Advanced modeling of the risk of loss of autonomy of elderly people

Leonie Le Bastard

THÈSE de DOCTORAT DE L'UNIVERSITÉ DE LYON

Opérée au sein de :

l'Université Claude Bernard Lyon 1

École Doctorale 486
Sciences Économiques et de Gestion

Spécialité de doctorat : Sciences de Gestion

Soutenue publiquement le 10/04/2024, par :

Léonie Le Bastard

Modélisation avancée du risque de dépendance des personnes âgées

---

Devant le jury composé de :

ARNOLD Séverine Professeure ordinaire, HEC Lausanne	**Rapporteure**
DEBON AUCEJO Ana María Professeure, Universitat Politècnica de Valencia	**Rapporteure**
CAMBOIS Emmanuelle Chercheure, Institut national d'études démographiques	**Examinatrice**
GUIBERT Quentin Maître de Conférences, Université Paris-Dauphine	**Examinateur**
PLANCHET Frédéric Professeur des Universités, Université Claude Bernard Lyon 1	**Président**
ROBERT Christian Professeur des Universités, Université Claude Bernard Lyon 1	**Directeur de thèse**
LOISEL Stéphane Professeur du CNAM, Conservatoire National des Arts et Métiers	**Co-Directeur de thèse**
TORRI Tiziana Encadrante de thèse en entreprise, SCOR SE	**Invitée**

Résumé

Les progrès de la médecine, couplés à l'avancée en âge de la génération Baby Boom et à la diminution des taux de fertilité, sont à l'origine d'une augmentation significative de la proportion de personnes âgées dans la population. On parle de vieillissement démographique. Ainsi, les pays développés font face à un enjeu majeur : celui de la perte d'autonomie des personnes âgées. La dépendance se définit par le besoin d'une tierce personne pour les activités de la vie quotidienne telles que manger, se laver, s'habiller, se déplacer ou encore être continent. Apparaissant majoritairement en raison de pathologies liées au vieillissement, cette perte d'autonomie engendre un coût financier important pour les individus et leurs familles. C'est dans ce contexte que les assureurs ont mis en place des produits d'assurance dépendance, permettant de couvrir une partie des frais liés à la dépendance, au-delà de l'Allocation Personnalisée d'Autonomie versée par la sécurité sociale en France. Cependant, la modélisation de ces produits est complexe en raison du peu de données dont disposent les assureurs sur ces produits développés majoritairement au début des années 2000. Aussi, cette thèse s'intéresse à la modélisation du risque de dépendance dans un contexte où les données sont limitées, impliquant l'étude de méthodes statistiques avancées afin d'exploiter au mieux l'information issue de chaque base de données.

Le Chapitre 1 présente le contexte général de l'assurance dépendance ainsi que les concepts mathématiques de bases utilisés dans le reste du manuscrit. L'objectif est ensuite de permettre au lecteur de positionner cette thèse, en présentant succinctement les problématiques abordées ainsi que les principales contributions.

Le Chapitre 2 s'intéresse à l'estimation de la mortalité aux âges avancés, pour lesquels les assureurs n'ont observé que très peu, voire pas du tout, d'individus depuis la création de leur produit d'assurance dépendance. Ce manque de données s'explique notamment par le peu d'historique des produits d'assurance dépendance, ainsi que par l'âge limite imposé à la souscription. L'estimation des taux de décès aux âges avancés nécessite alors des méthodes statistiques d'extrapolation. Dans le contexte de l'étude d'un produit d'assurance

dépendance, au minimum deux lois de mortalité sont à estimer : celle des autonomes et celle des dépendants. Ce chapitre propose une méthode d'extrapolation cohérente et simultanée de ces deux lois, en utilisant la connaissance sur la mortalité globale du portefeuille d'assurés.

Le Chapitre 3 est dédié à l'étude de méthodes statistiques permettant de concaténer des données issues de contrats n'utilisant pas des définitions uniformes de la dépendance. En effet, il n'existe pas de consensus commun entre l'ensemble des assureurs sur le degré de perte d'autonomie donnant lieu au versement de la rente, ainsi que sur les modalités des contrats telles que la période de franchise. Par conséquent, l'agrégation de données de plusieurs assureurs dans le but d'augmenter le volume de données, et ainsi améliorer l'estimation du risque, est complexe. Estimer les lois biométriques sur des données de plusieurs assureurs, sans prendre en compte la différence de définition, peut mener à une sous-estimation du risque pour l'assureur ayant les conditions de versement de rentes les moins strictes. Ce chapitre propose deux méthodes permettant de combiner l'information issue de deux types de contrats ayant des définitions différentes de l'état de dépendance. Nous appliquons ces méthodes à l'agrégation de données issues de deux produits d'assurance, dont l'un seulement dispose d'une période de franchise de trois mois.

Le Chapitre 4 aborde la problématique de la prise en compte, dans l'estimation de la mortalité des dépendants, de l'information sur la pathologie ayant causé la perte d'autonomie. Plusieurs études ont démontré que celle-ci impacte significativement la mortalité des individus en situation de dépendance. En raison d'un manque de données suffisantes et de la préférence des assureurs pour des modèles de mortalité simples, tout en maintenant de bonnes performances prédictives, les actuaires ne peuvent se permettre d'estimer une table de mortalité spécifique par pathologie. Regrouper celles semblables en termes de mortalité observée semble être un bon compromis. L'intensité de décès des individus dépendants étant fonction à la fois de leur âge mais également du temps écoulé depuis leur perte d'autonomie, la mortalité est représentée sous forme de surface. La complexité de ce regroupement réside alors dans le caractère bidimensionnel des lois de mortalité des dépendants, mais également dans la très faible représentation de certaines pathologies dans la base de données. Ainsi, en raison de l'absence de méthodes statistiques adaptées, de nombreux assureurs font appel à l'avis d'un expert pour constituer les groupes. Ce chapitre propose deux méthodes de clustering répondant à cette problématique, puis compare les performances prédictives des modèles de mortalité associés, à celles du modèle utilisant les groupes de pathologies construits par avis d'expert.

Mots clés : Assurance dépendance ; Modélisation actuarielle ; Mortalité ; Modèles multi-états ; Modèles semi-Markovien ; Extrapolation ; Agrégation de données ; Clustering.

Abstract

Medical advances, coupled with the ageing of the Baby Boom generation and a reduction in fertility rates, have led to a significant increase in the proportion of elderly people in the population. This phenomenon is called demographic ageing. Hence, developed countries are confronted with a major issue : the loss of autonomy among elderly individuals, defined as the inability to manage basic activities of daily living independently, such as eating, bathing, dressing, moving, or maintaining continence. Long-term care denotes a spectrum of services proposed to individuals who lost their autonomy. Mostly due to ageing pathologies, the loss of autonomy poses a considerable financial burden on individuals and their families. In this context, insurers offer long-term care insurance products, covering part of the cost generated by the loss of autonomy beyond public aid, such as the APA [1] in France. However, due to the short history of long-term care insurance products, which were mostly developed in the early 2000s, data remain scarce. As a result, modelling these products is complex. This PhD thesis focuses on modelling the loss of autonomy of elderly individuals in the context of data scarcity, requiring advanced statistical methods to make the best use of each available database.

In Chapter 1, we introduce the general context of long-term care insurance as well as the basic mathematical concepts used throughout the manuscript. We then present the issues addressed in this thesis as well as the main contributions.

In Chapter 2, we focus on estimating mortality rates at advanced ages, where insurers lack information. This scarcity is due to the relatively short history of LTC products and the age limit imposed at underwriting. As a result, statistical extrapolation methods are needed to estimate mortality rates at old ages. In the context of long-term care insurance product modelling, insurers must estimate at least two mortality functions : one associated with autonomous policyholders and one associated with disabled policyholders. In this chapter, we propose to complete the missing information on the mortality of autonomous and disabled insured people at advanced ages using information on the global mortality of

1. Allocation Personnalisée d'Autonomie

the portfolio, allowing for estimating and extrapolating both mortality laws simultaneously in a consistent manner.

Chapter 3 is dedicated to the use of statistical methods that enable the aggregation of experience data from several LTC insurance portfolios. While all LTC insurance products aim to cover the risk of loss of autonomy, the exact definition of the health state leading to a claim varies across different market and even within the same market. Furthermore, certain specificities in the contracts introduce disparity in the conditions for receiving the annuities between policyholders of different portfolios. As a result, aggregating data from multiple LTC portfolios with the intention of increasing the volume of data, and thus improving the accuracy of risk estimation, is challenging. Disregarding the heterogeneity of the definitions among contracts, can yield an underestimation of risk, especially for insurers with the least restrictive conditions. In this chapter, we propose two methods for combining the experience data of two portfolios with different disability definitions. These methods are applied to the aggregation of data from two products, one of which includes a deferred period, while the other pays the annuity upon recognition of the loss of autonomy.

In Chapter 4, we focus on the impact of the pathology responsible for the loss of autonomy on the mortality of a disabled insured individual. Several papers have shown that this information is a good predictor of mortality in LTC patients. However, due to the scarcity of data combined with the desire of insurers to have simple mortality models while maintaining good predictive performances for mortality, estimating a single mortality table for each pathology is complex and not optimal. Creating groups of pathologies with similar mortality rates seems a good compromise. As the mortality of disabled policyholders depends on both their attained age and the time elapsed since their loss of autonomy, we are in the context of surface clustering. The complexity of this aggregation arises from the two-dimensional aspect of mortality, as well as from the scarcity of data for certain rare pathologies. Thus, due to the absence of suitable statistical methods, many insurers rely on expert judgment to construct groups of pathologies. In this chapter, we present two methods to cluster pathologies into homogeneous groups in terms of mortality. We then compare the predictive performances of the related mortality models using groups constructed from each method, to that of the model relying on the expert judgment clusters.

Keywords : Long-Term Care Insurance ; Actuarial modelling ; Mortality ; Multi-state models ; Semi-Markovian models ; Extrapolation ; Data aggregation ; Clustering.

Remerciements

Je tiens tout d'abord à remercier Séverine Arnold et Ana María Debon Aucejo pour me faire l'honneur de rapporter ma thèse, ainsi que pour leurs précieux commentaires qui ont contribué à l'amélioration de ce manuscrit. Merci également à Emmannuelle Cambois, Quentin Guibert, et Frédéric Planchet d'avoir accepté de faire partie de mon jury de thèse.

J'exprime toute ma gratitude envers mes deux directeurs de thèse, Stéphane Loisel et Christian Robert, pour m'avoir accordé leur confiance en acceptant de me prendre sous leurs ailes. Votre soutien et vos précieux conseils ont été des éléments cruciaux pour le développement de ce travail. Je souhaite également vous exprimer ma reconnaissance pour votre accompagnement au cours de ces années de thèse, et pour avoir su répondre à mes nombreuses questions et me rassurer dans les périodes de doute.

Je tiens également à exprimer mes sincères remerciements à mes deux responsables de thèse en entreprise, Razvan Ionescu et Tiziana Torri. Vous avez été "le papa et la maman" de cette thèse, et je vous suis infiniment reconnaissante pour le soutien que vous m'avez apporté dans les moments plus difficiles.

Merci à SCOR qui m'a offert un cadre de travail propice au bon déroulement de la thèse. Merci à toutes les personnes qui ont contribué à mon épanouissement au sein de l'entreprise depuis mon arrivée en 2019. Vous êtes trop nombreux pour être tous cités, mais je remercie tout particulièrement mes collègues de l'équipe Biometric Risk Modelling. Un merci tout particulier à Albane pour les nombreuses discussions autour d'un café latté (avec plus ou moins de mousse selon l'humeur de la machine), ainsi que pour tous les échanges de musiques électro qui ont en bonne partie contribué à l'écriture de cette thèse en m'aidant à me mettre dans une bulle de concentration. Merci également à Thomas (et désolé pour les remises en question des modèles le vendredi soir), Julien, Agne, Xiao. Une pensée également à la team des doctorants, Antoine, Benno, Denis et William. Merci pour votre soutien et pour les échanges de tips autour de déjeuners un peu sectaires entre CIFRE. Merci également aux "anciens" de l'équipe, Guillaume Biessy et Ilan Cohen, pour m'avoir pris sous votre aile pour mon stage de Master 2 et pour m'avoir incité à me lancer dans

l'aventure de la thèse.

Un grand merci à tous les doctorants et postdocs du laboratoire SAF. Je n'étais pas souvent présente au labo, mais je garderais de bons souvenirs de mes passages. Une mention spéciale pour Tachfine, le champion incontestable du lancer de hâche. Je ne te serais jamais assez reconnaissante pour tout le soutien que tu m'a apporté au cours de cette dernière année de thèse. Les nombreuses discussions autour d'un verre de jus d'ananas m'ont énormément apportées, que ce soit pour des précieux conseils sur des aspects techniques, ou bien de simples tips de doctorants. Merci également à Behzad, Charles, Étienne, Ismaël, Karim, Natalya, Pierre C. (le plus rapide des CIFRE clandestins), et Rayane.

Évidemment, je ne serais pas celle que je suis sans le soutien permanent de ceux qui sont là depuis le début : mes parents et mon frère. Votre amour, votre soutien indéfectible et votre confiance en moi m'ont donné la détermination nécessaire pour mener à bien ce projet ambitieux. Étienne, ton courage et ton ambition ont toujours été source d'inspiration et de motivation pour moi. Merci de me pousser à être meilleure chaque jour.

