qui pourraient permettre d'adapter la méthodologie afin de mieux prendre en compte l'effet des covariables. Nous conclurons cette section en remarquant que les méthodes d'apprentissage statistique doivent être adaptées aux données censurées afin de résoudre les problématiques rencontrées en assurance vie. La méthode IPCW abordée ici n'est bien entendu qu'un exemple parmi d'autres mais nos travaux sur le sujet nous amènent à penser que des développements théoriques sur ce sujet sont nécessaire avant de pouvoir utiliser les méthodes d'apprentissage statistique pour l'étude du risque de dépendance. Ceci constitue un axe de recherche intéressant pour des travaux futurs mais dépasse le cadre de la présente thèse.

1.6 Organisation de la thèse

Le corps de la thèse est composé de trois chapitres traitant la problématique de la modélisation du risque de dépendance sous des angles différents. Ces chapitres peuvent être lus de manière indépendante. Néanmoins, ils participent d'une même démarche de maîtrise du risque et présentent de nombreux éléments méthodologiques communs.

1.6.1 Résumé du Chapitre 2

Dans ce chapitre, basé sur Biessy (2015b), nous cherchons à modéliser la trajectoire de personnes âgées dépendantes à l'aide d'un modèle comportant 4 états de dépendance. Pour ce faire, nous nous basons sur un jeu de données de l'Allocation Personnalisée d'Autonomie (APA) comprenant plus de 50 000 trajectoires en dépendance sur les années 2002 à 2005 incluses. Ce jeu de données ne permet pas d'évaluer des probabilités d'entrée en dépendance ou de décès des autonomes, aussi nous concentrons nous sur les transitions entre les états de dépendance et vers le décès. Le modèle présenté est semi-markovien en temps continu. L'approche retenue est paramétrique et consiste à modéliser directement le noyau semi-markovien à travers ses deux composantes : la probabilité de saut et la loi de durée connaissant le prochain état. La probabilité de saut est supposée indépendante de l'âge et du sexe. La loi de durée est par ailleurs supposée suivre une loi de Weibull dont le paramètre d'échelle dépend de l'âge, du sexe et également d'une variable de mélange qui est sensée refléter l'hétérogénéité de la population des dépendants. L'estimation des paramètres du modèle est effectuée grâce à la méthode du maximum de vraisemblance, en prenant en compte les nombreuses spécificités des données APA (troncature à gauche, censure à droite, et valeurs manquantes pour la transition vers le décès). Les résultats de l'estimation sont ensuite utilisés afin d'établir un coût du risque de dépendance associé, en se basant sur le principe d'équivalence ainsi que sur une méthode de simulation de type Monte Carlo. Enfin, une étude de sensibilité de la charge de sinistre par rapport à différents facteurs de risque est présentée.

1.6.2 Résumé du Chapitre 3

Ce chapitre est consacré à l'étude du modèle illness-death à 3 états sans retour vers l'autonomie. Il s'appuie sur des données de portefeuilles d'assurance avec plus de 20 ans d'historique. Ce chapitre propose une approche paramétrique pour l'estimation des trois lois intervenant dans le modèle dans un cadre semi-markovien en temps continu. Une forme paramétrique est proposé pour chacune des lois du modèle. La complexité de la transition de la dépendance vers le décès

conduit à introduire un mélange de loi dans l'expression de la mortalité des dépendants, sensé rendre compte de l'hétérogénéité au sein de la population des dépendants. Par ailleurs, en partant d'un système d'équations différentielles régissant l'évolution du modèle, nous faisons le lien entre les intensités du modèle *illness-death* et l'intensité de décès globale sur l'ensemble du portefeuille (indépendamment de l'état de dépendance de l'assuré). Nous utilisons le critère BIC afin de comparer les différents modèles paramétriques proposés. Au final, l'approche développée dans cet article possède l'avantage d'être entièrement automatisable. En effet, les estimateurs du maximum de vraisemblance proposés ne nécessitent pas de segmentation des données mais utilisent l'intégralité des trajectoires individuelles. Par ailleurs, le lissage et l'extrapolation des lois n'est pas nécessaire car déjà intégré dans la phase d'estimation des modèles paramétriques. Enfin, une comparaison graphique avec les résultats de méthodes non-paramétriques usuelles permet de conclure sur l'adéquation des modèles choisis pour le portefeuille considéré.

1.6.3 Résumé du Chapitre 4

Ce chapitre est consacré à une étude de la dépendance sous l'angle des pathologies. Il s'appuie sur un portefeuille contenant 14,000 sinistres pour lequel on dispose maintenant d'une classification selon 4 groupes de pathologies : le *cancer*, les *maladies neurologiques*, la *démence* et les *autres causes*. Nous introduisons un modèle avec 4 états de dépendance et en faisant le lien avec le modèle *illness-death*, nous obtenons une expression de la mortalité des dépendants tous groupes confondus en fonction des intensités d'entrée en dépendance et des intensités de décès associées à chaque groupe. Pour l'estimation des différentes intensités mentionnées ci-dessus, nous utilisons une méthode de maximum de vraisemblance local. Son principe consiste à estimer en chaque point l'intensité de transition en l'approchant localement à l'aide d'un polynôme dont les coefficients sont ainsi estimés à partir des observations voisines. Cette méthode a déjà été appliquée dans le contexte de l'assurance, mais toujours à des données agrégées sur des intervalles de temps discrets. Dans ce chapitre, nous proposons de l'appliquer directement aux trajectoires individuelles, à travers une modélisation en temps continu. La sélection des paramètres de lissage utilisés se fait à l'aide du critère AIC tout en contrôlant les résidus produits. Nous montrons qu'en estimant séparément par maximum de vraisemblance local les intensités de décès relatives à chaque groupe de pathologie, nous améliorons la qualité du résultat obtenu par rapport à l'estimation directe de la mortalité tous groupes confondus. Nous montrons également que l'évolution de l'intensité de décès des dépendants en fonction du temps passé en dépendance peut être expliquée en très grande partie par un effet de mélange des groupes de pathologie qui constituent la population des individus dépendants. Enfin, nous obtenons des statistiques permettant de comparer les pathologies à travers l'espérance de vie en dépendance associée ou la charge de sinistre pour l'assureur.

Chapitre 2

Long-Term Care Insurance: a multi-state semi-Markov model to describe the dependency process in elderly people.

Abstract

The pricing of today's long-term care insurance products relies on simple models where dependency is considered as a single homogeneous state. Because of aging population and rapid evolution in the field of medicine, it becomes paramount to get a clearer picture of the underlying risk. We believe it may only be achieved by taking into account several levels of dependency. A semi-Markov process is a multi-state process whose transition probabilities not only depend on the current state but also on the time spent in this state. This process has proven more flexible than the simple Markov process, and is core to numerous publications in the field of epidemiology. However its use in relation with long-term care insurance has remained mostly theoretical, mainly because of the lack of data available to insurers.

The present article aims at introducing the construction process of a semi-Markov model with 4 levels of dependency. This work is based on data from the French long-term care public aid: the "Allocation Personnalisée d'Autonomie" (APA). Firstly, we introduce the parameters used to model transitions between states. We then proceed to the calibration of those parameters, using a likelihood maximization method, while taking into account the peculiarities of the APA data set. Finally, we apply this model to the pricing of a fictive long-term care insurance product, using a Monte Carlo method.

2.1 Introduction

In developed countries, since the beginning of the 20th century, there has been a steady increase in life expectancy at birth of around one quarter every year, as a result of the rapid evolution in medical techniques. This, along with the aging of the baby-boomer generations has resulted in the expectation that the number of French people aged 65 or more will double by 2060 (Blanpain and Chardon, 2010). Among other consequences, it will cause a spike in the number of elderly dependent people (Lécroart, 2011).

Paradoxically enough, the dependency risk can be called a young risk, because, while mortality and longevity have been studied for more than a century, the first long-term care insurance products only appeared in the mid 1980s, products covering partial dependency being even more recent. As all products include a maximum age of subscription, the number of people who reached higher ages where the incidence of dependency becomes significant is quite low, and so is the number of claims. Therefore, the amount of data available to insurers is still very limited.

To price long-term care insurance products, most insurers use discrete time models with 3 states: autonomy, dependency and death (see figure 2.1). Most products today also cover partial dependency, providing a percentage of the benefit granted in total dependency. The pricing of such products is achieved by considering that they offer two distinct guarantees whose cost can be calculated using two separate 3-state models. This approach yields very robust results and allows the use of experience data. However, the underlying assumption is that partial and total dependency are two completely independent phenomena, with probabilities of becoming totally dependent being the same for both autonomous and partially dependent insured lives.

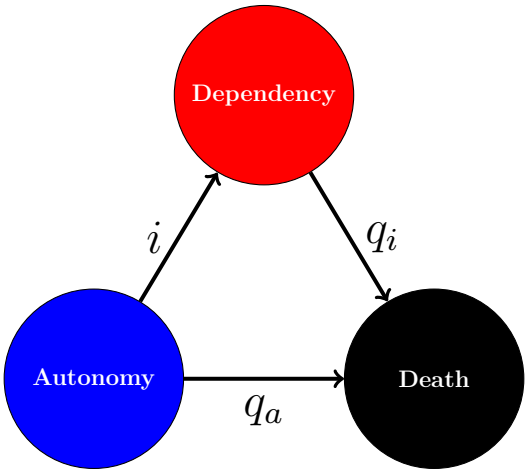


Figure 2.1: Simple model with 3 states: autonomy, dependency and death; i is the incidence rate and q_a (resp. q_i) the mortality rate for autonomous (resp. dependent) people.

To get a better understanding of the underlying dependency process, we believe it needs to be studied as a single multi-state process. The Markov process, for which transition probabilities only depend on the current state of the process, has already been used to this extent (Massonet, 2006; Déléglise et al., 2009). However, survival times in dependency depend heavily on the age of the individual at entry, but also on the time already spent in dependency, as very high death rates are observed during the first few months after entry. To take both phenomena into account, a classic Markov process is not enough and therefore a more flexible process is needed to get a model which matches insurers experience.

For a semi-Markov process, transition probabilities depend not only on the current state but also on the time spent in the current state. Such process has been extensively used in the field of epidemiology (Commenges, 2002) and yielded better results than Markov process when modeling complex phenomena like for example the evolution of HIV (Mathieu, 2006)

or follow-up of kidney transplant (Foucher et al., 2007). For long, the semi-Markov process has been identified as a powerful tool for long-term care insurance as well, in (Haberman and Pitacco, 1998; Denuit and Robert, 2007) and more recently in (Christiansen, 2012) while (Janssen and Manca, 2007) discussed numerical and computational issues. However, to the best of our knowledge, few papers focused on applications based on real data (one can however refer to Lepez, 2006) due to the unavailability of such data. As a consequence, issues that arise when working on censored data coming from longitudinal studies are rarely addressed, although developing methods to handle such data proves necessary to the application of semi-Markov models for insurance purposes.

This paper provides an application based on data from the French public aid for dependent elderly people: the APA: "Allocation Personnalisée d'Autonomie". We consider a model with 4 different states of dependency. To define those states we rely on the "Autonomie Gérontologue Groupes Iso-Ressources" (AGGIR) grid. The AGGIR grid aims to categorize people by groups of similar needs based on their level of dependency. It is used in France for the attribution of the APA. This grid describes 6 levels of dependency, from the more severe level Gir 1 to the less severe Gir 6. However, only people in states Gir 1 to Gir 4 may actually benefit from the public aid, hence we only consider these 4 levels of dependency in our model. A description for each level of dependency can be found in table 2.1. To determine to which group one belongs, the ability to perform 8 activities of daily living is assessed, in a similar way to definition used by most insurers around the world. However, in the case of the AGGIR grid, the degree of incapacity is also taken into account, and some activities have more weight than others. This translates into a complex algorithm (see Vetel et al., 1998).

Gir level	Associated definition
Gir 1	People confined to bed or to a wheelchair, and whose mental abilities are greatly impaired, who need constant care. Also people at the end of their lives.
Gir 2	People confined to bed or to a wheelchair, whose mental abilities are not impaired, and who require care for most activities of daily living or people with impaired mental abilities but able to move by themselves, who need permanent oversight.
Gir 3	People with mental autonomy and partial physical autonomy, who need help for cleaning and bathing several times a day.
Gir 4	People who can walk inside their home but who require help for cleaning, clothing and possibly transfers.
Gir 5	People who need occasional help for cleaning, cooking and houseclean.
Gir 6	People still autonomous for the main activities of daily living.

Table 2.1: Levels of the AGGIR grid from most severe (Gir 1) to least severe (Gir 6).

Most insurers use their own definition of dependency, based on activities of daily living (ADL), in order to make it easier for insured lives to understand and not to be impacted by future changes in the public definition. Those definitions and the AGGIR definition can nevertheless be compared to some extent. In our paper, the choice of an AGGIR-based definition is driven by the use of data from the French public aid. This data is gathered over the whole population, and the number of dependent people observed this way outweighs most insurers portfolio's, especially at higher ages. As this data is gathered using a very specific observation process, we develop a specific methodology to limit the associated observation bias. Besides, as the data only includes information about dependent people, incidence rates and mortality rates for autonomous people cannot be inferred from it and therefore need to be obtained from another source.

In this work, we focus on APA data which gathers the assessments of dependency states on individuals. Among other features, this data is right-censored and contains missing values. Our main assumptions on this data are that the stock effect observed can be removed by left truncating of the data, and that transitions times and evaluations times can be assimilated (see section 2.2 for more details). We fit a homogeneous semi-Markov model with four dependency states (plus death) relying on Weibull distribution laws that integrate Cox proportional hazard rates.

In the next section, we introduce the APA data and its peculiarities in terms of censoring and truncating. We discuss several assumptions that are necessary to process the data and use it in the following sections of the paper.

The third section then provides a definition for the semi-Markov process, and introduces the elements of our parametric model. For every transition, the duration is assumed to follow a Weibull law. The impact of sex and gender is then taken into account through a semi-proportional hazard model. The impact of pathologies, which are not observed in the data, but, we believe, explain the heterogeneity between trajectories, is modeled through a static frailty which takes only two value. The value of frailty is determined at entry in dependency with a probability which depends on both the gender and the age of entry of the individual, through a generalized linear model with a logistic link function. The impact of frailty on the duration law is also modeled through a semi-proportional hazard rate. At last, this section also provides an expression for the likelihood function associated with the model, which is used for the calibration of parameters.

The penultimate section presents the parameters estimated through the maximization of the likelihood function. An algorithm to generate trajectories from these estimations is developed, and used to get descriptive statistics about the modeled dependency process.

The last section introduces a specific methodology for the pricing of insurance products using the calibrated model. This method relies on Monte Carlo simulations as a closed formula for the premium is not available due to the complexity of the process. Using the Central Limit Theorem and the delta method, we then compute an upper-bound for the uncertainty on the estimated premium. This methodology is finally applied to the pricing of a fictive long-term care insurance product, with a quick analysis about reserves and sensitivity to different risk factors.

2.2 The APA data

2.2.1 Introducing the data

The APA: "Allocation Personnalisée d'Autonomie" is the French public aid for elderly dependent people. It has been introduced in 2002, and is only available to people aged 60 and more. People who want to benefit from the aid need to have their level of dependency evaluated by a public service team, and be assigned to group Gir 4 or more severe. Then, they agree with the team on a solution to cope with their dependency, and part of the cost is supported by the public aid, up to a maximum amount and depending on the people own resources.

The aid is managed locally by the French administrative area. Therefore each area gathers its own data. The data we were able to get are the same as in Lepez (2006). They have been gathered by 4 French administrative areas over the years 2002 to 2005. Only the individuals who have been granted the aid appear in the data.

Content of the data includes the following information

- The date of birth of the individual,
- The gender of the individual,

- The date of death of the individual, if death occurred during the observation period,
- The first date of evaluation which allowed the individual to benefit from the aid, with the result of this evaluation,
- Up to three subsequent evaluations with the result of those evaluations.

We observe on figure 2.2 (left) that the number of first evaluations in 2002 is way higher than during the subsequent years. This is due to the fact that the APA was created in 2002, hence many people had not taken any evaluation of their dependency level beforehand because they had no incentive to. We note that those people do not enter the APA all at once, but progressively, over the course of year 2002. This phenomenon is known as the "stock effect". On the other hand, figure 2.2 (right) shows that there is hardly any death observed before 2005. It appears that the information about deaths has only been collected from the year 2005. Consequently, deaths which occurred in 2002, 2003 or 2004 will be missing.

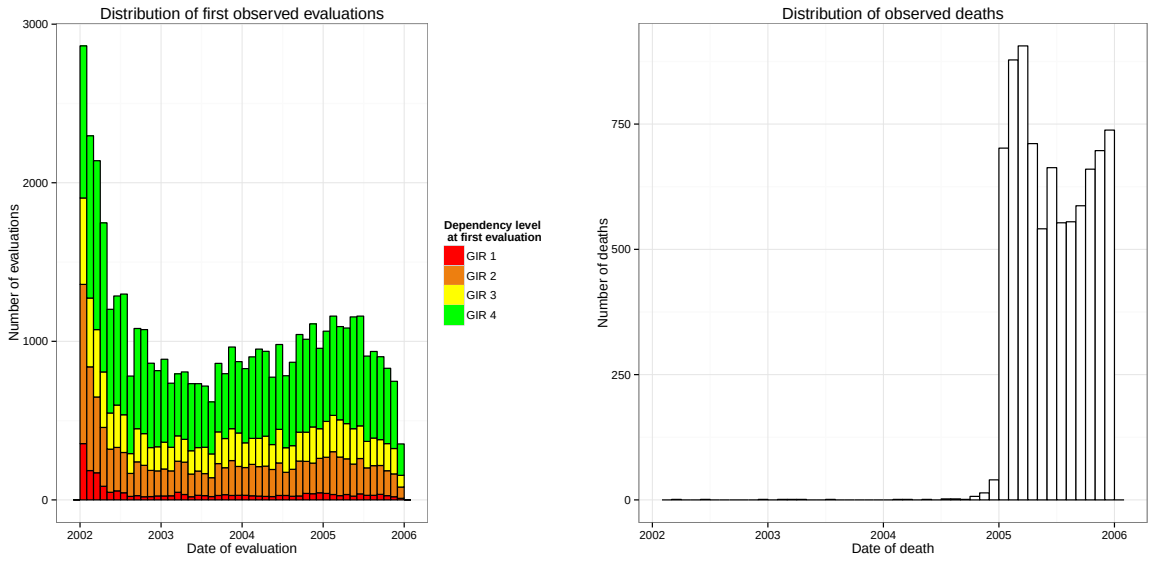


Figure 2.2: Left: distribution of first evaluations of Gir observed; right: distribution of observed deaths.

2.2.2 Discussion about the observation process

Before using the data in a model, we need to make and discuss several assumptions on the associated observation process. First of all, we note that it only contains evaluation dates whereas we are looking for the exact times at which transitions between states occurred. When we have two consecutive evaluations giving different results, we know that the transition occurred between the two evaluation times. Such phenomenon is called interval censoring. Methods to cope with interval-censored data have been developed in cases where the censoring process is non-informative (Foucher et al., 2007), for example when we have pre-scheduled evaluations, which is not the case for our process. In the case of an informative observation process, we need to be able to specify a model for the dependency between transitions and evaluations and misspecification of the model can alter the results (Chen et al., 2010).

For the APA data, however, evaluations can be requested by individuals or their usual doctor as they please, are free and can generally be obtained on short notice. We can therefore assume that, as soon as a transition occurs, the individual will request an evaluation, and consequently, it can be assumed that transitions can only occur at the evaluation times. Nevertheless, it should be noted that evaluations can still take place when no transitions has occurred. Besides, should the transition time and the evaluation time be very different, the results of the model

still holds in an insurance context. Indeed, the relevant information in this context would be the time at which the claim is filled, and therefore the same delay we observed between the transition time and the evaluation time would still be present in this case. Hence, for the remaining of the paper, we work under the assumption that transition only occur at evaluation times.

Another issue that needs to be discussed is that, at the start of the observation period, some individuals are already dependent. We have seen previously that, because of the "stock effect", any individual who entered the APA in 2002 may have already been dependent for some time already. This phenomenon, known as left-truncating, cannot be handled easily, except in the case of simple memoryless time-homogeneous Markov process. Therefore, we decide to remove individuals who became dependent before year 2003 from the data. This way, we obtain left truncated data which should not be affected by the "stock effect". This approach should not generate any bias, as the people who became dependent in 2002 should not be any different than their 2003 counterparts. However it reduces the number of data available, and the observation period is shortened from 4 to 3 years.

The same issue arises when we look at the age of the individuals at first evaluation. The APA is only granted to people aged 60 and more. Hence, for individuals entering the APA at 60, we don't know exactly for how long they have been dependent. Hence, we decide to remove individuals who were less than 61 at first evaluation from the data. As dependency before 60 is quite rare, the impact is very limited in this case.

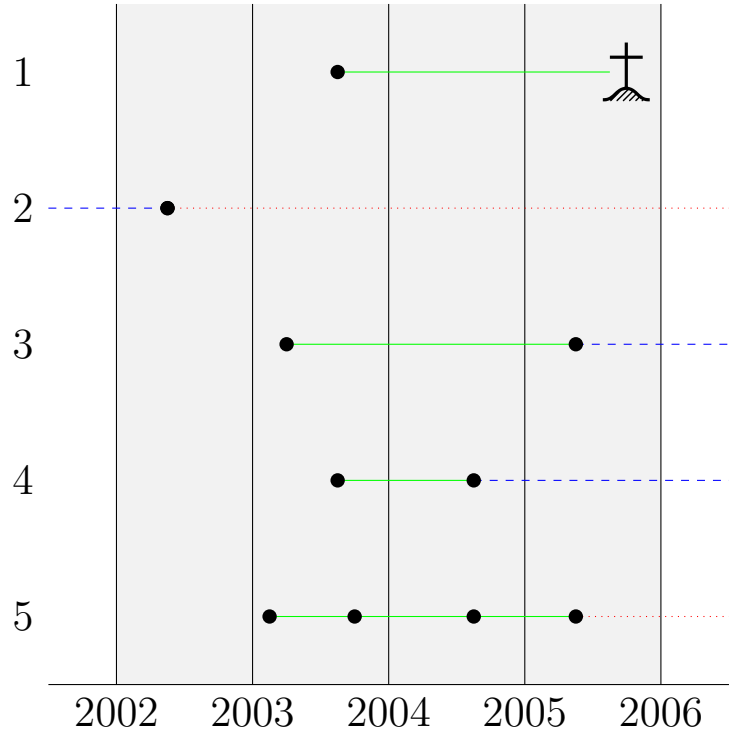
In the case where there are more than 4 evaluations for an individual, the subsequent evaluations are not recorded. This is actually a form of censoring because the only event that can be observed past the fourth evaluation is death. Information about death (resp. survival) of the individual after this point could still be taken into account, but we would not know the state of dependency of the individual at the moment of death (resp. at the end of the observation period). We consider instead that the observation period ends with the fourth evaluation. This could induce bias in the observation, but there are only few individuals for which 4 evaluations were observed, which may explain why this number was picked in the first place.

At the end of the observation period, due to the fact that death is not observed until 2005, we have several possible configurations:

- If death of the individual has been observed, then we have a full trajectory.

- If death of the individual has not been observed, and the last observed evaluation occurred in 2005, then we know that the individual is alive at the end of the observation period, because otherwise his death would have been recorded. The trajectory is then right censored, and the information about survival of the individual between his last observation and the end of the observation period should be taken into account through a term in the likelihood function.

- If death of the individual has not been observed, and the last observed evaluation occurred before year 2005, then the individual has either died before 2005, or he is still alive at the end of the observation period, because otherwise death during year 2005 would have been recorded. In both cases, as no new evaluation has been made, we know that no other transition may have happened. We call this phenomenon partial censoring, as only the death of the individual is actually censored. This last configuration may seem quite complex, but its probability can actually be expressed quite easily in terms of likelihood, as we will see in next section.



- 1: Fully observed trajectory, from the entry into dependency to the moment of death.
- 2: Discarded trajectory, since the entry into dependency occurs during year 2002.
- 3: Right-censored trajectory, with no missing information.
- 4: Partially censored trajectory, as death may have occurred but not been observed.
- 5: Frequency-censored trajectory, the observation stops after the fourth evaluation.

Figure 2.3: Evaluation process for the APA data: examples of trajectories. Plain (resp. dashed/dotted) lines correspond to observed (resp. censored/removed) parts of the trajectories in dependency.

Finally, let's note that in the data, there are cases where the dependency level of the individual improves between two consecutive evaluations. According to the definition used by French insurers, dependency is considered to be a consolidated and irreversible state. A temporary disability is not considered as dependency. Those improvements should thus be considered as errors in the evaluation diagnosis. Indeed, elderly people can have good days or bad days, which may cause the result of the evaluations to vary from one visit to another. Considering dependency as an irreversible process makes the estimation of the model much easier. Besides, from an insurance point of view, it is safer to consider there is no improvement, as an insured life will not be eager to declare any improvement in his health status that means no more benefit from the insurance. Unfortunately, the insurer has no way to detect such improvements as the cost of periodic controls would be way over their potential benefits. Consequently, in case of improvements, we consider that the state of dependency of the individual remains the same. Figure 2.3 provides an example of different kind of trajectories encountered and previously described.

Features of the observation process can be summarized as follows:

- It is left-truncated, according to both the calendar year and the age. People who are already dependent or dead on the 1st of January 2003, or are already dependent or dead at the age of 61 do not appear in the data.

- It is right-truncated. Only people who become dependent before the 31st of December 2005 appear in the data.
- It is partially censored. Death is not observed until the 1st of January 2005.
- It is frequency censored. The observation period ends after the 4th evaluation.
- It is right-censored. The observation period ends at the 31st of December 2005.

After processing the data, we have information about 31,731 dependent individuals, but only 9,270 observed transitions.

2.3 Description of the model

2.3.1 Introduction of the model

The model we present (see Figure 2.4) has 6 states: autonomy, death, and 4 different levels of dependency, Gir 1 being the most severe and Gir 4 the least severe (refer to table 2.1 for the description of those levels). Numbers will be associated with states: 5 for autonomy, 4 to 1 for Gir 4 to Gir 1 respectively and 0 for death. The model is unidirectional: transitions can only occur toward a more severe state of dependency or death. The dependency incidence rate $i(s)$ as well as the autonomous mortality rate $q^a(s)$, where s is the age of the individual, cannot be estimated from the APA data and we use exogenous information for those laws. In addition to the incidence rates, we also need to determine the distribution of the initial state of dependency, which will be estimated later in this section. This leaves us with 10 transitions for which we provide a semi-Markov model.

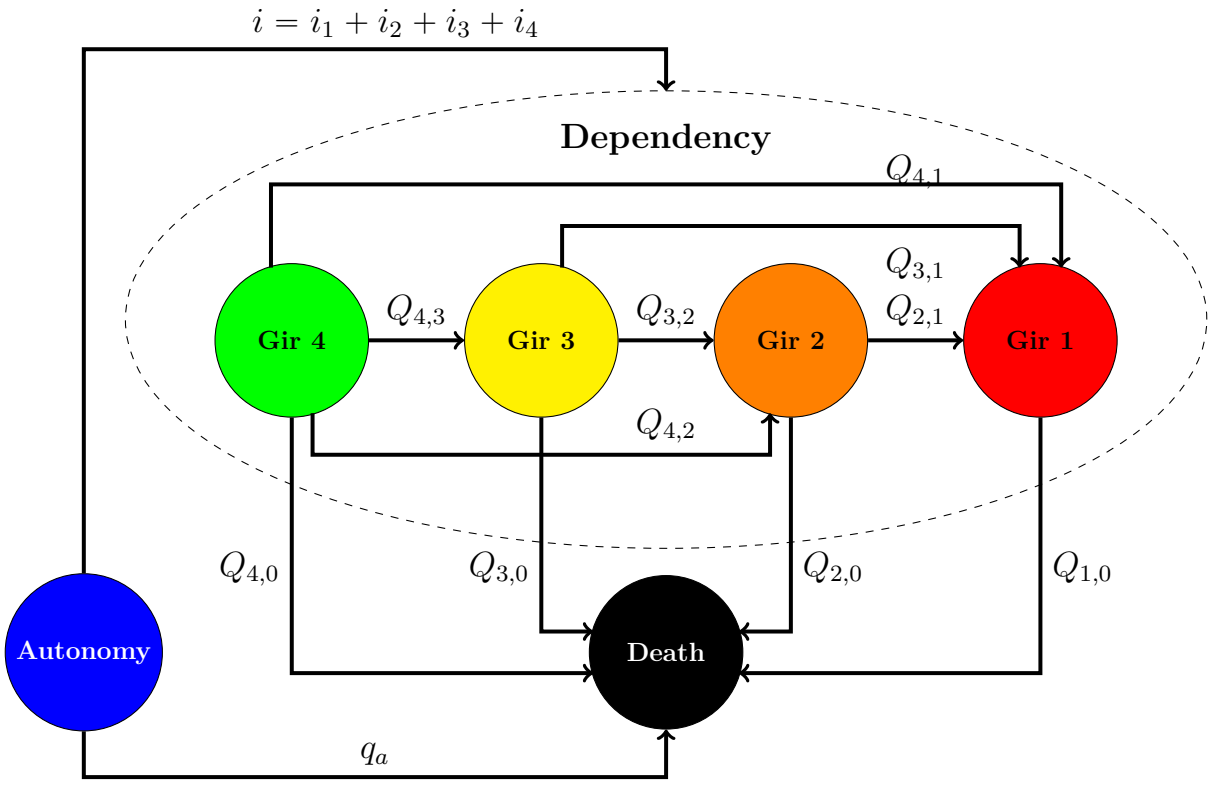


Figure 2.4: Semi-Markov model with 4 states of dependency. Probabilities of transitions originating from dependency states are defined using their semi-Markov kernel $Q_{i,j}$.

2.3.2 Elements of semi-Markov theory

Definition. *Let $Y = (Y_t)_{t\geqslant 0}$ be a right-continuous process which takes its values in a finite set of states $E \subset \mathbb{N}$. Let $X = (X_n)_{n\in\mathbb{N}}$ be the sequence of consecutive states visited by the process and $T = (T_n)_{n\in\mathbb{N}}$ the sequence of consecutive times at which changes in the value of Y occur. Y is called a semi-Markov process if (X, T) is a multidimensional Markov process, or more formally, for all $n \in \mathbb{N}$, $x > 0$ and $j \in E$*

$$\mathbb{P}(X_{n+1} = j,\ T_{n+1} - T_n \leqslant x \mid X_0,\ T_0,\ \ldots,\ X_n,\ T_n) = \mathbb{P}(X_{n+1} = j,\ T_{n+1} - T_n \leqslant x \mid X_n, T_n).$$

Furthermore, if those probabilities do not depend on T_n, then Y is called a time-homogeneous semi-Markov process.

The semi-Markov process keeps a memory of how long it has been in the current state. Nevertheless, similarly to the classical Markov process, it does not keep any memory about the previously visited states or transition times.

Definition. *Let Y be a time-homogeneous semi-Markov process, X (resp. T) the sequence of visited states (resp. transition times) associated with Y. We define*

- *The semi-Markov kernel*

$$\forall i, j \in E,\ \forall\ 0 \leqslant x, \quad Q_{i,j}(x) = \mathbb{P}(X_{n+1} = j,\ T_{n+1} - T_n \leqslant x \mid X_n = i).$$

- *The jump probabilities*

$$\forall i, j \in E,\ , \quad p_{i,j} = \mathbb{P}(X_{n+1} = j | X_n = i) = \lim_{x\to+\infty} Q_{i,j}(x).$$

- *The duration laws*

$$\forall i, j \in E, \forall\ x \geqslant 0, \quad F_{i,j}(x) = \mathbb{P}(T_{n+1} - T_n \leqslant x \mid X_n = i, X_{n+1} = j) = \begin{cases} \dfrac{Q_{i,j}(x)}{p_{i,j}} & \text{if } p_{i,j} > 0, \\ 0 & \text{otherwise.} \end{cases}$$

A semi-Markov process is entirely determined by its semi-Markov kernel and the initial states distribution. Besides, it can be noted that X is a discrete time Markov chain with values in E, whose transition probabilities are precisely the jump probabilities $p_{i,j}$.

We have the fundamental relation: $Q_{i,j}(x) = p_{i,j} \times F_{i,j}(x)$. To fully describe a semi-Markov process, we therefore need to model both the jump probabilities and the duration laws.

2.3.3 Model

Jump process

According to the previous definition, jump probabilities are indeed probabilities constrained by the following relations

$$\begin{cases} \forall i \neq j \in E,\ 0 \leqslant p_{i,j} \leqslant 1, \\ \forall i \in E,\ \displaystyle\sum_{j\neq i} p_{i,j} = 1. \end{cases}$$

Those probabilities will be estimated later alongside other parameters of the model.

For sake of clarity, we omit indexes of parameters corresponding to transitions $i \to j$ in the remain of this section.

Base duration laws

The base element of our model for duration law is the Weibull distribution. The hazard rate associated with this distribution is a single factor polynom, which degree depends on the shape parameters, which makes it very flexible. Besides, the distribution only consists of two parameters, and the survival function can easily be inverted, making it an excellent choice for optimization purposes. This distribution is commonly used in reliability theory, one can refer to Jiang and Murthy (1997) or Bucar et al. (2004), and was applied to model the dependency process by Lepez (2006).

The Weibull distribution can be described using either of those functions

- Survival function $S_0(x) = e^{-\sigma x^\nu}$,
- Density probability $f_0(x) = -\dfrac{\mathrm{d}S_0(x)}{\mathrm{d}x} = \sigma\nu x^{\nu-1}e^{-\sigma x^\nu}$,
- Hazard rate $h_0(x) = \dfrac{f_0(x)}{S_0(x)} = \sigma\nu x^{\nu-1}$,

with $\sigma, \nu > 0$.

The case $\nu = 1$ corresponds to an exponential distribution with constant hazard rate which brings us back to a time-homogeneous Markov model.

Integration of covariates

To take the covariates into account in the model, we make the assumption of proportional hazard rates, which was introduced by Cox, see for example (Cox and Oakes, 1984), and has since been used in a lot of publications. Our model will only consider two covariates: gender and age of entry in dependency. Gender is a binary covariate, so we only need to introduce a single parameter α for each transition in the model to account for its effect on hazard rates. On the other hand, age at entry in dependency can vary on a continuous scale. Nevertheless, we still decide to use a single parameter β so that if s is the age of entry in dependency, the hazard rate is multiplied by a factor $e^{\beta s}$. In this case, the choice of the exponential function is no longer neutral, the underlying assumption is that being one year older will have the same multiplicative effect on the hazard rate, regardless of the age s.

Finally, we have, for $x > 0$, $g \in \{1; 2\}$, $s > 0$

$$\begin{aligned} h_1(x|g,s) &= h_0(x)e^{\alpha g+\beta s} \\ S_1(x|g,s) &= S_0(x)^{\exp(\alpha g+\beta s)} \end{aligned}$$

where $\alpha, \beta \in \mathbb{R}$.

The proportional hazard model is widespread in survival data analysis, it is very simple to implement and requires few additional parameters. As we have proportional hazard for every transition in the process, our model is said to have semi-proportional hazard. It means that the propotional hazard assumption is less restrictive in our case that it would be otherwise.

Introduction of a transverse static frailty

While the previously introduced covariates should explain part of the heterogeneity in the trajectories, we believe that the pathologies, which are unobserved in the data, remain the main source of heterogeneity. A rough categorization of pathologies causing dependency would give us two groups, with on one hand, cancer, strokes and some other cardiovascular diseases, on the other hand, dementia, which is mainly caused by Alzheimer's disease, neurological diseases and arthrosis. The first group of diseases is associated with quick trajectories, while the second group goes in pair with a slower degenerative process which results in longer trajectories.

To take this heterogeneity into account, we introduce a transverse static frailty in the model. Frailty can be seen as an additional covariate whose value has an impact on the trajectory but the variable itself cannot be observed. We consider that each individual has its own frailty, which does not vary over time and impacts every transition the individual will undergo. Hence, it is both static and transverse. One of the main limitations of the semi-Markov model is that no information can be carried over to the next transition, and therefore the duration of transitions are uncorrelated. The introduction of frailty allows us to bypass this limitation. Instead of a propotional hazard through frailty, we could consider a hidden mixture model which is a more general case, but it would require additional parameters.

We assume that frailty u is distributed according to a Bernouilli law with parameter $\eta(g,s) \in]0\,;1[$ where g is the gender and s the age of entry in dependency. Furthermore, we use a generalized linear model with a logit link function to express the impact of g and s on $\eta(g,s)$

$$\log\left(\frac{\eta}{1-\eta}\right) = \eta_0 + \eta_1 \times g + \eta_2 \times s.$$

For each transition, the impact of frailty will be modeled by a single parameter γ through a proportional hazard rate e^{γ}, so this impact may be different for each transition. The conditional laws associated with this frailty u are

$$\begin{aligned}
h(x|g,s,u) &= e^{\gamma u} h_1(x|g,s) = \begin{cases} h_1(x|g,s) & \text{if } u=0, \\ e^{\gamma} h_1(x|g,s) & \text{if } u=1, \end{cases} \\
S(x|g,s,u) &= S_1(x|g,s)^{\exp(\gamma u)} \\
f(x|g,s,u) &= h(x|g,s,u) \times S(x|g,s,u) = e^{\gamma u} h_1(x|g,s) S_1(x|g,s)^{\exp(\gamma u)},
\end{aligned}$$

with $\gamma > 0$ to guarantee the identifiability of the model.

Summary of the model

Finally, our duration model is characterized by the following laws

$$\begin{aligned}
u &\sim \mathcal{B}\left(\frac{e^{\eta_0+\eta_1\times g+\eta_2\times s}}{1+e^{\eta_0+\eta_1\times g+\eta_2\times s}}\right), \\
\lambda(g,s,u) &= \sigma e^{\alpha g+\beta s+\gamma u}, \\
h(x|g,s,u) &= \nu\lambda(g,s,u)x^{\nu-1}, \\
S(x|g,s,u) &= \exp(-\lambda(g,s,u)x^{\nu}), \\
f(x|g,s,u) &= \nu\lambda(g,s,u)x^{\nu-1}\exp(-\lambda(g,s,u)x^{\nu}),
\end{aligned}$$

where

- g the gender and s the age of entry in dependency,
- $\sigma, \nu, \alpha, \beta, \gamma, \eta$ are parameters defined for each transition whose domains of definition are summarized in Table 2.2,
- η_0, η_1, η_2 are the parameters of the frailty, defined globally.

With the jump probabilities, we have a total of 59 parameters that need to be estimated. Figure 2.5 illustrates the roles the different parameters play in the determination of the duration law, and Table 2.2 their domain of definition and short descriptions as a reminder.

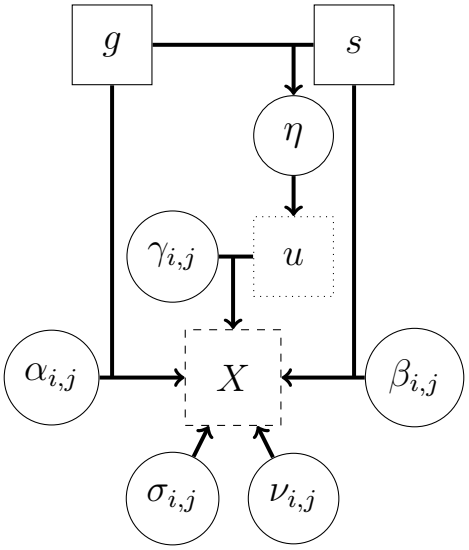


Figure 2.5: Role of parameters in the determination of the duration law X.

Parameter	Count	Domain of definition	Description
$p_{i,j}$	6	$0 \leqslant p_{i,j} \leqslant 1, \quad \sum_{k \neq i} p_{i,k} = 1.$	jump probabilities
$\sigma_{i,j}$	10	$\sigma_{i,j} > 0$	scale parameters of Weibull laws
$\nu_{i,j}$	10	$\nu_{i,j} > 0$	shape parameters of Weibull laws
$\alpha_{i,j}$	10	$\alpha_{i,j} \in \mathbb{R}$	impact of gender
$\beta_{i,j}$	10	$\beta_{i,j} \in \mathbb{R}$	impact of age of entry in dependency
$\gamma_{i,j}$	10	$\gamma_{i,j} > 0$	impact of frailty
η_0, η_1, η_2	3	$(\eta_0, \eta_1, \eta_2) \in \mathbb{R}^3$	distribution of frailty

Table 2.2: Summary of the different parameters used in the model.