À mes amis qui ont été là dans les bons et les mauvais moments, je vous suis reconnaissant pour votre amitié sincère et votre soutien inconditionnel. Les moments partagés ensemble (quoique plus rares en cette fin de thèse) à rire, refaire le monde, débattre, bien manger, ou juste papoter ont grandement contribué à mon bien être. Merci à la Team de Paris (Alexandre, Benjamin, Fabian, Juline, Pierre), à la Team "Korean Dream" (Adeline et Estelle), à la Team de l'INSA (Alex, Maxime, Emma et Muriel). Merci également à toi Sophie pour ton soutien et tes encouragements, qui m'ont aidée à surmonter les épreuves de la vie.

Merci également à tous ceux que j'ai rencontré sur le chemin de la vie, et notamment tous ceux que j'ai connu sur ou au bord d'une piste d'escrime. Je ne serais pas cette personne sans tout ce que m'a apporté cette passion. C'est en grande partie ce sport qui me suit depuis l'enfance qui m'a appris les valeurs de détermination, de persévérance, et de patience. Cette passion m'a également fait rencontrer des personnes incroyables, qui ont contribué à mon développement personnel, et sans qui je ne serais pas celle que je suis aujourd'hui : Gérard Vaillant et Sylvain Guyomard.

Table des matières

Chapitre 1

Introduction

Ce chapitre a pour objectif d'introduire le contexte général, ainsi que les concepts de base utilisés dans le cadre de la modélisation de produits d'assurance dépendance. Nous présentons ensuite les motivations ainsi que les contributions principales de chacun des chapitres de cette thèse, permettant au lecteur de positionner ce travail. Chaque chapitre de ce manuscrit dispose de sa propre bibliographie, située à sa fin.

1.1 Préliminaires/Contexte

Cette thèse a été effectuée dans le cadre d'une Convention Industrielle de Formation par la Recherche (CIFRE) au sein de l'entreprise SCOR [1], 6e réassureur mondial en termes de chiffres d'affaires. Très actif sur le marché de l'assurance dépendance, SCOR est exposé au risque de sous-estimation de la problématique de perte d'autonomie des personnes âgées. L'objectif de cette thèse, réalisée au sein du service de recherche et développement dédié à la modélisation des risques biométriques, est d'améliorer la modélisation des risques liés à la dépendance des personnes âgées.

1.1.1 Le risque de perte d'autonomie des personnes âgées

L'espérance de vie des hommes et des femmes a connu une augmentation significative au cours des dernières années grâce aux progrès de la médecine. Alors que, jusqu'aux années 1970, ces avancements étaient principalement axés sur la réduction de la mortalité infantile, ils ont contribué à réduire la mortalité aux âges avancés ces dernières années. En effet, selon les données issues de INSEE (2024), tandis que l'espérance de vie des hommes à la naissance a augmenté de 17% entre 1970 et 2023, leur espérance de vie résiduelle à 60 ans a enregistré une hausse de 46%. Ce même phénomène est observé chez les femmes dont l'espérance de vie a augmenté de 13% sur la même période, tandis que leur espérance de vie résiduelle à 60 ans a augmenté de 34% .

1. Société Commerciale de Réassurance

Nous faisons ainsi face ces dernières années à un vieillissement inévitable de la population dans les pays développés. Celui-ci s'explique par l'allongement de l'espérance de vie, couplé à la diminution des taux de fertilité dans le monde, et à l'avancée en âge de la génération Baby Boom née après la Seconde Guerre mondiale. Selon les projections de United Nations (2019), environ une personne sur six dans le monde sera âgée de plus de 65 ans en 2050, alors que ce ratio était d'une personne sur 11 en 2019.

Par ailleurs, l'espérance de vie augmente plus vite que l'espérance de vie en bonne santé, ayant pour conséquence une augmentation de la durée de vie en mauvaise santé des individus (DREES, 2018). Ainsi, le nombre de personnes âgées en situation de perte d'autonomie, également appelée état de dépendance, ne cesse d'augmenter.

La perte d'autonomie peut être liée à une multitude de pathologies, souvent liées au vieillissement, parmi lesquelles on retrouve le plus fréquemment la maladie d'Alzheimer et autres démences, le cancer, les maladies cardiovasculaires ou encore les maladies neurologiques telles que la maladie de Parkinson. Selon Fondation Vaincre Alzheimer (2019), environ 1 million de personnes sont touchées par la maladie d'Alzheimer en France, ce qui correspond à environ 8% des français de plus de 65 ans en 2020. Et cet effectif risque d'augmenter, pouvant atteindre 1.8 million d'individus atteints en 2050 selon France Alzheimer (2019). Ces personnes âgées n'ont parfois pas d'enfants, ou ces derniers vivent très loin, ce qui les oblige à recourir à une aide extérieure. Selon cette même source, cette perte d'autonomie engendre un coût annuel moyen par individu de 22 099 €, représentant ainsi un fardeau financier pour les individus et leurs familles.

L'augmentation des cas de dépendance concerne à la fois les hommes et les femmes, mais il est important de noter que, comme pour le risque de mortalité, les deux sexes ne sont pas égaux face à la dépendance. Ainsi, comme le montre INSEE (2023), à tout âge au-delà de 65 ans, la part de femmes en situation de perte d'autonomie est plus importante que celle des hommes. Ce phénomène sera abordé plus en détail dans la Section 1.2.4.

1.1.2 La prise en charge de la dépendance en France : le rôle de la Sécurité sociale

Depuis 2002, la France prévoit une aide publique permettant de couvrir une partie des frais liés à la perte d'autonomie des personnes âgées de plus de 60 ans. Il s'agit de l'Allocation Personnalisée d'Autonomie (APA), décrite sur le site du gouvernement en partenariat avec la Caisse Nationale de Solidarité pour l'Autonomie (CNSA, 2023). Cette aide, versée par le conseil départemental, existe sous 2 formes : l'APA à domicile, et l'APA en établissement. La première permet de financer les aides nécessaires au maintien à domicile (aide pour les soins, aide pour les repas, etc.), tandis que la seconde permet de financer une partie des frais en EHPAD (Établissement d'Hébergement pour Personnes Agées Dépendantes).

Trois conditions sont nécessaires pour bénéficier de cette aide de la Sécurité sociale :

- être âgé de plus de 60 ans,
- être dans une situation de perte d'autonomie, évaluée à l'aide de la grille AGGIR (Autonomie Gérontologie Groupe Iso Ressources), et
- résider en France de façon stable et régulière.

La grille AGGIR, créée conjointement par les médecins de la Sécurité sociale et la Société Française de Gérontologie en 1997, est décrite sur le site officiel de l'administration française (Service Public, 2024). Elle définit 6 niveaux de dépendance basés sur 10 variables représentatives de la perte d'autonomie physique et psychique, dites "discriminantes" :

1. Communiquer verbalement et/ou non verbalement, agir et se comporter de façon logique, et sensée par rapport aux normes admises par la société ;
2. Se repérer dans l'espace et le temps ;
3. Faire sa toilette ;
4. S'habiller, se déshabiller ;
5. Se servir et manger ;
6. Assurer l'hygiène de l'élimination urinaire et fécale ;
7. Se lever, se coucher, s'asseoir, passer de l'une de ces 3 positions à une autre ;
8. Se déplacer à l'intérieur du lieu de vie ;
9. Se déplacer en dehors du lieu de vie ;
10. Utiliser un moyen de communication à distance (téléphone, sonnette, etc.) dans le but d'alerter en cas de besoin.

Chacune de ses variables est évaluée selon une échelle comportant 3 niveaux (A, B ou C), basée sur 5 adverbes. La personne est-elle capable de faire l'action :

1. Seule ;
2. Spontanément ;
3. Totalement ;
4. Correctement ;
5. Habituellement ?

Un algorithme donnant des poids différents à chacune des variables, détermine ensuite le Groupe Iso Ressources du demandeur de l'APA (GIR 1, 2, 3, 4, 5 ou 6). Une description de chaque groupe est donnée par le Tableau 1.1. Seuls les GIR 1 à 4 ouvrent droit à l'allocation personnalisée d'autonomie.

GIR	Degrés de dépendance
GIR 1	- Personne confinée au lit ou au fauteuil, dont les fonctions mentales sont gravement altérées et qui nécessite une présence indispensable et continue d'intervenants - Ou personne en fin de vie
GIR 2	- Personne confinée au lit ou au fauteuil, dont les fonctions mentales ne sont pas totalement altérées et dont l'état exige une prise en charge pour la plupart des activités de la vie courante, - Ou personne dont les fonctions mentales sont altérées, mais qui est capable de se déplacer et qui nécessite une surveillance permanente
GIR 3	Personne ayant conservé son autonomie mentale, partiellement son autonomie locomotrice, mais qui a besoin quotidiennement et plusieurs fois par jour d'une aide pour les soins corporels
GIR 4	- Personne n'assumant pas seule ses transferts, mais qui, une fois levée, peut se déplacer à l'intérieur de son logement, et qui a besoin d'aides pour la toilette et l'habillage, - Ou personne n'ayant pas de problèmes locomoteurs, mais qui doit être aidée pour les soins corporels et les repas
GIR 5	Personne ayant seulement besoin d'une aide ponctuelle pour la toilette, la préparation des repas et le ménage
GIR 6	Personne encore autonome pour les actes essentiels de la vie courante

Table 1.1 – Caractéristiques des catégories GIR - (Service Public, 2024)

Bien qu'il n'y ait aucune condition de revenus pour bénéficier de cette aide, son montant dépend du niveau de revenus de l'individu. Le montant mensuel de l'APA à domicile ne peut excéder une limite dépendant du GIR, telle qu'indiquée dans le Tableau 1.2.

GIR	Montant mensuel maximum
GIR 1	1 914,04 €
GIR 2	1 547,93 €
GIR 3	1 118,61 €
GIR 4	746,54 €

Table 1.2 – Montants mensuels maximum de l'APA à domicile en fonction du GIR - (Service Public, 2024)

Toujours selon le site Service Public (2024), le reste à charge d'une personne percevant l'APA à domicile est de 90% pour des montants de ressources mensuelles supérieurs à 3 233,10 €.

Selon l'étude de la CNSA de 2019 (CNSA, 2019), le coût médian d'une chambre seule en

EHPAD pour une dépendance légère en région parisienne est de 2 004 €, atteignant même plus de 3 200 € à Paris et dans les Hauts-de-Seine. En cas de dépendance plus lourde, il faut ajouter à cela des frais journaliers médians de dépendance atteignant 22,50 €. Pour ces personnes, l'APA finance une partie du "tarif dépendance" de l'établissement concerné, en fonction du niveau de perte d'autonomie (GIR) du bénéficiaire. Cependant, l'essentiel des dépenses en établissement concerne le "tarif hébergement". Malgré l'allocation logement couvrant une partie de ce dernier, le reste à charge des bénéficiaires de l'APA en établissement reste très élevé. Selon DREES (2022), en 2019, le reste à charge moyen des résidents était de 1 957 € par mois avant la prise en compte de l'aide sociale à l'hébergement (ASH) dont une description détaillée est donnée sur le site officiel de l'administration française (Service Public, 2024).

La couverture de la Sécurité sociale est donc loin d'être suffisante pour couvrir les frais réels liés à la perte d'autonomie. Selon l'enquête d'opinion "Les Français et la dépendance" réalisée par Opinionay pour France Assureurs en 2021 (OpinionWay, 2021), plus de 2 français sur 5 ont été confrontés à des problèmes de dépendance liés à l'âge, personnellement ou via leur entourage. Environ un tiers d'entre eux affirment que cette perte d'autonomie a eu des conséquences sur leur foyer sur le plan financier. Il s'agit en effet d'une des plus grandes menaces pour la richesse d'une personne âgée en raison de l'insuffisance de la prise en charge publique. Les assureurs privés ont ainsi développé des produits spécifiques permettant de couvrir une partie des frais engendrés par la perte d'autonomie, en venant compléter l'aide versée par la Sécurité sociale.

1.1.3 Qu'est ce qu'un produit d'assurance dépendance ?

Un produit d'assurance est matérialisé par un contrat entre l'assureur et une ou plusieurs personnes, dans lequel l'assureur s'engage, en contrepartie du versement d'une prime d'assurance, à couvrir la survenue d'un risque aléatoire (Serge Braudo, 2024). Dans le cadre de l'assurance dépendance, le risque est la perte d'autonomie d'un individu ainsi que sa durée. Il peut s'agir d'un contrat individuel ou d'un contrat de groupe.

On retrouve deux familles de produits d'assurance dépendance : les produits avec prime nivelée viagère, utilisés pour les contrats individuels, et les produits avec prime de risque, principalement utilisés dans le cadre de contrats de groupes. En cas de contrat avec prime viagère nivelée, l'assuré est couvert viagèrement. En cas d'arrêt des paiements et au-delà d'une durée minimum de cotisation, souvent fixée à 8 ans, l'individu bénéficie d'une couverture réduite. La cotisation est fixée en début de contrat. Celle-ci est cependant révisable annuellement en fonction des résultats techniques du contrat et de l'évolution du risque de dépendance.

Dans le cadre d'un contrat avec prime de risque, l'adhérent verse en début d'année une

prime, le couvrant contre le risque de perte d'autonomie dans l'année en cours. L'assuré est ainsi couvert tant qu'il paye la cotisation. Dès lors que l'individu arrête le versement de primes, la couverture cesse quel que soit le nombre d'années cotisées. La cotisation est valable seulement pour l'année en cours, et est révisée annuellement en fonction de l'évolution du risque. Dans le cadre d'un produit de groupe, la prime est identique pour tous les membres et prend en compte la démographie du groupe.

Le risque étant aléatoire par définition, les individus ne peuvent être dépendants au moment de la souscription. Seuls ceux considérés autonomes par l'assureur peuvent s'assurer.

En France, la majorité des produits d'assurance dépendance individuels fonctionnent sur le principe suivant :

- L'assuré verse une prime annuelle ou mensuelle tant qu'il est autonome.
- En cas de perte d'autonomie avérée par l'assureur, il cesse de payer la prime et perçoit à la place une rente viagère annuelle ou mensuelle.
- Certains produits proposent, en plus de la rente, le versement d'un capital versé à la survenue de la perte d'autonomie. Celui-ci permet notamment de couvrir une partie des frais liés à l'adaptation du logement de l'assuré. Un service d'assistance peut également être proposé dans le contrat.

Dans certains pays, le versement de la prime et/ou de la rente peuvent être limités dans le temps.

1.1.4 La dépendance : un risque difficile à estimer

Tout comme l'APA, les produits d'assurance couvrant le risque de perte d'autonomie se sont essentiellement développés au début des années 2000. La perte d'autonomie intervient en majorité à un âge avancé, en moyenne à 83 ans selon le Ministère de la Santé et de la Prévention (Ministère de la Santé et de la Prévention, 2021). L'âge moyen à la souscription est quant à lui compris entre 60 et 64 ans (AG2R, 2023), correspondant à l'âge moyen de départ à la retraite. Par conséquent, les assureurs disposent à l'heure actuelle de peu de données aux âges élevés, rendant difficile l'estimation du risque.

Par ailleurs, les produits d'assurance dépendance constituent pour les assureurs un engagement à long terme. L'essentiel des contrats proposés en France garantissent, en cas de perte d'autonomie permanente et irréversible, le versement d'une rente jusqu'au décès. L'assureur doit par conséquent être vigilant sur l'estimation à la fois du risque de perte d'autonomie, ainsi que du risque de longévité des personnes dépendantes. En tant que réassureur, SCOR dispose de données provenant de plusieurs assureurs ayant des définitions parfois légèrement différentes. Il est donc difficile d'agréger différentes sources de données pour augmenter le volume des observations afin d'améliorer l'estimation du risque.

1.2 Description des produits d'assurance dépendance

Pour éviter les conséquences financières liées à la perte d'autonomie, les assureurs proposent des contrats prévoyant le versement d'un capital ou d'une rente en cas de dépendance avérée. Les caractéristiques principales de ces produits d'assurance dépendance ne sont pas uniformes entre les pays, ni même d'un assureur à un autre au sein d'un même marché.

1.2.1 La définition de la dépendance dans les contrats d'assurance

En France, les assureurs définissent la dépendance comme la perte permanente et définitive de l'autonomie. Il s'agit donc d'un état irréversible, se traduisant par la difficulté ou l'incapacité à effectuer seul des actes de la vie quotidienne tels que se déplacer, se nourrir, s'habiller, assurer son hygiène, etc. Comme pour l'allocation personnalisée d'autonomie (APA), les produits d'assurance dépendance privés définissent plusieurs degrés de perte d'autonomie, appelés dépendance légère, partielle ou totale. Ce degré est déterminé en fonction d'une grille choisie par l'assureur et définie dans le contrat.

En France, on distingue essentiellement 2 grilles. La grille AGGIR, également utilisée par la Sécurité sociale, et la grille AVQ souvent utilisée par les assureurs dans de nombreux pays.

Cette dernière se base sur l'évaluation du nombre d'Actes de la Vie Quotidienne (AVQ) dont la réalisation ne peut être effectuée même avec l'aide d'équipements adaptés, et nécessite l'aide d'une tierce personne de manière permanente pour l'effectuer. Plusieurs grilles fondées sur les AVQ existent selon le nombre d'actes de la vie quotidienne pris en compte par l'assureur. On retrouve essentiellement des définitions basées sur 4, 5 ou 6 AVQ qui sont les suivants :

- Faire sa toilette,
- Se déplacer,
- S'alimenter,
- Être continent,
- S'habiller,
- Effectuer ses transferts.

Le degré de dépendance est défini par le nombre d'actes que l'individu ne peut réaliser seul. La notation suivante est utilisée :

$$\begin{pmatrix}\text{Nombre d'actes ne pouvant}\\ \text{être réalisés seul}\end{pmatrix} \text{AVQ} \begin{pmatrix}\text{Nombre d'actes considérés}\\ \text{dans l'évaluation}\end{pmatrix}$$

Dans le cadre de la définition selon 4 AVQ, définition la plus utilisée, la dépendance totale

est déclarée à partir de 3 AVQ 4. L'individu est considéré comme partiellement dépendant à partir de 2 AVQ 4.

Un test complémentaire est généralement pratiqué par un neurologue ou un psychiatre pour les dépendances d'origine neuropsychiatrique. On retrouve notamment le "Mini Mental State Examination" MMSE introduit par Folstein (Folstein et al., 1975), consistant en une série de 30 questions permettant d'évaluer le niveau des troubles cognitifs du patient.

Il est important de noter qu'être dépendant selon la grille AGGIR n'implique pas la dépendance selon les AVQ et réciproquement. Ces deux définitions sont indépendantes. Par conséquent, un individu peut percevoir l'APA, mais ne pas être considéré comme dépendant par son assureur. Le versement de l'APA n'implique pas le versement de la rente de dépendance de l'assureur. Un assuré peut également percevoir une rente ou un capital via le contrat d'assurance, sans percevoir l'APA.

Le Tableau 1.3 compare les caractéristiques des deux grilles d'évaluation de la dépendance.

AGGIR	**AVQ**
Chaque variable (parfois découpée en sous-variables) possède 3 modalités (A, B ou C)	Système binaire, soit la personne peut effectuer l'AVQ soit elle ne peut pas
Pondération des actes	Actes équipondérés
Algorithme non transposable dans une notice	Plus lisible dans un contrat d'assurance
Bonne complétude	À compléter par le MMSE pour l'appréhension des dépendances d'origine neuropsychiatrique
Grille utilisée à l'échelle nationale	Grille utilisée dans de nombreux pays

TABLE 1.3 – Comparaison de la grille AGGIR et de la grille selon les AVQ

1.2.2 Caractéristiques de la souscription

Nous présentons dans cette section les caractéristiques principales des produits d'assurance dépendance, pouvant varier d'un contrat à un autre. Cette hétérogénéité rend complexe l'agrégation des données provenant de différents assureurs. Le Chapitre 3 présente une solution pour répondre à cette problématique.

Sélection médicale

Lors de la souscription d'une assurance individuelle, une étape de sélection médicale permet de s'assurer que l'individu est bien autonome, et éventuellement d'appliquer une surprime en cas de risque aggravé. Cette sélection, s'appliquant aux individus de plus de

50 ans, s'effectue généralement en deux étapes. Dans un premier temps, une déclaration de santé simple et rapide est remplie par le proposant. Celle-ci a pour but de détecter les éventuelles invalidités préexistantes ainsi que les traitements chroniques. Après analyse de ce questionnaire, le médecin-conseil de l'assureur peut, dans un second temps, soumettre le demandeur à un questionnaire plus détaillé ou lui imposer de passer une visite médicale. Cette deuxième étape, contenant des questions plus spécifiques, s'intéresse aux antécédents familiaux, aux pathologies existantes, ainsi qu'à l'environnement de l'individu. Les individus de moins de 50 ans ayant une invalidité ou Affection Longue Durée (ALD), sont également soumis à l'étape de la sélection médicale.

À la suite de la sélection médicale, l'assureur décide au choix : d'accepter la souscription au tarif de base, d'appliquer une surprime, ou de refuser la souscription.

Cette étape de sélection médicale implique ainsi l'état d'autonomie de l'individu à la date de souscription. Ainsi, on peut penser que l'ancienneté du contrat a un impact sur les intensités de transition vers la dépendance ou le décès. Il pourrait être intéressant de définir des intensités conditionnelles à l'âge de souscription. Cependant, cet effet s'atténue avec le temps depuis la souscription. Nous faisons l'hypothèse dans tout le manuscrit que l'âge de souscription n'impacte pas les intensités de transition. Ainsi, cet âge n'est pas pris en compte dans les modèles.

Limite d'âge

Les produits d'assurance individuels prévoient dans la plupart des cas une limite d'âge à la souscription. Celle-ci est notamment fixée à 74 ans en France chez AG2R La Mondiale ou encore La Banque Postale Prévoyance (LBPP (2020), AG2R La Mondiale (2019)). Predica, assurance du Crédit Agricole, autorise la souscription jusqu'à 75 ans inclus (PREDICA, 2018). En revanche, on ne retrouve aucune limite d'âge pour les contrats collectifs.

Délai de carence

Afin de limiter l'antisélection, certains contrats individuels prévoient un délai de carence. Ce délai correspond à une durée notée c suivant la souscription du produit durant laquelle l'entrée en dépendance provoque la nullité du contrat. Dans la majorité des produits d'assurance dépendance, cette durée c dépend de la cause de perte d'autonomie. Ainsi, l'assureur AG2R La Mondiale fixe un délai de carence d'un an en cas de maladie somatique, et de 3 ans en cas de démence, maladie d'Alzheimer et autres pathologies neurologiques et psychiatriques. En revanche, aucun délai de carence n'est appliqué en cas de perte d'autonomie en raison d'un accident (AG2R La Mondiale, 2019).

Franchise

Les produits d'assurance dépendance prévoient le versement d'une rente jusqu'au décès. Dans certains cas, la durée entre la perte d'autonomie et le décès est relativement courte,

impliquant des frais de dossiers élevés pour une faible somme versée à l'assuré. Par ailleurs, partant du constat que la majorité des sinistres courts peuvent être couverts par les économies des assurés, certains assureurs prévoient dans le contrat une période de franchise. Il s'agit d'une durée suivant l'entrée en dépendance durant laquelle la rente n'est pas versée. Dans la majorité des cas, celle-ci est inférieure à 3 mois. Cette franchise permet également de réduire le coût du produit pour l'assuré.

1.2.3 Tarification et provisionnement des produits d'assurance dépendance

On considère dans cette section un produit d'assurance dépendance simple versant, en cas de perte d'autonomie, un montant de rente annuelle R identique quel que soit le degré de dépendance de l'assuré. Les rentes sont versées en fin de période (annuité à terme échue), seulement si l'assuré est dépendant à la date de versement. Les primes sont versées en début de période (à terme d'avance).

La valorisation actuarielle d'un engagement nécessite l'estimation des flux futurs probables de trésorerie. Pour cela, chaque flux futur est pondéré par sa probabilité de survenance. Ceux-ci sont ensuite actualisés afin de refléter la valeur temps de l'argent. On parle alors de valeur actuelle probable (VAP). Par souci de simplification, on considère comme Dupourqué et al. (2019), un taux d'actualisation constant i, et on introduit le coefficient d'actualisation $\nu = \dfrac{1}{1+i}$.

1.2.3.1 Tarification

On appelle prime d'assurance la somme versée par l'assuré à l'assureur en contrepartie de la couverture face au risque. Ce montant est calculé pour chaque assuré, afin de refléter au mieux son risque. Contrairement à la majorité des activités commerciales, l'assurance est basée sur le principe de l'inversion du cycle de production. L'assureur fixe le prix du contrat avant de connaître son coût réel de revient. Il s'appuie alors sur des outils mathématiques afin d'évaluer le coût futur de celui-ci. Le calcul de la prime dépend de probabilités de passage d'un état de santé à un autre, appelées lois de transition. Dans le contexte d'un produit d'assurance dépendance, la tarification fait intervenir à la fois une loi d'incidence reflétant la probabilité de perte d'autonomie, ainsi que des lois de mortalités. Les probabilités de décès étant plus élevées pour les dépendants, la tarification nécessite l'utilisation de lois distinctes de mortalité selon l'état de santé de l'individu : autonome ou dépendant. Par ailleurs, nous verrons dans la suite qu'il est très souvent admis que la durée depuis la perte d'autonomie t d'un individu a un effet significatif sur sa probabilité de décès au cours de l'année. Cette durée t est également appelée "duration" dans l'ensemble de cette thèse.

Pour estimer le montant de prime d'un contrat souscrit par un individu d'âge x, l'assureur estime :

- la valeur actuelle probable de son engagement, notée Π_x, et
- la valeur actuelle probable de l'engagement de l'assuré, notée $\Pi_x^{'}$.

On appelle prime pure du contrat le montant permettant d'égaliser l'engagement de l'assureur et celui de l'assuré. La prime d'assurance ou cotisation prend quant à elle en compte : la prime pure, les frais de gestion, les taxes, les frais de commercialisation, ainsi que les charges financières.