2.3.4 Likelihood function

We denote by T_1 (resp. T_2) the beginning of the death observation period (resp. the end of the observation period). We recall that in our data, T_1 is the 1st of January 2005 and T_2 the 31st of December 2005. In addition, for each individual p, we introduce

- n_p the number of observed transitions,
- $X^p = (X_k^p)_{1 \leqslant k \leqslant n_p}$ the sequence of visited states,
- $t^p = (t_k^p)_{1 \leqslant k \leqslant n_p}$ the sequence of transition times,
- indexes (δ_1^p, δ_2^p) indicating if the trajectory is right censored or partially censored, where

$$(\delta_1^p, \delta_2^p) = \left(\mathrm{I}[X_{n_p}^p \neq 0, T_1 \leqslant t_{n_p} < T_2], \mathrm{I}[X_{n_p}^p \neq 0, t_{n_p} < T_1] \right),$$

- a vector of covariates $Z_p = (g_p, s_p)$ where g_p is the gender and s_p the age of entry in dependency.

.

The log-likelihood function has the following expression

$$l = \sum_{p=1}^{N} \log \left(\eta(g_p, s_p) \times l_p^1 + (1 - \eta(g_p, s_p)) \times l_p^0 \right),$$

$$l_p^u = \left(\prod_{k=1}^{n_p-1} \underbrace{C_{X_k^p, X_{k+1}^p}(t_{k+1}^p - t_k^p | g_p, s_p, u)}_{\text{observed transition}} \right) \times \underbrace{C^1_{X_{n_p}^p}(T_2 - t_{n_p}^p | g_p, s_p, u)^{\delta_1^p}}_{\text{right censoring}} \times \underbrace{C^2_{X_{n_p}^p}(T_1 - t_{n_p}^p, T_2 - t_{n_p}^p | g_p, s_p, u)^{\delta_2^p}}_{\text{partial censoring}}$$

where, for $i, j \in E$, $x > 0, x_1 \geqslant x_2 > 0$, $g \in \{1, 2\}, s > 0, u \in \{0, 1\}$, N is the number of observed individuals and

- $C_{i,j}(x|g, s, u) = p_{i,j} \times f_{i,j}(x|g, s, u)$ is a term associated with an observed transition. First term $p_{i,j}$ gives the probability of observing the transition $i \to j$ and $f_{i,j}(x|g, s, u)$ the conditional probability of this transition happening precisely after a duration x has passed,
- $C_i^1(x|g, s, u) = \sum_{j<i} p_{i,j} \times S_{i,j}(x|g, s, u)$ is a right-censoring term, whose value is the probability of remaining in state i for a duration x,
- $C_i^2(x_1, x_2|g, s, u) = p_{i,0} \times (1 - S_{i,0}(x_1|g, s, u)) + \sum_{j<i} p_{i,j} \times S_{i,j}(x_2|g, s, u)$ is a composed censoring term. First term $p_{i,0} \times (1 - S_{i,0}(x_1|g, s, u))$ is the probability of dying before a time x_1 has passed, and second term $\sum_{j<i} p_{i,j} \times S_{i,j}(x_2|g, s, u)$ the probability of remaining in state i for a duration x_2. As $x_1 < x_2$, both events are exclusive and their sum gives the likelihood associated with individuals for which death may have happened before T_1 and has not been observed.

2.4 Results and trajectories

In this section, we present results of the model described in the previous section.

2.4.1 Estimation of parameters

We need to estimate 59 parameters by maximizing a likelihood function gathering information about the 31,731 trajectories we extracted from the APA data. We use the Nelder-Mead algorithm (Nelder and Mead, 1965) to maximize the likelihood function. This algorithm is based on successive evaluations of the optimization function on the vertices of a simplex which evolves in accordance with the results of those evaluations. Geometrical transformations like reduction, extension, or reflection are applied to the simplex in order to explore the most promising parts of the solution space. This algorithm offers a very powerful alternative to the Newton-Raphson algorithm when the computation of derivatives is not possible, and the solution space too large to use evolutionary algorithms. An implementation of the Nedler-Mead algorithm is provided in the programming language `R` (R Core Team, 2016) trough the function constrOptim(), which allows for linearly constrained optimization.

As the optimization algorithm can converge toward a local optimum of the likelihood function, we perform 100 iterations of the algorithm with randomly generated values for the initial parameters, and keep the best solution found at the end. The likelihood of the solution can vary a lot between two iterations, which prevents the use of tests based on likelihood. However, it seems to always converge toward one of a few local optima. With 100 iterations, relative stability of the result is achieved, with the best five solutions yielding similar parameters results. The estimated parameters can be found in Table 2.3. We note that gender has a significant impact on the duration laws. Women ($g = 2$) have lower hazard rates than men ($g = 1$) for every transition, especially the transitions that lead to death. For example the hazard rate for the transition $4 \to 0$ is 2.5 times higher for men than for women. Consequently women survive longer than men in dependency. Besides, hazard rates also increase with age, for every

transition. The hazard rate for transition $4 \to 2$ is 4 times higher for someone aged 95 than for someone aged 65.

Transitions	$p_{i,j}$	$\sigma_{i,j}$	$\nu_{i,j}$	$\alpha_{i,j}$	$\beta_{i,j}$	$\gamma_{i,j}$	η_0	η_1	η_2
$4 \to 3$	0.27	0.0107	1.43	−0.23	0.044	0.13	0.93	−0.06	−0.04
$4 \to 2$	0.34	0.0043	1.43	−0.15	0.046	0.62			
$4 \to 1$	0.03	0.0005	1.65	−0.11	0.070	1.17			
$4 \to 0$	0.37	0.0413	1.39	−0.90	0.039	3.09			
$3 \to 2$	0.43	0.0375	1.43	−0.12	0.029	0.57			
$3 \to 1$	0.05	0.0136	1.59	−0.22	0.044	0.22			
$3 \to 0$	0.52	0.0439	1.23	−0.73	0.037	2.95			
$2 \to 1$	0.13	0.1279	1.49	0.06	0.008	0.21			
$2 \to 0$	0.87	0.0515	1.23	−0.82	0.037	3.38			
$1 \to 0$	1.00	0.0711	1.14	−0.61	0.036	3.64			

Table 2.3: Estimated values of parameters in the final model.

Estimated probability of frailty can be found in Figure 2.6. The probability of having a positive frailty decreases with age, from 25 % at 50 to 5 % at 100. This probability is very close for men and women. The impact of frailty is directly related to the severity of the transition. Indeed, frailty has very high impact on transitions toward death or a non-consecutive state, increasing hazard rate by up to 3700 % in the case of transition $1 \to 0$ but a lower impact on transitions between consecutive states, with only a 14 % increase for transition $4 \to 3$.

Figure 2.7 shows the duration law we obtain for the transition from Gir 4 to death. The individuals with frailty have very high hazard rates, and therefore they also die very quickly. This results in high hazard rates for the general population over the first year of dependency, with a decrease as the individuals with frailty die.

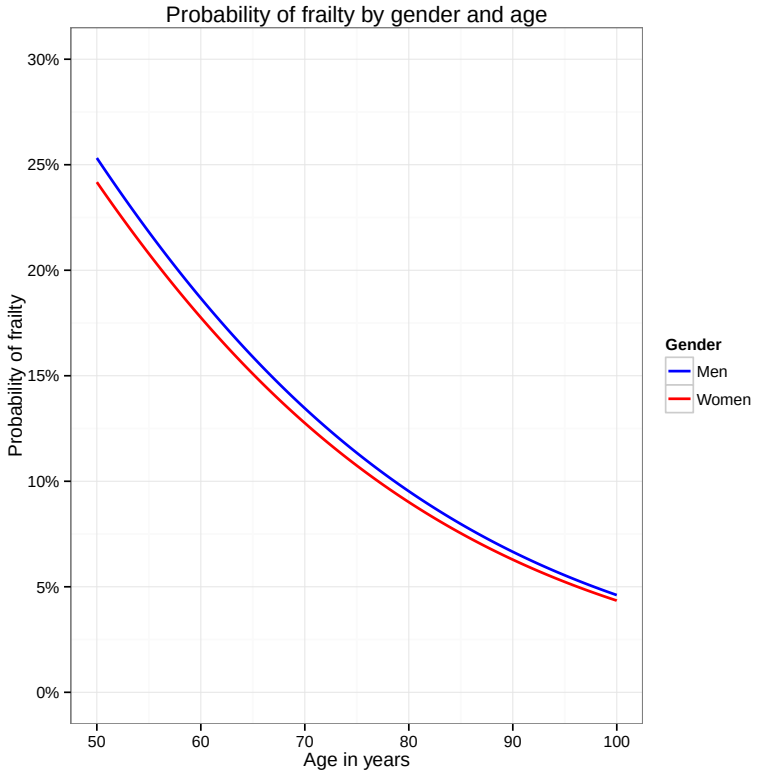


Figure 2.6: Estimated probability of frailty with respect to gender and age of entry.

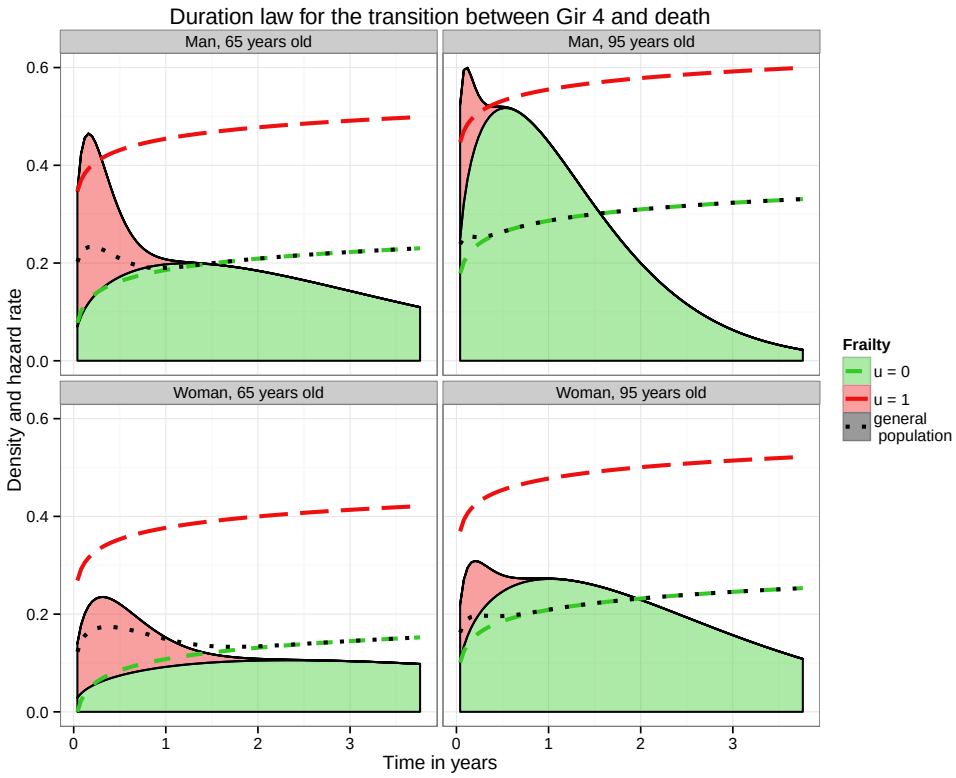


Figure 2.7: Duration law for the transition from Gir 4 to death; Grayed areas represent the associated density for people with and without frailty. Plain (resp. dashed/dotted) lines represent the hazard rate for the general population (resp. for people with/without frailty).

In addition, we also need to determine the distribution of the initial state of dependency. For an individual of gender g who becomes dependent at age s, we want to estimate the probability for the initial state of dependency to be Gir i for $i \in \{1; 2; 3; 4\}$.

For each gender and each age of entry in dependency between 65 and 95, we look at the state of entry in dependency for the people who became dependent at that age. It can be seen as a sample of a multinomial distribution of parameters the probabilities of entries in dependency at that age. We use empirical estimators to determine those probabilities. Furthermore, as we have at least 5 entries in dependency for each age and each state of dependency, the Fischer condition is met and it makes sense to compute normal confidence intervals for those rates.

For each state of dependency, we then perform smoothing of the empirical probabilities by age of entry using the unidimensional Whittaker-Henderson method, as described in Planchet and Thérond (2006), with parameters $h = 3$ and $z = 2$, and the weight of each probability being equal to the number of entries in dependency at that age. This choice of weights ensures that for each age, the sum of the smoothed probabilities for the different states of entry is still equal to 1, while this result would not hold for other types of interpolation as for example a generalized linear model. For an individual of gender g and age of entry in dependency s, the state of entry is determined using the estimated probabilities that we note $(e_i(g, s))_{i\in\{1,4\}}$ and which will be used latter for the simulation of trajectories.

Figure 2.8 highlights a significant difference between men and women. At age 65, men are more likely than women to directly enter a severe state of dependency. However, this trend shifts over time and the situation is reversed for ages over 90. This phenomenon can be interpreted by considering the underlying pathologies. On one hand, we know that men are more subject to cancers, which result in very severe dependency very quickly, and the incidence rate for dependency caused by cancer becomes lower with age. On the other hand, women are more affected by dementia, which become more frequent at higher ages. Early states of dementia are not recognized by the AGGIR grid as dependency as long as they are not associated with

functional limitations. When those limitations occur, the dependency state may already be very severe.

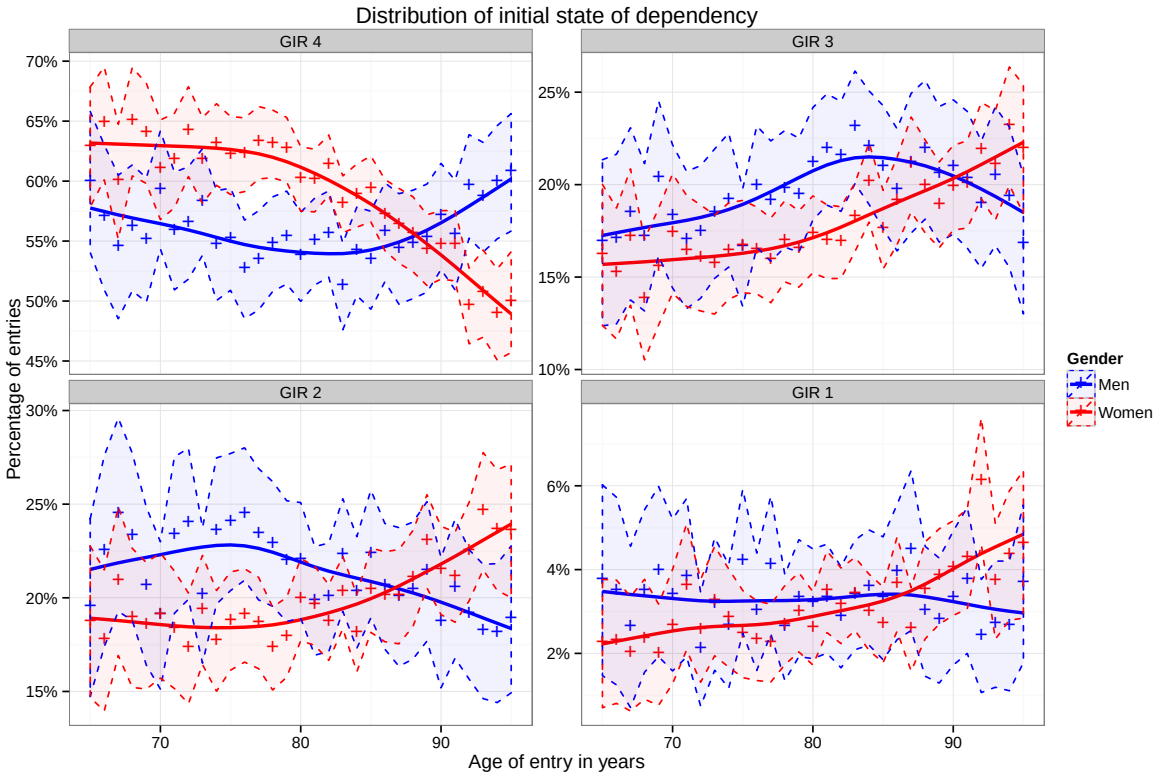


Figure 2.8: Distribution of initial states of dependency. Empirical probabilities (circles for men, triangles for women) with associated normal 95 % confidence intervals , and results of Whittaker-Henderson smoothing (plain line for men, dashed line for women).

2.4.2 Simulation of trajectories

In order to generate trajectories, in an algorithmically efficient way, we need to define several quantities, for individuals of gender $g \in \{1; 2\}$. Note that incidence rate and autonomous mortality rate are defined so that every year i is applied first and q_a is applied on the remaining autonomous people.

- $p_i^g(s, x)$ the probability at age s to become dependent at age $s + x$,
- $p_a^g(s, x)$ the probability at age s to die at age $s + x$ without ever becoming dependent,
- $p_i^g(s)$ the probability at age s to become dependent one day,
- $p_a^g(s)$ the probability at age s to die without ever becoming dependent,
- $p_{|i}(s, x)$ the probability at age s, knowing one will become dependent before dying, that the entry in dependency occurs at age $s + x$,
- $p_{|a}(s, x)$ the probability at age s, knowing one will die without ever becoming dependent, that the death occurs at age $s + x$,

where $s, x \in \mathbb{N}$. We note that $p_i^g(s) + p_a^g(s) = 1$ for every $s \in \mathbb{N}$.

Those quantities can be linked to the incidence and mortality rates

$$
\begin{aligned}
p_i^g(s,x) &= \left(\prod_{k=0}^{x-1}(1-i^g(s+k))(1-q_a^g(s+k))\right) i^g(s+x),\\
p_a^g(s,x) &= \left(\prod_{k=0}^{x-1}(1-i^g(s+k))(1-q_a^g(s+k))\right) (1-i^g(s+x)) \times q_a^g(s+x),\\
p_i^g(s) &= \sum_{x=0}^{\infty} p_i^g(s,x),\\
p_a^g(s) &= \sum_{x=0}^{\infty} p_a^g(s,x),\\
p_{|i}^g(s,x) &= \frac{p_i^g(s,x)}{p_i^g(s)},\\
p_{|a}^g(s,x) &= \frac{p_a^g(s,x)}{p_a^g(s)},
\end{aligned}
$$

where $s, x \in \mathbb{N}$.

Simulation algorithm

The trajectory of an individual p characterized by his gender g_p and his age s_p^0 at the start of the simulation consists of

- the number of visited states: n_p. As our dependency process is assumed to be unidirectional, and we have 4 states of dependency and death as a terminal state, we have $2 \leqslant n_p \leqslant 6$,
- a set of visited states: $X^p = (X_k^p)_{1 \leqslant k \leqslant n_p}$,
- a set of transition times: $t^p = (t_k^p)_{1 \leqslant k \leqslant n_p}$,

where t_0^p is the time at which the simulation starts, and $X_1^p = 5$, as we only consider individuals who are autonomous at the start of the simulation.

To simulate the trajectory of an individual p, we use the following algorithm

1. We set $X_1^p = 5$, and $t_1^p = s_p^0$.
2. With probability $p_a(s_p^0)$, the individual dies without becoming dependent. In this case, we set $X_2^p = 0$ and go to step 3. Otherwise the individual becomes dependent and we go to step 4.
3. The age of death is $t_2^p = s_p^0 + x + r$ where x is distributed according to the probabilities $p_{|a}(s_p^0, l)$ for $l \in \mathbb{N}$ and r is the fractional part of the year, uniformly distributed on $[0\,;1]$. The trajectory ends with the death of the individual and the algorithm stops, with $n_p = 2$.
4. The age of entry in dependency is $t_2^p = s_p^0 + x + r$ where x is distributed according to the probabilities $p_{|i}(s_p^0, l)$ for $l \in \mathbb{N}$ and r is the fractional part of the year, uniformly distributed on $[0\,;1]$. We note $s_p = t_2^p$ the age of entry in dependency and we set $k = 2$.
5. The probabilities $(e_i(g_p, s_p))_{i \in \{1,4\}}$ are used to determine the state of entry in dependency X_2^p.
6. The frailty u_p is set to 1 with probability $\eta(g_p, s_p)$ and to 0 otherwise.

7.
- The next state X_{k+1}^p is determined with respect to probabilities $p_{X_k^p,j}$ for all $j \in \{0,\dots,X_k^p-1\}$.
- To determine the time spent in state X_k^p, we first draw a random variable x distributed uniformly on $[0\,;1]$. We then set $t_{k+1}^p = t_k^p + y$ where y is defined below

$$y = \left[\frac{e^{-(\alpha g_p+\beta s_p+\gamma u_p)}}{\sigma}\ln\left(\frac{1}{1-x}\right)\right]^{\frac{1}{\nu}}.$$

  We deliberately omitted indexes in the previous formula for the sake of clarity. They should be X_k^p and X_{k+1}^p.
- We increment k. If the individual is still alive, i.e. $X_k^p \neq 0$, we repeat the steps of 7. Otherwise, the trajectory ends with the death of the individual and $n_p = k$.

2.4.3 Statistics on simulated trajectories

Figure 2.9 (left) gives the distribution of the age of entry in dependency, which only depends on exogenous data used for incidence and autonomous mortality rates. According to those laws, for a population of individuals aged 60, men become dependent at 84 on average and women at 87. Figure 2.9 (right) gives the distribution of survival time in dependency, regardless of age of entry. The density is very high during the first year, due to frailty. Besides, women survive in average 4 years in dependency while men survive a little less than 3. Those results could be compared to those presented in Debout (2010), a study of the APA data performed over 300,000 trajectories over a 6 years period. Figure 2.10 gives the life expectancy in dependency and the average time spent in each state, as a function of the age of entry. Despite the probability of frailty being lower at higher ages of entry, the life expectancy decreases with age, for both men and women.

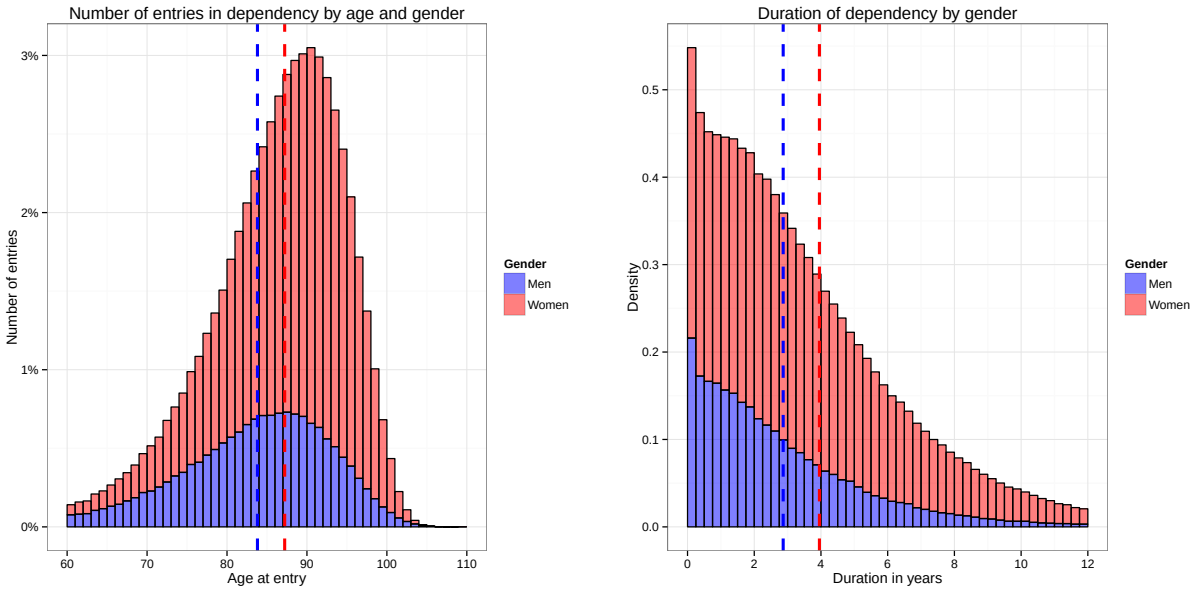


Figure 2.9: Left: distribution of entries in dependency by age as a percentage of total entries for a population of 1,000,000 at 60. Right: density for the distribution of survival time in dependency, computed on the same population. For both graphs, plain lines (resp. dashed lines) represent the mean value for men (resp. women).

Figure 2.10: Life expectancy, and its breakdown by state of dependency, with respect to the age of entry in dependency, for men (left) and women (right), based on 1,000,000 simulations.

2.5 Application to pricing

2.5.1 Pricing methodology

An estimator for the premium

We consider a long-term care insurance product characterized, on one hand by a sequence of periodic benefit cash flows B, and on the other hand by a sequence of periodic premium cash flows P such that B and P have the same periodicity. We assume that conditions for the payment of the benefit (resp. the premium) has been set in the product description. For a fixed amount of benefit, we define the pricing of the product as finding the corresponding amount of premium so that the expectancy of the discounted cash flow of premium matches the expectancy of the discounted cash flows of benefit.

For a sequence of periodic cash flows $F = (F_i)_{i\in\mathbb{N}}$, and a fixed actuarial rate $\tau \geqslant 0$, we define the associated Net Present Value (NPV)

$$\mathrm{NPV}(F) = \sum_{i=0}^{\infty} \frac{F_i}{(1+\tau)^i}.$$

If we further assume that the amount of every premium cash flow is either null or equal to a fixed amount p^*, which covers the case of single premium and level premiums product, the problem becomes finding p^* such that

$$p^* = \frac{\mathrm{NPV}(B)}{\mathrm{NPV}(\mathbb{I}_P)},$$

where $\mathbb{I}_P = \dfrac{P}{p^*}$ is the sequence of premium unit cash flows.

Most insurance models rely on a discrete time scale model, for which a closed formula for the premium p^* can be calculated. In a multi-states continuous scale model however, multiple integrals appear in the equivalent formula, for which an analytical solution does not exist. Therefore we have to rely on another method for the pricing.

We decide to use a Monte Carlo method, which relies on the simulation of trajectories in order to find an estimate which converges toward the right amount of premium. We use the following methodology:

- We generate n trajectories using the algorithm provided in the previous section.
- For each trajectory $k \in \{1; \dots; n\}$, we determine the NPV of both the benefit cash flows NPV_k^B and the premium unit cash flows NPV_k^P.
- We use the following estimator for the amount of premium

$$\widehat{p}_n = \frac{\frac{1}{n}\sum_{k=1}^{n} \text{NPV}_k^B}{\frac{1}{n}\sum_{k=1}^{n} \text{NPV}_k^P}.$$

According to the law of large numbers, this is a consistent estimator of p^*, i.e. $\widehat{p}_n \underset{n\to+\infty}{\to} p^*$ almost surely.

Uncertainty on premium estimation

For a sample of n trajectories, let us denote by $\widehat{\mu_B^n}$ and $\widehat{\sigma_B^n}$ (resp. $\widehat{\mu_P^n}$ and $\widehat{\sigma_P^n}$) the empirical estimators of mean and variance of NPV(B) (resp. NPV(P)) and $\widehat{\rho^n}$ the empirical estimator of the correlation between NPV(B) and NPV(P).

According to the central limit theorem, we have

$$\sqrt{n}\left[\begin{pmatrix}\widehat{\mu_B^n}\\ \widehat{\mu_P^n}\end{pmatrix} - \begin{pmatrix}\mu_B\\ \mu_P\end{pmatrix}\right] \overset{n\to\infty}{\to} \mathcal{N}(0,\Sigma) \quad \text{where } \Sigma = \begin{pmatrix}\sigma_B^2 & \rho\sigma_B\sigma_P\\ \rho\sigma_B\sigma_P & \sigma_P^2\end{pmatrix}.$$

Let us denote by g the function

$$\begin{array}{rccc} g: & \mathbb{R}\times\mathbb{R}^* & \longrightarrow & \mathbb{R} \\ & (x,y) & \longmapsto & \dfrac{x}{y}. \end{array}$$

We have $p_n = g(\widehat{\mu_B^n}, \widehat{\mu_P^n})$ and, for $(x,y) \in \mathbb{R}^*\times\mathbb{R}$, $\vec{\nabla} g(x,y) = \left(\frac{1}{y}, -\frac{x}{y^2}\right)$.

We have according to the Delta method

$$\sqrt{n}\left[\widehat{p}_n - p^*\right] \overset{n\to\infty}{\sim} \mathcal{N}\left(0, \vec{\nabla} g(\mu_B,\mu_P)^t \Sigma \vec{\nabla} g(\mu_B,\mu_P)\right).$$

with

$$\begin{aligned}\vec{\nabla} g(\mu_B,\mu_P)^t \Sigma \vec{\nabla} g(\mu_B,\mu_P) &= \begin{pmatrix}\frac{1}{\mu_P} & -\frac{\mu_B}{\mu_P^2}\end{pmatrix}\begin{pmatrix}\sigma_B^2 & \rho\sigma_B\sigma_P\\ \rho\sigma_B\sigma_P & \sigma_P^2\end{pmatrix}\begin{pmatrix}\frac{1}{\mu_P}\\ -\frac{\mu_B}{\mu_P^2}.\end{pmatrix}\\ &= \frac{1}{\mu_P^2}\left(\sigma_B^2 - 2\rho\frac{\mu_B}{\mu_P}\sigma_B\sigma_P + \frac{\mu_B^2}{\mu_P^2}\sigma_P^2\right).\end{aligned}$$

Slutsky's theorem ensures that the former convergence still holds when we replace the different quantities by their empirical estimators and therefore, for $\alpha \in \,]0\,;1[$, we have the following asymptotic upper-bound for the distance between the premium and its estimator

$$|\widehat{p}_n - p^*| \leqslant \sqrt{\widehat{\sigma_B^n}^2 - 2\widehat{\rho^n}\frac{\widehat{\mu_B^n}}{\widehat{\mu_P^n}}\widehat{\sigma_B^n}\widehat{\sigma_P^n} + \frac{\widehat{\mu_B^n}^2}{\widehat{\mu_P^n}^2}\widehat{\sigma_P^n}^2} \times \frac{\Phi^{-1}(1-\frac{\alpha}{2})}{\widehat{\mu_P^n}\sqrt{n}},$$

with an asymptotic level of confidence of $1-\alpha$, where Φ is the cumulative distribution function of the standard normal law.

2.5.2 Practical case

Product description

For this product, the claims are assessed using the AGGIR grid. Only states Gir 1 to Gir 4 are considered as dependency. A constant level premium is paid by the insured life, at the beginning of every month, as long as he is alive and autonomous. Should the insured life become dependent, the premium is no longer due, and benefit will be granted instead at the end of every month, while he is still alive. The amount of benefit depends on the state of dependency

- Gir 1: 1,300 €,
- Gir 2: 1,100 €,
- Gir 3: 800 €,
- Gir 4: no benefit.

An additional cash amount of 1650 € is also granted at the end of the first month of dependency, regardless of the dependency state. Besides, several additional features are added to the product. First of all, a deferral period of 3 months is fixed, so that no payment is made during the first three months spent in dependency, except for the 1650 €. Furthermore, an elimination period of 2 years is added to the product, with counter-insurance on the premium. It means that, should the insured life become dependent during the two years period after subscribing, the contract would be canceled and all premium paid would be refunded to the insured life. Finally, a technical interest rate of 2 % will be set.

Results of pricing

We determine the price of the previous product for several ages of subscription, based on 1,000,000 simulations, which gives us a relative uncertainty on the premium lower than 0.4 %. We compute a single price for both men and women, based on exogenous assumptions on the gender distribution in the initial portfolio. A summary of the results is provided in table 2.4.

Age of subscription	40 years	50 years	60 years	70 years	80 years
Monthly premium	17.42 €	25.14 €	38.36 €	63.18 €	119.12 €
95 % Confidence interval	± 0.07 €	± 0.09 €	± 0.14 €	± 0.22 €	± 0.40 €

Table 2.4: Premium for several ages of subscription.

Reserves

In long-term care insurance, we are mostly concerned about two categories of technical reserves, the reserve for premium and the reserve for claim.

On one hand, the incidence rate of dependency increases with age, and so does the associated risk. On the other hand, with level premiums, the amount of premium remains the same over the years. Therefore a reserve for premium should be constituted with incoming premiums to cope for this increase of the risk. The amount of reserve is defined as the difference between the net present value of the future benefit and the future payments, for insured lives who are not dependent yet.

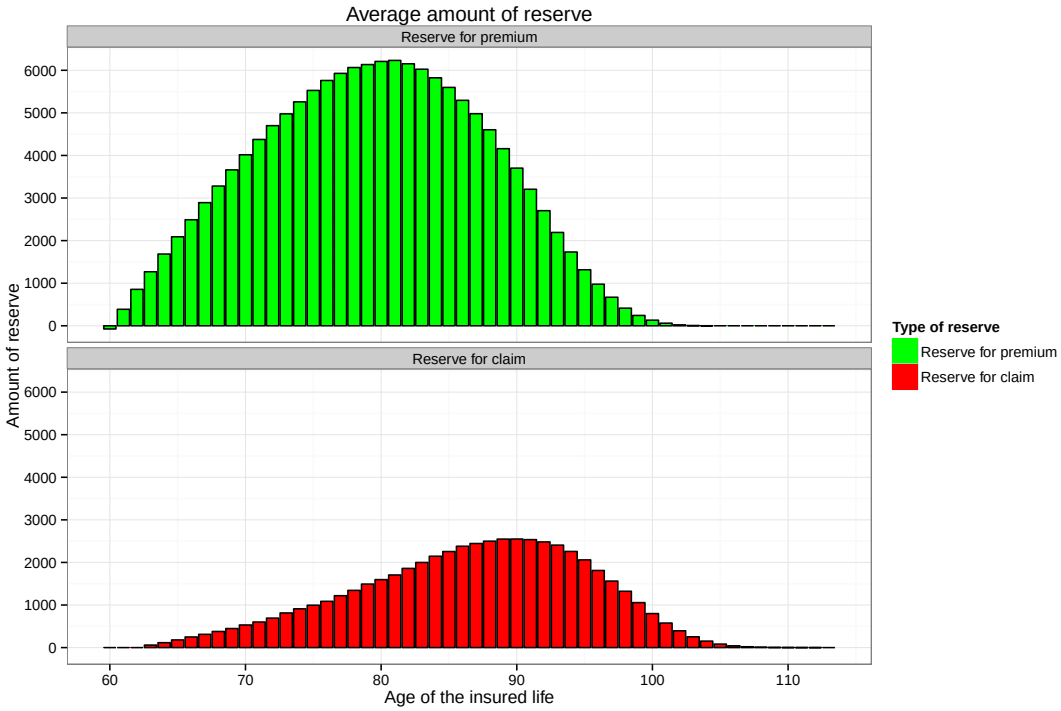


Figure 2.11: Projected amount of reserve for one insured live aged 60 at subscription, computed on a portfolio of 1,000,000 insured lives.

Whenever a claim occurs, a reserve for claim should be constituted to account for the future payments of benefit associated with this claim. The amount of reserve corresponds to the best estimate of the net present value of the benefits. In classic long-term care insurance models, the amount of reserve only depends on the gender, age of entry in dependency and time spent in dependency. With our model, the current state of dependency and the time spent in this state also give additional information and therefore should be used to get a more accurate estimation of the required amount of reserve.

Figure 2.11 provides the projected average amount of reserve required for one insured live aged 60 at subscription, computed at the time of subscription.

If the model was to be used for actual pricing, computation of actual reserves would need to be performed every year. A simulation method for already dependent people, similar to the one we introduced in the model section, should therefore be used. The only difference is that in this case, based on the history of the trajectory, we would first have to compute the probability of frailty, draw the frailty accordingly, and finally complete the trajectory.

Sensitivity to different risk factors

Shocks	Incidence rate	Autonomous mortality rate	Hazard rates in dependency	Interest rate
Value of shock	+ 10 %	- 10 %	- 10 %	- 50 bps
Resulting premium (38.36 € with no shock)	40.70 €	39.36 €	41.33 €	41.20 €
Relative variation	+ 6.1 %	+ 2.6 %	+ 7.9 %	+ 7.4 %

Table 2.5: Sensitivity to different risk factors on the premium at 60.

We study the impact of several factors of risk on the premium, such as incidence rate, mortality rate for autonomous people, hazard rates in dependency and interest rates. The impact of factors of risk such as incidence rate, mortality rate for autonomous people, hazard rates in dependency and interest rates are given in table 2.5, for a population of 1,000,000 insured lives aged 60 at subscription. It can be noted that the dependency risk and the longevity risks are positively correlated, as people surviving to higher ages means more premiums but also more people likely to become dependent, the second effect outweighing the first. Besides, long-term care insurance products have a very high sensitivity to the interest rate.

2.6 Discussion

In this paper, we presented the construction steps of a 4-state semi-Markov model for the dependency process, based on data from the French public aid, the APA: "Allocation Personnalisée d'Autonomie". Semi-Markov models have been widely described in the actuarial literature, but there are only few applications based on real long-term care insurance data, because the available data is very scarce. As a consequence, methods to deal with censored data, which is encountered in longitudinal studies, have rarely been described.

The model we developed accounts for the effect of covariates like gender and age of entry in dependency, through semi-proportional hazard rates. Heterogeneity caused by underlying pathologies, which are not observed in the data, is also taken into account through a static frailty. Estimation of parameters was performed using the maximum likelihood method, with the introduction of specific terms to deal with right censoring and missing dates of death in the data. We then provided an algorithm to generate trajectories as well as a Monte Carlo method for the pricing of long-term care insurance products. At last, we presented an application to a fictive product with a quick look at reserve and sensitivity to risk factors.

The data at our disposal provides information about 31,731 individuals, over an observation period of 3 years, with only 1 year for which death was observed. Nevertheless, the results we obtained from the model proved quite close to those presented in Debout (2010), a report based on a much larger sample of the APA data, gathering trajectories about 300,000 individuals over a 6 years period.

The underlying pathology is one of the main causes of heterogeneity in the trajectories. A study about incapacitating pathologies and their relative importance can be found in Monod-Zorzi et al. (2007). In our paper, those pathologies were not observed and we introduced a static Bernoulli frailty to account for their effect. In the future we plan on working on a portfolio which contains pathologies. This will allow us to get a better interpretation of our results, and see if it is necessary to use a more complex model for frailty with three or more levels.

Besides, we mentioned that, with a multi-state semi-Markov model, the amount of reserve for claim should be calculated while taking into account the current state of dependency and the time spent in this state. This could lead to a more accurate estimation of reserves? However, it requires us to be able to generate trajectories for people who are already dependent, and specific methods need to be developed.

At last, if we had access to more recent data from the APA, we would be able to test the adequacy of our model, especially for trajectories longer than 3 years which can only be partially observed in the current data set.

Chapitre 3

Continuous time semi-Markov inference of biometric laws associated with a Long-Term Care Insurance portfolio

Abstract

Unlike the mortality risk on which actuaries have been working for more than a century, the long-term care (LTC) risk is relatively new and as of today hardly mastered. Semi-Markov processes have been identified as an adequate tool to study this risk. Nevertheless, access to data is limited and the associated literature still scarce. Insurers mainly use discrete time methods directly inspired from the study of mortality in order to build experience tables. Those methods however are not perfectly suited for the study of competing risk situations.

The present article provides a theoretical framework to estimate biometric laws associated with a long-term care insurance portfolio. The presented method relies on a continuous-time semi-Markov model with three states: autonomy, disability and death. The process describing the state of disability is defined through its transition intensities. We provide a formula to infer the mortality of autonomous people from the mortality of the whole portfolio, on which we have more reliable knowledge. We then propose a parametric expression for the remaining intensities of the model. In particular, incidence in LTC is described by a logistic formula. Under the assumption that the disabled population is a mixture of two latent populations with respect to the category of pathology that caused LTC, we show that the resulting intensity of mortality in LTC takes a very peculiar form and depends on time spent in the LTC state. Estimation of parameters relies on the maximum likelihood method. Our parametric approach, while inducing model uncertainty, eliminates issues related to segmentation in age categories, smoothing or extrapolation at higher ages and thus proves very convenient for the practitioner. Finally, we provide an application using data from a real long-term care insurance portfolio.

3.1 Introduction

Disability among elderly people can be defined as a permanent state of inability to autonomously perform activities of daily living. It is mostly caused by diseases linked to ageing, such as dementia, neurological diseases, cardiovascular diseases and cancer. Disabled elderly people require regular care whose frequency increases with the severity of their status. While some people can rely at least partially on their family or their friends for help, others have to hire professional caregivers or join a nursing home, whose average cost exceeds 3,000 € a month. Despite public aids, this cost proves overwhelming for most pensioners. Therefore, to long-term care (LTC) is associated a financial risk to which most people are exposed. In France, part of this risk is transferred through private insurance contracts.

The long-term care risk is complex. Its study requires to take into account incidence in LTC as well as probabilities of death for both autonomous and disabled people, which are very different from another. This risk is directly related to ageing through pathologies, and longevity gains in the second half of the 20th century made it paramount. The very first long-term care insurance products appeared in the US during the 1980's and shortly after in France. Average age at subscribing for those products is close to 60 when the average age at which LTC occurs is near to 85. Therefore, even on older portfolios, the number of claims remains limited. Moreover, in France, insurers and public aids use different definitions to asses the level of required care. Those definitions, as well as insurers underwriting and claims policies often change over time. All those elements make data aggregation from several sources very difficult, which may explain the difficulty of getting a better knowledge of the risk.

Markov processes are such that their transition probabilities only depend on the current state of the process. A semi-Markov process is a generalization for which transition probabilities depend on both the current state and the time spent in the current state. One can find more details about those processes in Cinlar (1969). Multi-state models based on Markov and semi-Markov processes have led to many applications in the field of epidemiology. As the long-term care state is mainly caused by pathologies, those processes appear as natural candidates to study the long-term care risk. This framework has already been described for example in Haberman and Pitacco (1998) or Christiansen (2012). Several studies based on US national data have also been performed. One can refer to Robinson (1996), Pritchard (2006) or more recently Fong et al. (2015). On the other hand, studies based on portfolio data Guibert and Planchet (2014)as are very rare. Practitioners nevertheless played a key role in the knowledge of the LTC risk. One of the very first models on the French market was presented by SCOR (1995). Relying on a parametric approach, it highlights the exponential increase in the probabilities of incidence in LTC, and defines mortality in LTC (resp. autonomous mortality) as a linear function of the general population mortality, computed via an exogenous mortality table. With only 5 parameters required to model the whole process, it is remarkably simple. It is however based on the Markov assumption that mortality in LTC only depends on the age of the disabled life, and not on the time since the entry in LTC. The Markov assumption is still used today by many insurers as well as in recent academic papers like Pitacco (2015) or Fong et al. (2015), because it allows for simpler models. However, it does not reflect the reality of the long-term care process, for which mortality is much higher during the first year in LTC than for the subsequent year. For an insurance company, ignoring this feature of the risk can be very damaging. Indeed, it leads to greatly overestimating mortality in LTC based on the first-year mortality experience and therefore underestimating the required amount of reserve, which results in heavy losses in the future.

Semi-Markov processes have already been used for disability insurance, especially through the illness-death model as described in Pitacco (2014). However, one has to keep in mind that on one hand disability insurance only lasts until retirement age with a limited period for benefits. On the other hand, individual long-term care insurance relies on lifetime annuities

with no expiry date. Therefore, while a similar model may be used for both risks, issues related to extrapolation of biometric laws at higher ages and higher duration in the disabled state arise in the study of the long-term care risk. For the same reason, non-parametric methods based on Nelson-Aalen estimator (Klein, 1991) that have also been used to study the long-term care risk, for example in Guibert and Planchet (2015) still need to be associated with parametric methods for the extrapolation step.

In this article, we present a parametric approach relying on a continuous-time semi-Markov process, which is defined using its transition intensities. Compared to a discrete-time approach, it allows to get a more straightforward modeling of the process, while correctly taking into account the competing risks (disability and death). Section 2 introduces the model and derives an equation to express the autonomous mortality using general mortality and other intensities of the model. Benefits to use general mortality instead of autonomous mortality are discussed with more details. We then introduce the intensity for general mortality of the portfolio using a simple relational model as in Brass (1971). We propose a parametric expression for the intensity of incidence in LTC, based on the logistic form introduced by Perks (1932) for the study of mortality. We use a complex parametric model for the intensity of mortality in LTC, corresponding to a latent mixture model where we consider two homogeneous populations of disabled people, with two different levels of mortality. Estimation of parameters relies on the maximum likelihood method. We also introduce formulas for pricing and reserving based directly on the transition intensities. Section 3 provides an application of the model based on data from a real insurance portfolio. For each transition intensity, several models of increasing complexity are compared using the Bayesian Information Criterion (BIC). Comparison with empirical transition rates is also provided. Robustness of estimation is then assessed using a non-parametric quantile bootstrap method. Finally, Section 4 summarizes the results obtained and discusses limits and potential improvements of the model.

3.2 Model

3.2.1 Notations

For $x_0 \geqslant 0$, let us consider a continuous-time process $(Z_x)_{x \geqslant x_0}$ with values in the 3-state set $E = \{A, I, D\}$ of autonomy, LTC (or "illness"), death, respectively. Let us assume that Z is *càd-làg*. The index variable of the process Z is called age of the process. For $x \geqslant x_0$ let us denote by A_x (resp. I_x, D_x) the probability for the process to be in the state of autonomy (resp. LTC, death) at age x or more formally

$$\begin{aligned} A_x =& \mathbb{P}(Z_x = A | Z_{x_0} = A), \\ I_x =& \mathbb{P}(Z_x = I | Z_{x_0} = A), \\ D_x =& \mathbb{P}(Z_x = D | Z_{x_0} = A). \end{aligned}$$

Let us further assume that $Z_{x_0} = A$. Hence $A_{x_0} = 1$ and for all $x \geqslant x_0$, $A_x + I_x + D_x = 1$. We now assume that $(Z_x)_{x \geqslant x_0}$ is a non-homogeneous semi-Markov process and introduce the transition intensities, also called instantaneous transition probabilities. Transition intensities allow us to fully describe the behaviour of the process

$$\begin{aligned} \mu_a(x) =& \lim_{h \to 0} \frac{1}{h} \mathbb{P}\left(Z_{x+h} = D | Z_x = A\right), \\ \lambda(x) =& \lim_{h \to 0} \frac{1}{h} \mathbb{P}\left(Z_{x+h} = I | Z_x = A\right), \\ \mu_i(x,t) =& \lim_{h \to 0} \frac{1}{h} \mathbb{P}\left(Z_{x+t+h} = D | Z_{x^-} = A, Z_x = I, Z_{x+t} = I\right). \end{aligned}$$

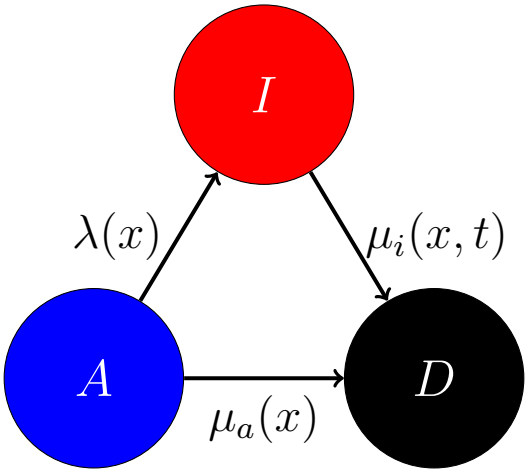


Figure 3.1: The 3 states continuous-time semi-Markov model and the associated transition intensities.

Those intensities are called respectively intensity of entry in LTC, intensity of autonomous mortality and intensity of mortality in LTC, with the latter intensity depending on both the age at onset of LTC and time spent in LTC. We consider that death is an absorbing state and that there is no transition allowed from LTC to autonomy. To understand this last assumption, one has to keep in mind that on the French long-term care insurance market, the LTC benefit is only granted when the disabled state is expected to be permanent. Therefore cases of return to the autonomy state are quite rare, compared to other markets where this is not the case. Furthermore, once the benefit is granted, the annuitant is not required to provide any proof that they are still disabled. Hence ignoring cases of return to the autonomy state does not introduce any inconsistency with the way the insurance products are priced, and it allows for simpler models. Given the limited amount of available data this proves very convenient. A representation of the model can be found on Figure 3.1.

Lemma 1. *Let $x \geqslant x_0$. The probability A_x (resp. I_x) to be in the autonomous (resp. disabled) state at age x may be expressed directly from the transition intensities of the model and we have*

$$A_x = \exp\left(-\int_{x_0}^{x} \left[\lambda(u) + \mu_a(u)\right] du\right), \tag{3.1}$$

$$I_x = \int_{x_0}^{x} \lambda(u) A_u \exp\left(-\int_{u}^{x} \mu_i(u, v-u)\, dv\right) du. \tag{3.2}$$

Proof. For $x \geqslant x_0$, $h \geqslant 0$, we have

$$\mathbb{P}(Z_{x+h} = A) = \left[1 - \mathbb{P}(Z_{x+h} = I | Z_x = A) - \mathbb{P}(Z_{x+h} = D | Z_x = A)\right] \times \mathbb{P}(Z_x = A)$$

and therefore

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{P}(Z_x = A) = -\left[\mu_a(x) + \lambda(x)\right]\mathbb{P}(Z_x = A).$$

As $A_{x_0} = 1$, this equation has a unique solution

$$A_x = \exp\left(-\int_{x_0}^{x} \left[\lambda(u) + \mu_a(u)\right] \mathrm{d}u\right). \tag{3.3}$$

For $x \geqslant x_0$, $t, h \geqslant 0$, we can write

$$\begin{aligned}\mathbb{P}(Z_{x+t+h} = I | Z_{x^-} = A, Z_x = I) =& \mathbb{P}(Z_{x+t+h} = I | Z_{x^-} = A, Z_x = I, Z_{x+t} = I) \\ & \times \mathbb{P}(Z_{x+t} = I | Z_{x^-} = A, Z_x = I).\end{aligned}$$

which gives us

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{P}(Z_{x+t} = I|Z_{x^-} = A, Z_x = I) = -\mu_i(x,t)\mathbb{P}(Z_{x+t} = I|Z_{x^-} = A, Z_x = I).$$

As

$$\mathbb{P}(Z_x = I|Z_{x^-} = A, Z_x = I) = 1$$

we obtain

$$\mathbb{P}(Z_{x+t} = I|Z_{x^-} = A, Z_x = I) = \exp\left(-\int_0^t \mu_i(x,u)\mathrm{d}u\right).$$

Then as we have the following decomposition

$$I_x = \int_{x_0}^{x} \mathbb{P}(Z_u = A)\mathbb{P}(Z_u = I|Z_{u^-} = A)\mathbb{P}(Z_x = I|Z_{u^-} = A, Z_u = I)\mathrm{d}u,$$

we get an expression of the probability to be disabled at age $x \geqslant x_0$

$$I_x = \int_{x_0}^{x} \lambda(u)A_u \exp\left(-\int_u^x \mu_i(u, v-u)\mathrm{d}v\right). \tag{3.4}$$

□

3.2.2 Link with general mortality

Let us consider the intensity of mortality for the aggregated population of autonomous and disabled (hereafter general mortality) defined by

$$\mu_g(x) = \lim_{h\to 0} \frac{1}{h} P\left(Z_{x+h} = D|Z_x \in \{A, I\}\right).$$

Figure 3.2 represents the fourth transition in our model, a transition between life and death for the general population.

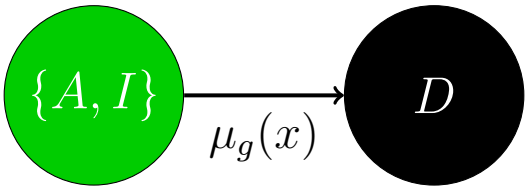


Figure 3.2: Intensity of transition for the general population.

Lemma 2. *For $x \geqslant x_0$ and $t \geqslant 0$, let us denote by $\Delta(x,t)$ the difference between the intensity of mortality in LTC and the intensity of autonomous mortality for the same current age, so that $\Delta(x,t) = \mu_i(x,t) - \mu_a(x+t)$. Then the intensity of autonomous mortality is solution of the following equation*

$$\mu_a(x) = \mu_g(x) - \frac{\int_{x_0}^{x} \lambda(u)\Delta(u, x-u)\exp\left(-\int_u^x \left[\Delta(u, v-u) - \lambda(v)\right]dv\right)du}{1 + \int_{x_0}^{x} \lambda(u)\exp\left(-\int_u^x \left[\Delta(u, v-u) - \lambda(v)\right]dv\right)du}. \tag{3.5}$$

Proof. Differentiating (3.1) and (3.2) gives us equations (3.6) and (3.7) below which describe the evolution of the probabilities A_x and I_x. Similarly, from the definition of μ_g we get equation (3.8). We obtain a system of 3 differential equations

$$\frac{\mathrm{d}}{\mathrm{d}x}A_x = -[\lambda(x) + \mu_a(x)]A_x, \tag{3.6}$$

$$\frac{\mathrm{d}}{\mathrm{d}x}I_x = \lambda(x)A_x - \int_{x_0}^{x} \lambda(u)A_u \exp\left(-\int_u^x \mu_i(u, v-u)\mathrm{d}v\right)\mu_i(u, x-u)\mathrm{d}u, \tag{3.7}$$

$$\frac{\mathrm{d}}{\mathrm{d}x}(A_x + I_x) = -\mu_g(x)(A_x + I_x). \tag{3.8}$$

Summing the evolution equations (3.6) and (3.7), then identifying with equation (3.8) yields

$$\mu_g(x)(A_x + I_x) = \mu_a(x)A_x + \int_{x_0}^{x} \lambda(u)A_u \exp\left(-\int_u^x \mu_i(u, v-u)\mathrm{d}v\right)\mu_i(u, x-u)\mathrm{d}u.$$

With simple algebra we get

$$\mu_a(x) = \mu_g(x)\left(1 + \frac{I_x}{A_x}\right) - \int_{x_0}^{x} \lambda(u)\frac{A_u}{A_x} \exp\left(-\int_u^x \mu_i(u, v-u)\mathrm{d}v\right)\mu_i(u, x-u)\mathrm{d}u.$$

Using (3.2) and (3.1), we obtain after a few simplifications

$$\mu_a(x) = \mu_g(x) - \int_{x_0}^{x} \lambda(u) \exp\left(-\int_u^x [\mu_i(u, v-u) - \lambda(v) - \mu_a(v)]\,\mathrm{d}v\right)[\mu_i(u, x-u) - \mu_g(x)]\,\mathrm{d}u.$$

Now, we replace the intensity of mortality in LTC using the formula

$$\mu_i(x, t) = \mu_a(x + t) + \Delta(x, t)$$

which gives us

$$\mu_a(x) = \mu_g(x) - \int_{x_0}^{x} \lambda(u) \exp\left(-\int_u^x [\Delta(u, v-u) - \lambda(v)]\,\mathrm{d}v\right)[\mu_a(x) - \mu_g(x) + \Delta(u, x-u)]\,\mathrm{d}u.$$

This finally leads to the result. □

Equation (3.5) allows us to use the general mortality instead of the autonomous mortality in the model, at the cost of the introduction of Δ, the difference between autonomous mortality and mortality in LTC. On one hand, mortality of autonomous people is complex to predict, because people can leave the state of autonomy either by becoming disabled or dying. Furthermore, the scope of autonomous people depends directly on the definition used for LTC. Therefore predicting the autonomous mortality requires to have intensive knowledge of the LTC process beforehand. On the other hand, general mortality has been studied for a long time by actuaries, demographers, biologists and is very well documented. One can therefore rely on reference mortality tables for ages where no portfolio data is available.

The formula does not give an analytic expression for the intensity of autonomous mortality in the most general case. Numerical methods can however be used to compute it. As will be seen in section 3.2.4, choosing an *ad hoc* model, the inner integrals in the formula take an analytic expression and numerical approximation is only required for the outer integrals. Intensity of general mortality appears directly in the equation, which is very convenient if we want to use an external reference for this intensity.

3.2.3 Data structure

Data issued from insurance portfolios generally consists of two databases. The first database gathers information on contributors and the second on annuitants. We also define the database of insured lives obtained by merging the two previous bases which will be used for the estimation of general mortality. From one portfolio to another, data quality and available information may vary a lot. In what follows, we assume both databases contain at least the variables of Table 3.1, listed as follows:

- DoB: date of birth of the individual,
- DoS: date of start. For contributors, it is the date of subscribing. For annuitants, the date of entry in LTC,
- DoE and CoE: Respectively the date of end and cause of end for the individuals. In the case where the observation ends because of death, we use code 1 for the cause, in the case of exit because of entry in LTC, we use code 2. For individuals still autonomous when the observation stops, trajectories are right-censored. We use code 0 and the date of exit is the date at which observation ends.

DoB	DoS	DoE	CoE
12/23/1941	11/10/1992	09/27/2006	2
06/14/1926	03/28/1997	12/31/2014	0
04/17/1937	04/27/1995	04/08/2003	1

Table 3.1: Example of a database of contributors.

Other covariates such as gender, type of residence (home or facility), marital status, cause of disability, amount of annuity bought or premium for substandard risk may be available and bring additional information. In what follows, we assume that only gender is available and we estimate a separate model for male and female.

The observation period must often be limited in some way:

- By removing the last year of individual exposure. With each database is associated a date of extraction, which is the date of the latest entry in the database. In practice, most claims are reported up to one year after their occurrence, which may result in some missing information during the last year of observation. It may therefore be a good idea to set a date for the end of the observation one year prior to the date of extraction, in order not to underestimate the number of events. For events that occurred during the last year of observation, the associated code must then be set to 0.
- By removing the first 3 years of individual exposure. On the french individual long-term care insurance market, there is usually an elimination period of 3 years for dementia and neurological pathologies which results in fewer claims during those 3 years. In order not to underestimate the incidence rate, we therefore remove the exposure for the first 3 years of observation for each individual.
- By shortening the observation period. When we study the behaviour of a population for a specific risk, it may change over time. There are several factors involved, such as changes in the definition of LTC, underwriting and claim management policy in addition to underlying changes of biometric laws. Shortening the observation period is therefore required in order to minimize those effects and set a good compromise between large volume and stability of the underlying risk over the period.

Once the data has been processed, we may easily compute quantities of interest which will be used in the estimation procedure

- The age of entry $x = \text{DoS} - \text{DoB}$,
- The age of exit $y = \text{DoE} - \text{DoB}$,
- The cause of exit $c = \text{CoE}$.

3.2.4 Parametric modelling of the intensities

In this section, we propose to rely on a parametric expression for each of the transition intensities in the model.

Intensity of general mortality

We want to assess the general mortality of our portfolio, which is seen as a specific population inside the French population. To do so, we rely on the database of insured lives as well as on an external mortality reference, using the Brass relational model as described in Brass (1971, 1974) or Hannerz (2001).

Let F be the cumulative distribution function associated with an intensity of mortality μ such that

$$F(x) = 1 - \exp\left(-\int_{x_0}^{x} \mu(u)\mathrm{d}u\right).$$

Then we define the cumulative distribution odds (CDO) by

$$\text{CDO}(x) = \frac{F(x)}{1 - F(x)}.$$

In his model, Brass relies on the assumption that the logarithm of the CDO associated with the mortality of a reference population and the mortality of a specific population are parallel curves. We denote by μ_g and F_g (resp. μ_g^{ref} and F_g^{ref}) the intensity of mortality and cumulative distribution function associated with the mortality of the specific (resp. reference) population. Under this assumption, the Brass estimator for the intensity of mortality of the specific population is

$$\widehat{\mu_g}(x) = \frac{\widehat{\beta}\mu_g^{ref}(x)}{1 - (1 - \widehat{\beta})F_g^{ref}(x)}$$

where $\widehat{\beta}$ is the solution of the equation

$$\sum_x D_x = \sum_x D_x^{ref} \frac{\widehat{\beta} N_x}{N_x^{ref}(1 - (1 - \widehat{\beta})F_g^{ref}(x))}$$

and D_x, N_x (resp. D_x^{ref}, N_x^{ref}) are the total number of deaths observed and the number of years lived between ages x and $x + 1$ by the specific population (resp. by the reference population). the Brass model only requires the estimation of a single parameter $\widehat{\beta}$. It gives an estimator for the intensity of mortality which converges smoothly towards the mortality reference at higher ages while predicting the same number of deaths as in the empirical data, given the empirical exposure.

Intensity of incidence in LTC

For the intensity of incidence in LTC, we consider the logistic model introduced in Beard (1959, 1971) and Perks (1932)

$$\lambda(x) = \frac{\exp(a_\lambda x + b_\lambda)}{1 + \exp(a_\lambda x + c_\lambda)} + d_\lambda \tag{3.9}$$

with $a_\lambda > 0$, $b_\lambda, c_\lambda \in \mathbb{R} \cup \{-\infty\}$ and $d_\lambda \geqslant 0$.

Experience from insurers shows that the intensity of incidence in LTC increases exponentially with respect to age (SCOR, 1995). At higher ages, data becomes scarcer. As LTC is linked to ageing, it is reasonable to expect that the behaviours of mortality and morbidity are quite similar and that an exponential or logistic form is suited to model incidence in LTC. The logistic model has already been used to this extent, e.g. in Rickayzen and Walsh (2002). Let us notice that the exponential models introduced in Gompertz (1825) and Makeham (1867) may be seen as limit cases of the logistic model, for which $c_\lambda = -\infty$.

For an individual p defined by their age of entry in the portfolio $x_p \geqslant 0$, their age of exit $y_p > x_p$ and the associated exit cause $c_p \in \{0, 1, 2\}$ the log-likelihood has the following expression

$$\begin{aligned}
l_p(\lambda) =& \log\left[\exp\left(-\int_{x_p}^{y_p}\lambda(u)\mathrm{d}u\right)\lambda(y_p)^{\delta^2_{c_p}}\right] \\
=& \delta^2_{c_p}\log(\lambda(y_p)) - \int_{x_p}^{y_p}\lambda(u)\mathrm{d}u \\
=& \delta^2_{c_p}\log\left[\frac{\exp(a_\lambda y_p + b_\lambda)}{1+\exp(a_\lambda y_p + c_\lambda)} + d_\lambda\right] - \frac{\exp(b_\lambda - c_\lambda)}{a_\lambda}\log\left[\frac{1+\exp(a_\lambda y_p + c_\lambda)}{1+\exp(a_\lambda x_p + c_\lambda)}\right] - d_\lambda(y_p - x_p),
\end{aligned}$$

where for $k, l \in \mathbb{N}$, $\delta^l_k = \begin{cases} 1 \text{ if } k = l, \\ 0 \text{ otherwise.} \end{cases}$

Intensity of mortality in LTC

Disability may be caused by a wide range of underlying pathologies. Unfortunately most of the time those pathologies are not available in the data. This results in heterogeneity among disabled people. In this section, we provide a simple parametric model which accounts for the heterogeneity caused by pathologies. In order to do this, we must rely on several strong assumptions. First of all, we assume that underlying pathologies can be divided into two main groups. On one hand we have pathologies associated with very high mortality such as terminal cancer mainly or more rarely respiratory diseases. For such diseases, life expectancy at the onset of LTC is within a few months. On the other hand, dementia, neurological or cardiovascular diseases have an associated life expectancy which is closer to 5 years. We further assume that among each group the population can be considered as homogeneous. We could consider three or more groups of pathologies but then inference of parameters would prove extremely difficult and this would be at the expense of robustness in the estimation procedure.

We then consider an additive model for the intensity of mortality in LTC, so that the mortality within each group is the sum of a common term (which may be for example the autonomous mortality at the current age) plus a term which only depends on the pathology group and the age at onset of LTC. The underlying assumption to this additive model of mortality is that people who become disabled have increased mortality from the pathology that caused disability but are still exposed to other causes of death. Also, considering that the additional mortality depends on the age at onset of LTC rather than on the current age may

seem very restrictive and the model may not be accurate for very high duration in the disabled state. However due to the very high mortality in the disabled state, cases of exceptional longevity should remain rare enough and the resulting impact quite limited. Under those assumptions, the resulting intensity of mortality in LTC takes a very peculiar form as we show in the following lemma.

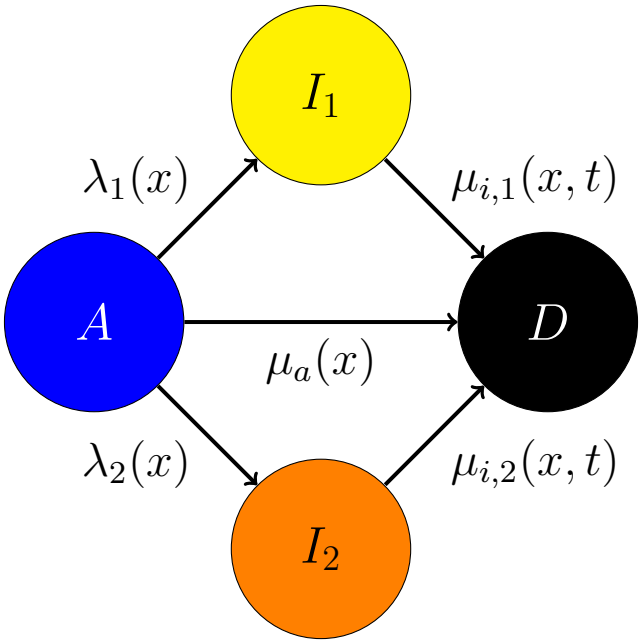


Figure 3.3: Model with 2 completely separate states of LTC.

Lemma 3. *Let us consider a model with 2 distinct states of disability I_1 and I_2, such that the respective transition intensities from autonomy to those states are λ_1 and λ_2 respectively and no transition is allowed between those states or back to autonomy (see Figure 3.3). Let us further assume that the intensity of mortality in state I_k can be written as*

$$\mu_{i,k}(x,t) = \mu_0(x,t) + \Delta_k(x)$$

with $x \geqslant x_0$, $t \geqslant 0$ where $\mu_0(x,t)$ is a common mortality term and $\Delta_k(x)$ a state-specific mortality term, for $k \in \{1,2\}$.

Then to ensure the embedding of the 3-states model in this model, the following relations must be satisfied for all $x > x_0$, $t \geqslant 0$

$$\lambda(x) = \lambda_1(x) + \lambda_2(x)$$

$$\mu_i(x,t) = \begin{cases} \mu_0(x,t) + \Delta_1(x) + \dfrac{\Delta_2(x) - \Delta_1(x)}{1 + \frac{\lambda_1(x)}{\lambda_2(x)} \exp([\Delta_2(x) - \Delta_1(x)]\, t)} & \text{if } \lambda_2(x) \neq 0 \\ \mu_0(x,t) + \Delta_1(x) & \text{otherwise.} \end{cases}$$

Proof. The first relation on the incidences in LTC is obvious, as well as the case where $\lambda_2(x) = 0$. For the second relation, let us denote by $\eta_k(x,t)$ the proportion of disabled people in state I_k among the population of people who became disabled at age $x \geqslant x_0$ and then survived for a time $t \geqslant 0$.

Let us define for $x \geqslant x_0$, $t \geqslant 0$, $k \in \{1,2\}$ and $h > 0$

$$\eta_k(x,t) = \frac{\mathbb{P}(Z_{x+t} = I_k | Z_{x^-} = A, Z_x \in \{I_1, I_2\})}{\mathbb{P}(Z_{x+t} \in \{I_1, I_2\} | Z_{x^-} = A, Z_x \in \{I_1, I_2\})}.$$

and

$$\eta_k(x,t,h) = \frac{\mathbb{P}(Z_{x+t} = I_k | Z_{x-h} = A, Z_x \in \{I_1, I_2\})}{\mathbb{P}(Z_{x+t} \in \{I_1, I_2\} | Z_{x-h} = A, Z_x \in \{I_1, I_2\})}.$$

On one hand, we have

$$\eta_k(x,t,h) \underset{h\to 0}{\to} \eta_k(x,t),$$

and on the other hand

$$\begin{aligned}
\eta_k(x,t,h) =& \frac{\mathbb{P}(Z_{x+t}=I_k, Z_{x-h}=A, Z_x\in\{I_1,I_2\})}{\mathbb{P}(Z_{x+t}\in\{I_1,I_2\}, Z_{x-h}=A, Z_x\in\{I_1,I_2\})}\\
=& \frac{\mathbb{P}(Z_{x+t}=I_k, Z_x=I_k, Z_{x-h}=A)}{\sum_{l=1}^{2}\mathbb{P}(Z_{x+t}=I_l, Z_x=I_l, Z_{x-h}=A)}\\
=& \frac{\mathbb{P}(Z_{x-h}=A)\mathbb{P}(Z_x=I_k|Z_{x-h}=A)\mathbb{P}(Z_{x+t}=I_k|Z_x=I_k, Z_{x-h}=A)}{\sum_{l=1}^{2}\mathbb{P}(Z_{x-h}=A)\mathbb{P}(Z_x=I_l|Z_{x-h}=A)\mathbb{P}(Z_{x+t}=I_l|Z_x=I_l, Z_{x-h}=A)}\\
=& \frac{\mathbb{P}(Z_x=I_k|Z_{x-h}=A)\mathbb{P}(Z_{x+t}=I_k|Z_x=I_k, Z_{x-h}=A)}{\sum_{l=1}^{2}\mathbb{P}(Z_x=I_l|Z_{x-h}=A)\mathbb{P}(Z_{x+t}=I_l|Z_x=I_l, Z_{x-h}=A)}\\
\underset{h\to 0}{\to}& \frac{\lambda_k(x)\exp\left(-\int_0^t \mu_{i,k}(x,u)\mathrm{d}u\right)}{\sum_{l=1}^{2}\lambda_l(x)\exp\left(-\int_0^t \mu_{i,l}(x,u)\mathrm{d}u\right)}\\
=& \frac{\lambda_k(x)\exp\left(-\int_0^t \mu_0(x,u)\mathrm{d}u\right)\times\exp\left[-\Delta_k(x)\times t\right]}{\sum_{l=1}^{2}\lambda_l(x)\exp\left(-\int_0^t \mu_0(x,u)\mathrm{d}u\right)\times\exp\left[-\Delta_l(x)\times t\right]}\\
=& \frac{\lambda_k(x)\exp\left[-\Delta_k(x)\times t\right]}{\sum_{l=1}^{2}\lambda_l(x)\exp\left[-\Delta_l(x)\times t\right]}.
\end{aligned}$$

By uniqueness of the limit we obtain

$$\eta_k(x,t) = \frac{\lambda_k(x)\exp\left[-\Delta_k(x)\times t\right]}{\sum_{l=1}^{2}\lambda_l(x)\exp\left[-\Delta_l(x)\times t\right]}.$$

Now the intensity of mortality for the population of disabled people is

$$\begin{aligned}
\mu_i(x,t) =& \sum_{k=1}^{2}\eta_k(x,t)\mu_{i,k}(x,t)\\
=& \sum_{k=1}^{2}\frac{\lambda_k(x)\exp(-\Delta_k(x)t)}{\sum_{l=0}^{2}\lambda_l(x)\exp(-\Delta_l(x)t)}\mu_{i,k}(x,t)\\
=& \mu_0(x,t) + \sum_{k=1}^{2}\frac{\lambda_k(x)\exp(-\Delta_k(x)t)}{\sum_{l=0}^{2}\lambda_l(x)\exp(-\Delta_l(x)t)}\Delta_k(x)\\
=& \mu_0(x,t) + \Delta_1(x) + \frac{\lambda_2(x)\exp(-\Delta_2(x)t)}{\lambda_1(x)\exp(-\Delta_1(x)t)+\lambda_2(x)\exp(-\Delta_2(x)t)}\left[\Delta_2(x)-\Delta_1(x)\right]\\
=& \mu_0(x,t) + \Delta_1(x) + \frac{\Delta_2(x)-\Delta_1(x)}{1+\frac{\lambda_1(x)}{\lambda_2(x)}\exp\left\{\left[\Delta_2(x)-\Delta_1(x)\right]t\right\}},
\end{aligned}$$

which proves the lemma.

□

In what follows, we assume that the assumptions of the lemma are satisfied and that the common mortality term is the intensity of mortality for autonomous people μ_a at the same current age $x+t$. Let us denote, for $x \geqslant x_0$

$$\theta(x) = \frac{\lambda_2(x)}{\lambda(x)}.$$

We now have

$$\mu_i(x,t) = \mu_a(x+t) + \Delta_1(x) + \frac{\theta(x)\left[\Delta_2(x) - \Delta_1(x)\right]}{\theta(x) + \left[1-\theta(x)\right]\exp\left\{\left[\Delta_2(x) - \Delta_1(x)\right]t\right\}}$$

and thus

$$\Delta(x,t) = \Delta_1(x) + \frac{\theta(x)\left[\Delta_2(x) - \Delta_1(x)\right]}{\theta(x) + \left[1-\theta(x)\right]\exp\left\{\left[\Delta_2(x) - \Delta_1(x)\right]t\right\}}. \tag{3.10}$$

The log-likelihood associated with survival in LTC for an individual p with an age of entry in LTC $x_p \geqslant 0$, an age of exit $y_p > x_p$ and the associated cause of exit $c_p \in \{0,1\}$ then takes the following expression:

$$\begin{aligned}
l_p(\mu_a,\Delta_1,\Delta_2,\theta) =& \log\left[\exp\left(-\int_{x_p}^{y_p}\mu_i(x_p,u-x_p)\mathrm{d}u\right)\mu_i(x_p,y_p-x_p)^{\delta^1_{c_p}}\right] \\
=& \delta^1_{c_p}\log(\mu_i(x_p,y_p-x_p)) - \int_{x_p}^{y_p}\mu_i(x_p,u-x_p)\mathrm{d}u \\
=& \delta^1_{c_p}\log\left(\mu_a(y_p) + \Delta_1(x_p) + \frac{\theta(x_p)\left[\Delta_2(x_p)-\Delta_1(x_p)\right]}{\theta(x_p)+\left[1-\theta(x_p)\right]\exp\left(\left[\Delta_2(x_p)-\Delta_1(x_p)\right]\left[y_p-x_p\right]\right)}\right) \\
& - \int_{x_p}^{y_p}\mu_a(u)\mathrm{d}u - \Delta_2(x_p)\left[y_p-x_p\right] \\
& + \log\left\{\theta(x_p) + \left[1-\theta(x_p)\right]\exp\left(\left[\Delta_2(x_p)-\Delta_1(x_p)\right]\left[y_p-x_p\right]\right)\right\}.
\end{aligned}$$

For a given μ_a, the previous log-likelihood may then be computed analytically, which allows for the estimation of $\Delta_1, \Delta_2, \theta$ and then Δ by plugging in equation (3.10) using maximum likelihood.

Intensity of autonomous mortality

The maximum likelihood method in the previous section requires to know the autonomous mortality beforehand. We therefore need to compute an intermediary estimate of the autonomous mortality whose sole purpose is the estimation of Δ. Indeed, the ultimate autonomous mortality is then computed thanks to Lemma 2.

Once again we rely on the logistic model introduced in Beard (1959, 1971) and Perks (1932)

$$\mu_a(x) = \frac{\exp(a_a x + b_a)}{1+\exp(a_a x + c_a)} + d_a \tag{3.11}$$

with $a_a > 0$, $b_a, c_a \in \mathbb{R} \cup \{-\infty\}$ and $d_a \geqslant 0$.

For an individual p defined by their age of entry in the portfolio $x_p \geqslant 0$, their age of exit $y_p > x_p$ and the associated exit cause $c_p \in \{0,1,2\}$ the log-likelihood has the following expression

similar to section 3.2.4

$$\begin{aligned} l_p(\mu_a) =& \log\left[\exp\left(-\int_{x_p}^{y_p}\mu_a(u)\mathrm{d}u\right)\mu_a(y_p)^{\delta^1_{c_p}}\right] \\ =& \delta^1_{c_p}\log\left[\frac{\exp(a_a y_p+b_a)}{1+\exp(a_a y_p+c_a)}+d_a\right]-\frac{\exp(b_a-c_a)}{a_a}\log\left[\frac{1+\exp(a_a y_p+c_a)}{1+\exp(a_a x_p+c_a)}\right]-d_a(y_p-x_p). \end{aligned}$$

3.2.5 Parameters estimation procedure

To estimate the parameters, we use the following procedure

1. We estimate the parameters for the intensity of general mortality $\widehat{\mu_g}$ by using the individuals of both databases and the Brass relational model (as in Brass, 1971) in order to get a robust estimate of the intensity of general mortality with convergence towards a reference mortality table at higher ages (see section 3.2.4).

2. We estimate the parameters for the intensity of incidence in LTC $\widehat{\lambda}$ (resp. a first-step estimate of the autonomous mortality $\widehat{\mu_a}^{(1)}$), using the contributors database and the Perks logistic model (as in Perks, 1932). More precisely $\widehat{\lambda}$ (resp. $\widehat{\mu_a}^{(1)}$) is the maximum likelihood estimator (MLE) constructed by summing over the individuals the log-likelihoods given in section 3.2.4 (resp. 3.2.4).

3. We estimate the parameters for the additional mortality in LTC $\widehat{\Delta}$ from $\widehat{\mu_a}^{(1)}$ and the annuitant database, using the MLE constructed by summing over the individuals the log-likelihoods given in section 3.2.4. Several parametric forms for $\widehat{\Delta}$ will be tested in the next section.

4. Thanks to equation (3.5), we compute the value of a second-step estimator for the intensity for autonomous mortality $\widehat{\mu_a}^{(2)}$, relying on $\widehat{\lambda}$, $\widehat{\Delta}$, $\widehat{\mu_g}$ and using numerical methods to approximate the outer integrals in (3.5). This second-step estimator should give more reliable results at higher ages where no data is available.

A summary of the procedure is provided in Figure 3.4.

Furthermore when we deal with complex models, it can be very interesting to compare them to some of their sub-models to see if the use of many parameters is justified. To this extent, we can rely on the *Bayesian Information Criterion (BIC)*. For a model m_i characterized by a number of parameters k_i and a log-likelihood function l_i maximized at θ_i, the expression of the criterion is as follows

$$BIC_i = -2l_i(\theta_i) + k_i\log(n),$$

where n represents the number of observed transitions in the expression of the likelihood. The choice of the coefficient in front of the number of parameters $\log(n)$ differs from the one made in the original Akaike's Information Criterion (AIC) where this coefficient is 1. Also, let us note that in the version of the criterion, n is the number of observed transitions and not the number of individuals as in the original criterion. The interest in introducing this modification in the case of censored data is discussed in Volinsky and Raftery (2000). Using the BIC, we are able to compare models, the model with the lower BIC being the "best" model. In the next section we use the BIC to challenge the use of more complex parametric models introduced in this section and lower the overall number of parameters.

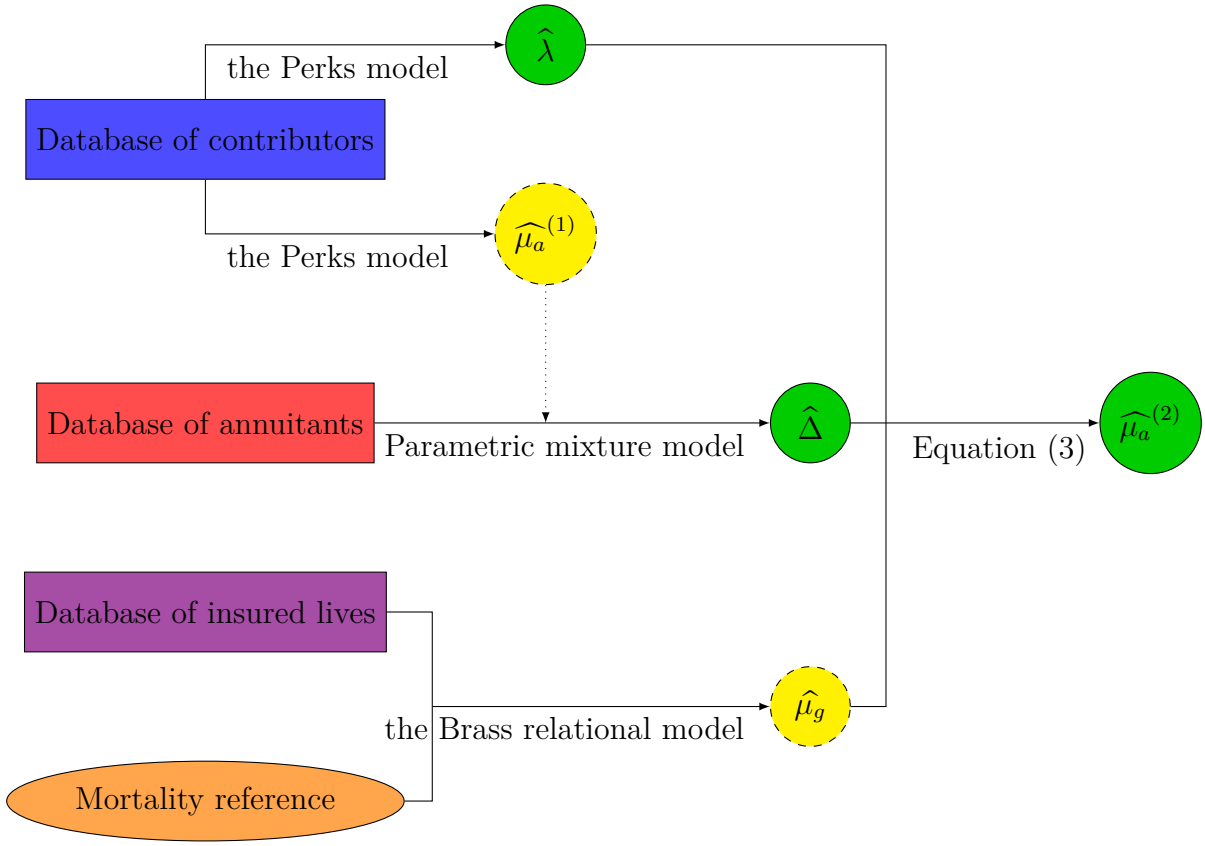


Figure 3.4: Procedure for the estimation of biometric laws. Dashed (resp. plain) circles represent intermediary (resp. final) estimates of biometric laws. Database of insured lives is obtained by merging the contributors and annuitants databases.