Nous nous concentrons tout d'abord sur l'estimation de la prime pure, dépendant de l'âge à la souscription x.

Cas d'un produit avec prime viagère nivelée

La valeur actuelle probable de l'engagement de l'assureur Π_x est donnée par la formule suivante :

$$\Pi_x = \int_0^{+\infty} \nu^u \underbrace{{}_up_x^a \, \lambda_{x+u}}_{\substack{\text{Probabilité de perte} \\ \text{d'autonomie à l'âge} \\ \text{exact } x+u}} R\mathrm{a}_{x+u,0}^d \, \mathrm{d}u, \tag{1.1}$$

où :

- R est le montant annuel de rente,
- λ_x représente le risque instantané de perte d'autonomie d'un individu d'âge x. Cette grandeur, couramment utilisée dans l'étude des modèles de survie, est introduite plus en détail dans la Section 1.4.2.
- ${}_up_x^a$ est la probabilité de survie dans l'état d'autonomie à u années d'un individu d'âge x,
- $\mathrm{a}_{x,t}^d$ représente la valeur actuelle probable d'une rente annuelle de 1€ versée à un individu d'âge exact x, dépendant depuis t années, et ce jusqu'à son décès.
- ${}_kp_{x,t}^d$ est la probabilité de survie à k années d'un individu d'âge x, dépendant depuis t années.

Dans le cas particulier où la durée depuis la perte d'autonomie t est égale à 0 ou correspond à une date de versement de rente (i.e $t \in \mathbb{N}$ en cas de versements annuels, $t = \frac{m}{12}, m \in \mathbb{N}$ en cas de versements mensuels), la formule de $\mathrm{a}_{x,t}^d$ est donnée par

$$\mathrm{a}_{x,t}^d = \begin{cases} \displaystyle\sum_{k=1}^{+\infty} \nu^k {}_kp_{x,t}^d & \text{en cas de versements annuels,} \\ \displaystyle\frac{1}{12}\sum_{k=1}^{+\infty} \nu^{k/12} {}_{k/12}p_{x,t}^d & \text{en cas de versements mensuels.} \end{cases} \tag{1.2}$$

Par souci de simplification, supposons le cas particulier où l'âge x de l'assuré correspond à l'âge de souscription ou à un âge exact de versement de cotisation. La prime étant versée en début de période par l'assuré sur une base annuelle ou mensuelle, la valeur actuelle probable de l'engagement de l'assuré $\Pi_x^{'}$ est donnée par

$$\Pi_x^{'} = P_x \ddot{a}_x^a, \tag{1.3}$$

où P_x représente le montant de prime pure, et

$$\ddot{a}_x^a = \begin{cases} \displaystyle\sum_{k=0}^{+\infty} \nu^k {}_k p_x^a & \text{en cas de versements annuels,} \\ \displaystyle\frac{1}{12} \sum_{k=0}^{+\infty} \nu^{k/12} {}_{k/12} p_x^a & \text{en cas de versements mensuels.} \end{cases} \tag{1.4}$$

$\ddot{a}_x^a$ représente la valeur actuelle probable d'une rente annuelle de 1€ versée à terme d'avance par un individu autonome d'âge exact x, et ce jusqu'à son décès.

En pratique, l'assureur peut être amené à évaluer la valeur actuelle probable de son engagement à une durée exacte t ne correspondant pas toujours à une date de versement de rente. De la même manière, l'estimation de la valeur actuelle probable de l'engagement de l'assuré peut se faire à un âge exact x ne correspondant pas à une date de versement de prime. La remarque suivante présente la généralisation des Équations 1.2 et 1.4 aux âges x et durations t exactes.

Remarque : Généralisation des VAP aux âges x et durations t exacts

Notons f la fréquence des versements (i.e $f = 1$ en cas de versements annuels, $f = 12$ en cas de versements mensuels, etc.). Soit x_0 représentant l'âge de souscription pour la valorisation de la prime $\ddot{a}_x^a$, ou l'âge d'entrée en dépendance pour la valorisation de la rente $a_{x,t}^d$. On considère ici le cas de la valorisation à un âge x ne correspondant pas nécessairement à une date de versement. La Figure 1.1 illustre les flux probables de trésorerie futurs. Par mesure de simplification, seuls les 3 prochains flux sont représentés.

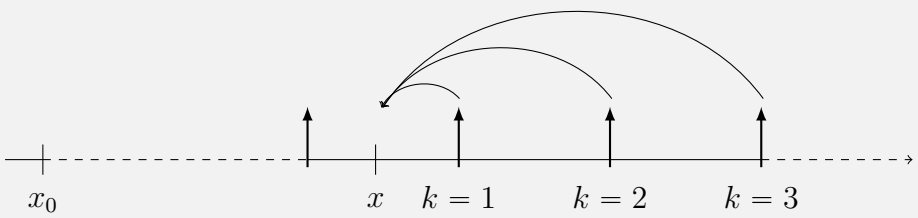


FIGURE 1.1 – Illustration des flux de trésorerie

Remarque : Généralisation des VAP aux âges x et durations t exacts (suite)

On peut facilement montrer que l'âge au dernier versement avant x est égal à

$$x_0 + \frac{\lfloor (x - x_0)f \rfloor}{f},$$

où $\lfloor \cdot \rfloor$ est la fonction partie entière.
Ainsi, on note $\tau(k) = \dfrac{\lfloor (x - x_0)f \rfloor + k}{f} - (x - x_0)$, le temps avant le k-ième versement.
Les valeurs actuelles probables sont alors données par

$$\mathrm{a}^d_{x,t} = \frac{1}{f} \sum_{k=1}^{+\infty} \nu^{\tau(k)}{}_{\tau(k)}p^d_{x,t}, \quad \text{et}$$
$$\ddot{\mathrm{a}}^a_x = \frac{1}{f} \sum_{k=0}^{+\infty} \nu^{\tau(k)}{}_{\tau(k)}p^a_x.$$

Dans le cas de la valorisation de la rente en dépendance, on remarquera que $x - x_0 = t$.

Par conséquent, la prime pure pour une souscription à l'âge x est égale à

$$P_x = \frac{\Pi_x}{\ddot{\mathrm{a}}^a_x}, \tag{1.5}$$

où $\ddot{\mathrm{a}}^a_x$ est donné par l'Équation 1.4 car la valorisation est effectuée à l'âge x de souscription.

Cas d'un produit avec prime de risque

Dans le cas d'un produit avec prime de risque, la prime ne couvrant que le risque de perte d'autonomie pendant l'année en cours, les formules d'engagement de l'assureur et de l'assuré sont simplifiées. L'assuré ne verse qu'une prime unique en début d'année. La VAP de son engagement est donc égale au montant de la prime. Par conséquent,

$$P_x = \Pi_x = \int_0^1 \nu^u \lambda_{x+u}\ \mathrm{a}^i_{x+u,0}\ \mathrm{d}u. \tag{1.6}$$

Impact des caractéristiques du contrat

Le tableau suivant présente l'impact sur les formules de tarification de quelques clauses fréquentes des contrats d'assurance dépendance, telles que le délai de carence, la franchise, ou encore le versement d'un capital K au moment de l'entrée dans l'état de dépendance.

GIR	Impact sur les engagements de l'assureur ou de l'assuré
Délai de carence c	$\Pi_x = \int_c^{+\infty} \nu^u {}_up_x^a \ \lambda_{x+u} \ R\mathrm{a}_{x+u,0}^d \ \mathrm{d}u$
Franchise fr **(en mois)**	$\mathrm{a}_{x,t}^d = \begin{cases} \sum_{k=1}^{+\infty} \nu^{k+fr/12} \ {}_{k+fr/12}p_{x,t}^d & \text{en cas de versements annuels,} \\ \frac{1}{12} \sum_{k=fr+1}^{+\infty} \nu^{k/12} \ {}_{k/12}p_{x,t}^d & \text{en cas de versements mensuels.} \end{cases}$
Versement d'un capital K **lors de l'entrée en dépendance**	$\Pi_x = \int_0^{+\infty} \nu^u {}_up_x^a \ \lambda_{x+u} \ \left(R\mathrm{a}_{x+u,0}^d + K\right) \mathrm{d}u$

TABLE 1.4 – Clauses fréquentes des produits d'assurance dépendance et leurs impacts sur la tarification

De la prime pure à la prime commerciale

La prime pure reflète uniquement le coût de l'engagement de l'assureur envers le bénéficiaire. Elle permet par définition de couvrir la valeur estimée des sinistres. Cependant, par la vente de produits d'assurance dépendance, l'assureur est également soumis à des frais de commercialisation du produit, tels que des frais de courtages, des frais de gestion de rentes, des charges financières ainsi que des taxes. Ceux-ci sont reportés sur la valeur de la prime réellement payée par l'assuré, appelée prime commerciale et notée P_x^c. Nous considérons ici seulement 2 chargements : une commission g (en %) prélevée sur le montant de prime commerciale pour financer les frais de commercialisation, ainsi que des frais de gestion de rente r (en %) appliqués au montant de rente. Le passage de la prime pure à la prime commerciale est donné par la formule suivante :

$$P_x^c = \frac{1+r}{1-g} P_x. \tag{1.7}$$

1.2.3.2 Provisionnement

Au delà de la tarification, l'assureur doit évaluer la valeur de ses engagements vis-à-vis de ses assurés afin de constituer des provisions suffisantes pour pouvoir les honorer en cas de sinistre. On présente ici les deux principales provisions constituées par l'assureur afin d'honorer ses engagements vis-à-vis des assurés.

La provision mathématiques ou provision pour risques croissants

La provision mathématique (PM), dont la définition est donnée par l'article R 343-3

du Code des Assurances (Dalloz, 2015), est la "différence entre les valeurs actuelles des engagements respectivement pris par l'assureur et par les assurés". L'intérêt de cette provision est de faire face à l'augmentation du risque avec l'âge, ce qui explique sa deuxième dénomination : provision pour risques croissants (PRC). En effet, en cas de prime viagère nivelée, les primes payées les premières années sont supérieures au risque encouru à cette période. Elles permettent ainsi de compenser le risque plus élevé aux grands âges, période à laquelle la prime annuelle versée est inférieure au risque réel encouru. La PM est constituée pour les assurés autonomes, et est égale à la somme des provisions individuelles constituées pour chaque assuré.

Pour un assuré d'âge x, ayant souscrit à l'âge x_0 et ayant une prime pure P_{x_0}, la provision mathématique est égale à

$$PM_x^{x_0} = (\Pi_x - P_{x_0}\ddot{a}_x^a)(1+r), \tag{1.8}$$

où r désigne les frais de gestion sur le montant de rente.

Par définition de la prime pure, la provision mathématique est nulle à l'âge de souscription, i.e. $PM_{x_0}^{x_0} = 0$.

La provision pour sinistres à payer

La provision pour sinistres à payer (PSAP) est quant à elle constituée pour les assurés dépendants. Elle correspond à la valeur actuelle probable des dépenses futures liées aux sinistres en cours. Comme pour la PM, la provision pour sinistre à payer est égale à la somme de provisions individuelles.

Considérons un assuré d'âge x dépendant depuis t années. Par mesure de simplification, nous considérons ici le cas particulier où la duration t est de la forme $t \in \mathbb{N}$ en cas de versements annuels de la rente, ou $t = \frac{m}{12}, m \in \mathbb{N}$ en cas de versements mensuels. La provision pour sinistres à payer est donnée par la formule suivante :

$$PSAP_{x,t} = R\mathrm{a}_{x,t}^d(1+r). \tag{1.9}$$

Pour plus de détails sur la tarification et le provisionnement des produits d'assurance dépendance, le lecteur pourra se référer à Deléglise et al. (2009) et Dupourqué et al. (2019). Ces calculs, nécessaires à la création et au pilotage d'un produit d'assurance dépendance, font ainsi intervenir des lois biométriques de passage d'un état de santé à un autre. Il est donc important pour l'assureur de bien les estimer afin d'évaluer au mieux le tarif, ainsi que les provisions nécessaires pour garantir son engagement vis-à-vis des assurés.

1.2.4 Prise en compte du sexe de l'assuré

Malgré la réduction de l'écart d'espérance de vie à la naissance entre sexes entre le milieu des années 1990 et aujourd'hui (Meslé, 2004), l'espérance de vie des femmes reste encore à l'heure actuelle plus élevée que celle des hommes (Seifarth et al., 2012). Selon une étude plus récente de l'INSEE [2], les femmes vivent en moyenne 5,9 années de plus que les hommes (INSEE, 2023). Par conséquent, ces dernières sont plus impactées par les maladies liées au vieillissement, sources de perte d'autonomie.

Ainsi, au delà des différences face au risque de mortalité, hommes et femmes ne sont pas égaux face à la dépendance, comme le montre notamment une étude effectuée par l'INED [3] (Bonnet et al., 2011). Cette étude s'intéresse à l'espérance de vie à 65 ans avec et sans dépendance, en comparant notamment le cas des hommes et des femmes. Les résultats de cette recherche, représentés sur la Figure 1.2, montrent que les années supplémentaires de vie des femmes par rapport aux hommes sont essentiellement des années avec difficultés pour effectuer les tâches domestiques et les activités de soins corporels, correspondant donc à des années en situation de dépendance. Plus récemment, Crimmins et al. (2019) montre une forte hétérogénéité de la morbidité et de la dépendance selon le sexe. À partir de données d'individus de plus de 50 ans, provenant de plusieurs pays, cette étude compare les performances en terme de capacité à effectuer les activités instrumentales de la vie quotidienne (IAVQ) des hommes et des femmes. Quel que soit le pays, les femmes présentent un risque significativement plus élevé de perte d'autonomie que les hommes.