3.2.6 Pricing and reserving

We consider a product where the autonomous insured lives pay a fixed amount of premium while they are autonomous. Should they become disabled, the premium is no longer due and they are entitled to an annuity instead. On the French long-term care insurance market most products rely on monthly premium and monthly benefit. For simplicity sake, we consider continuous time premiums and annuities instead, the difference with monthly quantities being extremely low. We denote by τ the continuous time actuarial interest rate used to compute discounted cash flows.

Let us introduce additional notation

$$A(x,y) = \mathbb{P}(Z_y = A | Z_x = A) = \frac{A_y}{A_x} = \exp\left(-\int_x^y [\mu_a(u) + \lambda(u)]\,\mathrm{d}u\right),$$

$$I_x(t,s) = \mathbb{P}(Z_{x+s} = I | Z_{x-} = A, Z_x = I, Z_{x+t} = I) = \exp\left(-\int_t^s \mu_i(x,u)\mathrm{d}u\right)$$

and

$$\overline{A}(x,y) = e^{-\tau(y-x)} A(x,y) = \exp\left(-\int_x^y [\mu_a(u) + \lambda(u) + \tau]\,\mathrm{d}u\right),$$

$$\overline{I}_x(t,s) = e^{-\tau(s-t)} I_x(t,s) = \exp\left(-\int_t^s [\mu_i(x,u) + \tau]\,\mathrm{d}u\right)$$

for $x_0 \leqslant x \leqslant y$ and $0 \leqslant t \leqslant s$ such that A (resp. I) is the survival probability in the state of autonomy (resp. in the disabled state) and $\overline{A}$ (resp. $\overline{I}$) the discounted survival probability in the aforementioned state.

We define the following quantities that are required for the pricing of the product:

- $P(x)$ the expected value of insured liabilities for an autonomous insured life with current age x for a 1 € yearly premium

$$P(x) = \int_x^\infty \overline{A}(x,u)\mathrm{d}u.$$

- $RFC(x,t)$ the expected value of insurer liabilities for a disabled insured life with an age x at the onset of LTC and a time t spent in the disabled state and a 1 € yearly annuity, also called reserve for claim

$$\mathrm{RFC}(x,t) = \int_t^\infty \overline{I}_x(t,u)\mathrm{d}u.$$

- $\Pi(x)$ the expected value of insurer liabilities for an autonomous insured life with current age x, associated with a 1 € yearly annuity

$$\Pi(x) = \int_x^\infty \lambda(u)\overline{A}(x,u)\mathrm{RFC}(u,0)\mathrm{d}u.$$

- The stability premium $p^*(x)$. It is the value of premium that matches insurer and insured liabilities at the time of subscribing. For an age x at subscribing we have

$$p^*(x) = \frac{\Pi(x)}{P(x)}.$$

- The reserve for premium (RFP) which is constituted for autonomous people. Its amount is equal to the expected value of future liabilities minus the expected value of premium. For an insured of age at subscribing x_s, current age x, the associated amount of reserve is

$$\mathrm{RFP}(x_s,x) = P(x)\left[p^*(x) - p^*(x_s)\right].$$

3.3 Results

In this section, we provide an example using aggregated data from several French long-term care insurance portfolios. The definition used for LTC is 3ADL4 which means that an insured life is considered disabled if he/she has permanently lost the ability to do on their own at least 3 out of the 4 activities of daily living defined by the contract: functional mobility, dressing, bathing, eating. The portfolio we consider contains a very large number of policies and covers a relatively long period. The date of extraction is 11/31/2014 for both contributors and annuitants. We remove the first 3 years spent by contributors in the portfolio and then consider a 12 year observation period between 1/1/2002 and 12/31/2013 for contributors and a 20 year observation period between 1/1/1994 and 12/31/2013 for annuitants. Database of contributors contains over 1.5 million years of exposure with 69.8 % of the lines being right censored. Database of annuitants contains close to 45,000 years of exposure and 29.4 % of right censored lines. Women account for 65.4 % of contributors and 66.7 % of annuitants. Separate models are estimated for men and women.

3.3.1 General mortality

We use the Brass relational model with data for the french population over the years 2010 to 2013 coming from the Human Mortality Database (University of California, Berkeley (USA) and Max Planck Institute for Demographic Research (Germany), 2015) that we choose as our mortality reference. At the same time, we compute empirical probabilities of death, using the Hoem estimator as described in Planchet and Thérond (2006), as well as 95 % confidence intervals under the normal approximation, over the age range where the Cochran criterion is satisfied. Figure 3.5 displays the logarithm of the cumulative distribution odds (CDO) for empirical probabilities and the mortality reference and the difference between them. As it is close to a straight line, the underlying assumption of the model is satisfied. Figure 3.6 represents the observed and reference mortality, as well as the mortality fitted using the Brass relational model. The latter mortality is close to the observed mortality for ages where enough data is available and then smoothly converges toward the reference at higher ages, where no data is available.

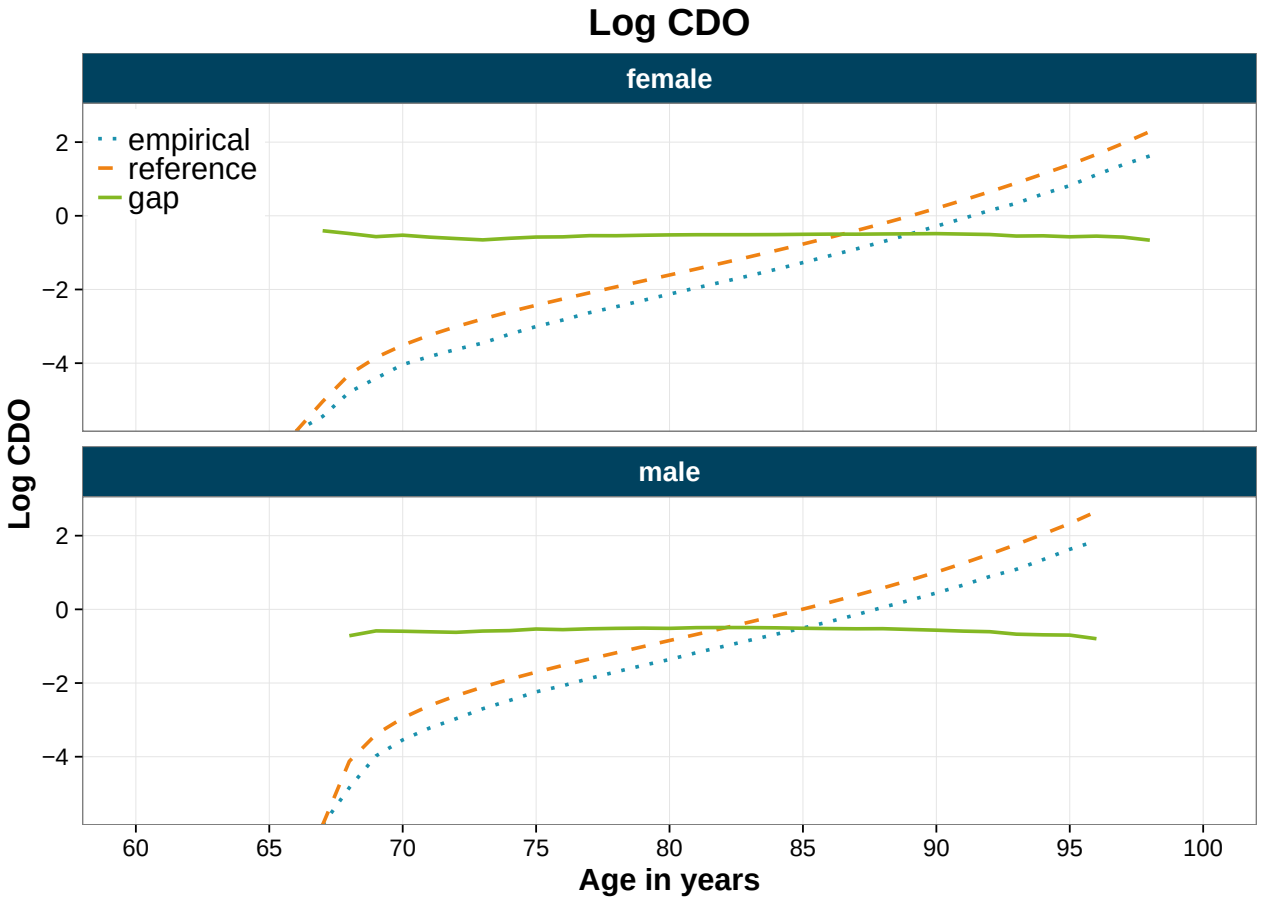


Figure 3.5: Logarithm of Cumulative Distribution Odds (CDO) for observed mortality (dotted), reference mortality (dashed) and their difference (plain).

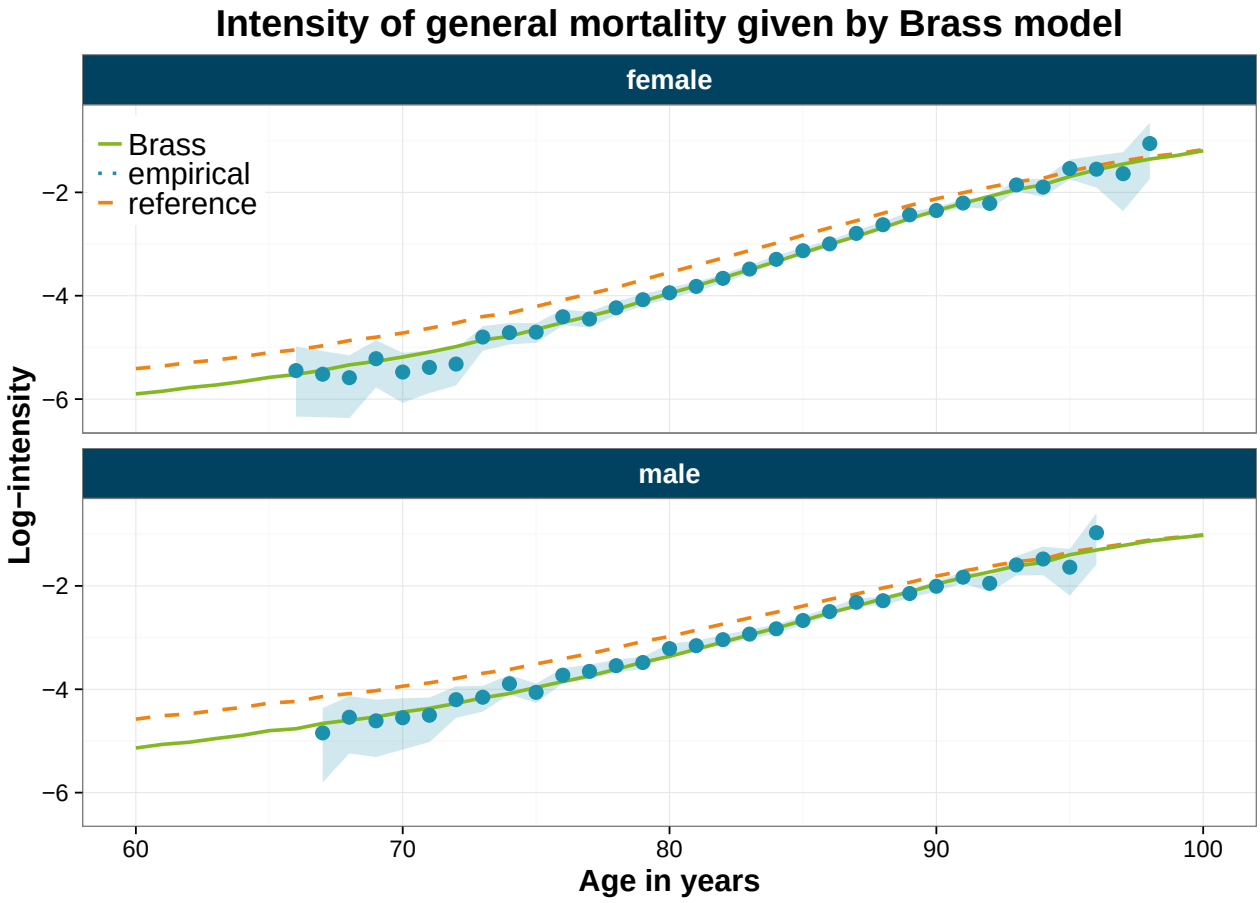


Figure 3.6: Intensity of mortality estimated from the data (dots) with 95 % confidence intervals, from the mortality reference (dashed) and given by the Brass relational model (plain).

3.3.2 Incidence in LTC

Model	Intensity	l(males)	BIC(males)	l(females)	BIC(females)
Gompertz	$\lambda(x) = e^{a_\lambda x + b_\lambda}$	- 25,099.12	50,223.64	- 52,380.25	104,788.21
Makeham	$\lambda(x) = e^{a_\lambda x + b_\lambda} + d_\lambda$	- 25,099.12	50,232.11	- 52,380.25	104,797.44
Beard	$\lambda(x) = \frac{e^{a_\lambda x + b_\lambda}}{1 + e^{a_\lambda x + c_\lambda}}$	- 25,094.61	**50,223.09**	- 52,326.58	**104,690.09**
Perks	$\lambda(x) = \frac{e^{a_\lambda x + b_\lambda}}{1 + e^{a_\lambda x + c_\lambda}} + d_\lambda$	- 25,093.74	50,229.81	- 52,325.83	104,697.83

Table 3.2: Value of log-likelihood l and BIC of previously introduced models for the incidence in LTC.

The results for the estimation of incidence in LTC can be found in Table 3.2. the Gompertz (resp. the Beard) model performs better than the Makeham (resp. the Perks) model according to the BIC, which means that the use of an extra parameter which represents an initial level of incidence present at all ages is not required. In addition, the Beard logistic model is a better fit to the data than the Gompertz exponential model. One can come to this conclusion by looking at Figure 3.7 which represents the empirical incidence as well as the inferred incidence for the Gompertz and the Beard models. The empirical incidence in LTC increases exponentially at

first but at higher ages there is a slowing down in this increase, more pronounced for females than for males, which makes the Beard logistic model a better fit.

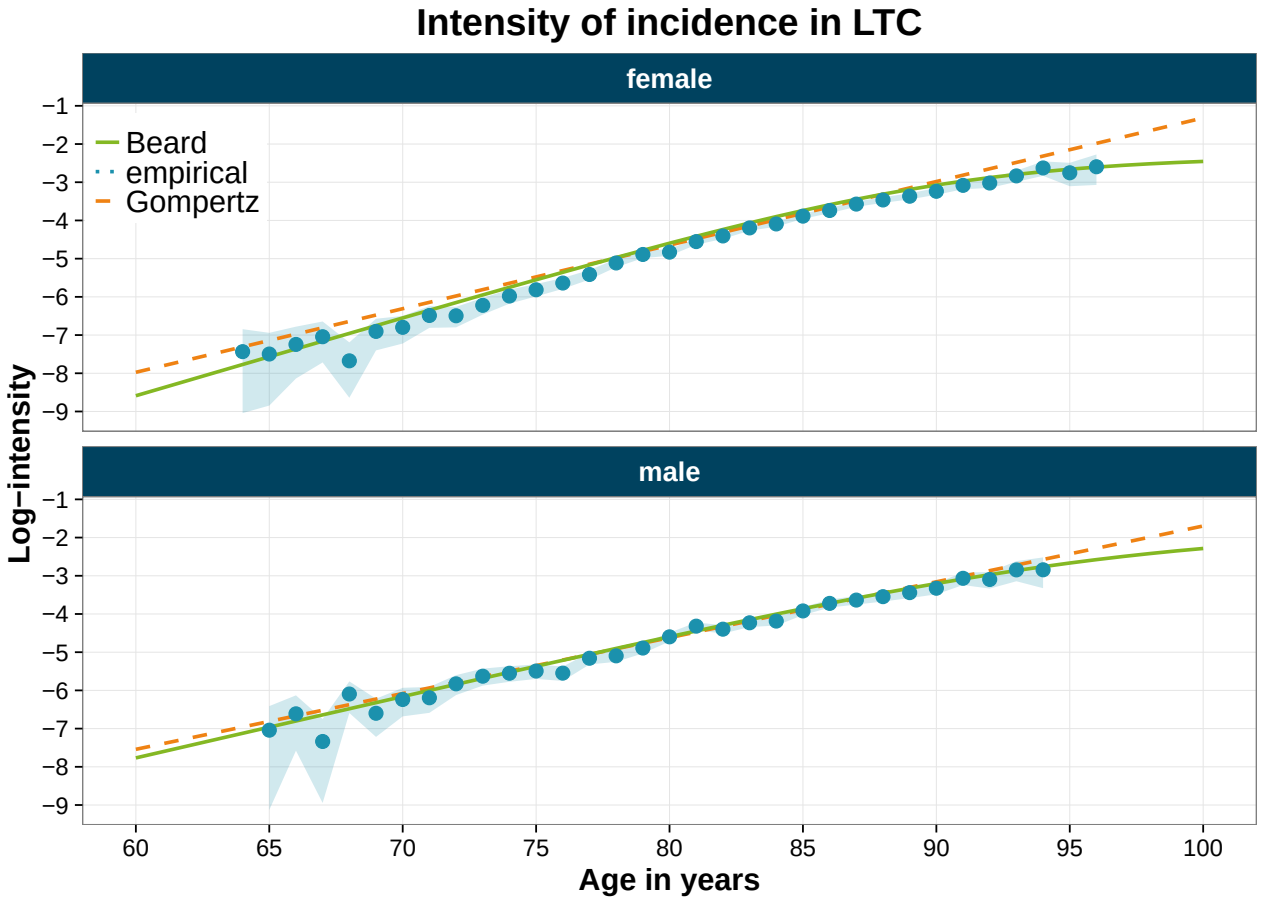


Figure 3.7: Estimates of the incidence in LTC. Dots and ribbon represent empirical estimates with 95 % confidence intervals. Plain (resp. dashed) line represents the Gompertz (resp. the Beard) model fitted to the data.

3.3.3 Mortality in LTC

We rely on the results from section 3.2.4 and define the mortality in LTC by providing a parametric model for Δ_1, Δ_2 and θ. In this section, we only focus on a handful of well known models that in our opinion are the most obvious candidates. Furthermore, for the sake of simplicity, we only consider models where Δ_1 and Δ_2 take the same parametric form. For $\Delta_1(x)$ and $\Delta_2(x)$, we consider constant, the Gompertz and the Makeham exponential models as well as the Beard and the Perks logistic models (so 5 different models in total). For $\theta(x)$, we have the additional constraint that we should have $0 \leqslant \theta(x) \leqslant 1$ for all ages. We consider a constant model then 4 logistic models with increasing degrees of freedom. Indeed, the full logistic law has 4 parameters and therefore 4 degrees of freedom. By setting the ultimate values for $x = -\infty$ and $x = +\infty$ respectively to 0 and 1, we obtain a logistic model with only 2 parameters. We may relax either of those constraints by introducing additional parameters $0 \leqslant \alpha \leqslant \beta \leqslant 1$ so that α (resp. β) is the ultimate value of $\theta(x)$ when $x = -\infty$ (resp. $x = +\infty$). Hence, we estimate $5 \times 5 = 25$ combinations of models. Results are available in Table 3.3 in the Appendix.

Figure 3.8 represents each of the models on the angle of the number of parameters and maximum log-likelihood. As the BIC is a linear combination of the two aforementioned components, contour curves of increasing BIC correspond to parallel lines of increasing intercept in this representation. The model with the best BIC is such that there is no other model in the

upper half two-dimensional space delimited by the associated contour curve. Distance from any model to this contour curve is proportional to the difference in BIC between that model and the best model. The criterion selects model 9 for males (models 7 and 10 being close contenders) and model 10 for females. All those models rely on the Gompertz law for the specific mortality terms Δ_1 and Δ_2. As regards θ, model 10 uses the full 4 parameters logistic model while model 9 only uses 3 parameters, the asymptotic value for the prevalence of high mortality (group 2) pathologies at lower ages being set to 100%. From this point all results are based on the parameters inferred for models 9 for males and 10 for females.

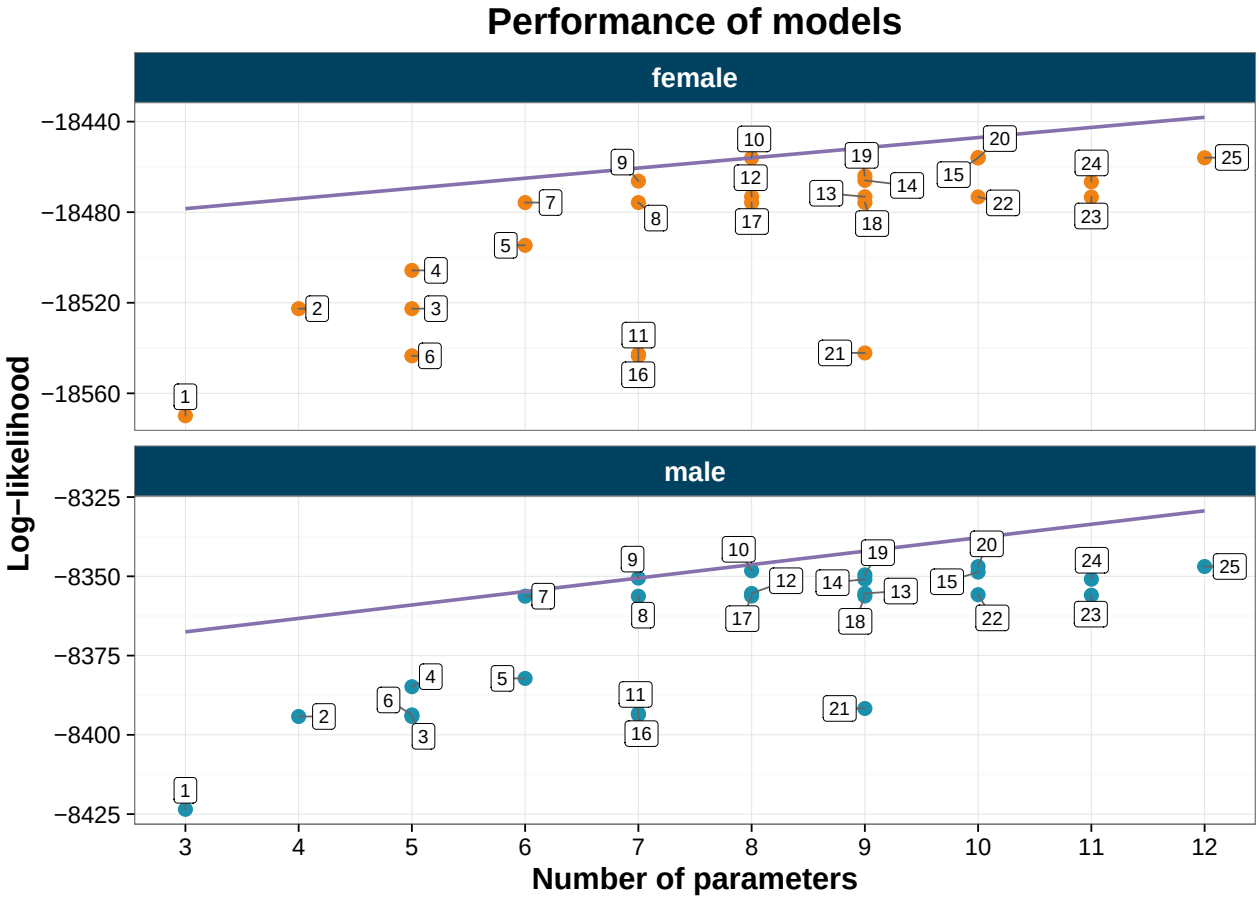


Figure 3.8: Representation of the models (dots) in the plan of log-likelihood and number of parameters. The plain line represents the contour curve for the model with the highest BIC.

Figure 3.9 displays the value of $\theta(x)$ which represents the prevalence of high mortality pathologies among newly disabled people inferred by the model. This prevalence decreases with age and is much higher for males (70 % at age 60 and 17 % at age 90) than for females (40 % at age 60 and 6 % at age 90). Figure 3.10 represents the specific mortality terms $\Delta_1(x)$ and $\Delta_2(x)$ and the resulting mortality term for the newly disabled $\Delta(x, 0)$, which is the weighted mean of $\Delta_1(x)$ and $\Delta_2(x)$ with weights $1 - \theta(x)$ and $\theta(x)$ respectively. We observe that $\Delta_2(x)$ is way higher than $\Delta_1(x)$. Besides the initial mortality $\Delta(x, 0)$ decreases with age until 85 then remains stable. Let us remind that for higher durations, $\Delta(x, t)$ converges toward the lower value between $\Delta_1(x)$ and $\Delta_2(x)$ as the weight of the population with higher mortality in the mixture decreases to 0. Those results seem compatible with our interpretation in terms of cancer for the group of high mortality pathologies and dementia as well as cardiovascular and neurological diseases for the other group. However, one should keep in mind that pathologies are not actually observed in the data and Figures 3.9 and 3.10 only represents the underlying distribution inferred by the model.

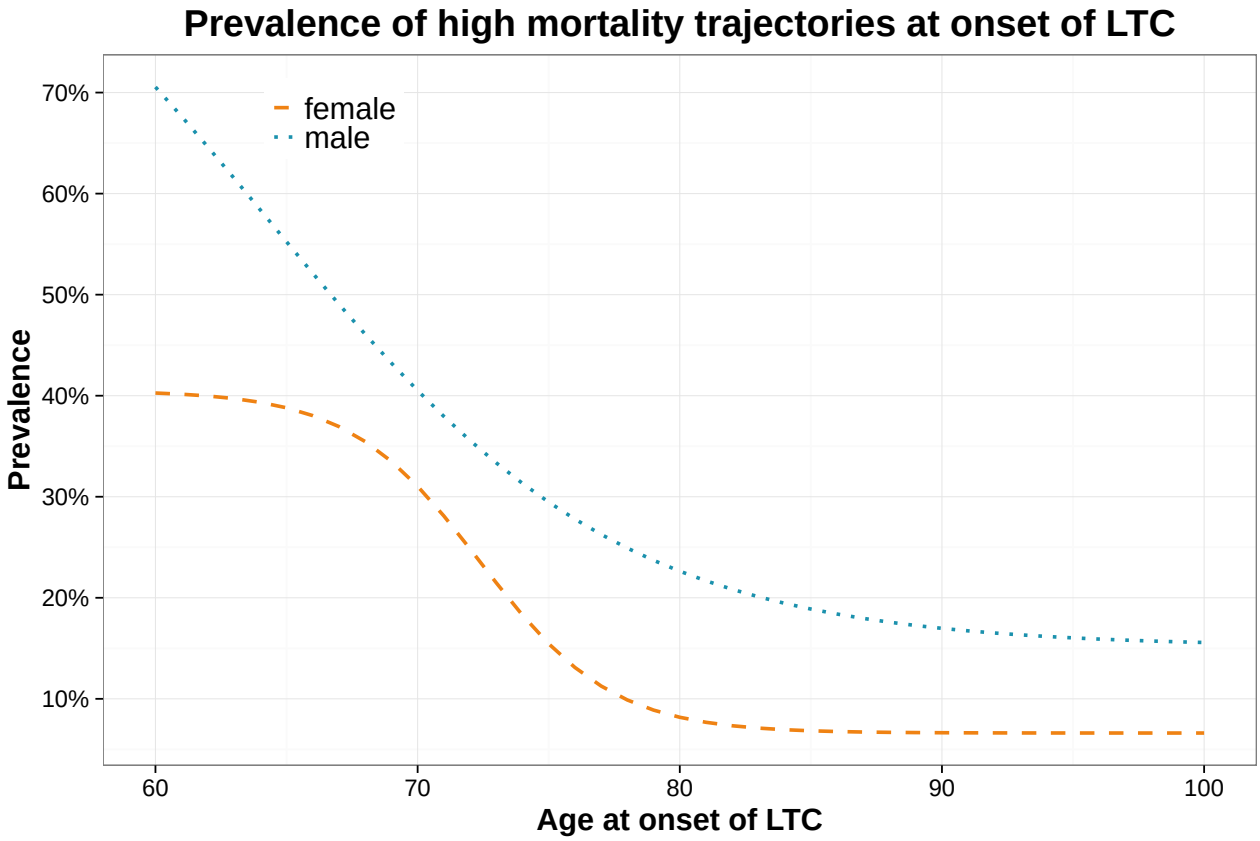


Figure 3.9: Inferred prevalence of high mortality trajectories in the population of newly disabled.

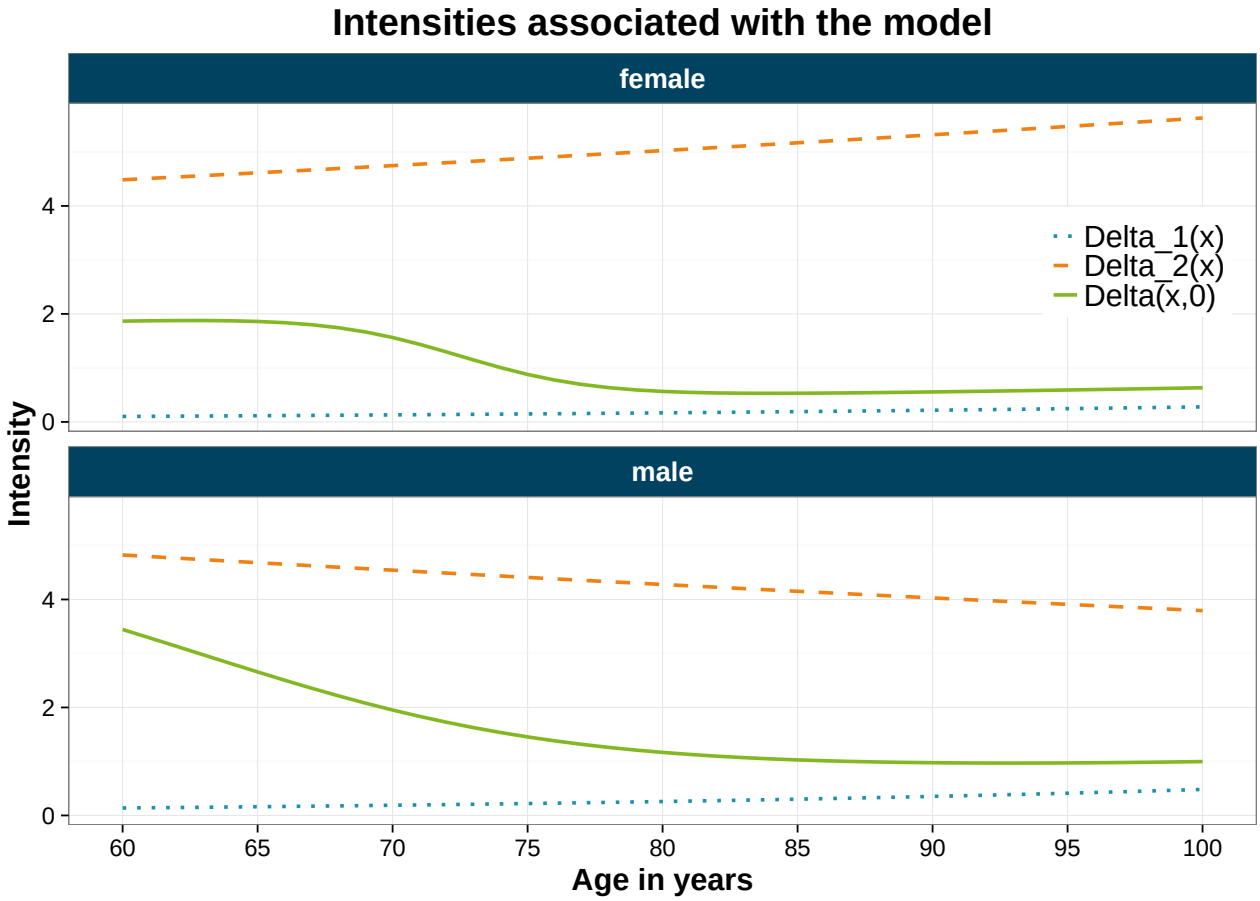


Figure 3.10: Specific mortality terms for both populations in the mixture (dotted and dashed lines), and resulting mortality at the onset of LTC (plain line).

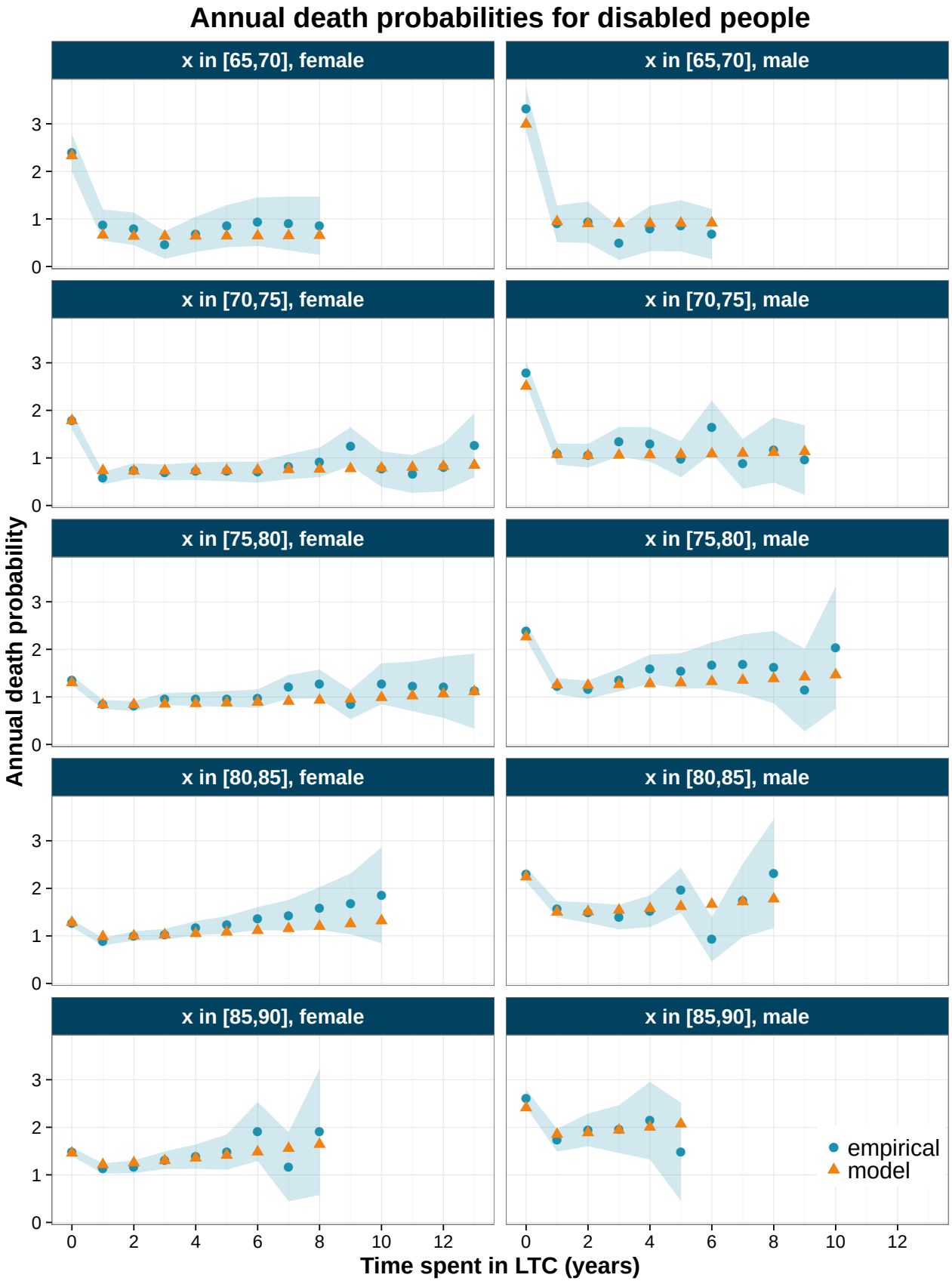


Figure 3.11: Consecutive death probabilities for disabled people according to the model (triangles) with empirical probabilities (circles) and associated 95 % confidence intervals. The y-scale has been re-normalized to preserve confidentiality of results.

Figure 3.11 represents annual death probabilities associated with the empirical data on one hand and given by the model on the other hand. We compute empirical annual probabilities by grouping disabled people, according to their age of entry in LTC with 5-year age bands between

65 and 90. For each age band, we then compute annual death probabilities by duration under the assumption that the intensity of mortality is constant over intervals of one year for the duration. We represent 95 % confidence intervals for those probabilities under the normal approximation (as in Planchet and Thérond, 2006). We also compute annual death probabilities given by the model for individuals of ages 67.5, 72.5,..., 87.5 at onset of LTC. Confidence intervals are still very wide, especially for men as well as at lower/higher age at onset of LTC and/or high duration in LTC. Nonetheless, the annual probabilities computed using the model appear to match the empirical probabilities very well for both males and females when data is available. It appears that by taking into account a mixture component in the model, we were able to reproduce the evolution of death probabilities with respect to time spent in LTC. Nevertheless, in each component of the mixture, time spent in LTC only appears in the autonomous mortality term, through the current age $x + t$.

3.3.4 Autonomous mortality

Figure 3.12 represents the initial intensity of autonomous mortality we get from the first-step estimator $\widehat{\mu_a}^{(1)}$ and the refined intensity from the second-step estimator $\widehat{\mu_a}^{(2)}$. The refined intensity remains close to the empirical intensity for females but for males there are some divergences which can be explained by the lower volume of data.

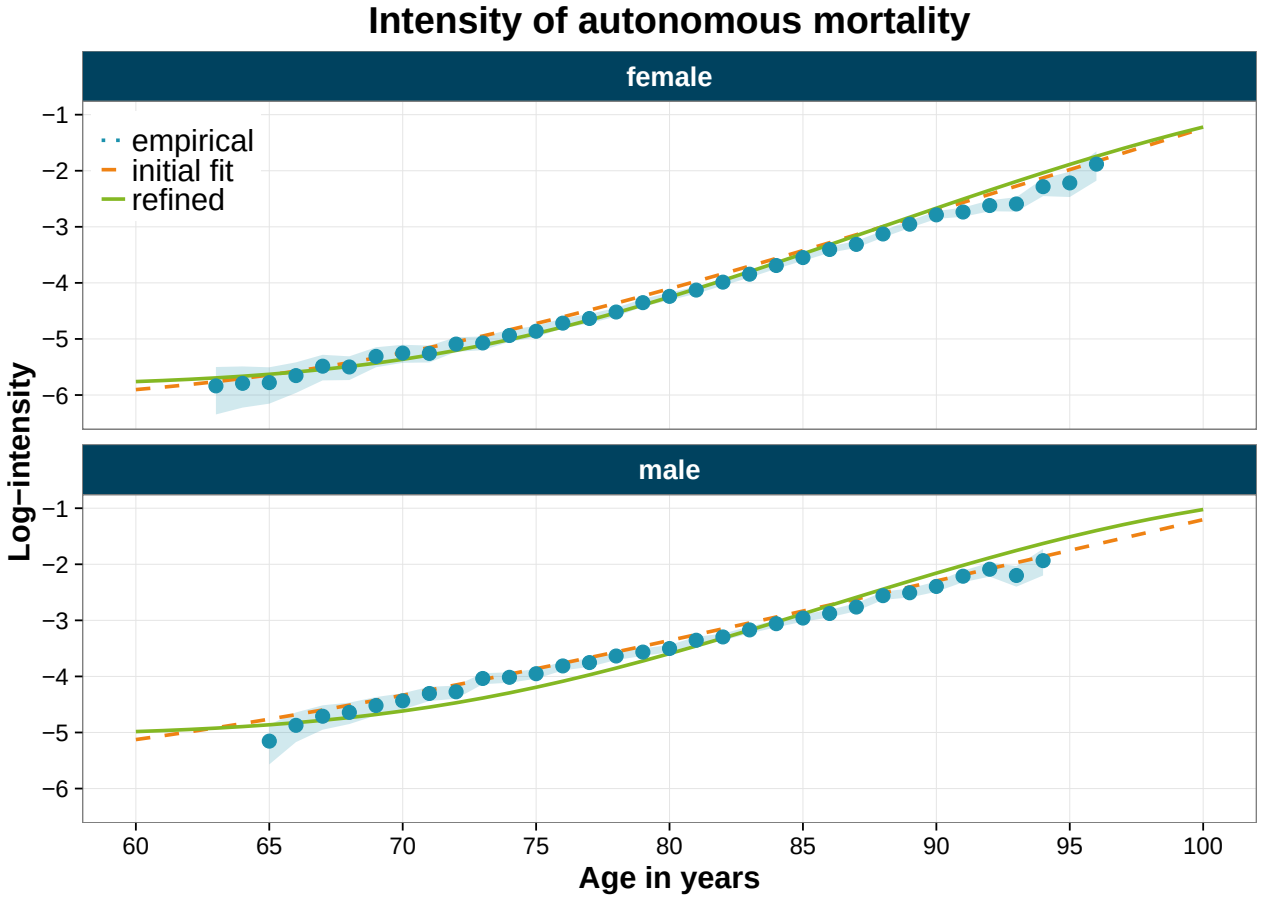


Figure 3.12: Intensity of autonomous mortality. Dots: empirical rates; Dashed line: direct fit of the Perks model; Plain line: refined intensity from equation (3.5).

3.3.5 Summary of intensities and prevalence of LTC

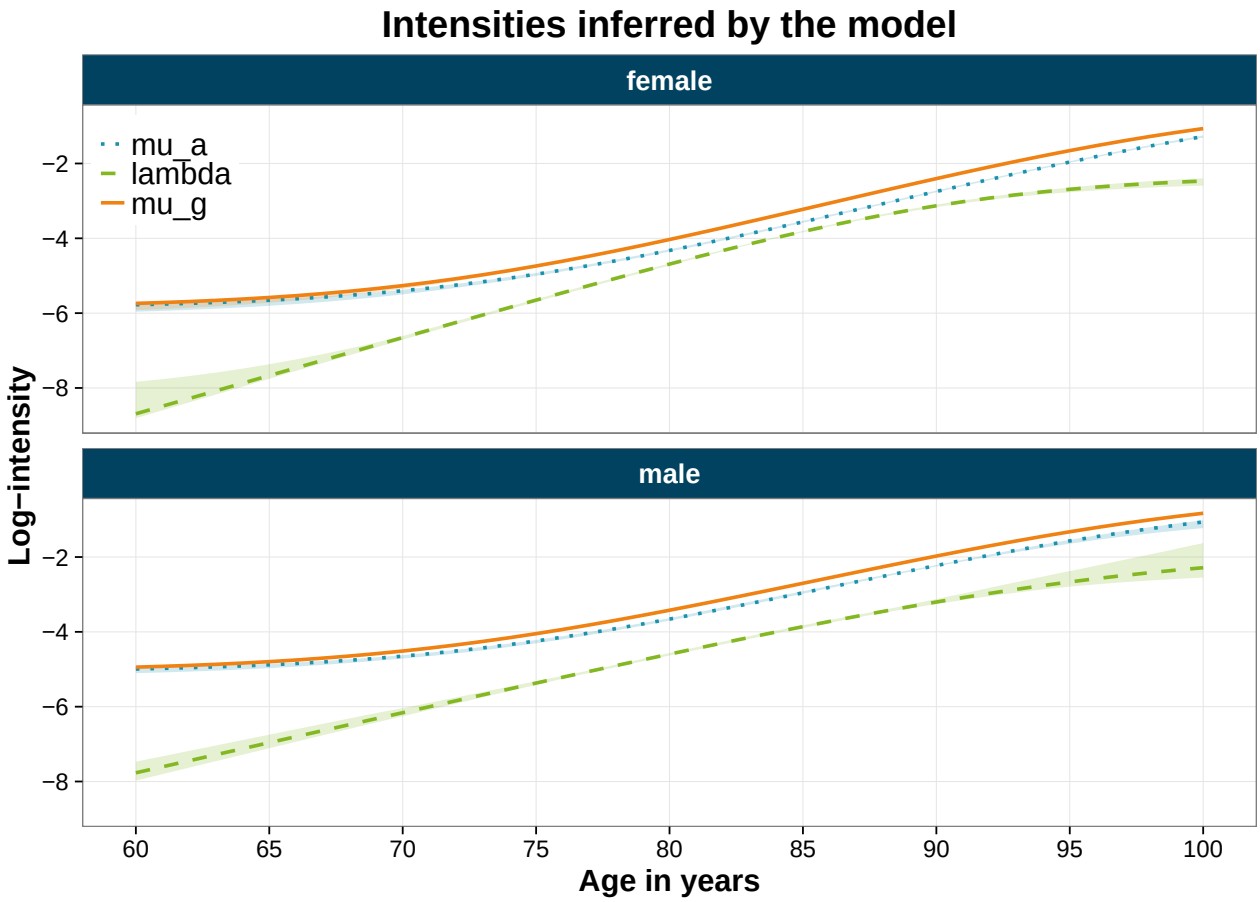


Figure 3.13: Intensities of transition with 95 % confidence intervals.

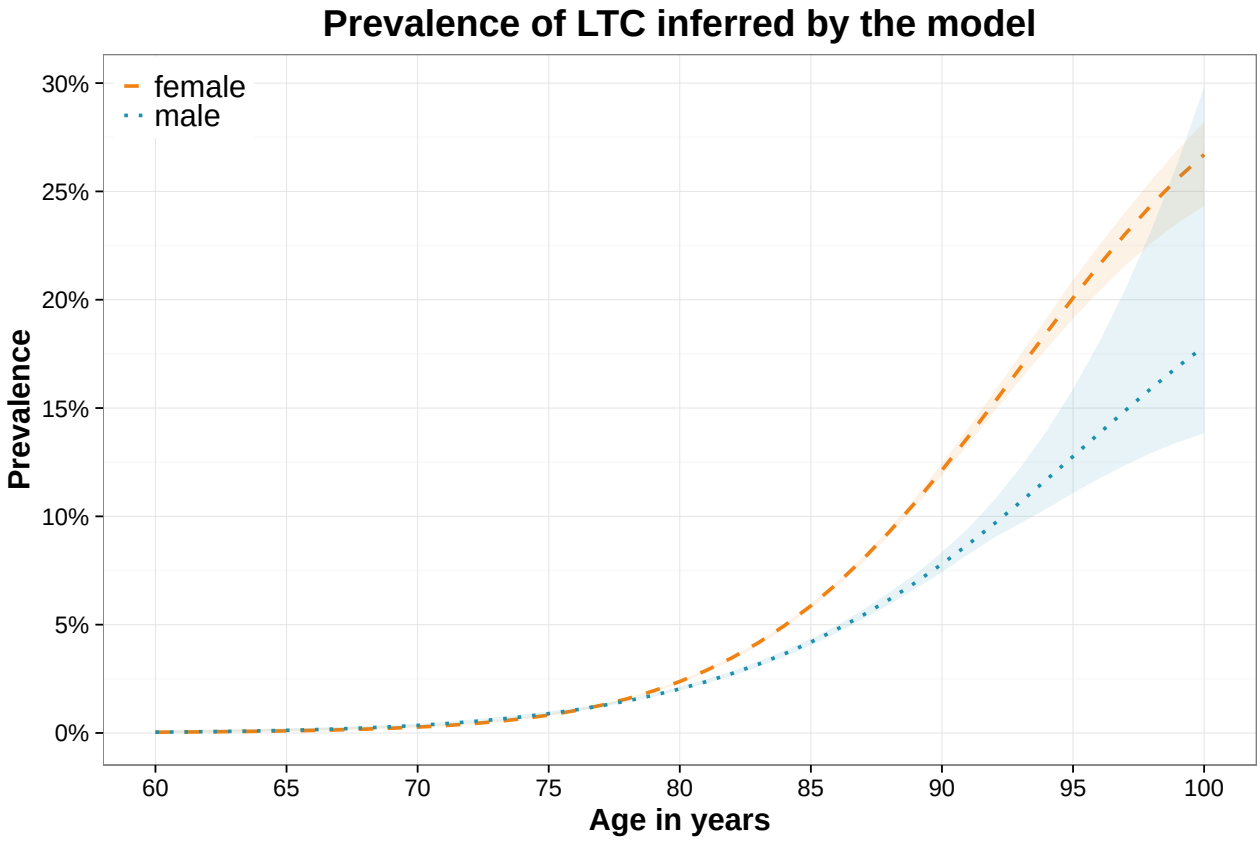


Figure 3.14: Prevalence of LTC in the general population with 95 % confidence intervals.

In order to assess the robustness of the estimation performed, we use a non-parametric quantile bootstrap method. From the initial database of insured lives, we build 200 new samples by drawing, with replacement, as many individuals as in the initial observation database. For each sample, we then run all the steps of the estimation procedure, including the choice of the best model according to the BIC. We then use the inferred parameters to compute the final intensities of transition as well as the prevalence of LTC. Finally, for each age, we select the 2.5 % and 97.5 % quantiles of the empirical distribution of those quantities in order to get bootstrap confidence interval.

Figure 3.13 represents the intensity of mortality for the general population, the intensity of mortality for autonomous people as well as the incidence in LTC. Confidence intervals are very tight for autonomous and general mortality. For the incidence in LTC they are larger, especially at lower or higher ages, and for males as the data is scarcer. Figure 3.14 represents the prevalence of LTC among the general population inferred by the model. The prevalence increases almost exponentially with respect to age at first, with a slowing down at higher ages. Prevalence is initially close for males and females, but from the age of 80 it becomes much higher for females. Overall the confidence intervals are very large especially for males at higher ages where the number of survivors is limited.

3.3.6 Results of pricing and reserving

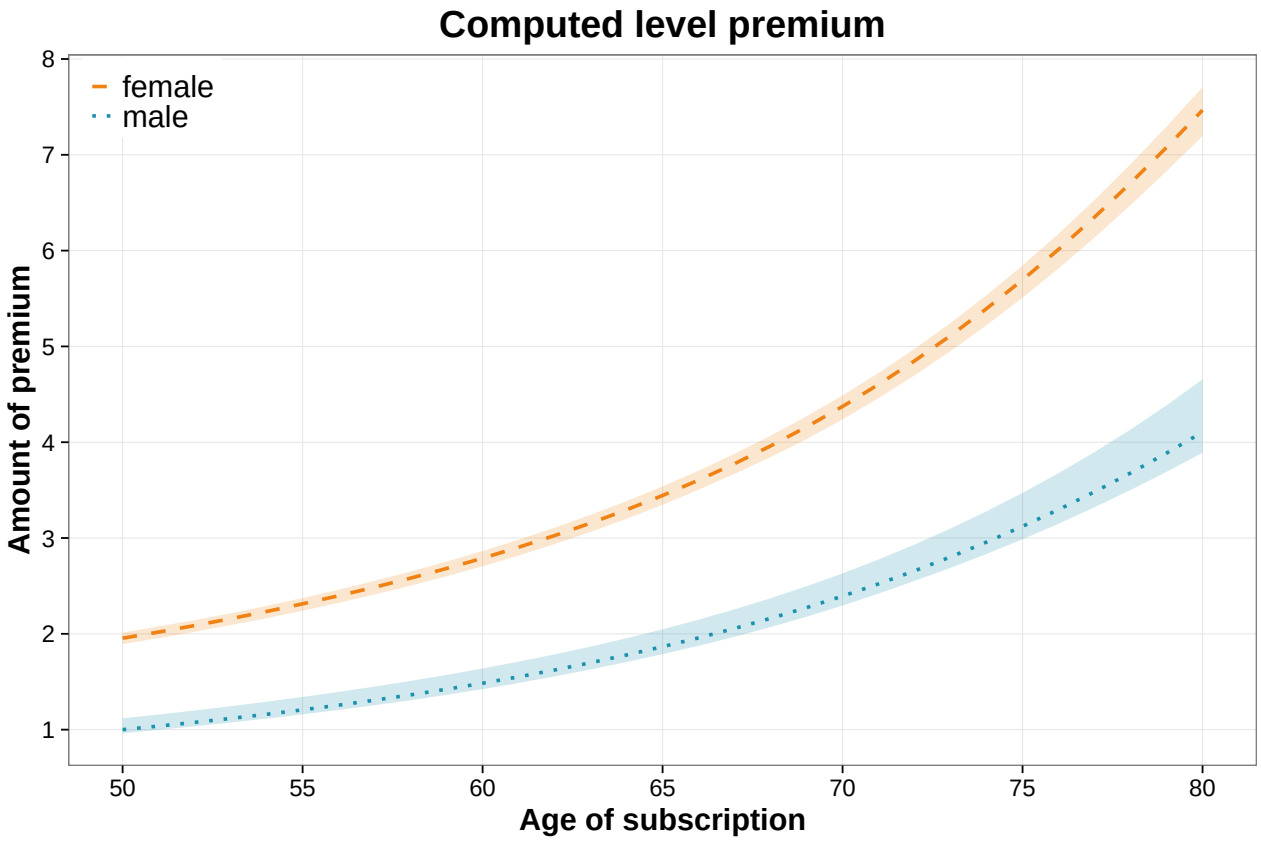


Figure 3.15: Amount of monthly premium required according to the model, with 95 % confidence intervals obtained by bootstrap. The y-scale has been re-normalized to preserve confidentiality of results.

We consider a long-term care insurance product where autonomous policyholders pay a monthly level premium, whose amount is set based on their age of subscription. Should they become disabled, they would stop paying the premium and instead receive a monthly annuity of 1,000 € until they die. We use an actuarial interest rate of 1 % for the pricing of the product. Figure 3.15 shows the required level of premium according to the model for ages at subscribing

from 50 and 80, as well as confidence intervals obtained by bootstrap, using the methodology described in the previous section. The premium increases exponentially with age and is is twice as expensive for women than for men. Confidence intervals are relatively tight given the uncertainty on the underlying biometric laws. The method therefore proves quite robust.

We also compute the average reserve of premium on Figure 3.16. We define it as the product between the probability $A(x_s, x)$ for the individual to remain autonomous between the age of subscription x_s and the current age x and the associated amount of reserve for premium $\text{RFP}(x_s, x)$ at that age. The reserve for premium reaches a maximum between ages 78 and 88, depending on the age of subscription and then decreases when the cost associated with the claims starts to outweigh the amount of premium. We then compute the average reserve for claim on Figure 3.17. We define as the product between the survival probability in LTC $I_x(0, t)$ at the age of claim x for the given duration t and the associated amount of reserve for claim $\text{RFC}(x, t)$. This reserve decreases by duration as the number of survivors does. The initial amount of reserve for claim reaches its maximum for claim inception under 60 for women and between 70 and 80 for men. Indeed, for males, the incidence of cancer is very high for ages under 70. Therefore men under 70 have very high death probabilities for the first year following the onset of LTC while men over 80 have very high death probabilities for the subsequent years, because mortality from other causes of death get higher with ages. For women, this second phenomenon carries more weight, and the most expensive claims are made before age 60.

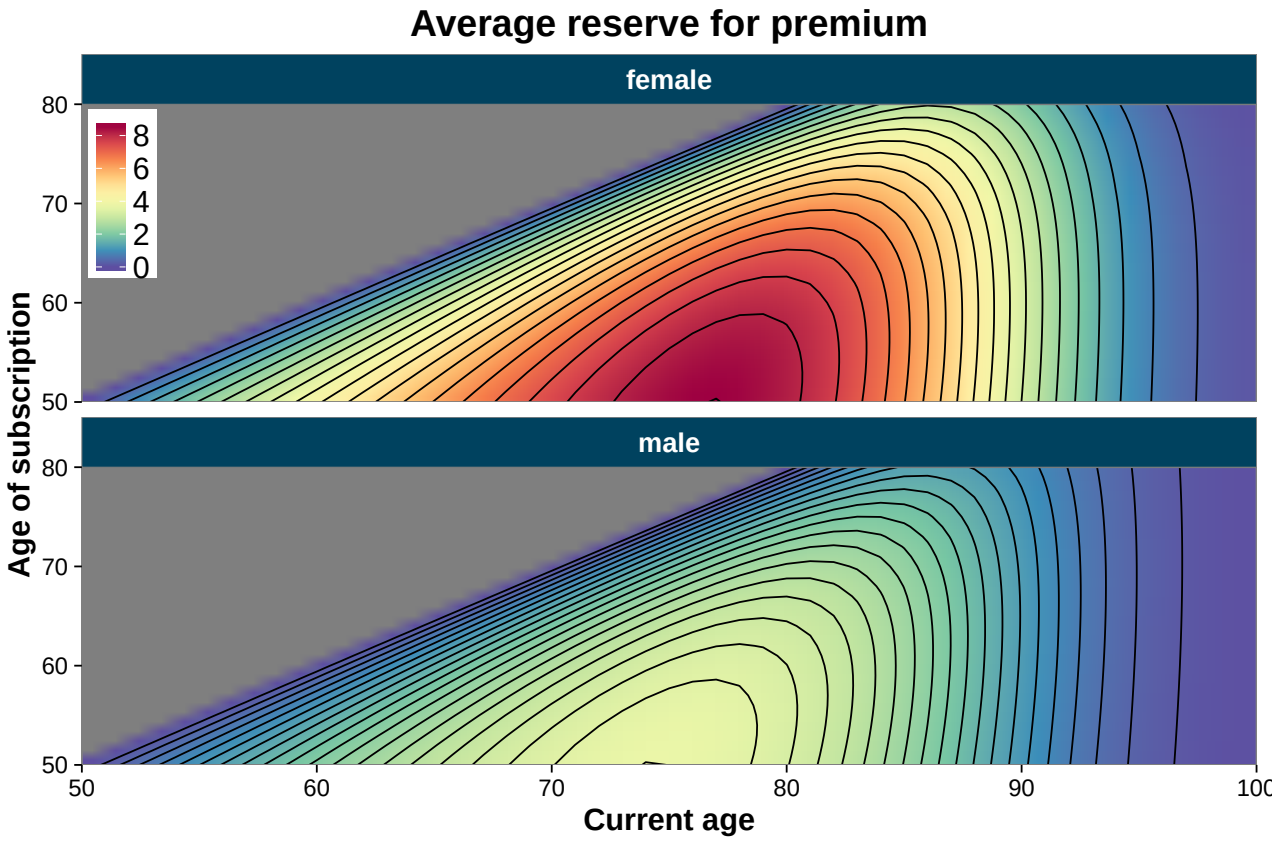


Figure 3.16: Expected value of reserve for premium by age at subscribing and current age, assessed at subscribing. The z-scale has been re-normalized to preserve confidentiality of results.

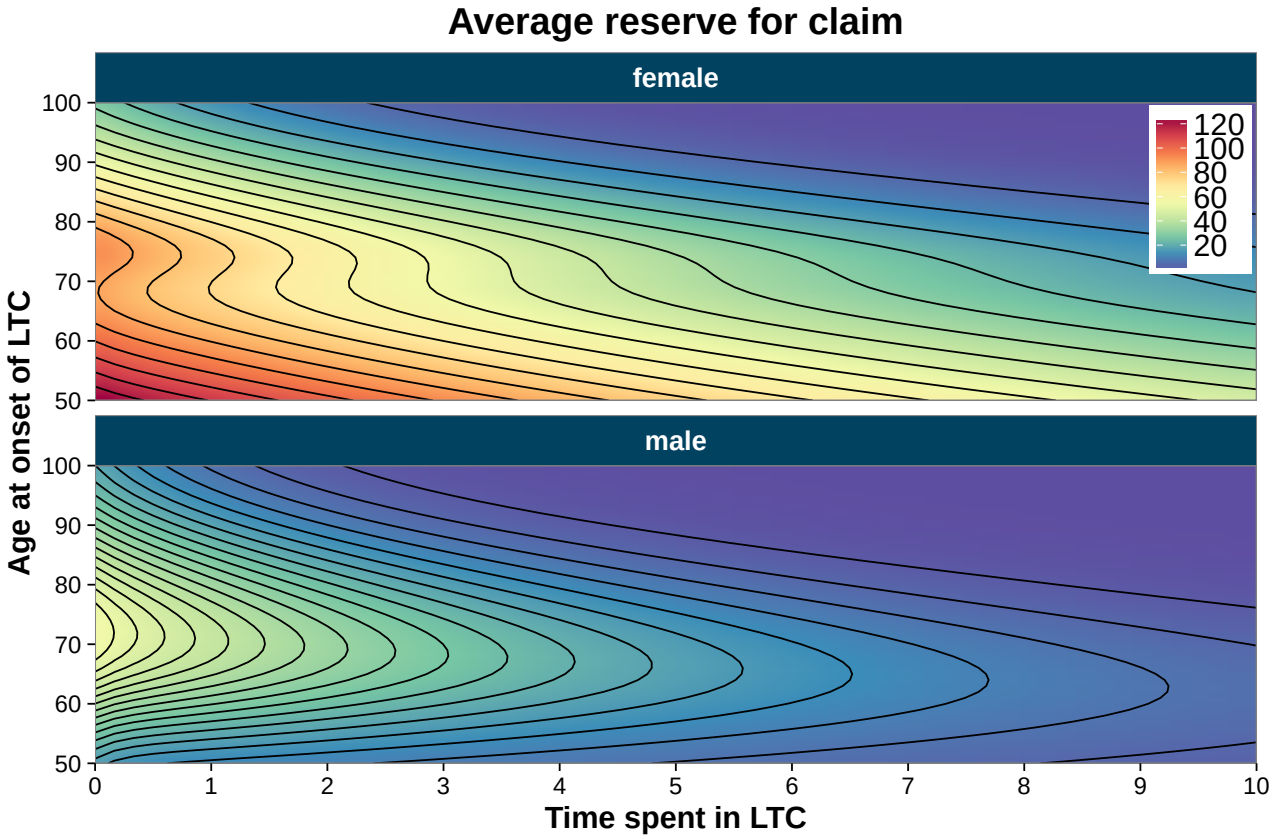


Figure 3.17: Expected value of reserve for claim by age at entry in LTC and time spent in LTC, assessed at claim inception. The z-scale has been re-normalized to preserve confidentiality of results.

3.4 Discussion

In this paper, we introduce a method to estimate biometric laws associated with a long-term care insurance portfolio. This method relies on a continuous time semi-Markov model, as opposed to discrete-time methods used by practitioners in most countries (with the notable exception of Denmark, see Ramlau-Hansen (1991)). This model relies on 3 transition intensities: incidence in LTC, autonomous mortality and mortality in LTC. We suggest parametric models for the transition intensities. The Brass relational model is used for the intensity of general mortality and the Perks logistic model is used for incidence in LTC, as well as for the first-step estimate of autonomous mortality. As regards mortality in LTC we introduce a mixture model. The aim is to model the underlying heterogeneity in the population caused by the very different pathologies that may lead to LTC. Inference of parameters relies on the maximum likelihood method. We then introduce a formula to include general mortality of the portfolio in the model (on which we expect to have more reliable knowledge) and use it to get a second-step estimator of the autonomous mortality, which should prove more reliable at higher ages. We also provide adequate formulas for continuous-time pricing and reserving based directly on the transition intensities. Let us remind that there is only few data available at higher ages on for high duration in LTC. Therefore parametric methods are compulsory to extrapolate biometric laws at higher ages. Using parametric models from the start is very convenient for the practitioner as it allows to derive biometric laws in a single step, while non-parametric methods requires to find adequate age bands to perform empirical estimations, smooth the empirical probabilities and finally extrapolate the results for higher ages.

We then apply our methodology to data from a real long-term care insurance portfolio. Empirical probabilities demonstrates that mortality during the first year following the onset

of LTC is way higher than for the subsequent years. A semi-Markov model which takes into account both the age at the onset of LTC and the duration in the LTC state is therefore required in order to explain this phenomenon. By taking into account heterogeneity in the trajectories through a mixture model, we obtain such a model for the mortality in LTC which proves very close to empirical estimations. This may indicate that most of the effect of duration on the mortality actually comes from the heterogeneity of causes.

In the present article, we take into account several potential sources of error. As we use a parametric approach, there is a significant risk of modeling error that we try to mitigate by comparing the results of the model with the empirical annual probabilities obtained using a classic non-parametric approach. We also consider several sub-models and remain parsimonious in the number of parameters we introduce by using the Bayesian Information Criterion to compare models. Furthermore, the robustness of estimation is also assessed using a non-parametric quantile bootstrap method.

The parametric form we introduce for mortality in LTC is based on the assumption that pathologies can be sorted in two main groups of homogeneous mortality. This assumption may be tested by focusing on the study of the pathologies causing LTC. Data containing information about pathologies is however extremely scarce and kept private by most insurers. Another limit to our estimation approach is that it is stationary and does not consider that biometric laws are changing over time. The estimation of drifts would indeed prove very difficult because of the limited observation period, and lack of consistency in definition of LTC as well as changes in underwriting and claim management policies over time. Also, most products in France allow the insurer to increase the level of premium in order to account for drifts in the underlying risk. While this may justify not to consider any trend in the model, a sensitivity approach would in any case prove very useful. We could consider several scenarios for the improvement of incidence and mortality rates and look at the impact on the insurer technical result. Nevertheless, to the best of our knowledge, neither the data nor the theoretical framework associated with this issue exist. Finally, the model only considers one level of LTC, when most individual LTC products currently sold provide several levels of benefits according to the severity of the disability state. Extending the model to consider several levels of LTC as in Lepez et al. (2013) or Biessy (2015b) would therefore prove very useful. Once again, finding adequate data to perform estimation of parameters is very challenging.

Appendix

Model	Δ_1, Δ_2	θ	k	l(males)	BIC(males)	l(females)	BIC(females)
1	Constant	Constant	3	- 8,423.49	16,872.43	- 18,569.90	37,166.71
2	Constant	Logistic$(0,1)$	4	- 8,394.21	16,822.36	- 18,522.60	37,081.07
3	Constant	Logistic$(0,\beta)$	5	- 8,394.21	16,830.84	- 18,522.60	37,090.04
4	Constant	Logistic$(\alpha,1)$	5	- 8,384.81	16,812.05	- 18,505.70	37,056.24
5	Constant	Logistic(α,β)	6	- 8,382.20	16,815.30	- 18,494.63	37,043.07
6	Gompertz	Constant	5	- 8,393.70	16,829.81	- 18,543.44	37,131.71
7	Gompertz	Logistic$(0,1)$	6	- 8,356.27	16,763.44	- 18,475.74	37,005.29
8	Gompertz	Logistic$(0,\beta)$	7	- 8,356.27	16,771.93	- 18,475.74	37,014.26
9	Gompertz	Logistic$(\alpha,1)$	7	- 8,350.55	**16,760.50**	- 18,466.26	36,995.30
10	Gompertz	Logistic(α,β)	8	- 8,348.25	16,764.38	- 18,456.02	**36,983.79**
11	Makeham	Constant	7	- 8,393.26	16,845.91	- 18,542.66	37,148.10
12	Makeham	Logistic$(0,1)$	8	- 8,355.39	16,778.65	- 18,473.17	37,018.08
13	Makeham	Logistic$(0,\beta)$	9	- 8,355.39	16,787.14	- 18,473.17	37,027.04
14	Makeham	Logistic$(\alpha,1)$	9	- 8,350.91	16,778.18	- 18,465.94	37,012.60
15	Makeham	Logistic(α,β)	10	- 8,348.66	16,782.18	- 18,456.04	37,001.75
16	Beard	Constant	7	- 8,393.62	16,846.64	- 18,543.43	37,149.63
17	Beard	Logistic$(0,1)$	8	- 8,356.21	16,780.30	- 18,475.74	37,023.23
18	Beard	Logistic$(0,\beta)$	9	- 8,356.22	16,788.81	- 18,475.74	37,032.20
19	Beard	Logistic$(\alpha,1)$	9	- 8,349.54	16,775.44	- 18,463.98	37,008.67
20	Beard	Logistic(α,β)	10	- 8,346.92	16,778.69	- 18,455.87	37,001.41
21	Perks	Constant	9	- 8,391.68	16,859.73	- 18,542.14	37,165.00
22	Perks	Logistic$(0,1)$	10	- 8,355.79	16,796.42	- 18,473.26	37,036.20
23	Perks	Logistic$(0,\beta)$	11	- 8,355.95	16,805.23	- 18,473.30	37,045.24
24	Perks	Logistic$(\alpha,1)$	11	- 8,350.91	16,795.15	- 18,466.62	37,031.88
25	Perks	Logistic(α,β)	12	- 8,346.90	16,795.62	- 18,455.96	37,019.53

Table 3.3: Value of log-likelihood l and BIC of models for mortality in LTC.

Chapitre 4

A semi-Markov model with pathologies for Long-Term Care Insurance

Abstract

Most Long-Term Care (LTC) Insurance products rely on definitions for functional disability based on the Activities of Daily Living (ADL). While functional disability may reflect the level of care required by the insured life, it is not on its own a good predictor of lifespan in LTC, which strongly depends on the underlying pathology responsible for disability. Indeed, cancer and respiratory diseases are associated with extremely short lifespan while dementia and neurological diseases make for much longer claims. Pathologies are therefore responsible for heterogeneity in the data, which makes estimation of mortality in LTC a difficult issue. As a consequence the associated literature is still scarce.

In this paper, we study the mortality in LTC associated with 4 different groups of pathologies: *cancer*, *dementia*, *neurological diseases* and *other causes* based on data from a French LTC portfolio. We consider a semi-Markov framework, where mortality in LTC depends on both age at claim inception and time already spent in LTC. We first derive the incidence rate in LTC and mortality rate associated with each group of pathologies and for both males and females. To do so, we rely on local likelihood methods that we apply directly to transition intensities of the model. We then combine those transition intensities to get a second-step estimator of the overall mortality in LTC, which proves more accurate than a direct estimate regardless of the pathology. Finally our results indicates that the peak of mortality following entry in LTC observed in the data is mostly due to the *cancer* group.

4.1 Introduction

Long-Term Care (LTC) Insurance providers often base their guarantees on the notion of Activities of Daily Living to describe the level of functional disability reached by their insured lives. The main advantages of this approach is that it is easy to understand for the insured life, relatively easy to assess for the insurer and it has strong links with the level of care required by the insured life. Therefore, to better describe the process of LTC, a natural approach is to consider models that take into account several levels of functional disability. Multi-states models have been identified early as an adequate tool to study the risk, as in Haberman and Pitacco (1998) or Denuit and Robert (2007). However, LTC products are relatively recent, and the available data is still scarce. Alternatively, Czado and Rudolph (2002) provide an interesting study of the impact of covariates such as gender and level of care on the risk based on a German LTC portfolio and using Cox proportional hazard model (as in Cox, 1972). While Markov models allow for an impact of the age of the insured life on the risk, they generally ignore the impact of the duration in the LTC state. Nevertheless, Helms et al. (2005) partially overcome this limit by considering number of years already spent in LTC as a covariate in a Markov framework using the Aalen-Johansen estimator (one can refer to Aalen and Johansen, 1978, for a description of the estimator) also based on a German insurance portfolio. The semi-Markov framework, where mortality in LTC depends on both age of entry and duration in LTC, has only been applied very recently to LTC. Using data from the French public aid for LTC, Lepez et al. (2013) and Biessy (2015b) rely on a multi-state model for the mortality in LTC with 4 states of functional disability. However, the short observation period and specificities of the methodology used to gather the data makes estimation very difficult.

Functional disability is only a descriptive indicator of the insured life health status while LTC has a variety of causes, the most frequent being cancer, dementia, neurological diseases, cardiovascular diseases, muscular and skeletal diseases. Among less frequent causes one can find respiratory diseases, blindness, depression, accidents, diabetes and cirrhosis. This results in a large amount of heterogeneity within the disabled population and lifespan in LTC strongly depends on the underlying cause. Therefore causes should be included in LTC models when data is available. Guibert and Planchet (2014) study the incidence rates associated with 4 groups of pathology, based on a French LTC portfolio. Using the same data, Guibert and Planchet (2015) computes non-parametric estimate of mortality rates in LTC for each group of pathology using the Aalen-Johansen estimator. Data containing information about pathologies is however extremely rare. When this information is missing, Biessy (2015a) suggests to introduce nonetheless a mixture model and derives a parametric model for the mortality in LTC which proves a reasonable fit to the data considered.

In this article, we propose to estimate the incidence rates and the mortality in LTC associated with 4 different groups of pathology: *cancer*, *dementia*, *neurological diseases* and *other causes*. We then recombine the results in order to get a second-step estimator of the global mortality in LTC. We rely on data from a French LTC portfolio with close to 20,000 observed claims. We consider the framework of the illness death model with no return (see Pitacco, 2014, for an extensive study of the illness-death model and its uses in health and life insurance) and consider that mortality in LTC is a function of both age at entry and time spent in LTC, as in Biessy (2015a). To estimate the mortality in LTC, we propose to rely on local likelihood methods first introduced in Tibshirani and Hastie (1987). The interested reader can refer to the excellent book from Loader (1999) for an exhaustive description of those methods and their various applications. An application to the estimation of mortality in LTC is presented in Tomas and Planchet (2013), who compare local likelihood methods with adaptive bandwidth to non-adaptive local likelihood or B-splines methods (one can refer to Eilers and Marx, 1996, for a description of B-splines). In this article, we propose to apply local likelihood methods for the direct estimation of continuous time transition intensities rather than for the smoothing of

empirical estimates. While, this approach is partially described in Chapter 7 of Loader (1999), it has received very little interest and to the best of our knowledge has not yet been used for insurance portfolio studies. A strong constraint of this approach is that an access to individual data is required. Also, a small extension of the original model provided by Loader (1999) is required to handle left-truncated data.

Section 2 of the article provides a general description of the data at hand, and introduces the framework for the estimation of incidence and mortality in LTC. We first consider a multi-state model with autonomy, death and 4 states of LTC corresponding to the 4 groups of pathologies available in the data and provide a formula to compute the overall mortality in LTC from incidence and mortality rates for each group. Section 3 introduces the local likelihood method and apply it for the estimation of autonomous mortality, incidence in LTC and the 2-dimension surface of mortality in LTC. Section 4 highlights some of the results obtained about the impact of causes on the risk and results on the second-step estimate of mortality we get by recombining all causes. Lastly, section 5 summarizes the results of the article and discusses limits and potential improvements of our approach.

4.2 Data and model

4.2.1 Data at hand

The data used in this article comes from a French LTC portfolio. The associated policy is triggered when insured lives have lost the ability to perform on their own at least 3 of the 4 Activities of Daily Living as given in the product definition: eating, bathing, clothing and functional mobility. We consider an observation period from 1998 to 2015 included. For each individual, the following information is available: date of birth, date of subscription, date of entry in LTC, cause of entry in LTC, date of death and gender. Causes of entry in LTC are clustered in 4 groups. The first group contains cases of dementia, mainly caused by Alzheimer disease. The second group gathers neurological disease, among them Parkinson disease as well as sclerosis. The third group contains cases of cancer: tumors and lymphocytes. Lastly, the fourth group gathers claims that do not fit in the previous 3 categories. Major causes for LTC that have not been mentioned previously include cardiovascular diseases such as infarction or stroke, muscular and skeletal diseases such as arthrosis or arthritis. Cases where the need for LTC arises from multiple pathologies also fall in this category. Other minor causes for LTC include accidents, depression, blindness, cirrhosis, HIV, diabetes. We therefore expect a lot of residual heterogeneity in this last group. The distribution of pathologies in claims can be found in Table 4.1. *Dementia* appears to be the most frequent cause of claim for both genders followed by *neurological diseases* and *other causes*. *Cancer* is less frequent but the amount of available data is still large with more than 900 uncensored claims for each gender.

Pathology	Number of claims		Censoring rate	
	male	female	male	female
Dementia	2,584	7,007	23.5 %	37.5 %
Neurological	1,362	2,194	19.9 %	30.3 %
Cancer	982	1,024	4.5 %	6.3 %
Other	1,207	2,752	17.1 %	28.2 %
Total	**6,135**	**12,977**	**18,9 %**	**33,3 %**

Table 4.1: Distribution of claims.

4.2.2 Notation

In this section we first introduce the 3-state illness-death model used to describe the LTC process. We then consider a multi-state process with p states of LTC corresponding to p different causes. The application is based on the previously introduced data which corresponds to the case $p = 4$. Finally we provide a way to derive the all-causes mortality in LTC in the former model from estimates of transitions intensities in the latter.

The illness-death model

For $x_0 \geqslant 0$, let us consider a continuous-time process $(Z_x)_{x \geqslant x_0}$ with values in the 3-state set $E = \{A, I, D\}$ of autonomy, LTC (or "illness"), death respectively. Let us assume that Z is *càd-làg* and that $Z_{x_0} = A$.

We now assume that $(Z_x)_{x \geqslant x_0}$ is a non-homogeneous semi-Markov process, that death is an absorbing state and that there is no transition allowed from LTC to autonomy. We introduce the transition intensities, also called instantaneous transition probabilities

$$\begin{aligned}
\mu_a(x) =& \lim_{h \to 0} \frac{1}{h} \mathbb{P}\left(Z_{x+h} = D | Z_x = A\right), \\
\lambda(x) =& \lim_{h \to 0} \frac{1}{h} \mathbb{P}\left(Z_{x+h} = I | Z_x = A\right), \\
\mu_i(x,t) =& \lim_{h \to 0} \frac{1}{h} \mathbb{P}\left(Z_{x+t+h} = D | Z_{x^-} = A, Z_x = I, Z_{x+t} = I\right).
\end{aligned}$$

Those intensities are called respectively intensity of incidence in LTC, intensity of autonomous mortality and intensity of mortality in LTC, with the latter intensity depending on both the age at onset of LTC (also called age at claim) and time spent in LTC. Figure 4.1 provides a representation of the model. The interested reader can refer to Biessy (2015a) for a more detailed study of the model.

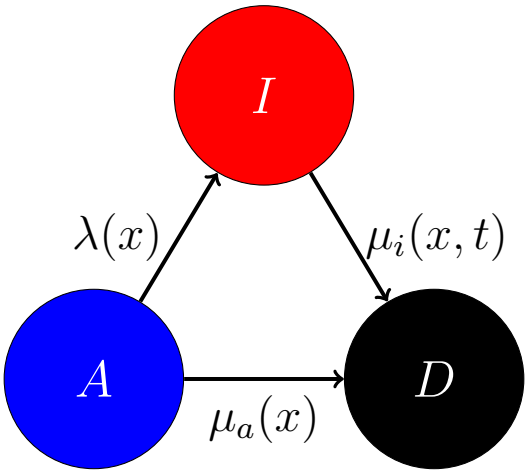


Figure 4.1: The 3 states continuous-time semi-Markov model with its transition intensities.

Multi-state model with pathologies

For $x_0 \geqslant 0$, let us consider another continuous-time process $(Z_x)_{x \geqslant x_0}$ with values in the discrete set $E = \{A, I_1, \ldots, I_p, D\}$ of autonomy, LTC (or "illness") with cause 1,. . . ,p at claim and death respectively. Let us assume that Z is *càd-làg* and that $Z_{x_0} = A$.

We now assume that $(Z_x)_{x \geqslant x_0}$ is a non-homogeneous semi-Markov process and consider that death is an absorbing state and that there is no transition allowed from LTC to autonomy. Besides, let us notice that each state of LTC corresponds to a different group for the cause

of claim. Therefore there is no transition between states of LTC. We introduce the transition intensities of the model

$$\begin{aligned}
\mu_a(x) =& \lim_{h\to 0} \frac{1}{h}\mathbb{P}\left(Z_{x+h} = D | Z_x = A\right),\\
\lambda_k(x) =& \lim_{h\to 0} \frac{1}{h}\mathbb{P}\left(Z_{x+h} = I_k | Z_x = A\right),\\
\mu_{i,k}(x,t) =& \lim_{h\to 0} \frac{1}{h}\mathbb{P}\left(Z_{x+t+h} = D | Z_{x^-} = A, Z_x = I_k, Z_{x+t} = I_k\right).
\end{aligned}$$

for $k \in \{1, \ldots, p\}$.

Those intensities are called respectively intensity of entry in LTC with cause k, intensity of autonomous mortality and intensity of mortality in LTC for cause k. A representation of the model is given by Figure 4.2.