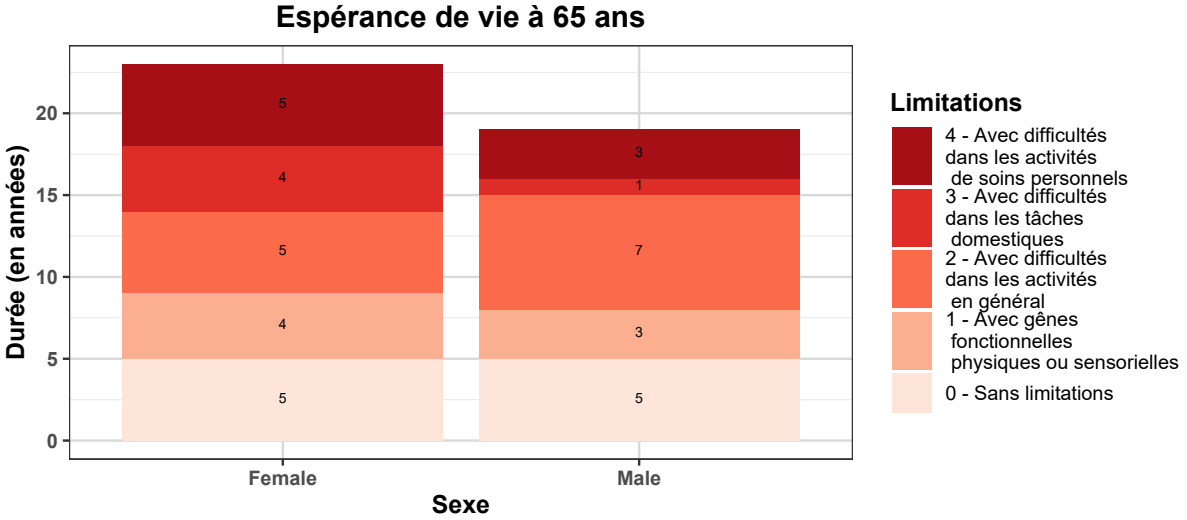


FIGURE 1.2 – Étude INED : Espérance de vie à 65 ans avec et sans limitations (en 2008)

2. INSEE : Institut National de la Statistique et des Études Économiques
3. INED : Institut National d'études Démographiques. Établissement de recherche public spécialisé dans l'étude des populations

La distribution des maladies responsables de la perte d'autonomie varie selon le sexe. On observe ainsi que le cancer, ayant une forte surmortalité en comparaison à celle associée aux maladies neurodégénératives, est plus représenté chez les hommes que chez les femmes. Ces dernières entrent en dépendance en majorité pour des maladies liées au vieillissement pour lesquelles la surmortalité par rapport à la population générale est moins marquée que celle du cancer. Ainsi, les femmes auront tendance à rester plus longtemps dans l'état de dépendance que les hommes. Par conséquent, la prime à l'équilibre devrait être plus élevée pour les femmes que pour les hommes.

Cependant, depuis l'arrêté du 18 décembre 2012 relatif à l'égalité entre les hommes et les femmes en assurance (Ministère de l'Economie et des Finances, 2012), les contrats d'assurance dépendance vendus dans l'Union Européenne ne peuvent plus appliquer une prime différente selon le sexe de l'assuré. L'assureur se trouve alors dans l'obligation de faire une hypothèse sur la distribution hommes/femmes dans son portefeuille d'assurés afin de déterminer la prime pure permettant d'équilibrer l'engagement de l'assureur et de l'assuré. Cette hypothèse de répartition est cruciale dans l'étape de tarification. Une sous-estimation de la proportion de femmes dans le portefeuille a pour conséquence une sous-estimation de la prime, et donc par conséquent de potentielles fortes pertes pour l'assureur ayant sous-estimé le risque. Rien n'empêche cependant l'assureur de prendre en compte le sexe de l'assuré pour le provisionnement, afin d'obtenir une estimation plus précise de ses réserves. Cet arrêté sur l'égalité entre hommes et femmes en assurance ne s'applique pas aux contrats d'assurance vendus aux États-Unis. Ceux-ci prévoient ainsi une prime plus élevée pour les femmes.

1.3 Données disponibles en assurance dépendance

Le risque de dépendance est un risque pour lequel peu de données sont disponibles. En pratique, en France, on retrouve principalement 2 sources : les données publiques de l'APA, et les données récoltées au sein des différentes compagnies d'assurance. Les données peuvent être soit à la maille individuelle soit agrégées. Dans le premier cas, la base contient à minima une ligne par individu. Dans le second cas, les individus disposant des mêmes caractéristiques sur les variables prises en compte dans l'étude sont regroupés en une seule ligne. L'ensemble des données utilisées dans cette thèse sont initialement à la maille individuelle.

1.3.1 Structure des données

Si ce n'est pas déjà le cas, les données peuvent aisément être séparées en plusieurs bases, permettant de refléter l'expérience dans chacun des états (une pour l'autonomie, et une par degré de dépendance selon le nombre de niveaux considérés dans l'étude). La base

des cotisants, notée BC retrace l'expérience des autonomes. La ou les base(s) de rentiers, notée(s) BR_k retrace(nt) l'expérience dans l'état de dépendance k.

Le début et la fin de la période d'observation sont notés τ_d et τ_f.

La base BC, permettant d'estimer le risque de décès des autonomes ainsi que le risque de perte d'autonomie, contient les informations suivantes :

- la date de naissance, notée DoB (Date of Birth) ;
- le sexe g ;
- la date de souscription s'il s'agit de données d'un assureur, notée DoU (Date of Underwriting) ;
- la date de fin d'observation dans l'état d'autonomie, notée DoE (Date of End), telle que

$$DoE = \min(\text{date de décès}, \text{date de perte d'autonomie}, \text{date de rachat}, \tau_f) ;$$

- la cause de fin d'observation c, qui peut être le décès, la perte d'autonomie, le rachat, ou la censure si l'individu est autonome à la date τ_f.

Le terme "rachat" est utilisé pour représenter l'arrêt de paiement des cotisations. Dans ce cas, l'assuré sort du portefeuille de l'assureur et son état de santé n'est plus suivi.

On note DoS (Date of Start) la date de début d'observation, pouvant différer de la date de souscription, telle que

$$DoS = \max(DoU, \tau_d).$$

Soit un individu présent dans la base de données des autonomes BC. La Figure 1.3 représente son expérience de sa naissance à sa sortie de l'état d'autonomie. La période d'observation (entre DoS et DoE) est mise en évidence en vert.

La ou les bases de rentiers BR_k, permettant d'estimer le risque de décès dans l'état de dépendance de niveau k, ainsi que le risque d'aggravation du niveau de dépendance, contient les informations suivantes :

- la date de naissance, notée DoB ;
- le sexe g ;
- la date d'entrée dans l'état de dépendance k, notée DoI (Date of Incidence) ;
- la date de fin d'observation dans l'état de dépendance k, notée DoE, telle que $DoE = \min(\text{date de décès, date d'entrée dans un niveau supérieur de dépendance}, \tau_f)$;
- la cause de fin d'observation c, qui peut être le décès, l'entrée dans un autre niveau de dépendance plus élevé, ou la censure si l'individu est dans l'état k à la date τ_f.

La date de début d'observation dans l'état k, notée DoS, peut différer de la date d'entrée dans cet état. Elle est donnée par

$$DoS = \max(DoI, \tau_d).$$

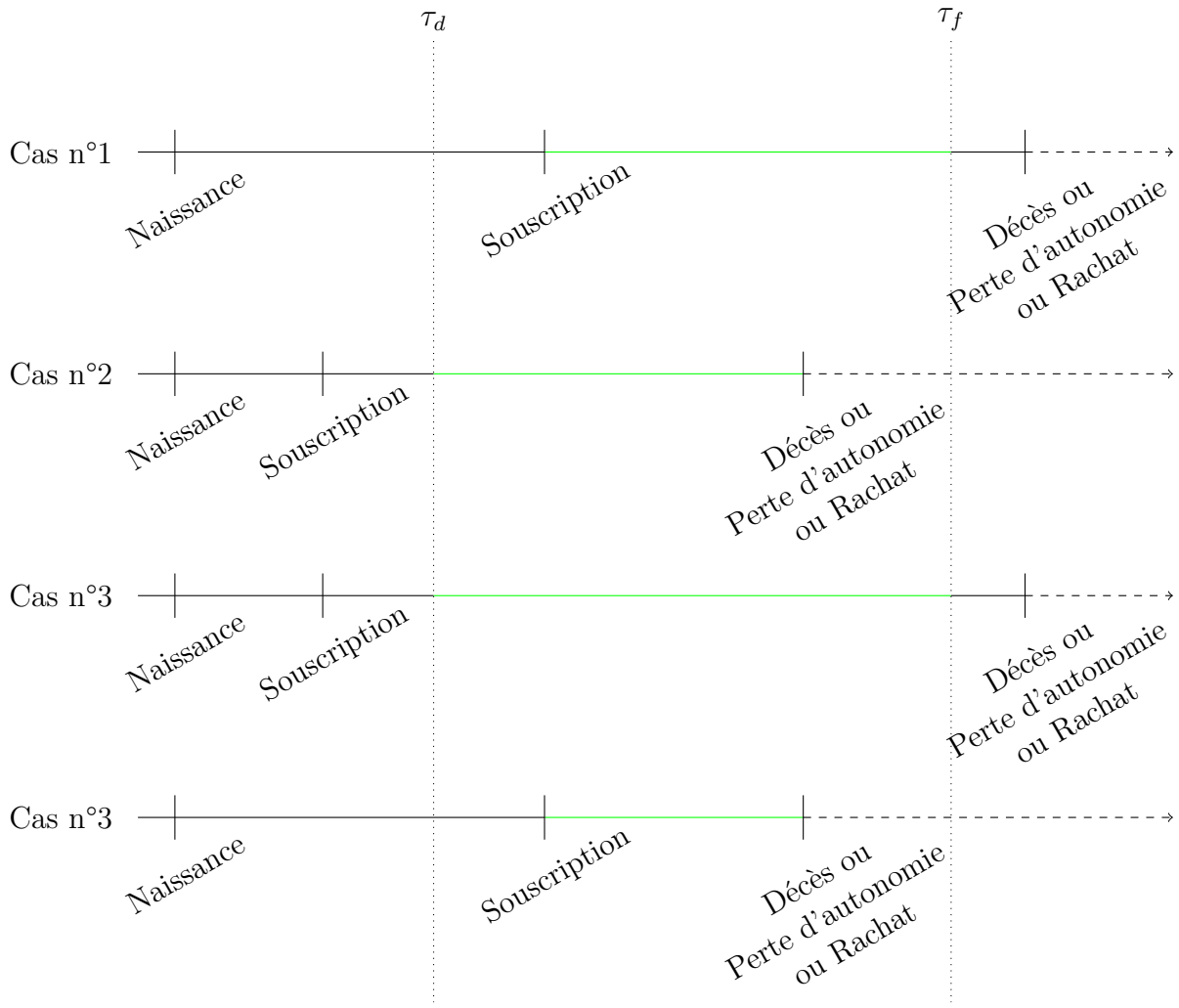


FIGURE 1.3 – Observation d'un individu dans l'état d'autonomie

On notera que l'on retrouve cette structure de données dans la thèse de Biessy (2016b) également effectuée au sein de SCOR.

Certains produits, notamment à l'étranger, considèrent comme possible l'amélioration de l'état de santé de l'individu, voire son rétablissement. Dans ce cas, la cause de sortie de la base BR_k peut également être l'amélioration de l'état de santé. C'est notamment le cas dans les données utilisées dans le Chapitre 4, décrites dans la Section 1.3.3.

1.3.2 Problématiques liées aux données

Lors de l'analyse de données, plusieurs problématiques peuvent émerger. Tout d'abord, l'information peut ne pas être complète du fait des phénomènes de troncatures et de censures. La date de début d'observation de l'individu notée DoS correspond en effet à une troncature à gauche. L'état de l'individu avant cette date n'est pas connu. Par ailleurs, la date de fin d'observation pouvant correspondre à un décès ou à une perte d'autonomie, peut également correspondre à une date de perte de suivi de l'état de santé de l'individu en cas de rachat, ou de fin de période d'observation si l'individu est vivant à la fin de la période de suivi. Dans ces deux derniers cas, on parle de censure à droite. L'état de l'individu après cette date est inconnu. La censure à droite peut être soit aléatoire, dans le cas d'un rachat par exemple, ou bien fixe, dans le cas de la fin de période d'étude τ_f. Pour plus de détails sur les phénomènes de troncatures et de censures dans l'analyse de données de survie, le lecteur pourra se référer à Kleinbaum and Klein (1996) et Huber-Carol (1994).