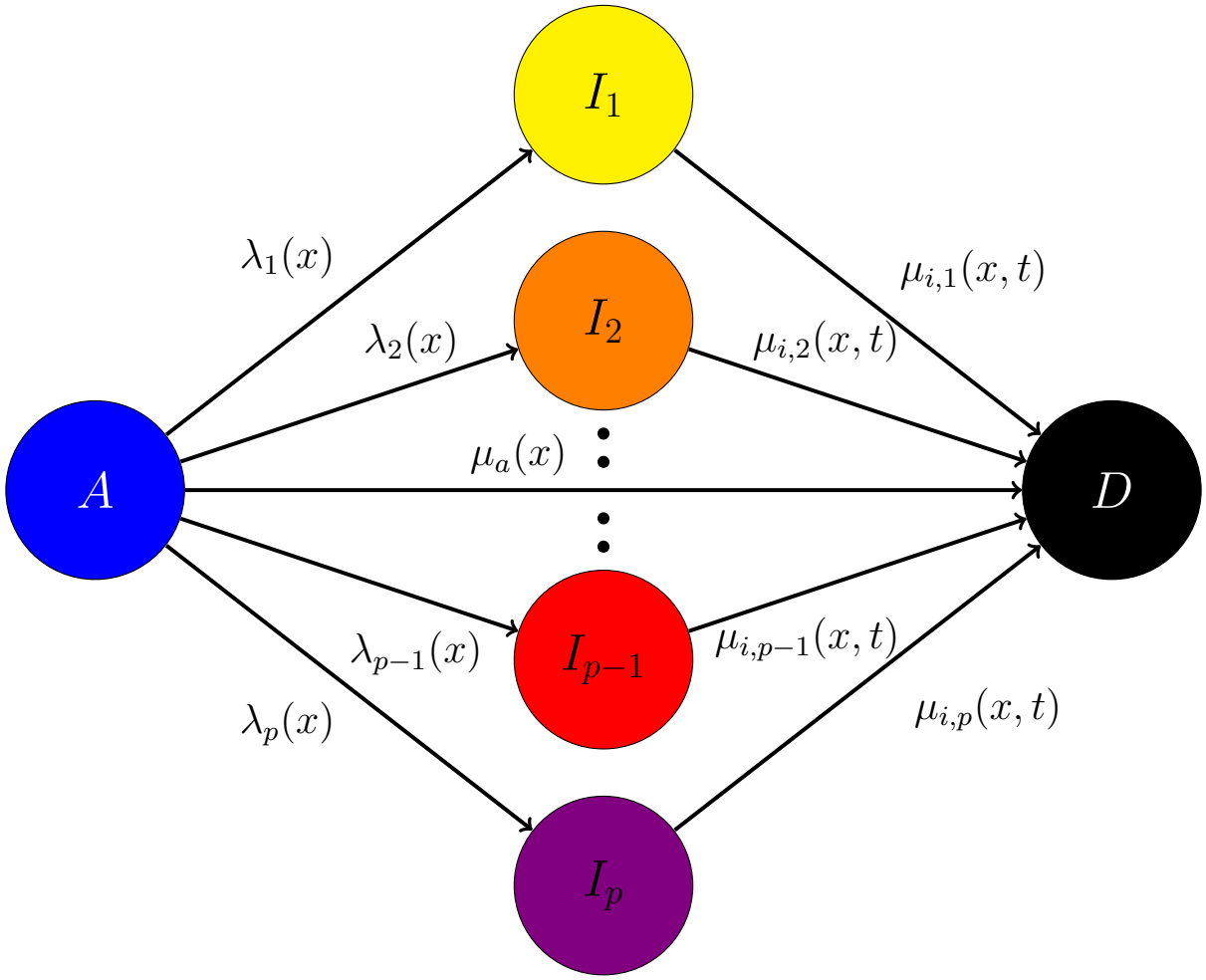


Figure 4.2: The continuous-time semi-Markov model with several groups of pathologies.

Link between both models

In this section, we introduce a relation which links the mortality in LTC of the previous 3-state model to intensities of transition of the multi-state model with pathologies.

Lemma 4. *Let us consider the two models defined above such that under the same notation* $I = I_1 \cup \cdots \cup I_p$ *and for* $(j,k) \in \{1, \ldots, p\}$ *such that* $j \neq k$, $I_j \cap I_k = \varnothing$. *Then to ensure the*

compatibility between both models, the following relations must be satisfied for all $x > x_0$, $t \geqslant 0$

$$\begin{aligned}\lambda(x) =& \sum_{k=1}^{p} \lambda_k(x) \\ \mu_i(x,t) =& \sum_{k=1}^{p} \eta_k(x,t)\mu_{i,k}(x,t),\end{aligned}$$

where for $x \geqslant x_0$, $t \geqslant 0$, $k \in \{1, \dots, p\}$

$$\eta_k(x,t) = \frac{\lambda_k(x) \exp\left(-\int_0^t \mu_{i,k}(x,u)\,du\right)}{\sum_{j=1}^{p} \lambda_j(x) \exp\left(-\int_0^t \mu_{i,j}(x,u)\,du\right)}$$

4.3 Local likelihood

In this section, we present the local likelihood methods that we use to build smooth estimates for each of the transition intensity in the previous models. In a nutshell, for a given function of interest and a given set of points, local likelihood approximates the function in the neighbourhood of each point by a different polynomial function. Coefficients are inferred using a localized version of the maximum likelihood estimator (MLE), which differs from the original maximum likelihood estimator through the addition of a kernel function whose value decreases with the distance to the fitting point. Local likelihood possesses some of the strengths associated with a parametric approach, while being more flexible. Local likelihood was initially introduced in Tibshirani and Hastie (1987) and extended to density estimation by Loader (1996). Chapter 7 of Loader (1999) briefly introduces the estimation of hazard rates using local likelihood, which is also implemented in the **locfit** package from the same author, on the statistical programming software **R** (R Core Team, 2016). While hazard rate estimation is very similar to density estimation, we believe that providing explicit formulas may prove very useful nonetheless.

4.3.1 Uni-dimensional local likelihood for right-censored left-truncated data

In this section we are interested in the estimation of a hazard rate function $\mu(x)$ which only depends on the current age x of the individual. In our model, $\mu(x)$ may be the autonomous mortality or the incidence rate in LTC. Each trajectory consists of x_i the age at which the individual enters the observation, y_i the age at which the individual leaves the observation and c_i an indicator of whether the trajectory is censored before the event of interest occurs (i.e. $c_i = 1$ in case of right-censoring and 0 otherwise). As in most life insurance portfolio studies, observations are left-truncated and right-censored. Let us notice that the **locfit** package does not handle left-truncated observations.

Principle of the estimation

For an age x and for u close to x, $\log \mu(u)$ may be locally approximated by a polynomial function i.e. for $d \in \mathbb{N}$ there exists $a = (a_0, a_1, \dots, a_d)^T$ such that

$$\begin{aligned}\log \mu(u) =& a_0 + a_1(u-x) + \cdots + a_d(u-x)^d + \mathrm{o}((u-x)^d) \\ =& \langle a, A(u-x) \rangle + \mathrm{o}((u-x)^d)\end{aligned}$$

where $A(u) = (1, u, \dots, u^d)^T$.

We first introduce the local log-likelihood function

$$\mathcal{L}_x(a) = \sum_{i=1}^{n}(1-c_i)W\left(\frac{y_i-x}{h(x)}\right)\langle a, A(y_i-x)\rangle - \int_{x_{min}}^{y_{max}} N(u)W\left(\frac{u-x}{h(x)}\right)e^{\langle a,A(u-x)\rangle}\mathrm{d}u \qquad (4.1)$$

where

$$N(u) = \sum_{i=1}^{n} \mathbf{1}\{x_i < u \leqslant y_i\}$$

is the number of individuals at risk at age u, W is a kernel function, h a bandwidth function, $x_{min} = \min_{i\in\{1,\dots,n\}} x_i$ and $y_{max} = \max_{i\in\{1,\dots,n\}} y_i$.

An estimate $\hat{a} = (\hat{a}_0, \dots, \hat{a}_d)^T$ of a is then obtained by maximizing the local likelihood function

$$(\hat{a}_0, \dots, \hat{a}_d)^T = \operatorname*{argmax}_a \mathcal{L}_x(a).$$

Finally an estimate $\hat{\mu}(x)$ of $\mu(x)$ is given by

$$\hat{\mu}(x) = \exp(\hat{a}_0)$$

Local likelihood methods require the setting of several parameters

- The degree d of the polynomial function fitted locally,
- The kernel function W used in the localized MLE,
- The bandwidth function h.

The degree of the polynomial function used for the fit has a significant impact on the estimation. Higher degrees provide more flexible fits, the downside being that the volatility of such fit is higher. Local constant fit corresponding to $d = 0$ usually leads to poor results in the tail of the distribution, while there is usually no reason to use fits of degree greater than 3. Compared to the other two components, the choice of the kernel has a rather limited impact on the estimation, as long as the kernel belongs to one of the common families (one can refer to Wand and Jones, 1995, for an introduction on those kernels functions). In what follows we use the Epanechikov kernel defined as

$$W(u) = \mathbf{1}\{|u| < 1\}\left(1-u^2\right)^2.$$

Epanechikov kernel function is optimal in the sense that any weight function producing the same asymptotic bias has larger asymptotic variance. One may refer to Epanechnikov (1969) for a proof of this result. At last, the choice of the bandwidth is of utmost importance and several methodologies exist. Therefore it deserves its own section.

Bandwidth choice

For an age x, the bandwidth function $h(x)$ determines the number of trajectories used in the estimation of $\mu(x)$. The most obvious choice of bandwidth is a constant bandwidth independent from x. This approach is however very limited, as intuitively in regions where data is scarce, a larger bandwidth should be selected to limit the variance of $\mu(x)$. At the other end of the scope of complexity, one may find more sophisticated adaptive bandwidths methods as the intersection of confidence intervals methodology presented in Chichignoud (2010) or methods which try to minimize a local validation criterion as presented in Chapter 10 of Loader (1999).

The bandwidth that we use in what follows is a slightly modified version of the nearest neighbours bandwidth presented in Chapter 2 of Loader (1999). In the initial method, for a

given α the bandwidth $h_\alpha(x)$ is obtained by sorting the vector of distances between x and each of the y_i by increasing order and selecting the j-th element of the vector where $j = \lceil n\alpha \rceil$. We bring a slight modification to this method by only considering distances from x to uncensored observations. This way we make sure that for low values of α, there is still a reasonable number of non-zero components in the left term of the localized likelihood.

Parameter selection

For each value of the degree of the fit d and for each fraction of nearest neighbours used for bandwidth selection α we have a local likelihood estimator $\hat{\mu}_{d,\alpha}$ of the hazard rate μ. In this section we introduce tools to compare those estimators and select adequate values for smoothing parameters.

We denote for $k \in \mathbb{N}$

$$M_k(x, d, \alpha) = -\int_{x_{min}}^{y_{max}} N(u) \left[W\left(\frac{u-x}{h_\alpha(x)} \right) \right]^k A(u-x) A(u-x)^T \exp\left(-\hat{\mu}_{d,\alpha}(u)\right) \mathrm{d}u.$$

While there is no guarantee that $M_k(x, d, \alpha)$ may be inverted from a theoretical point of view, in practice except for very low values of α it always proved to be the case.

We define the influence function

$$\mathrm{infl}_{d,\alpha}(x) = W(0) e_1^T M_1^{-1}(x, d, \alpha) e_1$$

as well as the degrees of freedom

$$\nu(d, \alpha) = \sum_{i=1}^{n} (1 - c_i) \mathrm{infl}_{d,\alpha}(y_i)$$

where for $j \in \{1, \ldots, d+1\}$, e_j is a vector of size $d+1$ with its j-th component equal to 1 and other components equal to 0.

For $i \in \{1, \ldots, n\}$, $(1 - c_i)\mathrm{infl}_{d,\alpha}(y_i)$ may be interpreted as the sensitivity of the estimate $\hat{\mu}_{d,\alpha}(y_i)$ to the event, observed or not, (y_i, c_i). Let us notice that while censored events do not contribute to the estimation, the remain of the trajectory still plays a role through the number of individuals at risk $N(u)$. The degrees of freedom measure the flexibility of the fit. They may therefore be seen as an equivalent of the number of parameters in parametric models.

We already have a measure of the quality of the fit through the log-likelihood function

$$l(\mu) = \sum_{i=1}^{n} (1 - c_i) \log \mu(y_i) - \int_{x_{min}}^{y_{max}} N(u) \mu(u) \mathrm{d}u$$

and a measure of the complexity through the degrees of freedom. Models with higher degrees of freedom are likely to better replicate the features observed in data, some of which may not correspond to real phenomena. The selected model should therefore provide a good compromise between quality of the fit and number of degrees of freedom. A natural approach for model selection is then to consider a linear combination of those two components which gives the Akaike information criterion (AIC)

$$\mathrm{AIC}(d, \alpha) = -2l(\hat{\mu}_{d,\alpha}) + 2\nu(d, \alpha).$$

Model selection is then performed by minimizing the AIC. For a given degree d, we may therefore compute the value of the AIC on a grid of values for α and represent the value of the AIC as a function of ν as recommendeds in Chapter 10 of Loader (1999). In cases where several models exhibit very close values of AIC, Loader (1999) recommend to select the model with the lowest degree of freedom.

Pseudo-residuals

A sizeable difficulty while working with hazard rates is that we do not actually observe the hazard rate. Thus defining residuals and running diagnoses may prove difficult. One way to solve this problem would be to build empirical estimates of hazard rate, for example by dividing the number of events during intervals of the form $[x+1)$ by the number of years lived by the portfolio over the same interval. However, such estimates are very volatile on the tail of the distribution, where few data is available. Instead, an alternative solution consists of replacing empirical values by the output of a fit with a much smaller bandwidth. Such a fit should be well defined, have a bias which could be neglected and have a correct performance on the tail of the distribution. A local constant fit is not suited for this purpose as performance in the tail is very poor. We suggest to use a local linear fit corresponding to $d = 1$ and select a bandwidth using the α nearest neighbours methodology with $\alpha = 0.05$.

We define Pearson residuals as follows

$$r(x) = \frac{\hat{\mu}_{d,\alpha}(x) - \hat{\mu}_{1,0.05}(x)}{\sqrt{\text{var}(\hat{\mu}_{1,0.05}(x))}}$$

where

$$\text{var}(\hat{\mu}_{d,\alpha}(x)) = e_1^T M_1^{-1}(x,d,\alpha) M_2(x,d,\alpha) M_1^{-1}(x,d,\alpha) e_1$$

Those residuals may help us assess the quality of the fit. A good fit should present residuals with uniformly distributed signs as well as absolute values inferior to 2 most of the time.

Computational aspects

This section briefly discusses the implementation of the local likelihood method. As the implementation is roughly the same as in the **locfit** package in **R**, the interested reader will find a more detailed discussion in Chapter 12 of Loader (1999).

Estimating $\mu(x)$ for some x using local likelihood is computationally expensive. Indeed it requires to find the maximum of the localized likelihood function. To do so, one often searches for the point where the first order derivative is equal to 0. Except for the local constant fit ($d = 0$) for which a closed formula yields the solution, one has to rely on numerical methods. Our implementation is based on Newton-Raphson algorithm. At each step of the algorithm, the likelihood and its first order and second order derivatives are computed, which requires numerical approximation of the integral. To compute the likelihood of equation 4.1, one has to apply local likelihood as many times as the number of uncensored observations in the data, which proves very difficult to manage in a reasonable amount of time. Instead, we only use the local likelihood on a limited number of carefully chosen points, and then use an interpolation method to approximate $\mu(x)$ at any point of interest.

To define the points where we actually use local likelihood, we grow an adaptive tree using the method described in Chapter 12 of Loader (1999). We start with two boundaries that include all the observations, such as x_{min} and y_{max} then we create new points by splitting in half the intervals between two consecutive points. In each case, the point $(x+y)/2$ is added to the list if $|y - x| > c \min(h(x), h(y))$ where c is a tuning parameter. After some back-testing on the data we decide to set c at 0.01.

Once we have our grid of points, we implement Newton-Raphson algorithm to maximise the localized likelihood. First and second order of the derivatives are given by the formulas

$$\frac{\partial \mathcal{L}_x}{\partial a}(a) = \sum_{i=1}^{n} (1 - c_i) W\left(\frac{y_i - x}{h(x)}\right) A(y_i - x) - \int_{x_{min}}^{y_{max}} N(u) W\left(\frac{u - x}{h(x)}\right) A(u - x) e^{\langle a, A(u-x)\rangle} \mathrm{d}u$$

and

$$\frac{\partial^2 \mathcal{L}_x}{\partial a^2}(a) = -\int_{x_{min}}^{y_{max}} N(u) W\left(\frac{u-x}{h(x)}\right) A(u-x) A(u-x)^T e^{\langle a, A(u-x)\rangle} \mathrm{d}u.$$

At each step of the algorithm, we compute the localized likelihood and its first and second order derivatives. In Newton-Raphson algorithm, the next-step estimate $k+1$ is obtained from the estimate at step k using the following formula

$$a^{(k+1)} = a^{(k)} - \left(\frac{\partial^2 \mathcal{L}_x}{\partial a^2}(a)\right)^{-1} \frac{\partial \mathcal{L}_x}{\partial a}(a).$$

To ensure convergence, we instead use a so-called damped version of the original algorithm

$$a^{(k+1)} = a^{(k)} - \frac{1}{2^j}\left(\frac{\partial^2 \mathcal{L}_x}{\partial a^2}(a)\right)^{-1} \frac{\partial \mathcal{L}_x}{\partial a}(a)$$

where j is the smallest natural number such that

$$\mathcal{L}_x(a^{(k)}) \leqslant \mathcal{L}_x\left(a^{(k)} - \frac{1}{2^j}\left(\frac{\partial^2 \mathcal{L}_x}{\partial a^2}(a)\right)^{-1} \frac{\partial \mathcal{L}_x}{\partial a}(a)\right).$$

Finally, the initial value of the parameter vector is set as the solution for the constant local fit

$$a^{(0)} = \frac{\sum_{i=1}^{n}(1-c_i) W\left(\frac{y_i - x}{h(x)}\right)}{\int_{x_{min}}^{y_{max}} N(u) W\left(\frac{u-x}{h(x)}\right) \mathrm{d}u} e_1.$$

For the interpolation, we rely on a cubic interpolation method. For two fitting points v_0 and v_1, the interpolated value of $f(x)$ for $v_0 \leqslant x \leqslant v_1$ is given by

$$f(x) = \phi_0(\lambda) f(v_0) + \phi_1(\lambda) f(v_1) + (v_1 - v_0)\left[\psi_0(\lambda) f'(v_0) + \psi_1(\lambda) f'(v_1)\right]$$

where

$$\begin{aligned}
\lambda &= \frac{x - v_0}{v_1 - v_0} \\
\phi_0(\lambda) &= (1-\lambda)^2(1+2\lambda) \\
\phi_1(\lambda) &= \lambda^2(3-2\lambda) \\
\psi_0(\lambda) &= \lambda(1-\lambda)^2 \\
\psi_1(\lambda) &= \lambda^2(1-\lambda).
\end{aligned}$$

Interpolation is used not only for the transition intensity $\hat{\mu}$ but also for the influence function $\mathrm{infl}(x)$ and the variance of the estimate $\mathrm{var}(\hat{\mu}(x))$. Derivative of those functions may be approximated by their local slopes

$$\begin{aligned}
\hat{\mu}'(x) &\simeq \hat{a}_1 \exp(\hat{a}_0) \\
\mathrm{infl}'(x) &\simeq W(0) e_1^T M_1^{-1}(x) e_2 \\
\mathrm{var}'(\hat{\mu}(x)) &\simeq \hat{\mu}^2(x) e_1^T M_1^{-1}(x) M_2(x) M_1^{-1}(x) e_2
\end{aligned}$$

Let us notice that linear interpolation should still be used in the case of local constant fitting ($d = 0$) as derivatives are not available in that case.

4.3.2 Bi-dimensional local likelihood for right-censored data

In this section we focus on the mortality in LTC which depends on both the age at claim inception x and the time already spent in the LTC state t. As in the uni-dimensional case, each trajectory consists of x_i the age at which the individual enters the observation, y_i the age at which the individual leaves the observation and c_i an indicator of whether the trajectory is censored before the event of interest occurs. We also denote by $t_i = y_i - x_i$ the time spent in LTC when the individual leaves the observation sample.

For the sake of clarity, let us focus on the log-quadratic case which corresponds to $d = 2$. For each couple (x, t) and for (u, v) close to (x, t), $\log \mu(u, v)$ may be locally approximated by a polynomial function i.e. for $d \in \mathbb{N}$ there is $a = (a_0, a_1, \ldots, a_5)^T$ such that

$$\begin{aligned}\log \mu(u,v) =& a_0 + a_1(u-x) + a_2(v-t) + a_3(u-x)^2 + a_4(u-x)(v-t) \\ &+ a_5(v-t)^2 + \mathrm{o}((u-x)^2 + (v-t)^2) \\ =& \langle a, A(u-x, v-t)\rangle + \mathrm{o}((u-x)^2 + (v-t)^2)\end{aligned}$$

where $A(u, v) = (1, u, v, u^2, uv, v^2)^T$.

We first introduce the local log-likelihood function

$$\begin{aligned}\mathcal{L}_{x,t}(a) =& \sum_{i=1}^{n} (1 - c_i)\, W\left(\frac{\rho(x_i - x, t_i - t)}{h(x,t)}\right) \langle a, A(x_i - x, t_i - t)\rangle \\ &- \sum_{i=1}^{n} \int_0^{t_i} W\left(\frac{\rho(x_i - x, u - t)}{h(x,t)}\right) e^{\langle a, A(x_i - x, u - t)\rangle} \mathrm{d}u\end{aligned}$$

where ρ is a distance function, W is a kernel function and h is a bandwidth function.

An estimate $\hat{a} = (\hat{a}_0, \ldots, \hat{a}_5)^T$ of a is then obtained by maximizing the local likelihood function so that

$$(\hat{a}_0, \ldots, \hat{a}_5)^T = \underset{a}{\operatorname{argmax}}\ \mathcal{L}_{t,x}(a).$$

Finally an estimate $\hat{\mu}(x, t)$ of $\mu(x, t)$ is given by

$$\hat{\mu}(x, t) = \exp(\hat{a}_0)$$

There are a few noticeable differences with the uni-dimensional case. As h depends on both the age at claim inception x and the time spent in LTC t, we have to define the distance between two points (x, t) and (u, v). A natural choice would be the euclidean distance. However, both coordinates do not play a similar role and different weights should be considered. We normalize each coordinate by dividing it by the empirical standard deviation for this component observed on the sample and define the distance ρ as

$$\rho(u - x, v - t) = \sqrt{\left(\frac{u-x}{\hat{\sigma}(X)}\right)^2 + \left(\frac{v-t}{\hat{\sigma}(T)}\right)^2}$$

where $X = \{x_1, \ldots, x_n\}$, $T = \{t_1, \ldots, t_n\}$ and $\hat{\sigma}$ is the estimated standard deviation on the database.

Another sizeable difference is of computational nature. In the uni-dimensional case, it was possible to account for the number of individuals at risk through the addition of $N(u)$ and compute numerically a single integral. In the bi-dimensional case, age at claim inception impacts mortality and we must compute one integral for each trajectory. Regarding other aspects, the bi-dimensional case is very similar to the unidimensional case. Results obtained in the next two sections are based on the implementation of local likelihood from the **locfit** package.

Finally, as in the uni-dimensional case, we define pseudo-residuals thanks to a reference. In the bi-dimensional case, we decide to use a local constant fit and consider the distance to the 100-th nearest neighbour as the local bandwidth, which corresponds to $d = 0$ and $\alpha = 100/n$. This choice of bandwidth, while providing a fit with a low amount of smoothing, ensures that there is enough observations within the bandwidth to create regular mortality surfaces and avoid visual disturbances.

4.4 Inference of transition probabilities

In this section, we apply the local likelihood methods defined previously to infer the transition probabilities in the model. Autonomous mortality, incidence in LTC for each cause and overall incidence in LTC are inferred using the uni-dimensional method with our own implementation of the algorithm while mortality in LTC for each cause and overall mortality in LTC is computed using the algorithm implemented in the **locfit** package on **R**.

4.4.1 Autonomous mortality

Figure 4.3 represents the intensity of autonomous mortality smoothed using local likelihood with a logarithmic scale for ages between 60 and 95. Information about the selected smoothing parameters may be found in Table 4.2. The autonomous mortality is constantly higher for males than females and for both genders it increases exponentially with respect to age. Fits of degrees 2 and 3 with a large bandwidth were selected which indicates that the shape of the curve is quite complex but also regular. Figure 4.4 represents the residuals for the fit. Except at the border of the graphs where data is scarce, residuals have uniformly distributed signs and their absolute value rarely exceeds 2. From a graphical point of view, the fit does not appear to over-smooth the data.

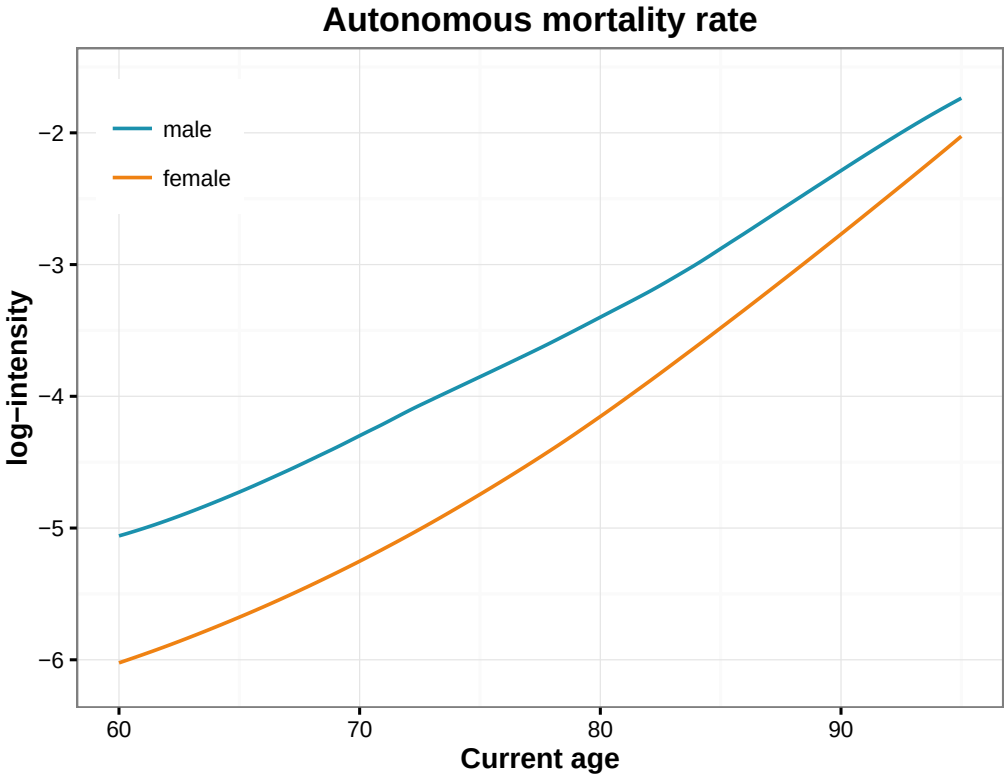


Figure 4.3: Intensity of autonomous mortality estimated using local likelihood: logarithmic scale.

Gender	d	α	ν	AIC
male	3	0.85	6.76	155,532.4
female	2	0.95	4.36	181,422.4

Table 4.2: Smoothing parameters selected for autonomous mortality.

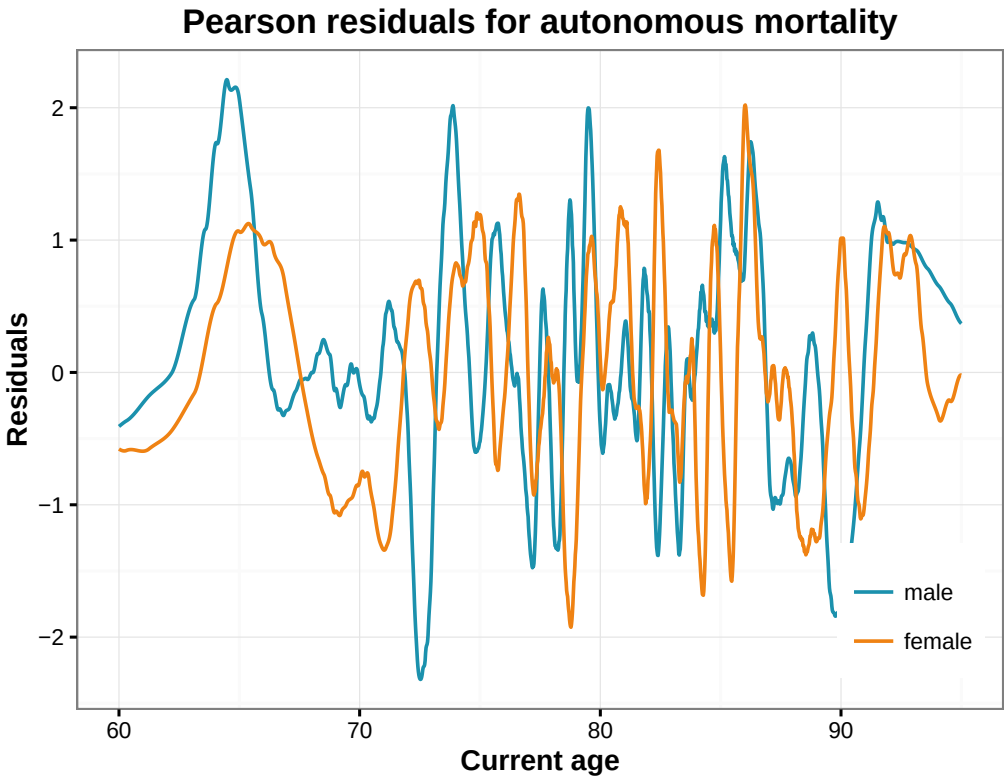


Figure 4.4: Pearson residuals for autonomous mortality.

4.4.2 Overall incidence in LTC

Figure 4.5 represents the intensity of incidence in LTC smoothed using local likelihood with a logarithmic scale for ages between 60 and 95 as well. Information about the selected smoothing parameters may be found in Table 4.3. Incidence in LTC is initially higher for males than for females but the situation is reversed for ages above 82. Degree 3 was selected for both genders, with an average bandwidth for males, which results in a complex curve with many degrees of freedom, while for females a large bandwidth was selected and the incidence curve appears very smooth. Figure 4.8 represents the residuals for the fit. Once again, except at the borders where data is scarce, residuals have uniformly distributed signs and their absolute value rarely go above 2.

Gender	d	α	ν	AIC
male	3	0.55	9.67	60,036.5
female	3	1	4.82	121,155.5

Table 4.3: Smoothing parameters selected for overall incidence in LTC.

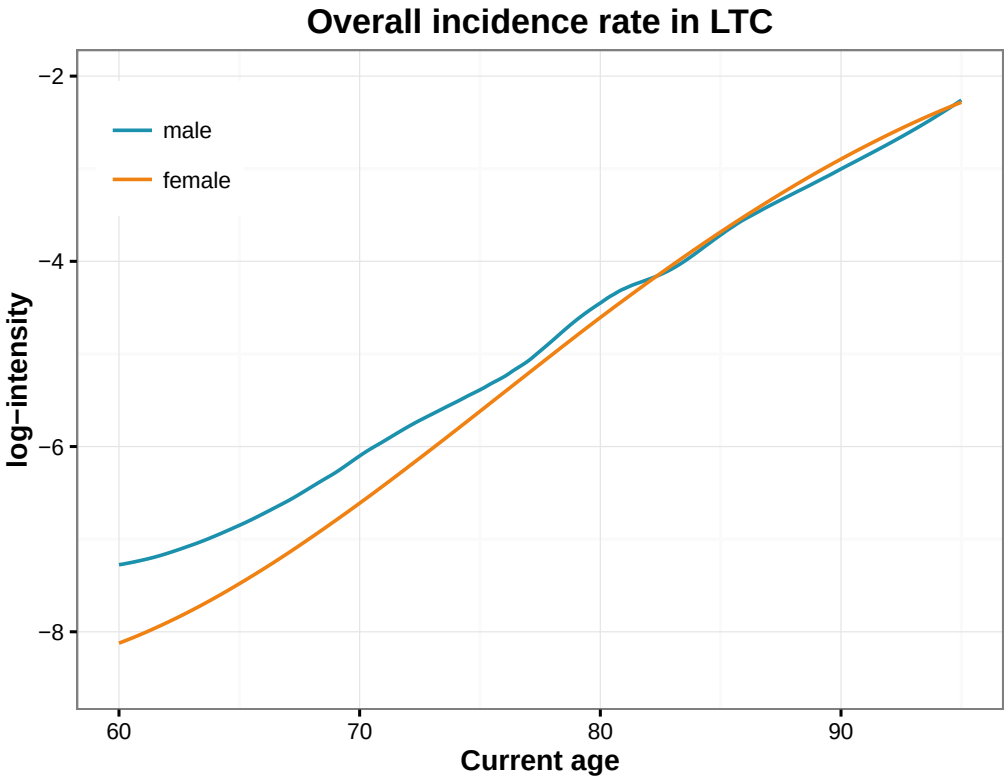


Figure 4.5: Intensity of overall incidence in LTC, estimated using local likelihood: logarithmic scale.

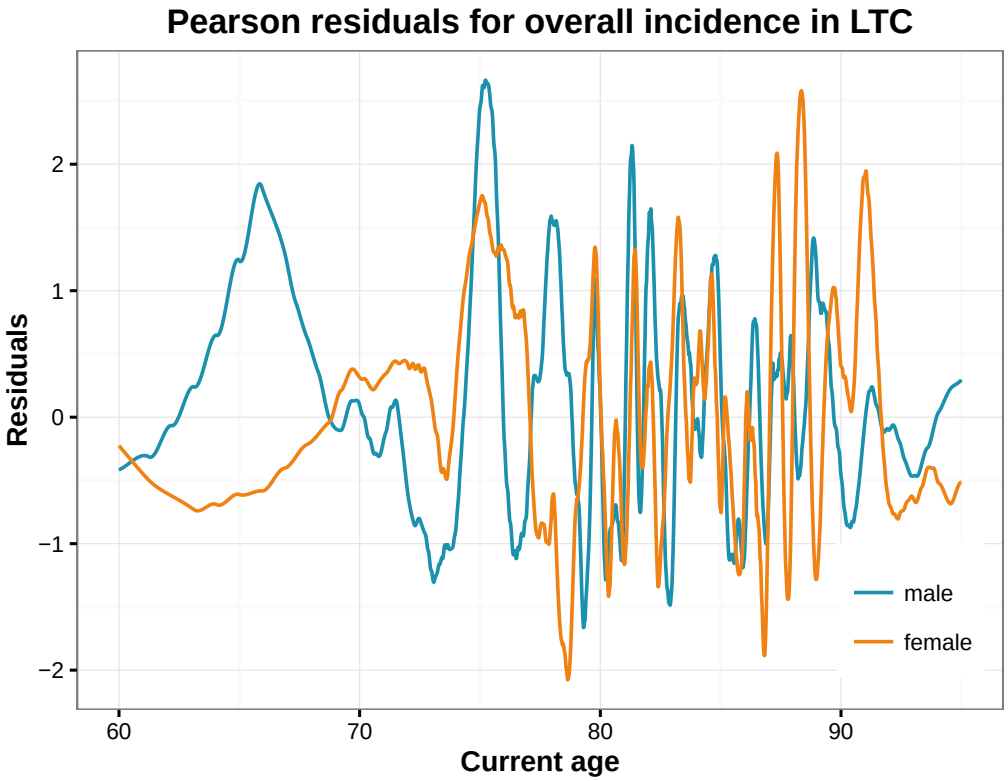


Figure 4.6: Pearson residuals for overall incidence in LTC.

4.4.3 Incidence in LTC by group

Figure 4.7 represents the intensity of incidence in LTC for each of the 4 groups of pathologies present in the data. Information about the selected smoothing parameters may be found in Table 4.4. For both genders, *cancer* is the most frequent disease at age 60, while *dementia* and *other diseases* have the higher incidence rates from age 80 onward. Males exhibit higher incidence rates for *cancer* and *neurological diseases*. Incidence rates for *dementia* and *other diseases* are lower for females than males at age 60 but higher from age 80 onward. Figure 4.6 represents the residuals for the fit. Once again, except for lower ages where data is scarce, signs are uniformly distributed and absolute value of the residuals rarely go above 2.

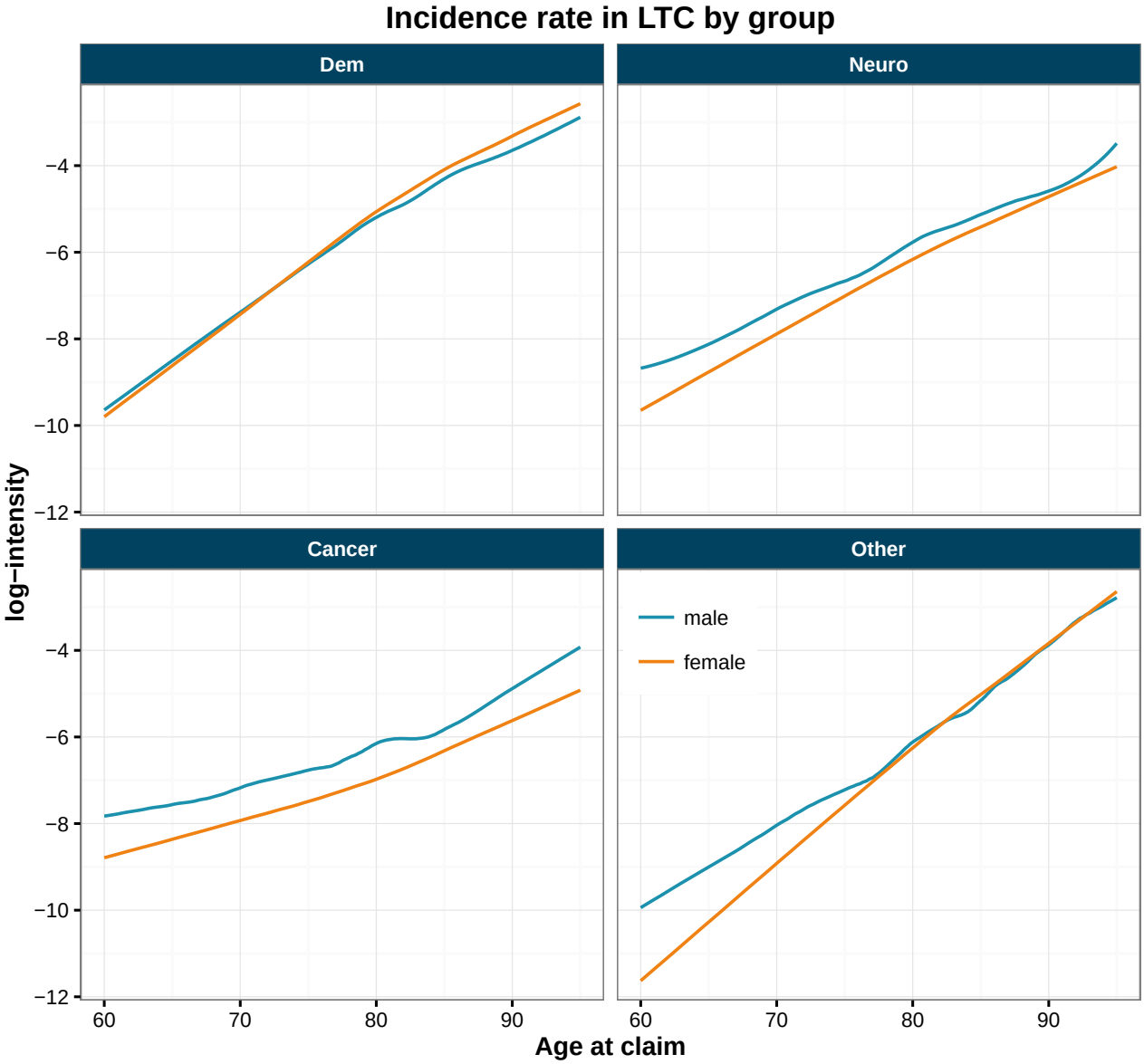


Figure 4.7: Intensity of incidence in LTC by group, estimated using local likelihood: logarithmic scale.

Gender	Group	d	α	ν	AIC
male	dementia	1	0.4	6.33	32,468.2
	neurological	3	0.65	8.85	19,576.1
	cancer	1	0.35	6.92	14,792.0
	other	1	0.25	8.96	17,110.1
female	dementia	1	0.5	5.30	82,897.1
	neurological	1	0.9	3.09	32,077.8
	cancer	1	0.85	3.26	17,077.3
	other	1	0.8	3.51	37,620.1

Table 4.4: Smoothing parameters selected for incidence in LTC by group.

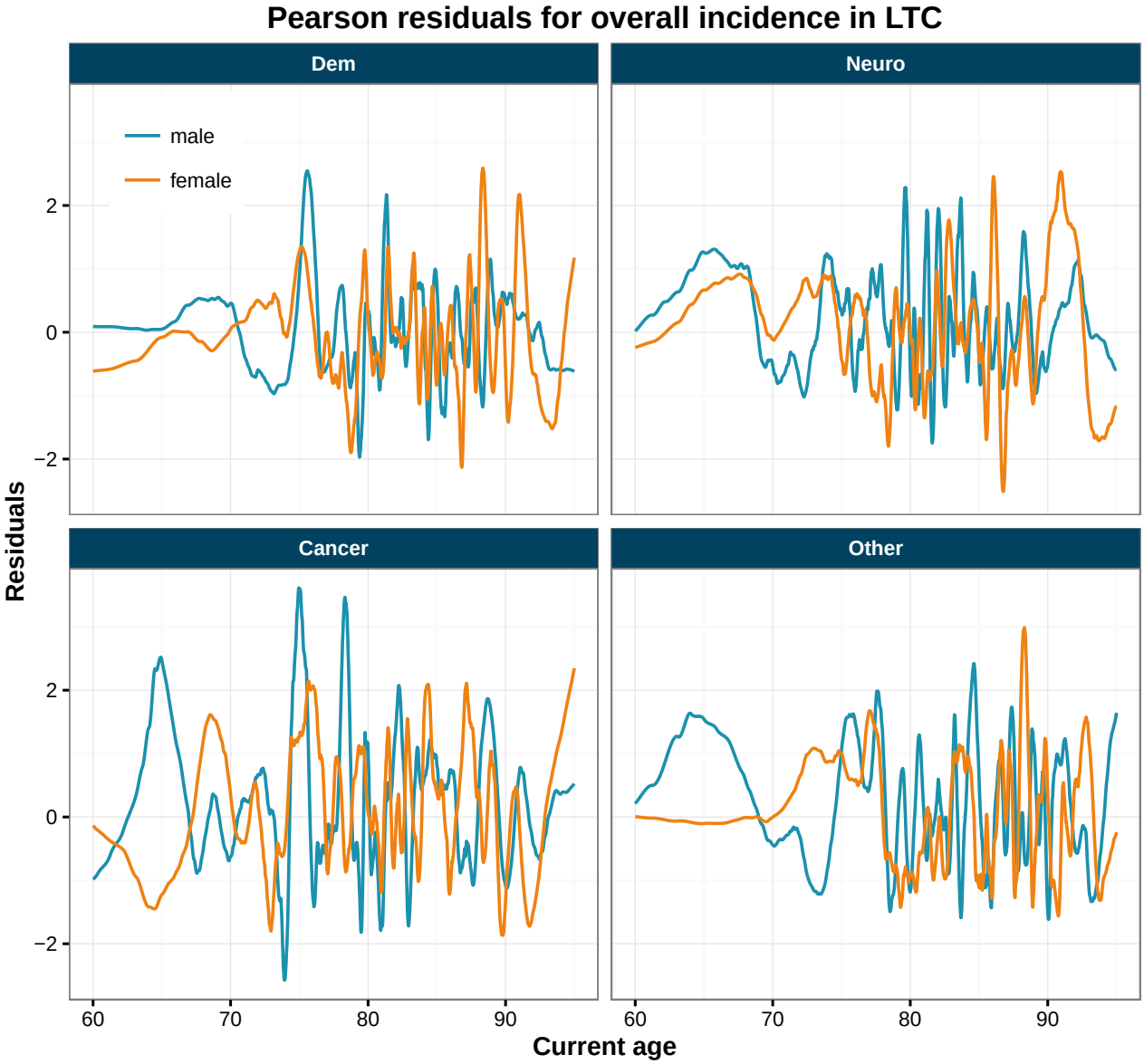


Figure 4.8: Pearson residuals for incidence in LTC by group.

4.4.4 Overall mortality in LTC

Figure 4.9 represents the mortality surface obtained by using the reference fit defined previously for x between 70 and 95 and for durations t lower than 8 years.

Mortality surfaces for males and females appear to present very similar features despite the level of mortality being constantly lower for females than for males. The initial level of mortality at claim inception is very high, and the mortality decreases very quickly with respect to duration t in LTC then increases slowly for higher durations. The impact of age at claim x on mortality seems non trivial.

Figure 4.10 represents the fit obtained using selected optimal parameters that may be found in Table 4.5. Compared to the reference fit, the mortality surfaces appear over-smoothed, especially for low durations. This is confirmed by Figure 4.11 which represents pseudo-residuals associated with the optimal fit. Indeed, those residuals show a lot of structure for low durations. The initial mortality appears to have been underestimated for durations below 1 and overestimated for durations between 1 and 2.

Gender	d	α	ν	AIC
male	1	0.6	7.71	6,801.3
female	1	0.5	8.80	22,502.5

Table 4.5: Smoothing parameters selected for overall mortality in LTC.

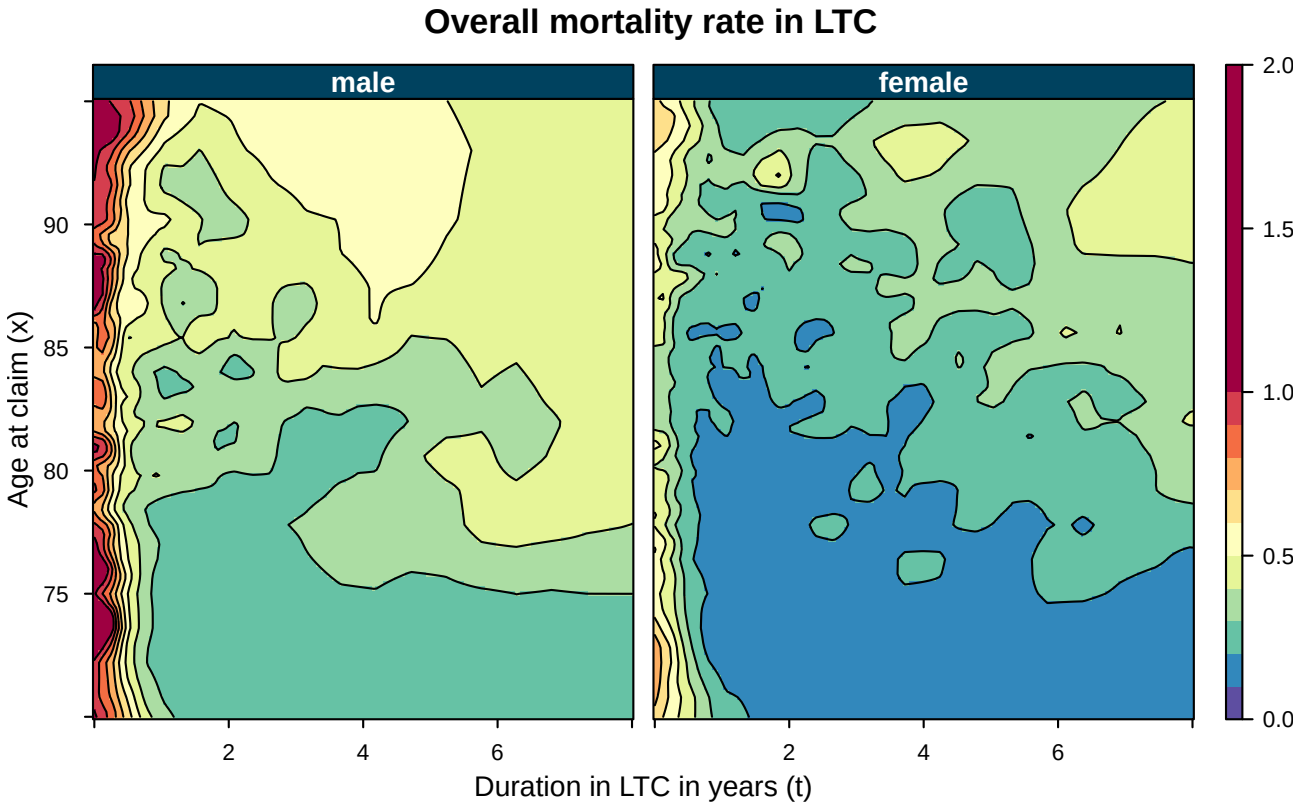


Figure 4.9: Intensity of mortality in LTC obtained by using constant local fitting and distance to the 100-th nearest neighbours as bandwidth.

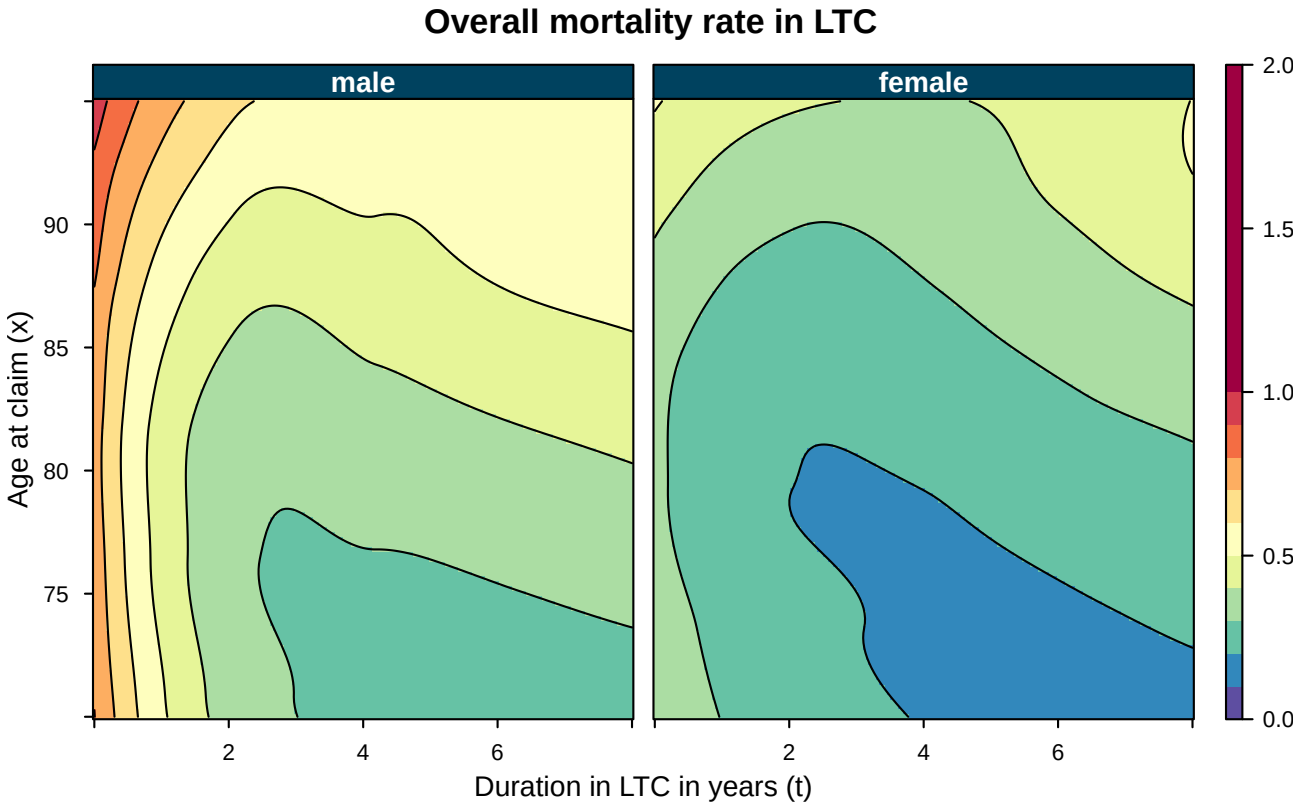


Figure 4.10: Intensity of mortality in LTC obtained using local likelihood with optimal smoothing parameters.

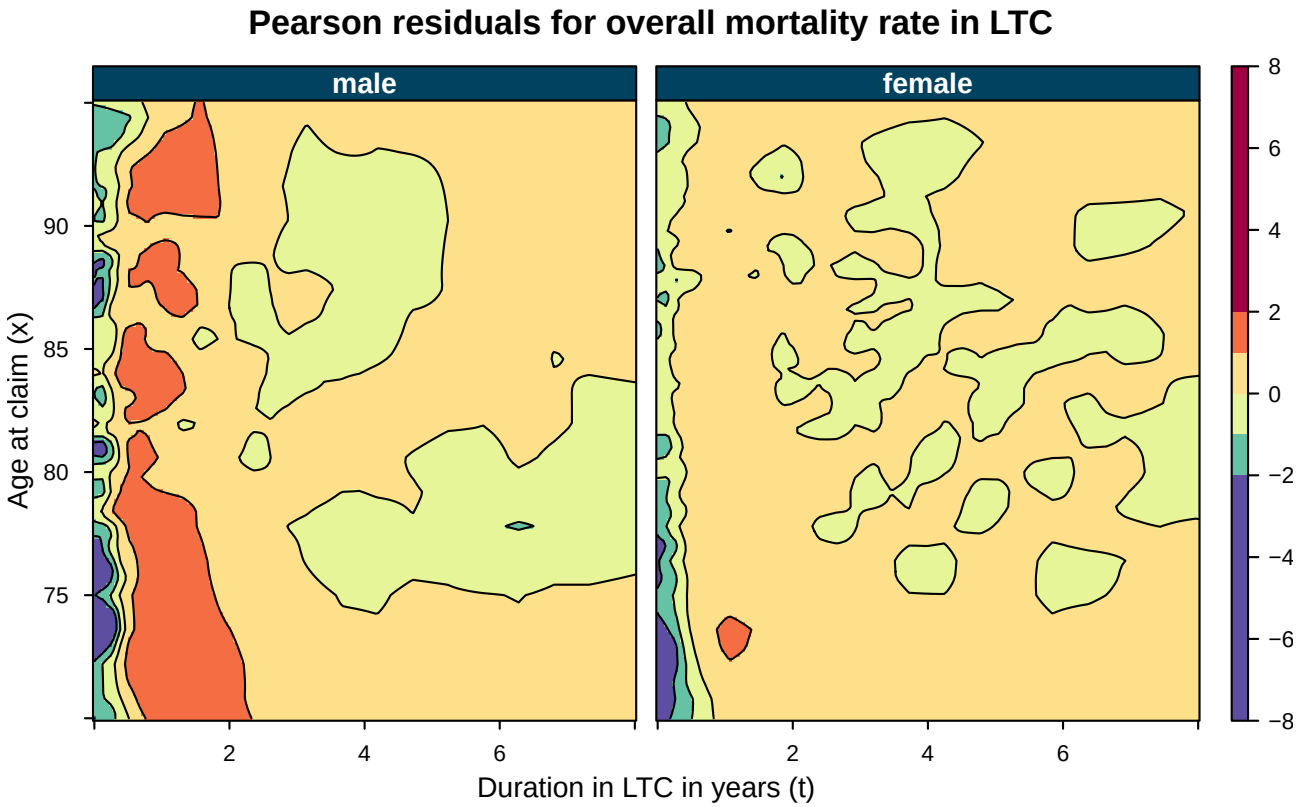


Figure 4.11: Pearson residuals associated with overall mortality in LTC.

4.4.5 Mortality in LTC by group

Figure 4.12 represents the surface of mortality in LTC for each group of pathology, obtained by applying a local constant fit with the 100-th nearest neighbours bandwidth. We observe significant differences between groups. First of all, mortality for *cancer* is extremely high during the first few months following claim inception. It is also the only group of pathology for which mortality does not increase with respect to age at claim. For *dementia*, mortality increases with respect to both age at claim and duration, while for the two remaining causes, mortality increases with age at claim but decreases with respect to duration. Let us notice that for other causes, especially for males, we observe a very high mortality for ages at claim inception beyond 85. Such decline may result from heterogeneity in the data. Similarly, we may believe that there is still heterogeneity in the category *other causes*. Our guess would be that infarction and stroke are responsible for the increased mortality near claim inception, which could explain why higher ages and males are mostly concerned. At last, for higher duration, mortality for *neurological diseases* and *other causes* starts to increase again which may correspond to the impact of the ageing process.

Gender	Group	d	α	ν	AIC
male	dementia	1	0.95	4.55	4,089.4
	neurological	1	0.75	6.26	2,397.5
	cancer	0	0.15	13.27	- 2,150.1
	other	1	0.45	9.30	1,013.3
female	dementia	1	0.9	5.06	13,300.2
	neurological	1	0.5	8.72	4,317.1
	cancer	0	0.15	13.00	- 1,679.5
	other	1	0.5	8.69	4,299.1

Table 4.6: Smoothing parameters selected for mortality in LTC by group.

Figure 4.13 represents the surface of mortality for each group obtained using selected optimal parameters that may be found in Table 4.6. Features of the initial surfaces seem to have been very well preserved. Fits of degree 1 have been selected as well as very large bandwidth except for the *cancer* group. Mortality surface obtained therefore appear very regular. Figure 4.19 represents the residuals for the fit. Those residuals do not present a lot of structure, except once again for *cancer*.

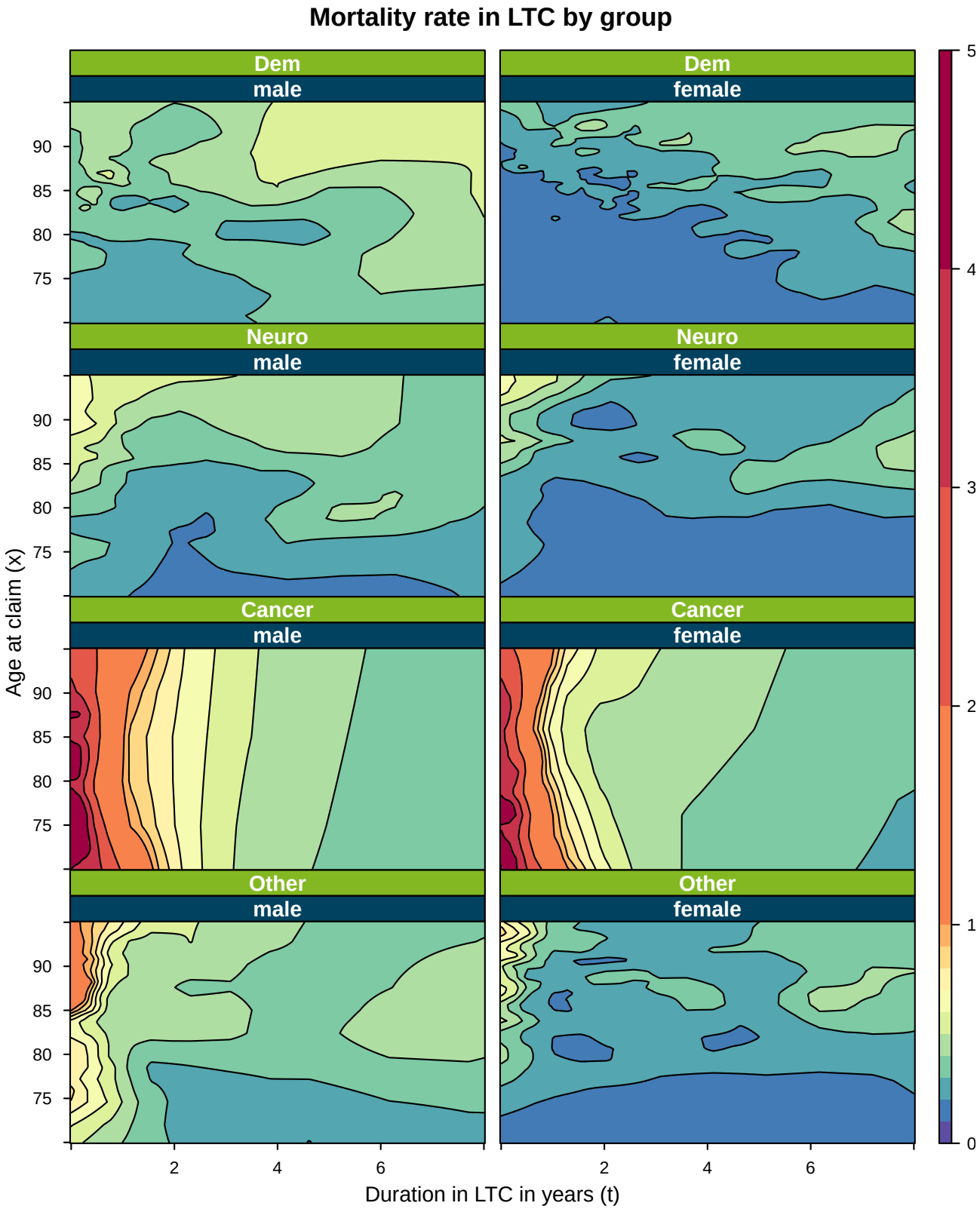


Figure 4.12: Cause-specific intensity of mortality in LTC obtained by using constant local fitting and distance to the 100-th nearest neighbours as bandwidth.

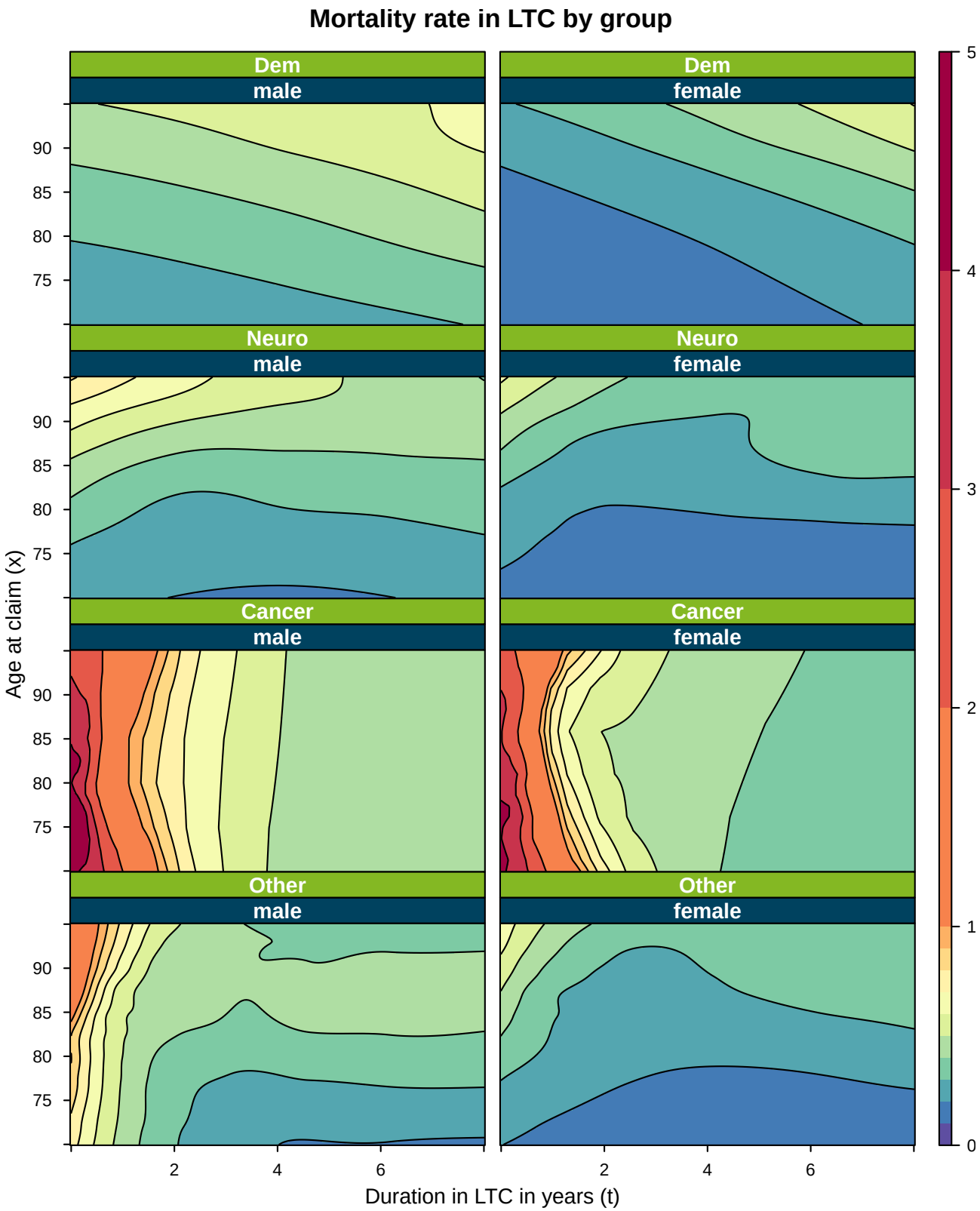


Figure 4.13: Cause-specific intensity of mortality in LTC obtained using local likelihood with optimal smoothing parameters.

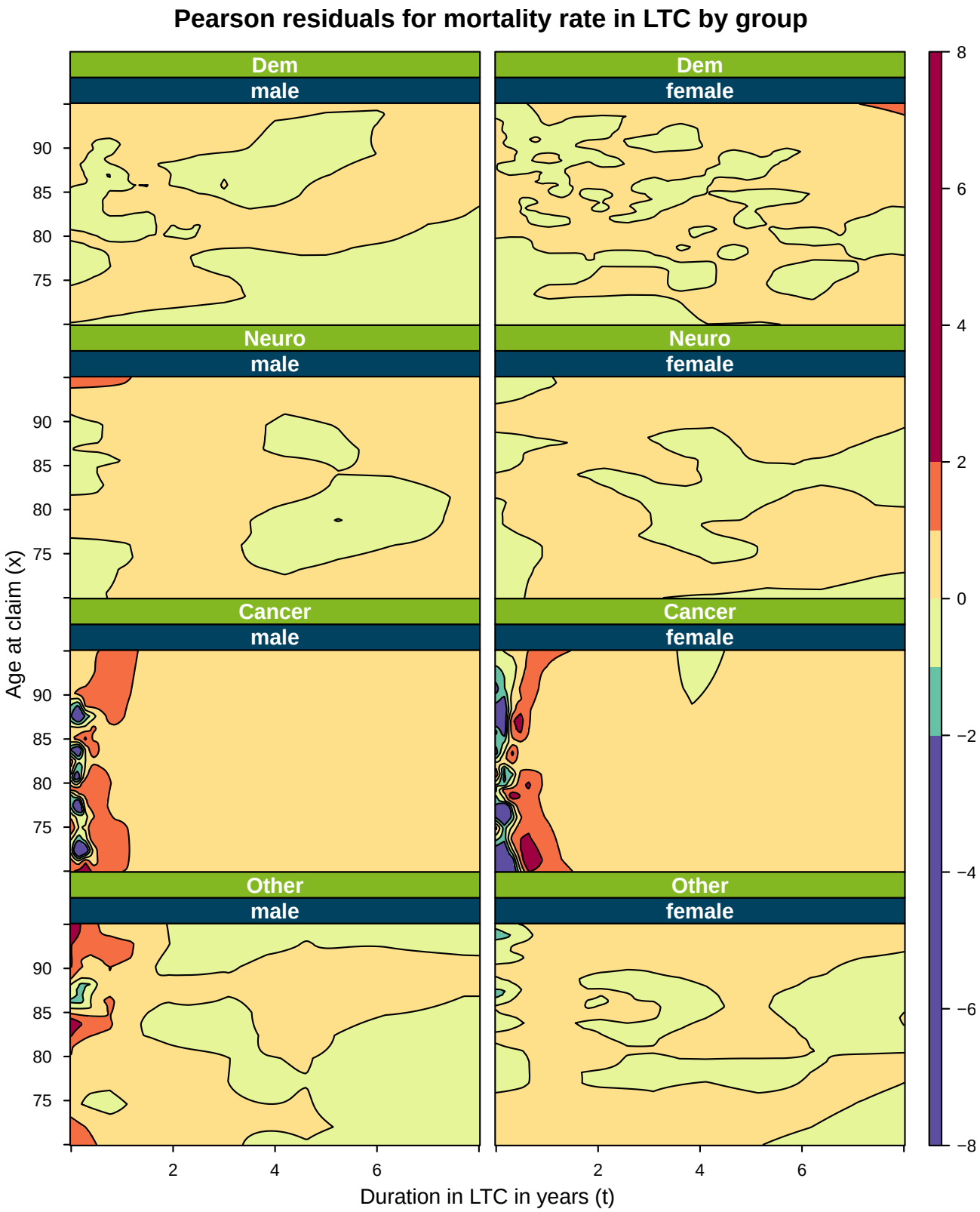


Figure 4.14: Pearson residuals associated with mortality in LTC by group.

4.5 Consequences for the LTC insurer

Using local likelihood, we were able to estimate the different transition intensities for both the illness-death model of Figure 4.1 and the multi-state model with several groups of pathology of Figure 4.2. In this section we first derive results relative to each group of pathologies and then use Lemma 1 to compute a second-step estimate of overall mortality in LTC. Lastly we look at the weight of each group in the disabled population and the contribution of those groups to overall mortality in LTC.

4.5.1 Results relative to individual groups of pathologies

Let us set $x_0 = 50$ and introduce the following notations

$$A(x) = \exp\left(-\int_{x_0}^{x} [\mu_a(u) + \lambda(u)]\,\mathrm{d}u\right),$$

$$I_{x,k}(t) = \exp\left(-\int_{0}^{t} \mu_{i,k}(x,u)\mathrm{d}u\right)$$

corresponding to the probability of remaining autonomous between ages x_0 and x and the probability of surviving while in LTC for an age of entry x and a pathology of group k between durations 0 and t, for $x_0 \leqslant x$, $0 \leqslant t$ and $k \in \{1, \ldots, p\}$.

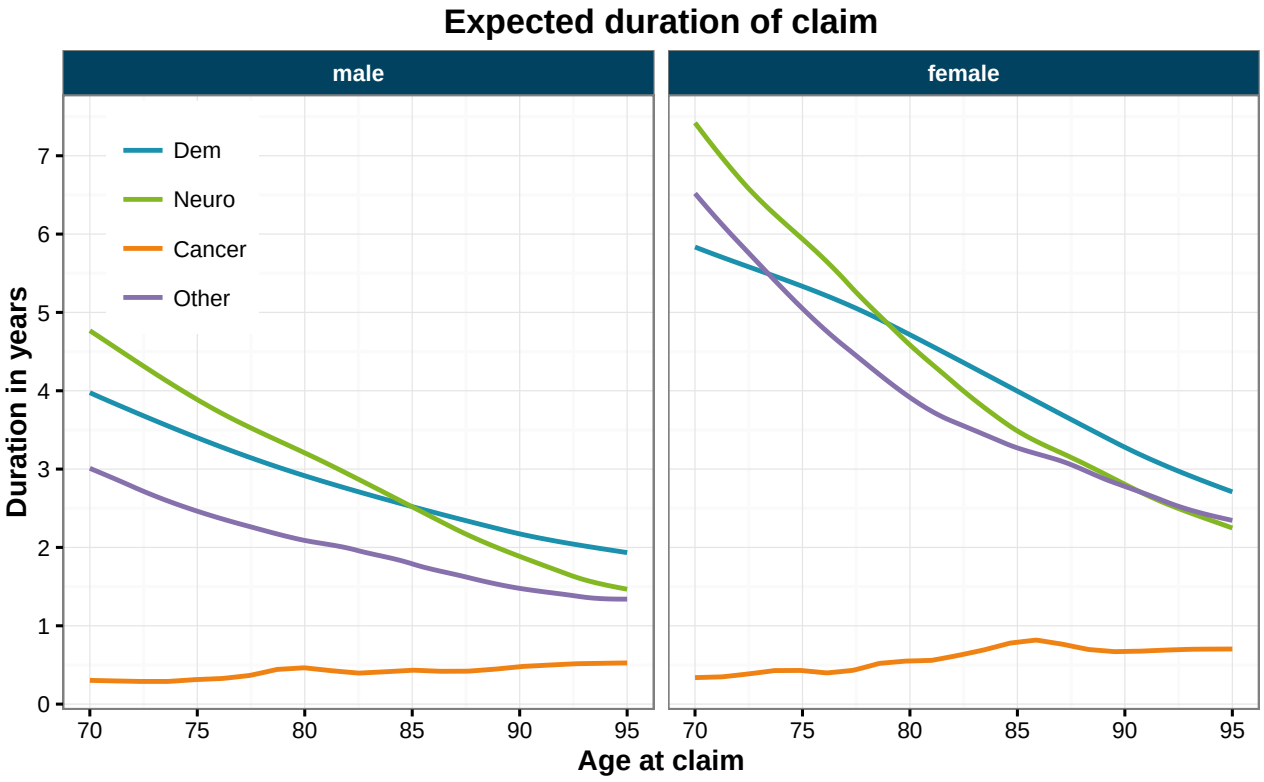


Figure 4.15: Expected duration of claim by age at claim inception, gender and pathology group.

Figure 4.15 represents the expected duration of claims $\mathrm{ED}_k(x)$ according to age at claim, gender and pathology group

$$\mathrm{ED}_k(x) = \int_{0}^{\infty} I_{x,k}(u)\mathrm{d}u.$$

Cancer is an outlier, with an associated duration of claim far below the level of other groups. It is also the only group of pathology for which expected duration of claim increases with respect to age at claim, when other groups exhibit a sharp decrease. Expected duration of claim is constantly lower for males than females for every group of pathology no matter the age at claim inception.

Using autonomous mortality in addition to the specific incidence and mortality rates by group, we are able to project the evolution of a population of initially autonomous insured lives from age 50 onwards.

Figure 4.16 represents the number of open claims $\mathrm{NC}_k(x)$ as a percentage of the initial autonomous population at age 50

$$\mathrm{NC}_k(x) = \int_{x_0}^{x} \lambda_k(u) A(u) I_{u,k}(x-u) \mathrm{d}u.$$

For an insurer, the area under each curve directly relates to the cost associated with each pathology, altough the actual cost also depends on the technical rate used by the insurer. For both genders, cost related to *cancer* may be neglected. *Dementia* is the more expensive cause for both males and females followed by *other causes*. Finally *neurological diseases* correspond to a higher fraction of total claim cost in males than in females. Maximum amount of open claims is reached at age 90 for males and 92 for females.

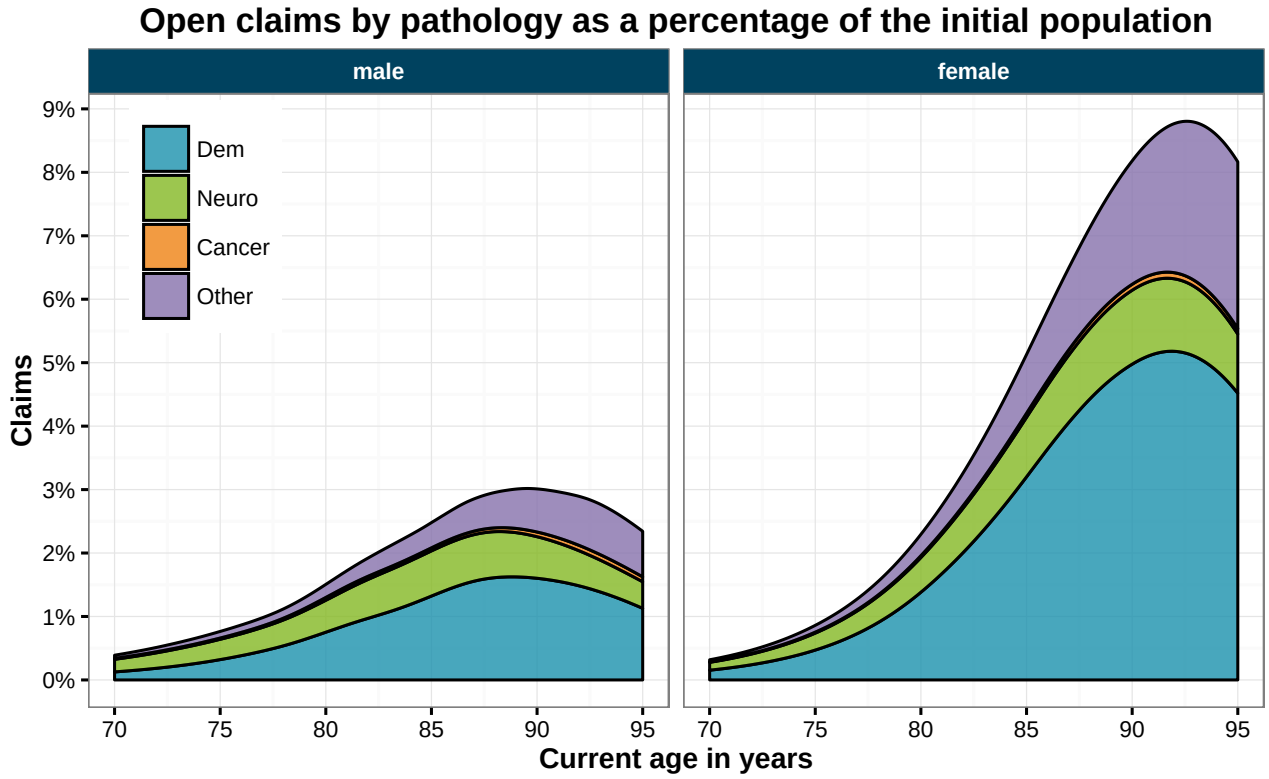


Figure 4.16: Projected open claims by pathology as a percentage of the initial population at age 50.

Figure 4.17 represents the prevalence $\mathrm{P}_k(x)$ of each pathology group in the portfolio

$$\mathrm{P}_k(x) = \frac{\mathrm{NC}_k(x)}{A(x) + \sum_{j=1}^{p} \mathrm{NC}_j(x)}.$$

Prevalence of cancer remains very low, which should come at no surprise given the high mortality rates associated with it. Prevalence of *neurological diseases* starts higher than other causes at

age 70 but increases more slowly. This prevalence is about the same for males and females. On the other hand, prevalence of *dementia* and *other causes* in females is much higher than prevalence in males. This is mostly due to the lower mortality observed for females in LTC.

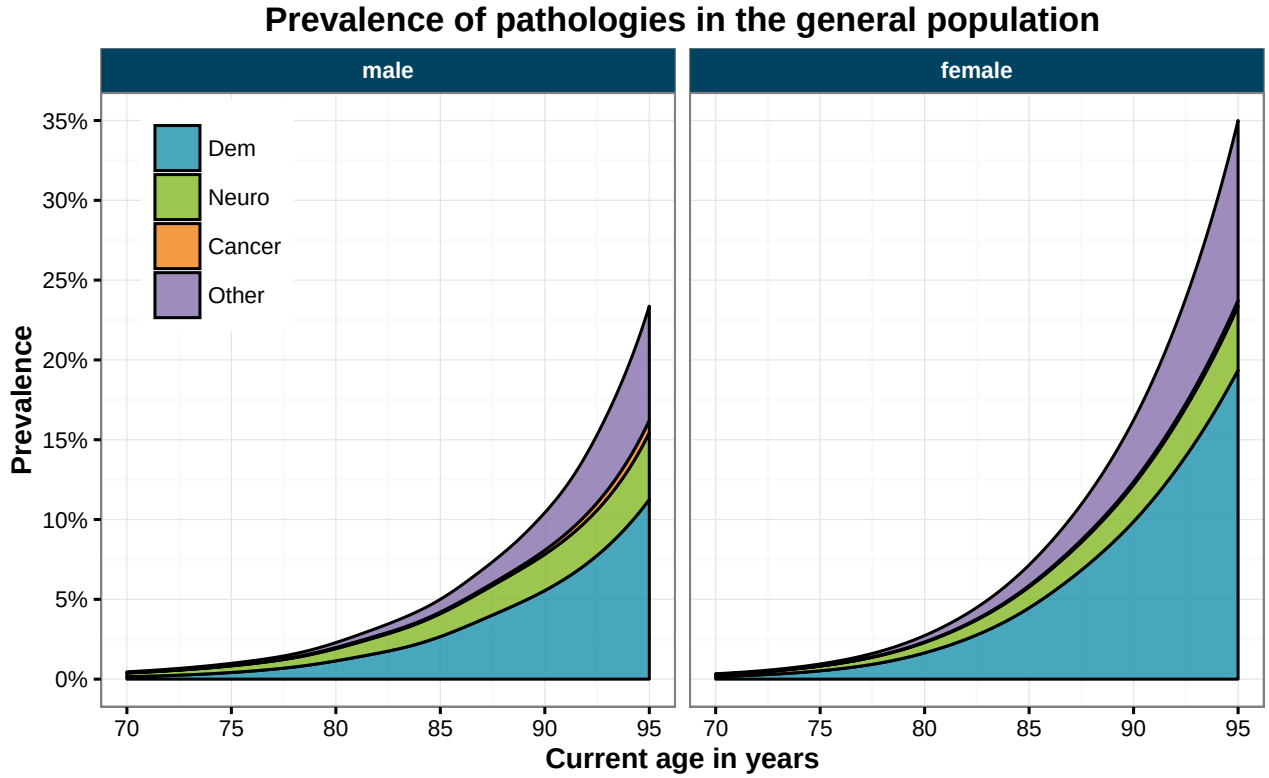


Figure 4.17: Projected prevalence of pathologies in the portfolio.