La présence d'une franchise dans le contrat d'assurance dépendance peut également poser des problèmes dans les données. Considérons un assuré décédant au cours de la période de franchise. Dans la grande majorité des cas, sa perte d'autonomie ne sera pas renseignée dans la base de données. Son décès est alors renseigné comme un décès en autonomie. Par conséquent, seuls les assurés survivant durant toute la période de franchise sont présents dans la base des rentiers, et aucun décès n'est observé pendant la période de franchise. Il s'agit ici d'une troncature à gauche.

La présence d'un délai de carence dans le contrat implique également une troncature à gauche. Pour rappel, le délai de carence dépend généralement de la pathologie responsable de la perte d'autonomie. L'entrée en dépendance n'étant possible qu'après le délai de carence écoulé, l'incidence est très faible les premiers mois suivant la souscription du contrat. Une perte d'autonomie au cours du délai de carence met fin au contrat, ayant pour conséquence une fin de suivi de l'état de santé de l'individu.

Une autre problématique à laquelle peut être confronté l'assureur dans l'analyse de ses données vient de la multitude de définitions. En effet, certains assureurs proposent au choix, de se couvrir uniquement contre la dépendance totale, ou bien contre la dépendance partielle et totale. Dans ce cas, la perte partielle d'autonomie n'est pas renseignée pour un

assuré couvert uniquement contre la perte totale d'autonomie. Au sein d'une même base de données, on se retrouve en présence d'assurés observés avec des définitions différentes de la dépendance. Dans cette situation, une des solutions consiste à séparer les observations en sous-bases de données homogènes en terme de définition. Les lois biométriques sont alors estimées indépendamment sur chacune d'entre elles. Dans un contexte de rareté des observations en assurance dépendance, cette solution ne semble pas viable pour les assureurs qui ne peuvent pas se permettre d'estimer le risque de manière indépendante pour chacune des définitions. Cette problématique est traitée dans le Chapitre 3 de cette thèse. Nous y proposons une méthodologie permettant de prendre en compte l'hétérogénéité des définitions de la dépendance entre contrats d'une même base de données.

Certains assurés observés au cours de la période d'étude peuvent avoir décidé de résilier leur contrat en arrêtant le paiement des primes. En France, selon l'ancienneté du contrat, deux situations sont possibles. Dans le premier cas, l'assuré met un terme au contrat avant 8 ans de cotisations. Il perd alors tous ses droits et ne perçoit aucune rente en cas de perte d'autonomie postérieure à la résiliation. L'état de santé de l'individu n'est plus suivi à partir de la résiliation. Il s'agit alors d'une censure à droite. Dans le deuxième cas, l'assuré met un terme à son contrat après 8 ans de cotisations. Il peut alors bénéficier d'une rente partielle en cas de perte d'autonomie après la résiliation. On parle de mise en réduction du contrat. Dans cette situation, l'état de santé de l'individu est encore suivi après résiliation. Il apparaît alors dans la base des cotisants jusqu'à son décès, sa perte d'autonomie, ou la fin de la période d'étude. Cependant, on remarque très souvent une sous-incidence des assurés dits "réduits". Cela peut en partie s'expliquer par un oubli de déclarer la perte d'autonomie. En effet, ces personnes souvent âgées, ainsi que leur famille, ne payant plus de cotisations, oublient l'existence de cette couverture partielle. À titre d'exemple, la Figure 1.4 utilisant les données utilisées dans le Chapitre 3 montre que seulement 0,8% des assurés réduits ont déclaré une perte d'autonomie, contre 3,8% des autres assurés. Il est important de prendre cette information en compte, au risque de sous-estimer le risque de perte d'autonomie du portefeuille. Une des solutions, optée dans le reste de cette thèse, est de traiter la mise en réduction comme une censure à droite. Ainsi, les assurés réduits ne sont plus suivis à partir de la date de mise en réduction.

Enfin, la pathologie n'ayant aucun impact sur le montant mensuel de la rente versée à l'assuré, elle n'est que très rarement enregistrée par les assureurs. Il est donc difficile d'obtenir des bases de données contenant cette information, bien qu'il ait été montré auparavant, notamment par Biessy (2016a), que celle-ci a un impact sur la durée moyenne de versement des rentes. Le provisionnement devrait donc prendre en compte la pathologie afin de mieux constituer notamment la provision pour sinistres à payer (PSAP).

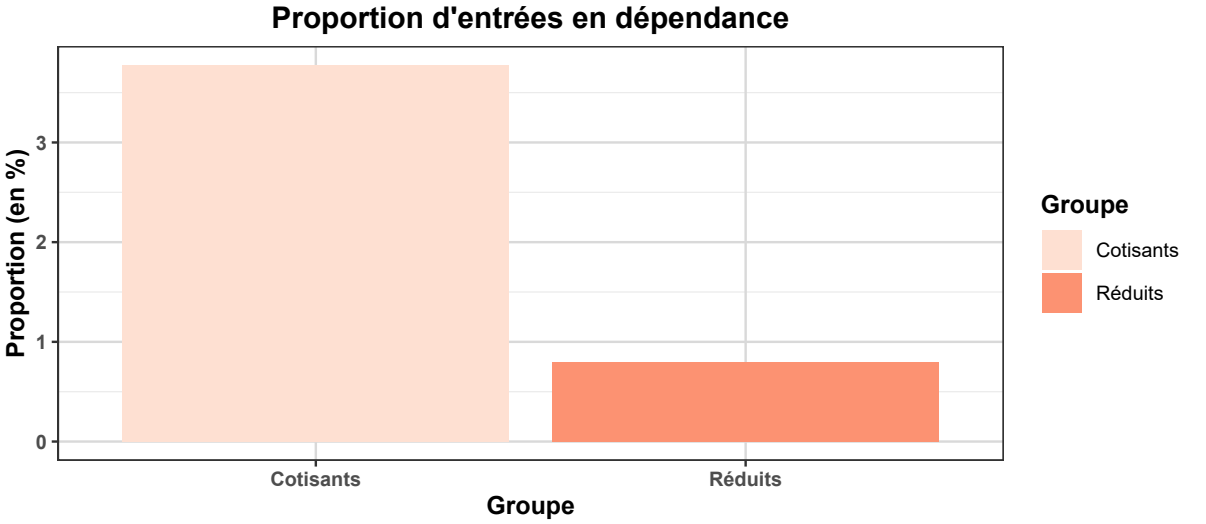


FIGURE 1.4 – Comparaison de l'incidence des assurés mis en réduction et des cotisants

1.3.3 Présentation des données utilisées dans le cadre de cette thèse

Cette thèse a mobilisé plusieurs jeux de données provenant de plusieurs assureurs. Tous ont nécessité une étape de nettoyage afin de s'assurer de l'homogénéité de la période d'observation, et de la cohérence et validité des dates d'évènements telles que les dates de souscription, de perte d'autonomie et de décès.

Le Chapitre 2 utilise une base de données provenant de 5 grands portefeuilles français d'assurance dépendance. L'ensemble des produits associés définissent la dépendance selon la grille AGGIR [4] décrite en Section 1.1.2. Malgré la souscription de certains individus à la couverture additionnelle contre la dépendance partielle, seule la dépendance totale (GIR12) est considérée dans cet article. La méthode développée dans cet article requiert l'estimation de la mortalité générale du portefeuille. Cette dernière, évaluée sur le portefeuille de dépendance, a nécessité une extrapolation aux grands âges par manque d'observations. Pour cela, nous faisons l'hypothèse d'une convergence de la mortalité du portefeuille vers la mortalité française provenant de "Human Mortality Database (HMD)" sur la période 2016 à 2018 disponible à l'adresse www.mortality.org.

Le Chapitre 3 fait appel à un jeu de données provenant d'un grand assureur français utilisant la définition selon la grille AGGIR. L'ensemble des assurés étant couverts à la fois contre la dépendance partielle et totale, l'unique état de dépendance considéré dans cet article regroupe l'ensemble des assurés ayant un niveau de dépendance GIR1 à GIR4. La méthode présentée dans cet article est appliquée au problème d'agrégation de portefeuilles d'assurance lorsque l'un des deux considère une période de franchise. En cas de franchise, il est courant que l'assureur ne déclare la perte d'autonomie qu'à la date de fin de période

4. Autonomie gérontologie groupes iso-ressources

de franchise en cas de survie. Pour les besoins de l'article, ce seul et unique portefeuille est aléatoirement séparé en deux, afin d'appliquer une période de franchise fictive sur l'un des deux. Cette séparation permet de considérer ensuite l'agrégation de deux portefeuilles homogènes ayant des définitions différentes de l'état de dépendance.

Le Chapitre 4 repose sur des données issues d'une caisse de maladie à l'étranger. Contrairement aux produits d'assurance proposés en France, le critère définitif de la perte d'autonomie n'est pas nécessaire pour percevoir la rente de dépendance. Ainsi, l'amélioration de l'état de santé de l'assuré est possible. Dans le cadre de l'étude de la mortalité en dépendance, le retour à l'autonomie est alors traité comme une censure à droite. Par ailleurs, le versement de la rente est limité à une durée de 5 ans, impliquant une sortie du portefeuille et donc une perte de suivi de l'assuré. Au-delà de ce délai, les données de mortalité en dépendance sont censurées. Cet article se concentrant sur la mortalité des dépendants, seule la base des rentiers est utilisée.

1.4 Outils théoriques

Après avoir présenté le contexte et défini le risque de dépendance, nous présentons les outils théoriques permettant de l'évaluer. Nous présentons dans un premier temps la théorie des modèles de durées. Nous présentons tout d'abord le cas simple d'un modèle de vie et de mort, avant de généraliser aux modèles en présence de risques concurrents, c'est-à-dire lorsque plusieurs causes peuvent être responsables de la sortie de l'état actuel de l'individu. Dans le cadre de la modélisation de produits d'assurance de personnes, il est courant de faire appel aux modèles multi-états. Après avoir présenté ces modèles et introduit les grandeurs associées permettant de quantifier et modéliser le risque, nous présentons dans la troisième section les méthodes d'estimation des probabilités de transition entre états.

1.4.1 Estimation du risque : la théorie des modèles de durées

Les modèles de durées permettent de modéliser le temps avant la survenue d'un évènement. Ils sont notamment utilisés en assurance ou en biologie pour étudier le temps restant avant la mort d'un individu ou organisme biologique, ou encore en électronique par exemple pour modéliser le temps avant la défaillance d'un composant. La variable d'intérêt est le temps noté $T > 0$.

1.4.1.1 Le cas simple du processus de vie et de mort

Plaçons-nous dans le cas simple d'un processus de vie et de mort, où T est une variable continue représentant le temps de survie d'un individu.

On appelle fonction de survie, la fonction $S(\cdot)$ telle que

$$S(x) = \mathbb{P}(T > x). \tag{1.10}$$

Propriété 1 *La fonction $S(\cdot)$ est décroissante, $S(0) = 1$ et $\lim_{x \to +\infty} S(x) = 0$.*

Il s'agit de l'opposé de la fonction de répartition de T. Par conséquent, la densité $f(x)$ est liée à la fonction de survie par l'équation suivante :

$$f(x) = -S'(x). \tag{1.11}$$

La fonction de hasard, appelée intensité de décès lors de la modélisation d'un processus de mortalité, est notée μ. Elle est parfois aussi notée λ dans certains ouvrages ou articles. Elle permet de décrire la transition entre l'état de vie et l'état de mort, telle que représentée par la Figure 1.5.

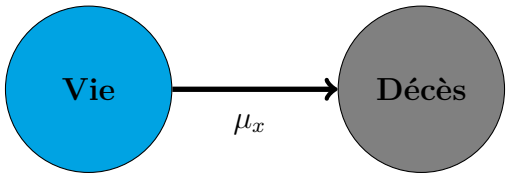


FIGURE 1.5 – Modélisation d'un processus de vie et de mort

Elle est définie par l'équation suivante, décrivant également son lien avec la fonction de survie $S(\cdot)$:

$$\mu_x = \lim_{h \to 0} \frac{\mathbb{P}(T < x + h | T \geqslant x)}{h} = -\frac{S'(x)}{S(x)}. \tag{1.12}$$

L'intensité de décès μ_x peut s'interpréter comme la probabilité de décès par unité de temps au voisinage de x. On l'appelle ainsi parfois "risque instantané de décès".

Il résulte de l'Équation 1.12 que

$$S(x) = \exp\left(-\int_0^x \mu_u \mathrm{d}u\right). \tag{1.13}$$

La fonction de survie conditionnelle $S_x(t)$ représente la probabilité qu'un individu d'âge x survive au moins t années. Autrement dit, il s'agit de la probabilité qu'il soit vivant à l'âge $x + t$. Elle est donnée par

$$\begin{aligned} S_x(t) = \mathbb{P}(T > x + t | T > x)) &= \frac{S(x+t)}{S(x)} \\ &= \exp\left(-\int_x^{x+t} \mu_u \mathrm{d}u\right). \end{aligned} \tag{1.14}$$

Une alternative à la variable aléatoire T pour décrire le processus de vie et de mort, est le processus stochastique $\{\mathcal{X}_x\}_{x>0} \in \{\text{Vie}, \text{Décès}\}$ représentant l'état occupé par l'individu à l'âge x, tel que $\mathcal{X}_0 = \text{Vie}$. Pour tout $x \geqslant 0$, l'évènement $\{T > x\}$ est équivalent à l'évènement $\{\mathcal{X}_x = \text{Vie}\}$.