4.5.2 A second-step estimate of mortality in LTC

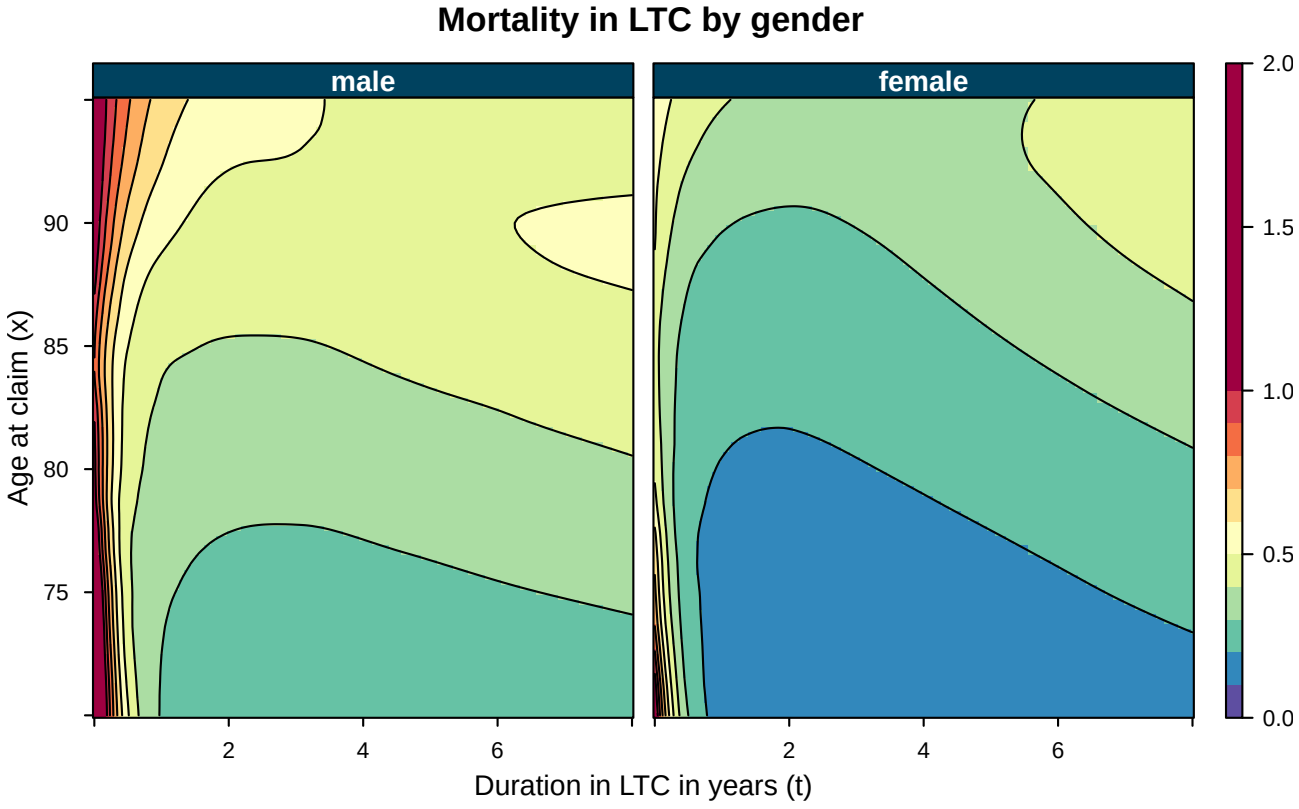


Figure 4.18: Intensity of mortality in LTC obtained by combining results from each group.

Finally, we combine the incidence and mortality for each group to get an overall mortality surface thanks to Lemma 1. Figure 4.18 represents the mortality surface obtained by combining the smoothed incidence rates in LTC for each group and the associated smoothed mortality rates in LTC. This second-step estimate reproduces the features of the reference fit with 100 nearest neighbours more fairly than the first-step estimate obtained by smoothing directly the overall mortality. Indeed the high mortality rates following claim inception are preserved and contour curves are matched more closely. If we take a look at the residuals represented on Figure 4.19, we see that with the exception of some features in the region of low age at claim and low duration, unlike the first-step estimate they no longer seem to present any structure.

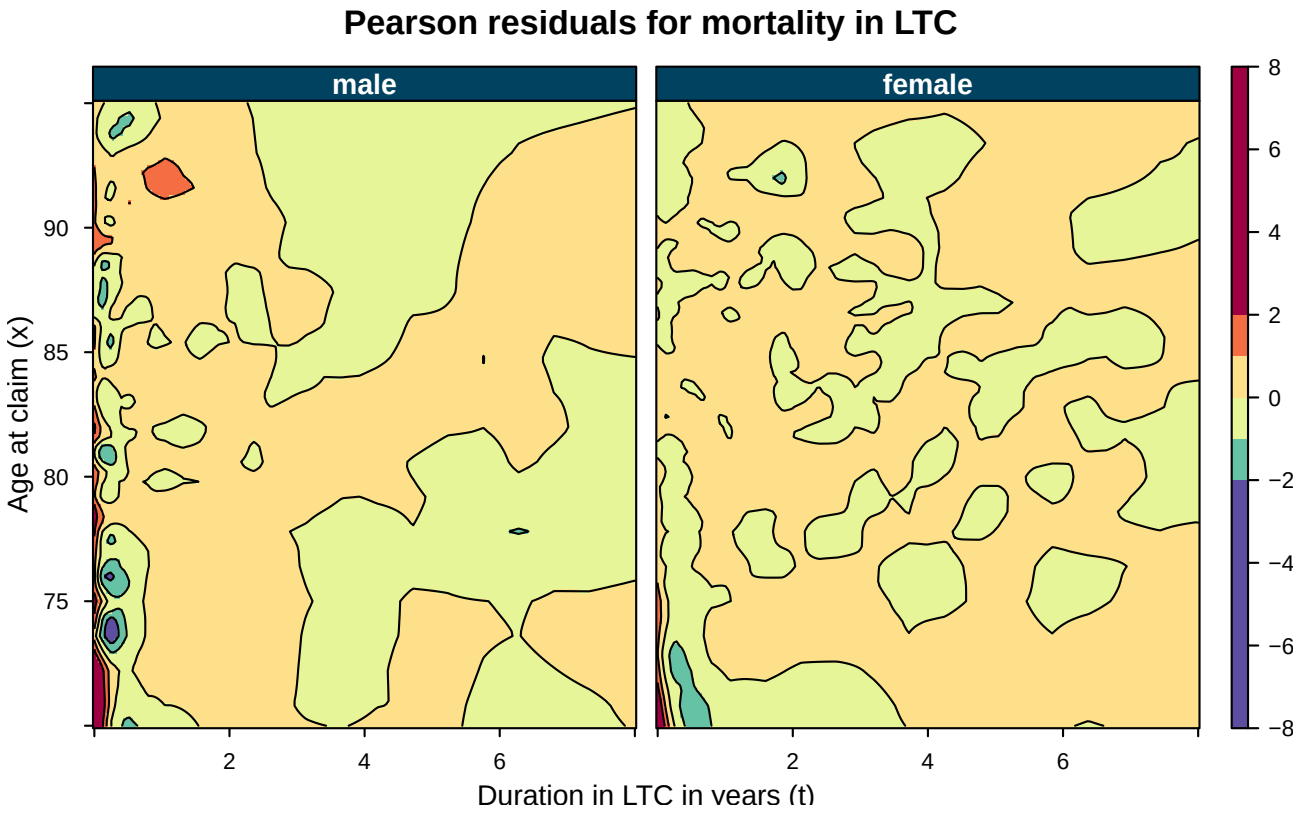


Figure 4.19: Pearson residuals associated with combined mortality in LTC.

4.5.3 Weight in the disabled population and contribution to mortality in LTC

The shape of the curve of mortality in LTC may be explained by the distribution of groups within the disabled population. Figure 4.20 represents the evolution of each group weight in this distribution as duration goes by, for several ages at claim. The fraction of disabled for a pathology group k corresponds to the quantity $\eta_k(x,t)$ introduced in Section 2. Due to the high mortality associated with it, *cancer* vanishes quickly from the disabled population. The proportion of *dementia* slightly declines with duration, especially for males, but remains significant. Proportions of *neurological diseases* and *other causes* tend to increase with duration, *neurological diseases* being more frequent for lower ages and *other causes* for higher ages at claim inception. Figure 4.21 represents the contribution of each pathology to the mortality in LTC $\eta_k(x,t)\mu_{i,k}(x,t)$. At duration 0, overall mortality in LTC is very high, due to the high mortality associated with *cancer*. This phenomenon is more substantial for males and for lower age at claim inception as the initial weight of *cancer* in the population is higher. At higher durations, contribution to mortality converges toward distribution of pathologies in the population.

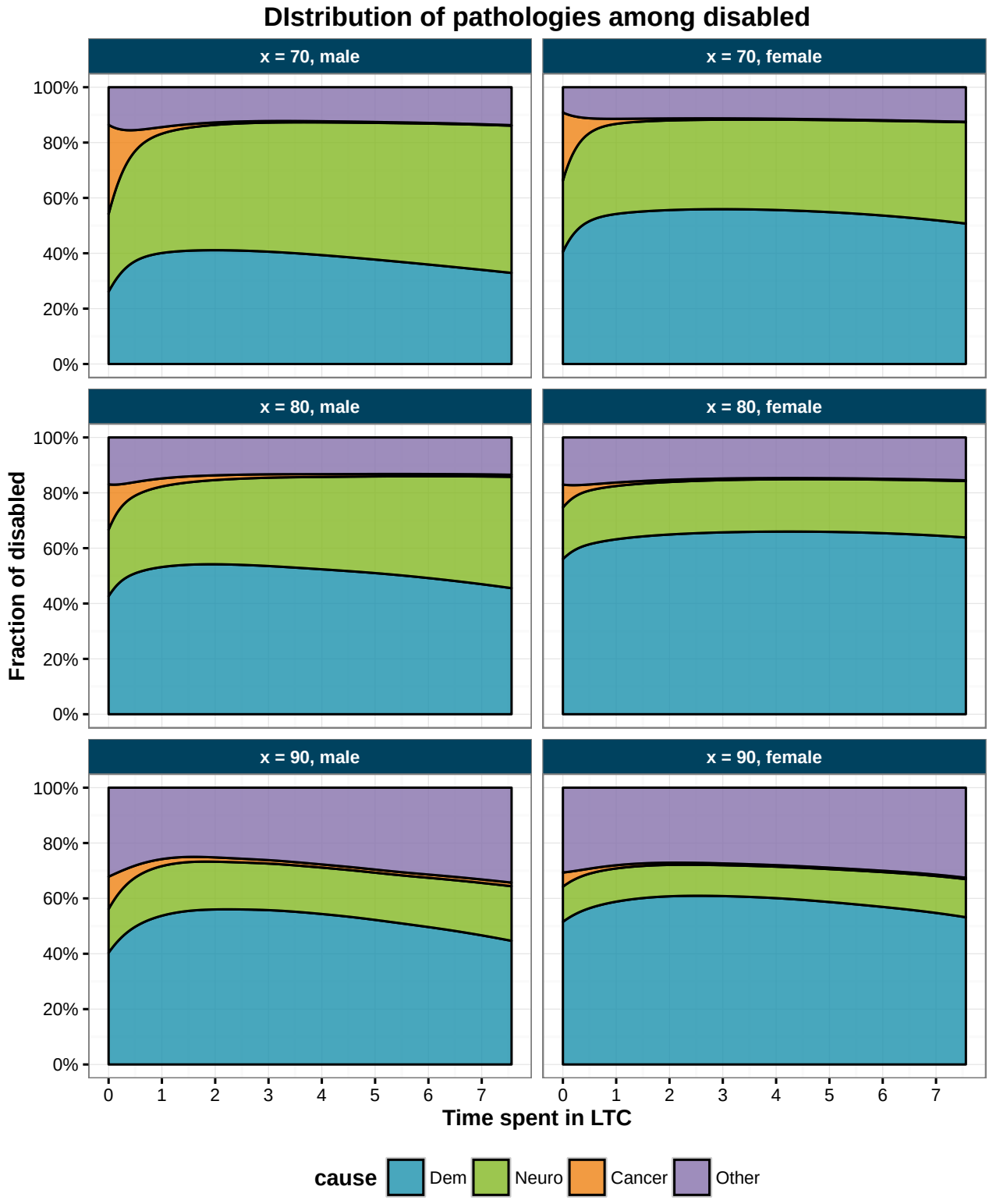


Figure 4.20: Distribution of pathologies among disabled.

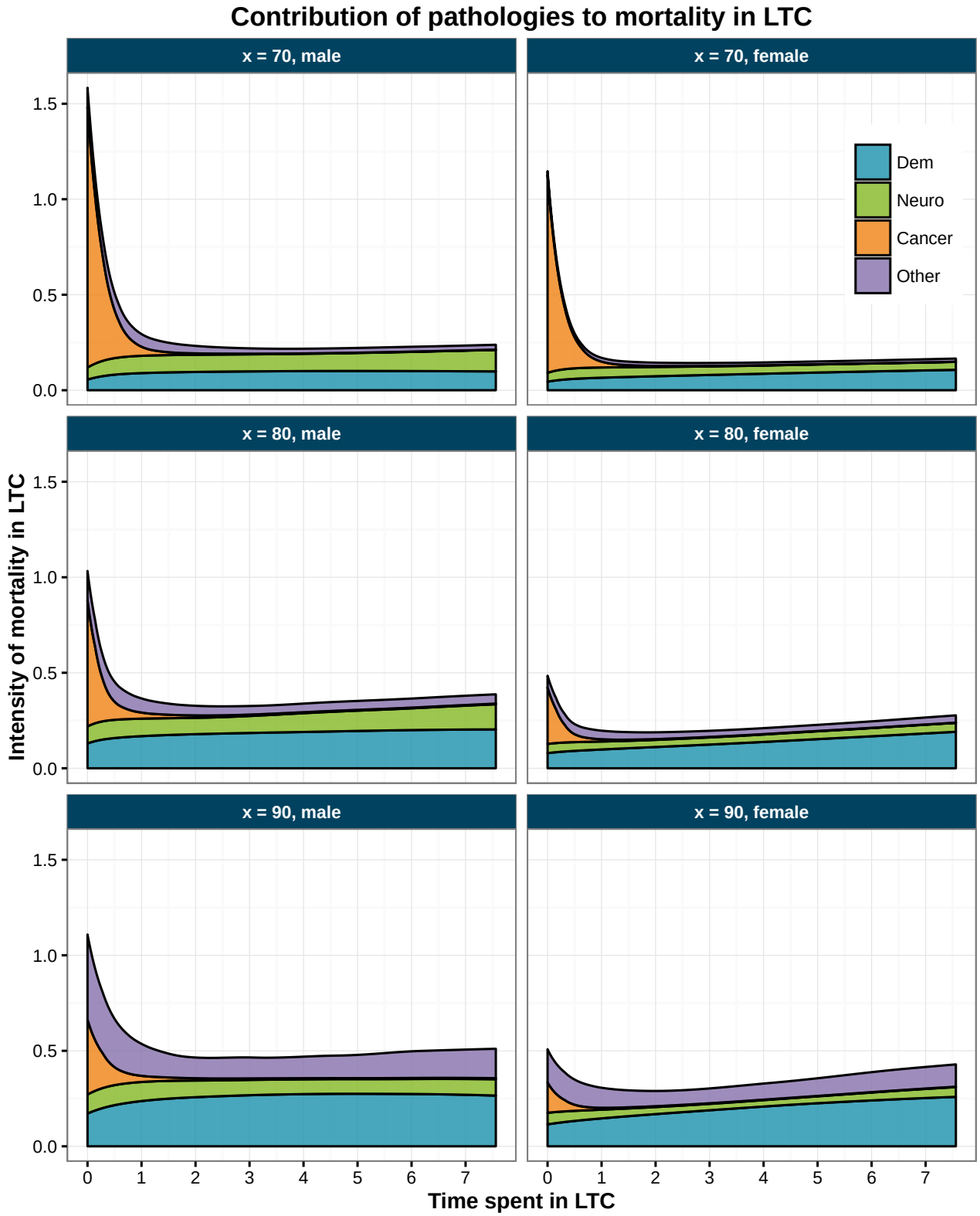


Figure 4.21: Contribution of pathologies to overall mortality in LTC.

4.6 Discussion

In this article, we study the mortality in Long Term Care (LTC) associated with 4 different groups of pathologies: *cancer*, *dementia*, *neurological diseases* and *other causes*. We rely on data from a French LTC portfolio. We introduce a continuous time semi-Markov framework where the mortality in LTC depends on both the age at claim inception and the time spent in the LTC state. We consider two models an illness-death model with 3 states autonomy, LTC

and death and a multi-state model with autonomy, death and 4 states of LTC which correspond to the 4 groups of pathologies observed in the data. Both models are defined using transition intensities: autonomous mortality, incidence in LTC and mortality in LTC (for all causes or for a given pathology group). We provide an expression of the overall mortality in LTC which depends on the incidence and mortality rates in LTC for each group thus linking both models. We introduce local likelihood methods that we use for direct inference of transition intensities in the model. Local likelihood methods are presented in the very didactic book from Loader (1999) and is implemented in the **locfit** package on the statistical software **R** (R Core Team, 2016). We rely on this implementation for the inference of mortality in LTC. Regarding the estimation of autonomous mortality and incidence rates in LTC, however, formulas need to be extended to account for left-truncated data and we rely on our own implementation of the method in that case. Local likelihood requires the specification of smoothing parameters such as the degree of the local fit and the bandwidth function. We rely on an adaptive bandwidth method where the bandwidth is defined as the distance to the k-th nearest neighbours. We test multiple combination of smoothing parameters and perform model selection based on Akaike information criterion (AIC). Once the inference of transition intensities has been completed, we derive a second-step estimate of overall mortality in LTC by combining incidence and mortality in LTC for each group of pathologies.

In the light of our results, impact of pathology groups on mortality in LTC proves substantial. Due to a sizeable difference in the initial level of mortality for *cancer* compared to other groups, a direct smoothing of the all-causes mortality in LTC leads to over-smooth the mortality surface in the region of low durations in LTC. On the other hand, smoothing separately the mortality in LTC associated with each group of pathology and combining the transition intensities results in a second-step estimate of the overall mortality in LTC allows us to overcome this difficulty. Studying mortality associated with each group of pathologies also provides useful information for the insurer, such as average duration of claim and cost associated with each group. At last, our results indicates that the singular shape of the curve for mortality in LTC with the high mortality right after claim inception is mostly due to the *cancer* group. This strongly advocates for estimating separately the mortality in LTC due to *cancer* and the mortality for other groups. When information about pathology is not available, one may still introduce a mixture model as in Biessy (2015a).

In this study, pathologies are gathered into 4 groups. Access to data with a higher level of details would allow to isolate some relatively frequent causes for LTC that are currently in the group of *other causes* such as cardiovascular diseases or muscular or skeletal diseases. Similarly, information about the type of cancer might be helpful to better model the mortality in LTC in the first few months following claim inception. Besides, the local likelihood estimates we provided could be used as a preliminary step in the setting-up of a parametric model. Indeed, while the local-likelihood estimate we obtain is useful on its own, a parametric model may yield better results when applied to a smaller portfolio with limited data. To this extent, our local-likelihood estimate makes for an interesting reference for comparison purposes. In addition, the degrees of freedom indicate how many parameters should ideally be used in the parametric model, though it may prove difficult due to the high amount of degrees of freedom in the selected fits. At last, we used the nearest neighbour methodology to select the bandwidth. While being an adaptive bandwidth method, it is still very basic and more complex methods among those presented in Tomas and Planchet (2013) may perform better.

Conclusion et perspectives

Dans le cadre de cette thèse, nous avons modélisé la perte d'autonomie des personnes âgées à l'aide d'un processus semi-markovien. Cette démarche est avant tout motivée par l'observation empirique de l'impact majeur du temps passé en dépendance sur la mortalité des individus dans cet état. Cette dernière est notamment beaucoup plus élevée lors de l'année qui suit l'entrée en dépendance. Ce phénomène a tout d'abord été observé au cours du Chapitre 2 sur les données publiques de l'APA. Au cours du Chapitre 3, le même phénomène est observé sur des données de portefeuille pour lesquelles les probabilités empiriques de décès des dépendants sont cette fois calculées de manière indépendante à l'aide d'un estimateur non-paramétrique discrétisé. En vertu de ces résultats, l'utilisation d'un modèle markovien simple ne prenant pas en compte le temps passé dans l'état de dépendance ne permet pas de capturer ce phénomène. Aussi une modélisation plus fine est nécessaire. Cela justifie l'approche semi-markovienne qui, est de notre point de vue, la manière la plus simple de prendre en compte le temps passé en dépendance dans l'estimation du risque.

Dans le Chapitre 2, nous utilisons en premier lieu une loi de Weibull afin de modéliser les lois de durée associées aux transitions entre états de dépendance et de la dépendance vers le décès. Afin d'augmenter le pouvoir prédictif de notre modèle, nous introduisons ensuite un mélange de deux lois de Weibull pour refléter l'hétérogénéité au sein des trajectoires en dépendance. Cette hétérogénéité est selon notre interprétation induite par les différentes pathologies qui mènent à l'état de dépendance. L'ajout d'une variable de mélange permet, dans le cas des transitions de la dépendance vers les états de dépendance les plus sévères ou le décès, d'augmenter sensiblement la vraisemblance du modèle. Selon le BIC et *a fortiori* pour l'AIC, ce choix se traduit par une amélioration de la qualité du modèle.

Au cours du Chapitre 3, nous reprenons l'idée d'introduire une variable de mélange et l'appliquons dans le cadre simplifié du modèle *illness-death* à 3 états. Cette fois-ci nous sommes à même de formuler des conclusions plus catégoriques. L'utilisation du mélange améliore sensiblement la qualité du modèle, et les résultats obtenus sont en outre très proches des probabilité empiriques de décès calculées à l'aide de méthodes non-paramétriques. De surcroît, dans ce modèle, le temps passé en dépendance n'intervient dans chaque composante du mélange qu'à travers un terme commun correspondant à la mortalité des autonomes qui peut être négligé sauf pour des durées passées en dépendance très longues qui sont rarement observées en pratique. Cela signifie que l'impact du temps passé en dépendance peut être modélisé de manière satisfaisante par un mélange dont chacune des composantes ne dépend pas (ou très peu) du temps passé en dépendance. L'effet de durée est de fait dû en très grande partie à un effet mélange. Ce résultat est ainsi utilisé dans le Chapitre 3 pour proposer une forme paramétrique de l'intensité de décès des dépendants.

Le Chapitre 4 permet entre autres d'étudier l'impact des pathologies sur le risque. Les données utilisées contiennent en effet plus de 14 000 sinistres dont les causes sont rassemblés en 4 groupes de pathologies : *cancer*, *maladies neurologiques*, *démence* et *autres causes*. Une estimation de la mortalité des dépendants associée à chacun de ces groupes permet de mettre en

évidence une différence d'ordre de grandeur entre la mortalité du groupe *cancer* et celle associée aux trois autres groupes. Une reconstitution de la mortalité tous groupes confondus à partir des lois d'incidence et de mortalité propre à chacun des groupes permet enfin de visualiser la contribution de chaque groupe à la mortalité tous groupes confondus des dépendants. Là encore, on retrouve le phénomène de surmortalité chez les dépendants de première année décrit ci-avant. Grâce à l'information sur les groupes de pathologies, il est cette fois possible d'imputer avec certitude ce phénomène au mélange du groupe *cancer* avec les trois autres groupes. Cela permet de valider les résultats obtenus lors des deux chapitres précédents et notamment le choix de deux groupes ainsi que leur interprétation à l'aide des pathologies. Cela constitue l'un des résultats les plus importants de la thèse.

Les autres apports de cette thèse sont principalement d'ordre méthodologique. Dans le Chapitre 2, nous présentons un algorithme pour la simulation de trajectoires multi-états qui peut être utilisé à des fins de tarification ou de calcul des provisions dans le cas où le calcul direct des engagements devient laborieux du fait du grand nombre de transitions possibles entre les états. Le Chapitre 3 propose une approche paramétrique complète pour l'estimation des lois, dont les différentes étapes peuvent être automatisées, ce qui peut s'avérer d'un grand intérêt pratique pour l'assureur. Elle permet également de contrôler le niveau de mortalité global du portefeuille en le comparant à une table de référence. Enfin, le Chapitre 4 s'intéresse à des méthodes de vraisemblance locale pour estimer les intensités du modèle et notamment l'intensité de décès des dépendants. Notre principale contribution réside dans l'application de ces méthodes aux trajectoires individuelles à travers une modélisation en temps continu plutôt qu'à des données agrégées associées à une modélisation en temps discret comme dans Tomas and Planchet (2013). Du point de vue théorique, cette approche est une extension des travaux de Loader (1996) pour l'estimation de densités à l'aide de la vraisemblance locale. Les grandes lignes de cette approche sont décrites dans Loader (1999) mais à notre connaissance elle n'a donné lieu à aucune autre application dans le domaine de l'assurance vie. L'implémentation de cette méthode nous permet d'aboutir à une représentation graphique de la mortalité des individus dépendants à l'aide de cartes de chaleurs. Ces dernières constituent selon nous un outil de représentation graphique adapté à la représentation du risque et néanmoins très peu utilisé.

Les produits dépendance vendus aujourd'hui possèdent plusieurs niveaux de garanties en fonction de la sévérité de l'état de dépendance. La tarification de ce type de produit nécessite ainsi une modélisation multi-états du risque avec la prise en compte d'autant d'états de dépendance que de niveaux de garantie définis par le produit. Dans le Chapitre 2, nous avons réalisé une étude de la mortalité des dépendants à travers un modèle à 4 états de dépendance. Cependant, en raison de la période d'observation limitée des données, de nombreux partis pris de modélisation ont dû être retenus. Aussi, les deux chapitres suivants se sont focalisés sur un modèle à 3 états avec un seul niveau de dépendance, sur lesquels des approches de modélisation plus raffinées ont pu être envisagées et testées. La prise en compte de deux états de dépendance dans ces modèles pourrait ainsi constituer une première extension des travaux proposés. A l'heure actuelle, le principal obstacle à cette approche est la rareté des données. En effet, la prise en compte d'un second niveau de dépendance suppose soit de disposer de données publiques comme celles de l'APA, mais sur une période plus longue, soit d'utiliser des données d'un produit récent comportant 2 niveaux de garanties. Dans ce dernier cas, la période d'observation sera malheureusement très limité, et le nombre de sinistres observés sera par conséquent très faible. La nécessité de disposer pour la tarification de nouveaux produits de données qui ne pourront être obtenues que de nombreuses années après le lancement du produit est l'un des paradoxes les plus emblématiques de l'assurance dépendance.

Dans les différents modèles proposés, on a fait l'hypothèse implicite que le risque étudié était invariant au cours du temps. On peut donner plusieurs justifications pour ce choix. En premier lieu, les données dont l'on dispose s'étendent sur une période limitée, avec un volume

limité qui ne permettrait pas en temps normal d'estimer une tendance pour l'évolution du risque. Par ailleurs, sur la vie d'un produit d'assurance, des facteurs propres à l'assureur mais externes au risque comme la sélection des assurés ou la gestion de sinistres peuvent avoir un impact sur le niveau du risque et ainsi interférer avec l'estimation des tendances. En outre, comme les portefeuilles étudiés sont des portefeuilles fermés, limités à la fois sur le plan des âges à la souscription et des années de souscription, il est difficile d'observer deux assurés du même âge à des dates éloignées dans le temps, ce qui rend l'estimation des tendances malaisé. Ainsi les assureurs français ne considèrent en règle générale pas de tendance dans leur modèle de tarification des produits dépendance. Cela se justifie par la possibilité prévue par le produit d'augmenter le niveau des primes afin de compenser une éventuelle dérive du risque. Néanmoins, d'un point de vue commercial, une telle augmentation n'est pas toujours souhaitable du fait du risque de réputation qu'elle peut engendrer pour l'assureur. Aussi la prise en compte de la dimension prospective du risque dans les modèles, si les données le permettent, est souhaitable et pourrait constituer un axe de développement porteur pour des travaux futurs.

Bibliographie

Aalen, O. O. and S. Johansen (1978). An empirical transition matrix for non-homogeneous markov chains based on censored observations. *Scandinavian Journal of Statistics*, 141–150.

Beard, R. (1959). Note on some mathematical mortality models. In : G.E.W. Wolstenholme and M. O'Connor (eds.).*The Lifespan of Animals. Ciba Foundation Colloquium on Ageing.* pp. 302-311. Little, Brown, Boston.

Beard, R. (1971). Some aspects of theories of mortality, cause of death analysis, forecasting and stochastic processes. In : *Biological Aspects of Demography* (ed. W. Brass). London : Taylor and Francis.

Biessy, G. (2015a). Continuous time semi-Markov inference of biometric laws associated with a Long-Term Care Insurance portfolio. Technical report, hal-01220564.

Biessy, G. (2015b). Long-term care insurance : A multi-state semi-Markov model to describe the dependency process in elderly people. *Bulletin Français d'Actuariat 15(29)*, 41–73.

Blanpain, N. and O. Chardon (2010). Projections de population à l'horizon 2060. INSEE Première, 1320.

Brass, W. (1971). Mortality models and their uses in demography. *Transactions of the Faculty of Actuaries 33*, 123–142.

Brass, W. (1974). Perspectives in population prediction : Illustrated by the statistics of england and wales. *Journal of the Royal Statistical Society 137(4)*, 532–583.

Breiman, L., J. Friedman, C. J. Stone, and R. A. Olshen (1984). *Classification and regression trees*. CRC press.

Bucar, T., M. Nagod, and M. Fajdiga (2004). Reliability approximation using finite Weibull mixture distributions. *Reliability Engineering and System Safety 84*, 241–251.

Chen, B., G. Y. Yi, and R. J. Cook (2010). Analysis of interval-censored disease progression data via multi-state models under a nonignorable inspection process. *Statistics in Medicine 29(11)*, 1175–1189.

Chichignoud, M. (2010). *Performances statistiques d'estimateurs non-linéaires*. Ph. D. thesis, Citeseer.

Christiansen, M. C. (2012). Multistate models in health insurance. *Advances in Statistical Analysis 96(2)*, 155–186.

Cinlar, E. (1969). Markov renewal theory. *Advances in Applied Probability 1*, 123–187.

Commenges, D. (2002). Inference for multi-state models from interval-censored data. *Statistical Methods in Medical Research 11(2)*, 167–182.

Cox, D. R. (1972). Regression models and life-tables. In *Breakthroughs in statistics*, pp. 527–541. Springer.

Cox, D. R. and D. Oakes (1984). *Analysis of survival data*. Monographs on Statistics and Applied Probability. Chapman & Hall, London.

Czado, C. and F. Rudolph (2002). Application of survival analysis methods to long-term care insurance. *Insurance : Mathematics and Economics 31*(3), 395–413.

Debout, C. (2010). Durée de perception de l'Allocation personnalisée d'autonomie (APA). Direction de la Recherche, des Études, de l'Évaluation et des Statistiques, Document de travail.

Denuit, M. and C. Robert (2007). *Actuariat des Assurances de Personnes*. Economica.

Déléglise, M. P., C. Hess, and S. Nouet (2009). Tarification, provisionnement et pilotage d'un portefeuille dépendance. *Bulletin Français d'Actuariat 9(17)*, 70–108.

Eilers, P. H. and B. D. Marx (1996). Flexible smoothing with b-splines and penalties. *Statistical science*, 89–102.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications 14*(1), 153–158.

Fong, J. H., A. W. Shao, and M. Sherris (2015). Multistate actuarial models of functional disability. *North American Actuarial Journal 19*(1), 41–59.

Foucher, Y., M. Giral, J.-P. Souilillou, and J.-P. Daures (2007). A semi-Markov model for multi-state and interval-censored data with multiple terminal events. Application in renal transplantation. *Statistics in Medicine 26(30)*, 5381–5393.

Gompertz, B. (1825). On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Phil. Trans. R. Soc. Lond. 115*, 513–583.

Guibert, Q. and F. Planchet (2014). Construction de lois d'expérience en présence d'évènements concurrents : Application à l'estimation des lois d'incidence d'un contrat dépendance. *Bulletin Français d'Actuariat 14*(27), 5–28.

Guibert, Q. and F. Planchet (2015). Non-parametric inference of transition probabilities based on aalen-johansen integral estimators for semi-competing risks data : application to ltc insurance. In *Conference of the LIFE Section of the International Actuarial Association*.

Haberman, S. and E. Pitacco (1998). *Actuarial Models for Disability Insurance*. Chapman and Hall/CRC, 1st edition.

Hannerz, H. (2001). An extension of relational methods in mortality estimation. *Demographic Research 4(10)*, 337–368.

Hardy, M. R., C. D. C. Dickson, and H. R. Waters (2011). Supplementary notes for actuarial mathematics for life contingent risks version 2.0.

Helms, F., C. Czado, and S. Gschlößl (2005). Calculation of ltc premiums based on direct estimates of transition probabilities. *Astin Bulletin 35*(02), 455–469.

Janssen, J. and R. Manca (2007). *Semi-Markov risk models for finance, insurance and reliability*. Springer Science & Business Media.

Jiang, R. and D. Murthy (1997). Two sectionals models involving three Weibull distributions. *Quality and Reliability Engineering international 13*, 83–96.

Joly, P., D. Commenges, C. Helmer, and L. Letenneur (2002). A penalized likelihood approach for an illness–death model with interval-censored data : application to age-specific incidence of dementia. *Biostatistics 3*(3), 433–443.

Klein, J. P. (1991). Small sample moments of some estimators of the variance of the kaplan-meier and nelson-aalen estimators. *Scandinavian Journal of Statistics*, 333–340.

Lécroart, A. (2011). Projections du nombre de bénéficiaires de l'APA en France à l'horizon 2040-2060 - Sources, méthode et résultats. Direction de la Recherche, des Études, de l'Évaluation et des Statistiques, Document de travail.

Lebarbier, E. and T. Mary-Huard (2006). Une introduction au critère BIC : fondements théoriques et interprétation. *Journal de la SFdS 147*, 39–57.

Lepez, V. (2006). *Trajectoires en Dépendance des personnes âgées : modélisation, estimation et application en assurance vie*. Mémoire d'actuariat, Centre d'Etudes Actuarielles.

Lepez, V., S. Roganova, and A. Flahault (2013). A semi-Markov model to investigate the different transitions between states of dependency in elderly people. In : Colloquium of the International Actuarial Association, Lyon.

Limnios, N. and G. Oprisan (2012). *Semi-Markov processes and reliability*. Springer Science & Business Media.

Loader, C. (1999). *Local regression and likelihood*. Statistics and Computing. Springer-Verlag, New York.

Loader, C. R. (1996, 08). Local likelihood density estimation. *Ann. Statist. 24*(4), 1602–1618.

Lopez, O. (2011). Nonparametric estimation of the multivariate distribution function in a censored regression model with applications. *Communications in Statistics-Theory and Methods 40*(15), 2639–2660.

Lopez, O., X. Milhaud, and P.-E. Thérond (2015). Tree-based censored regression with applications to insurance. Technical report, hal-01141228.

Lopez, O., V. Patilea, I. Van Keilegom, et al. (2013). Single index regression models in the presence of censoring depending on the covariates. *Bernoulli 19*(3), 721–747.

Makeham, W. M. (1867). On the law of mortality. *Journal of the Institute of Actuaries 13*, 325–358.

Massonet, B. (2006). L'assurance dépendance – Estimation des matrices de transition – Modélisation. In *Proceedings of ICA Paris*.

Mathieu, E. (2006). *Modélisations multi-états markoviennes et semi-markoviennes. Applications à l'état de sante des patients atteints par le virus du SIDA*. Thèse, Université de Montpellier.

Monod-Zorzi, S., L. Seematter-Bagnoud, C. Büla, S. Pellegrini, and H. Jaccard Ruedin (2007). Maladies chroniques et dépendance fonctionnelle des personnes âgées : données épidémiologiques et économiques de la littérature. Observatoire suisse de la santé, Document de travail.

Nelder, J. A. and R. Mead (1965). A simplex method for function minimization. *The Computer Journal 7*, 308–313.

Perks, W. (1932). On some experiments in the graduation of mortality statistics. *Journal of the Institute of Actuaries 63(1)*, 12–57.

Pitacco, E. (2014). *Health Insurance - Basic Actuarial Models*. Springer International Publishing.

Pitacco, E. (2015). Actuarial values for long-term care insurance products. a sensitivity analysis. *Working Paper*, ARC Centre of Excellence in Population Ageing Research.

Planchet, F. and P. Thérond (2006). *Modèles de durée. Applications actuarielles*. Economica.

Pritchard, D. J. (2006). Modeling disability in long-term care insurance. *North American Actuarial Journal 10*(4), 48–75.

R Core Team (2016). *R : A Language and Environment for Statistical Computing*. Vienna, Austria : R Foundation for Statistical Computing.

Ramlau-Hansen, H. (1991). Distribution of surplus in life insurance. *ASTIN Bulletin 21*, 57–71.

Rickayzen, B. D. and D. E. P. Walsh (2002). A multi-state model of disability for the United Kingdom : Implications for future need for long-term care for the elderly. *British Actuarial Journal 8*, 341–393.

Robinson, J. (1996). A long-term care status transition model. The Old-Age Crisis - Actuarial Opportunities : The 1996 Bowles Symposium, Georgia State University.

Satten, G. and M. R. Sternberg (1999). Fitting semi-Markov models to interval-censored data with unknown initiation times. *Biometrics 55(2)*, 507–513.

SCOR (1995). L'assurance dépendance. Dossiers SCOR, SCOR Tech.

Smith, M. (1993). *Neural networks for statistical modeling*. Thomson Learning.

Smola, A. J. and B. Schölkopf (2004). A tutorial on support vector regression. *Statistics and computing 14*(3), 199–222.

Tibshirani, R. and T. Hastie (1987). Local likelihood estimation. *Journal of the American Statistical Association 82*(398), 559–567.

Tomas, J. and F. Planchet (2013). Multidimensional smoothing by adaptive local kernel-weighted log-likelihood : Application to long-term care insurance. *Insurance : Mathematics and Economics 52*(3), 573–589.

University of California, Berkeley (USA) and Max Planck Institute for Demographic Research (Germany) (2015). Human mortality database. Data available at www.mortality.org or www.humanmortality.de.

Vetel, J., R. Leroux, and J. Ducoudray (1998). AGGIR, practical use, geriatric autonomy group resources needs. Soins Gerontol.

Volinsky, C. T. and A. E. Raftery (2000). Bayesian information criterion for censored survival models. *Biometrics 56(1)*, 256–262.

Wand, M. P. and M. C. Jones (1995). *Kernel smoothing*, Volume 60 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, Ltd., London.

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre : Modélisation semi-markovienne de la perte d'autonomie chez les personnes âgées : application à l'assurance dépendance

Mots clefs : Processus semi-markovien, assurance dépendance, données censurées, vraisemblance locale, modèle de mélange

Résumé : Défi majeur aux sociétés modernes, la perte d'autonomie chez les personnes âgées, connue également sous le nom de dépendance se définit comme un état d'incapacité à effectuer seul tout ou partie des Actes de la Vie Quotidienne (AVQ). Elle apparaît dans la grande majorité des cas sous l'effet des pathologies chroniques liées au vieillissement. Devant les coûts importants liés à cet état, les assureurs privés ont développé une offre destinée à compléter l'aide publique.
Pour quantifier le risque, un modèle multi-états est utilisé et se pose alors la question de l'estimation des probabilités de transition entre les états (l'autonomie, le décès ainsi qu'un ou plusieurs niveaux de dépendance. Sous l'hypothèse de Markov, ces dernières dépendent uniquement de l'état actuel, une hypothèse trop restrictive pour rendre compte de la complexité du processus de dépendance. Dans le cadre semi-markovien plus général, ces probabilités dépendent également du temps passé dans l'état actuel.
Au cours de cette thèse, nous étudions la nécessité d'une modélisation semi-markovienne du processus. Nous mettons en évidence l'impact du temps passé en dépendance sur les probabilités de décès. Nous montrons par ailleurs que la prise en compte de la diversité induite par les pathologies permet d'améliorer sensiblement l'adéquation du modèle proposé aux données étudiées. Plus encore, nous établissons que la forme particulière de la probabilité de décès en fonction du temps passé en dépendance peut être expliquée par le mélange des groupes de pathologies qui constituent la population des individus dépendants.

Title : Semi-Markov modeling of the loss of autonomy among elderly people: application to Long-term Care Insurance

Keywords : semi-Markov process, Long-Term-Care Insurance, censored data, local likelihood, mixture model

Abstract : A sizable challenge to modern societies, Long-Term Care (LTC) in elderly people may be defined as a state of incapacity to perform autonomously part of the Activities of Daily Living (ADL). In most cases, long-term care is caused by pathologies linked to aging. To cope with the sizeable costs linked to this state, private insurers have developed products in top of the public aid.
To quantify the long-term care risk, multi-state models are used for which transition probabilities between states (autononomy, death and one to several levels of LTC) need to be inferred. Under the Markov assumption, those probabilities only depend on the current state, this assumption being too restrictive in regards of the complexity of the underlying risk. In a semi-Markov framework, those probabilities also depends on the time spent in the current state.
In this thesis, we emphasis the need for the semi-Markov modeling. We demonstrate the impact of time spent in LTC on death probabilities. Besides, we exhibit that taking into account the diversity induced by pathologies leads to sizable improvements in the fit of the model to experience data. Furthermore, we highlight that the peculiar shape taken by death probabilities as a function of time spent in LTC may be explained by the mixture of pathology groups among the disabled population.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France