Par conséquent

$$S(x) = \mathbb{P}\left(\mathcal{X}_x = \text{Vie}\right), \quad \text{et} \tag{1.15}$$

$$S_x(t) = \mathbb{P}\left(\mathcal{X}_{x+t} = \text{Vie} | \mathcal{X}_x = \text{Vie}\right). \tag{1.16}$$

L'intensité de décès est alors donnée par l'équation suivante :

$$\mu_x = \lim_{h \to 0} \frac{\mathbb{P}\left(\mathcal{X}_{x+h} = \text{Décès} | \mathcal{X}_x = \text{Vie}\right)}{h}. \tag{1.17}$$

La probabilité de survie conditionnelle $S_x(t)$ est également souvent notée ${}_tp_x$ par les actuaires.

1.4.1.2 Modélisation en présence de plusieurs causes de sorties

La section précédente considère le cas simple à seulement 2 états représentant la vie et la mort. Cependant, de nombreuses situations peuvent amener à considérer plusieurs évènements provoquant la sortie de l'état initial. La modélisation de la mortalité par cause de décès est une des situations les plus parlantes dans lesquelles nous sommes en présence de risques dits concurrents.

Supposons le cas de M causes de sortie de l'état initial 0, tel que représenté par la Figure 1.6. Soit T_c le temps de passage de l'état 0 à l'état c. La survie dans l'état 0 est conditionnelle au fait de survivre à l'ensemble des causes de sortie. Ainsi, le temps de sortie de l'état 0 est $T = \min(T_1, \ldots, T_M)$. Par ailleurs, sous l'hypothèse d'indépendance des temps $(T_c)_{c \in \{1,\ldots,M\}}$,

$$\mathbb{P}(T > t) = \prod_{c=1}^{M} \mathbb{P}(T_c > t). \tag{1.18}$$

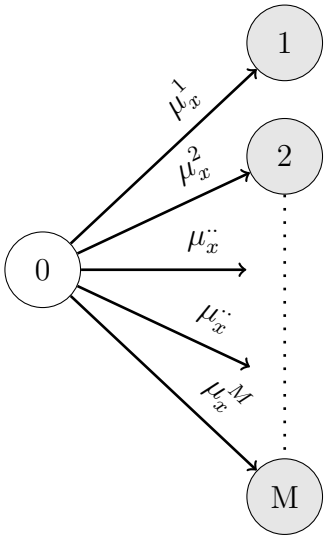


FIGURE 1.6 – Modélisation en présence de risques concurrents

L'intensité de transition associée à la cause c est donnée par l'Équation 1.19 suivante :

$$\mu_x^c = \lim_{h \to 0} \frac{\mathbb{P}(T_c < x + h, T = T_c | T_c > x)}{h} \tag{1.19}$$

La définition équivalente faisant intervenir le processus stochastique $\{\mathcal{X}_x\}_{x>0}$ est donnée par

$$\mu_x^c = \lim_{h \to 0} \frac{\mathbb{P}\left(\mathcal{X}_{x+h} = c | \mathcal{X}_x = 0\right)}{h} \tag{1.20}$$

Dans le cas de risques concurrents, on peut aisément montrer que l'intensité de sortie de l'état 0 i.e. l'intensité associée à la variable aléatoire T, est la somme des intensités $\mu_x^c, c \in \{1, \ldots, M\}$, i.e. :

$$\mu_x = \lim_{h \to 0} \frac{\mathbb{P}(T < x + h, |T > x)}{h} = \sum_{c=1}^{M} \mu_x^c. \tag{1.21}$$

Ainsi,

$$\mathbb{P}(T > t) = \exp\left(-\int_0^t \sum_{c=1}^{M} \mu_u^c \mathrm{d}u\right) = \prod_{c=1}^{M} \underbrace{\exp\left(-\int_0^t \mu_u^c \mathrm{d}u\right)}_{\mathbb{P}(T_c > t)}, \tag{1.22}$$

ce qui est cohérent avec l'Équation 1.18.

Par ailleurs, la probabilité de passage de l'état 0 à l'état c entre x et $x + t$ est donnée par

$$\mathbb{P}\left(\mathcal{X}_{x+t} = c | \mathcal{X}_x = 0\right) = \int_0^t \underbrace{\exp\left(-\int_x^{x+u} \mu_y \mathrm{d}y\right)}_{\mathbb{P}(T > x+u | T > x)} \mu_{x+u}^c \mathrm{d}u. \tag{1.23}$$

Pour une introduction plus complète de la théorie des modèles de durées, le lecteur pourra se référer à Kleinbaum and Klein (1996).

1.4.2 Modélisation du risque avec modèle multi-états

Les modèles multi-états sont souvent utilisés pour modéliser les contrats en assurance de personnes (Guibert, 2015). Ils sont notamment couramment utilisés pour étudier les produits incapacités, invalidés ou encore dépendance. Chaque état du modèle représente un état de santé de l'individu, ou statut de son contrat d'assurance. Le modèle permet ainsi de décrire l'historique de l'ensemble des états occupés par l'assuré sur toute la période d'étude, appelée trajectoire dans la suite. Hoem (1969) est le premier à avoir recours aux chaînes de Markov pour décrire les états successifs du contrat d'un assuré. Comme montré dans Czado and Rudolph (2002), la mortalité en dépendance dépend à la fois de l'âge

atteint, mais aussi de la durée depuis le diagnostic de perte d'autonomie, appelée duration. Ainsi, de nombreuses études proposent d'utiliser le modèle semi-Markovien, comme dans Biessy (2017); Fuino and Wagner (2018); Soetewey et al. (2022); Xuanyuan and Xuanyuan (2023). Contrairement aux modèles Markoviens, les modèles semi-Markoviens, introduits par Levy (1954), permettent de prendre en compte le temps passé dans l'état.

Le lecteur pourra se référer à Christiansen (2012) pour une revue de l'utilisation des modèles multi-états, Markoviens ou semi-Markoviens, pour la modélisation actuarielle des contrats d'assurance de personnes.

Les Chapitres 2 et 3 de cette thèse font appel au modèle à 3 états appelé modèle "Illness-Death". On considère donc dans la suite de cette section le cas d'un modèle à 3 états en temps continu, tel que représenté par la Figure 1.7. Cependant, une bonne compréhension des outils décrits dans la suite permet de généraliser aisément les formules à un modèle à plus de 3 états.

Comme la majorité des assureurs français, nous considérons dans cette section que le retour à l'autonomie est impossible. Cette hypothèse raisonnable de non retour à l'autonomie dans les modèles simplifie grandement les calculs. Cependant, dans le cadre de la modélisation d'un produit considérant comme possible le rétablissement et pour lequel un grand nombre de transitions de l'état de dépendance vers l'autonomie est observé, il peut être intéressant de considérer un modèle plus complexe, généralisant les formules suivantes.

À l'image de la section précédente, l'état de l'individu est décrit par le processus stochastique $\{\mathcal{X}_x\}_{x>0}$ tel que $\mathcal{X}_x$ décrit l'état occupé à l'âge x. Dans le cas de la modélisation d'un produit d'assurance dépendance, on note respectivement A et D les états d'autonomie et de dépendance. Le troisième état correspondant à l'état "Décès", il s'agit d'un état dit absorbant. On dit que l'état E est un état absorbant lorsque :

$$\{\mathcal{X}_x = E\} \implies \{\mathcal{X}_u = E\}, \forall u > x.$$

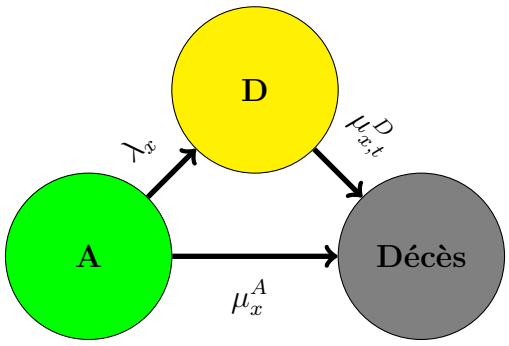


FIGURE 1.7 – Modélisation multi-états d'un produit d'assurance dépendance

À partir des définitions de la Section 1.4.1.2, les intensités de perte d'autonomie et de

décès autonomes sont respectivement données par

$$\lambda_x = \lim_{h \to 0} \frac{\mathbb{P}\left(\mathcal{X}_{x+h} = D | \mathcal{X}_x = A\right)}{h}, \quad \text{et} \tag{1.24}$$

$$\mu_x^A = \lim_{h \to 0} \frac{\mathbb{P}\left(\mathcal{X}_{x+h} = \text{Décès} | \mathcal{X}_x = A\right)}{h}. \tag{1.25}$$

Par ailleurs, dans le cas d'un modèle semi-Markovien, le taux instantané de décès des dépendants $\mu_{x,t}^D$ est donné par l'équation suivante :

$$\mu_{x,t}^D = \lim_{h \to 0} \frac{\mathbb{P}\left(\mathcal{X}_{x+h} = \text{Décès} | \mathcal{X}_x = D, \mathcal{X}_{(x-t)} = D, \mathcal{X}_{(x-t)^-} = A\right)}{h}. \tag{1.26}$$

En généralisant les résultats obtenus dans la Section 1.4.1.1, la probabilité pour une personne d'âge x dépendante depuis t_1 années, de survivre au moins t_2 années de plus, notée ${}_{t_2}p_{x,t_1}^D$, est donnée par l'équation suivante :

$${}_{t_2}p_{x,t_1}^D = \exp\left(-\int_0^{t_2} \mu_{x+u,t_1+u}^D \mathrm{d}u\right). \tag{1.27}$$

L'inégalité entre hommes et femmes, face à la fois au risque de perte d'autonomie et à la mortalité, justifie l'utilisation de lois de transition entre états différentes selon le sexe. Tout au long de cette thèse, cette inégalité est prise en compte en calibrant séparément et indépendamment les lois biométriques de chaque sexe. La modélisation étant indépendante, on considère un modèle multi-états par sexe. Ainsi, pour des raisons de lisibilité et de simplification, nous omettons le sexe g dans la suite de cette section, pour la formulation des taux de transition entre états.

Enfin, comme expliqué dans la Section 1.2.2, l'ensemble des modèles utilisés dans cette thèse supposent un effet négligeable de l'âge à la souscription sur les intensités de transition entre états. L'âge à la souscription n'apparaît pas dans l'estimation des intensités.

1.4.3 Estimation du risque

Les outils mathématiques utilisés pour modéliser un produit d'assurance dépendance font ainsi intervenir des grandeurs qu'il est nécessaire d'estimer pour décrire le risque. La description des modèles multi-états de la section précédente montre que les intensités de transition définissent entièrement le risque. Ainsi, la connaissance des intensités associées à chaque transition est suffisante pour modéliser entièrement le produit d'assurance.

1.4.3.1 Maximum de vraisemblance et intensités brutes

La méthode du maximum de vraisemblance est une méthode permettant d'inférer les paramètres d'un modèle en maximisant la vraisemblance des observations. La première étape consiste tout d'abord à écrire la fonction de vraisemblance associée aux données. La vraisemblance associée à un seul individu, notée L_k, représente la probabilité d'observer exactement sa trajectoire, sachant le modèle présenté dans la Section 1.4.2 et les lois de transitions associées. Sous l'hypothèse d'indépendance des individus du portefeuille, la vraisemblance de l'ensemble de la base de données, notée L, correspond au produit des vraisemblances individuelles, i.e. :

$$L = \prod_{k=1}^{N} L_k, \tag{1.28}$$

où N représente le nombre d'individus dans la base de données.

Cette vraisemblance fait intervenir les paramètres inconnus de notre modèle. Les paramètres permettant de maximiser la vraisemblance totale L, sont alors appelés les estimateurs du maximum de vraisemblance.

Pour chaque individu, on note :

- AoS l'âge au début de la période d'observation,
- AoI l'âge à la date de perte d'autonomie en cas d'entrée dans l'état D,
- AoD l'âge au décès si observé, et
- AoC l'âge de fin d'observation en cas de censure à droite (cf. Section 1.3.2).

Les âges AoI, AoD, AoC sont égaux à $+\infty$ en cas de non observation de l'évènement associé.

Les différentes trajectoires possibles d'un individu du portefeuille d'assurance, ainsi que les vraisemblances L_k associées, sont données dans la Table 1.5. La théorie des modèles de durées permet d'établir et de généraliser ces formules à d'autres modèles. Le lecteur pourra notamment se référer à Asanjarani et al. (2022) pour une bonne compréhension de la construction des fonctions de vraisemblance dans le cadre de modèles Markoviens et semi-Markoviens.

État initial	**Suite d'évènements observés**	**Vraisemblance L_i**
A	**Sous. → Dep. → Décès**	$\underbrace{\exp\left(-\int_{AoS}^{AoI}(\mu_u^A+\lambda_u)\mathrm{d}u\right)\lambda_{AoI}}_{L_i^A}\times$ $\underbrace{\exp\left(-\int_{AoI}^{AoD}\mu_{u,u-AoI}^D\mathrm{d}u\right)\mu_{AoD,AoD-AoI}^D}_{L_i^D}$
	Dep. → Décès	
	Sous. → Dep.	$\underbrace{\exp\left(-\int_{AoS}^{AoI}(\mu_u^A+\lambda_u)\mathrm{d}u\right)\lambda_{AoI}}_{L_i^A}\times$ $\underbrace{\exp\left(-\int_{AoI}^{AoC}\mu_{u,u-AoI}^D\mathrm{d}u\right)}_{L_i^D}$
	Dep.	
	Sous. → Décès	$\underbrace{\exp\left(-\int_{AoS}^{AoD}(\mu_u^A+\lambda_u)\mathrm{d}u\right)\mu_{AoD}^A}_{L_i^A}$
	Décès	
	Sous.	$\underbrace{\exp\left(-\int_{AoS}^{AoC}(\mu_u^A+\lambda_u)\mathrm{d}u\right)}_{L_i^A}$
	Ø	
D	**Décès**	$\underbrace{\exp\left(-\int_{AoS}^{AoD}\mu_{u,u-AoI}^D\mathrm{d}u\right)\mu_{AoD,AoD-AoI}^D}_{L_i^D}$
	Ø	$\underbrace{\exp\left(-\int_{AoS}^{AoC}\mu_{u,u-AoI}^D\mathrm{d}u\right)}_{L_i^D}$

TABLE 1.5 – Vraisemblance associée à chaque individu selon les transitions observées pendant la période d'étude

Pour l'ensemble des trajectoire possibles, présentées dans la Table 1.5, la vraisemblance peut être séparée en 2 parties, telle que

$$L_k = L_k^A \times L_k^D, \tag{1.29}$$

où :

- L_k^A représente la vraisemblance associée à l'expérience dans l'état d'autonomie,
- L_k^D est associée à l'expérience dans l'état de dépendance.

Il en découle que la vraisemblance totale, notée L, peut également être décomposée en deux termes, telle que

$$L = L^A \times L^D. \tag{1.30}$$

Par conséquent, sous l'hypothèse d'indépendance de la mortalité en dépendance, de la mortalité des autonomes et de la loi d'incidence, maximiser la vraisemblance totale L est équivalent à maximiser séparément L^A et L^D.

Hypothèse. 1.4.1

L'incidence et la mortalité des autonomes (λ_x et μ_x^A) sont constantes par morceaux en fonction de l'âge. La mortalité en dépendance ($\mu_{x,t}^D$) est constante par morceaux par âge et par durée depuis la perte d'autonomie.

On note $(x_i^A)_{(i\in\{1,\dots,M_x^A\})}$ et $(x_i^D)_{(i\in\{1,\dots,M_x^D\})}$ les points de subdivision de la variable d'âge pour l'expérience en autonomie et en dépendance respectivement. On note $(t_j^D)_{(j\in\{1,\dots,M_t^D\})}$ les points de subdivision de la variable de durée depuis la perte d'autonomie (duration). Les risques instantanés de perte d'autonomie et de décès des autonomes sont supposés constants sur les intervalles $[x_i^A, x_{i+1}^A]$. De même, le risque instantané de décès des dépendants est supposé constant sur les surfaces de la forme $[x_i^D, x_{i+1}^D] \times [t_j^D, t_{j+1}^D]$.

Sous l'Hypothèse 1.4.1 d'intensités constantes par morceaux, la vraisemblance de l'ensemble des observations dans l'état d'autonomie L^A est égale à

$$L^A = \prod_{i=1}^{M_x^A} \exp\left(-\mu_{x_i^A}^A \cdot e_{x_i^A}^A\right) \left(\mu_{x_i^A}^A \cdot e_{x_i^A}^A\right)^{d_{x_i^A}^A} \exp\left(-\lambda_{x_i^A} \cdot e_{x_i^A}^A\right) \left(\lambda_{x_i^A}^A \cdot e_{x_i^A}^A\right)^{n_{x_i^A}} \times Cst, \tag{1.31}$$

où :

- $e_{x_i^A}^A$ représente la somme des expositions centrales au risque de l'ensemble des individus, entre x_i^A et x_{i+1}^A,
- $d_{x_i^A}^A$ représente le nombre de décès observés en autonomie entre x_i^A et x_{i+1}^A,
- $n_{x_i^A}$ représente le nombre de pertes d'autonomie observées entre x_i^A et x_{i+1}^A,
- Cst est une constante.

On appelle exposition centrale d'un individu dans l'état E, sur l'intervalle $[x_a, x_b]$, la fraction de temps pendant laquelle cet individu est observé dans l'état E sur cet intervalle.

L'Équation 1.31 montre que la vraisemblance associée à l'expérience dans l'état d'autonomie est égale, à une constante près, au produit de vraisemblances de :

- M_x^A variables aléatoires de Poisson $D_{x_i}^A$ représentant les nombres de décès par tranche d'âge $[x_i^A, x_{i+1}^A]$, et
- M_x^A variables aléatoires de Poisson $N_{x_i}^A$ représentant les nombres d'entrée en dépendance par tranche d'âge $[x_i^A, x_{i+1}^A]$.

Ainsi, on retrouve couramment dans la littérature (Renshaw and Haberman, 1995; Dupourqué et al., 2019; Biessy, 2019) l'hypothèse de distribution de Poisson des nombres de décès notés $(D_{x_i}^A)_{i\in\{1,\dots,M_x^A\}}$, ainsi que des nombres de perte d'autonomies notés $(N_{x_i}^A)_{i\in\{1,\dots,M_x^A\}}$, tels que

$$\begin{aligned} D_{x_i^A}^A &\sim Poisson\left(\mu_{x_i^A}^A \cdot e_{x_i^A}^A\right), \forall i \in \{1,\dots,M_x^A\}, \text{et} \\ N_{x_i^A}^A &\sim Poisson\left(\lambda_{x_i^A} \cdot e_{x_i^A}^A\right), \forall i \in \{1,\dots,M_x^A\}. \end{aligned}$$

De même, sous l'Hypothèse 1.4.1, la vraisemblance associée à l'expérience dans l'état de dépendance, notée L^D, est donnée par l'équation suivante :

$$L^D = \prod_{i=1}^{M_x^D}\prod_{j=1}^{M_t^D} \exp\left(-\mu_{x_i^D,t_j^D}^D \cdot e_{x_i^D,t_j^D}^D\right)\left(\mu_{x_i^D,t_j^D}^D \cdot e_{x_i^D,t_j^D}^D\right)^{d_{x_i^D,t_j^D}^D} \times Cst, \tag{1.32}$$

égale à une constante près à la vraisemblance sous l'hypothèse de distributions de Poisson des nombres de décès observés en dépendance. Ainsi, l'estimateur du maximum de vraisemblance de la mortalité des dépendants peut être obtenu en supposant que

$$D_{x_i^D,t_j^D}^D \sim Poisson\left(\mu_{x_i^D,t_j^D}^D \cdot e_{x_i^D,t_j^D}^D\right), \forall i \in \{1,\dots,M_x^D\}, \forall j \in \{1,\dots,M_t^D\}.$$

Les intensités brutes, obtenues par maximisation des vraisemblances L^A et L^D vis-à-vis des paramètres $\left(\mu_{x_i^A}^A\right)_{(i\in\{1,\dots,M_x^A\})}$, $\left(\lambda_{x_i^A}\right)_{(i\in\{1,\dots,M_x^A\})}$ et $\left(\mu_{x_i^D,t_j^D}^D\right)_{\left(i\in\{1,\dots,M_x^D\},j\in\{1,\dots,M_t^D\}\right)}$, sont ainsi données par

$$\widehat{\mu_{x_i^A}^A} = \frac{d_{x_i^A}^A}{e_{x_i^A}^A}, \tag{1.33}$$

$$\widehat{\lambda_{x_i^A}} = \frac{n_{x_i^A}^A}{e_{x_i^A}^A}, \tag{1.34}$$

$$\widehat{\mu_{x_i^D,t_j^D}^D} = \frac{d_{x_i^D,t_j^D}^D}{e_{x_i^D,t_j^D}^D}. \tag{1.35}$$

1.4.3.2 Utilisation des modèles linéaires généralisés

L'hypothèse de distribution de Poisson des nombres de décès et entrées en dépendance, justifiée dans la Section 1.4.3.1, nous permet d'utiliser les modèles linéaires généralisés pour l'estimation des intensités. Commençons par faire quelques rappels sur les modèles linéaires généralisés, couramment appelés GLMs pour l'acronyme de Generalized Linear Models (Nelder and Wedderburn, 1972; McCullagh, 2019).

Comme son nom l'indique, les GLMs sont une extension des modèles linéaires. Ces derniers s'appuient sur l'hypothèse forte que la variable à expliquer Y, sachant les variables explicatives X, suit une loi normale. Les GLMs quant à eux, permettent le choix de la distribution de la variable Y parmi l'ensemble des lois de probabilités appartenant à la famille exponentielle, présentée dans Barndorff (1978). Une distribution de probabilité appartient à cette famille, si sa densité peut s'écrire sous la forme

$$f(y_i;\theta_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right), \tag{1.36}$$

où $a(\cdot)$, $b(\cdot)$ et $c(\cdot,\cdot)$ sont des fonctions, ϕ représente un paramètre de dispersion, et θ est appelé paramètre naturel.

On retrouve notamment dans cette famille les lois normales, Gamma, log-normales, exponentielles, ou encore la loi de Poisson, pour n'en citer que quelques-unes. Les GLMs permettent ainsi de modéliser des variables discrètes, contrairement à la régression linéaire qui suppose que la variable de réponse est continue.

Par ailleurs, tandis que dans le modèle linéaire l'espérance conditionnelle de la variable de réponse Y sachant X est une combinaison linéaire de X, les GLMs permettent plus de liberté en supposant une relation faisant intervenir une fonction de lien $G(\cdot)$, telle que

$$\mathbb{E}[Y|X] = G^{-1}(X\beta), \tag{1.37}$$

où β désigne le vecteur de coefficients à estimer. La fonction de lien $G(\cdot)$ permet ainsi de préciser le lien entre la variable à expliquer Y et les variables explicatives X. Enfin, tandis que les modèles linéaires supposent l'homoscédasticité des résidus, les GLMs permettent de modéliser des variables dont la variance est une fonction de la moyenne (hétéroscédasticité).

Les GLMs permettent de modéliser la variance de la variable à expliquer comme étant fonction de sa prédiction, ce qui est utile en cas d'hétéroscédasticité.

On note $Y_1, Y_2, \ldots, Y_n$ les n observations de la variable à expliquer.

Hypothèse. 1.4.2

Tout comme les modèles linéaires, les GLMs supposent l'indépendance des variables $Y_i, i \in \{1, \dots, n\}$.

Un GLM donné dépend principalement des 3 paramètres suivants :

- la loi de probabilité,
- la fonction de lien $G(\cdot)$, et
- le prédicteur linéaire noté η, tel que $\eta = X\beta$.

Bien que la fonction $G(\cdot)$ soit au libre choix de l'utilisateur, chaque loi de probabilité de la famille exponentielle dispose d'une fonction de lien spécifique appelée fonction de lien canonique, disposant de propriétés mathématiques intéressantes décrites dans Nelder and Wedderburn (1972). La fonction de lien canonique associée à une distribution de la famille exponentielle est la fonction telle que

$$G(\mathbb{E}[Y|X]) = \theta. \tag{1.38}$$

Par ailleurs, il peut être utile dans certains cas d'appliquer des transformations numériques aux variables explicatives afin de considérer des relations plus complexes entre les variables explicatives et la variable de réponse Y. Des effets d'interactions entre variables peuvent également être pris en compte dans le prédicteur linéaire. Cela est particulièrement intéressant si l'on souhaite estimer un coefficient de régression β_j associé à la variable X_j différent selon la modalité prise par une seconde variable X_l. Enfin, les modèles linéaires généralisés permettent de considérer, dans le prédicteur linéaire η, une covariable supplémentaire ayant un coefficient fixé, ne nécessitant pas d'estimation. Ce terme supplémentaire est appelé offset. Dans ce cas, le prédicteur linéaire devient

$$\eta = X\beta + \text{offset}. \tag{1.39}$$

La méthode la plus courante pour estimer les coefficients β est la méthode du maximum de vraisemblance. Cependant, il n'existe pas dans le cas général de formule fermée pour l'estimation de β. Ainsi, l'estimation repose très souvent sur des algorithmes d'optimisation, de type Iterative Reweighted Least Square (IRWLS) (Green, 1984). Dans le cas où toutes les variables explicatives sont catégorielles, Brouste et al. (2020) et Brouste et al. (2022) proposent une formule fermée des estimateurs de β, améliorant ainsi considérablement les temps de calcul.

Dans le cadre de la modélisation de la dépendance, Y correspond aux nombres de décès ou aux nombres d'entrées en dépendance. Les lois de probabilités les plus couramment

utilisées dans ce contexte sont les lois de Poisson et binomiale (Hunt and Blake, 2021).

Par souci de simplification, nous développons dans cette section uniquement le cas de la modélisation des nombres de décès en autonomie notés $D^A_{x^A_i}$, fonctions de l'âge x^A_i.

Dans le cadre d'une modélisation utilisant la loi binomiale,

$$\mathbb{E}[D^A_{x^A_i}] = {}^0e^A_{x^A_i} q^A_{x^A_i}, \tag{1.40}$$

où ${}^0e^A_{x^A_i}$, appelée exposition initiale, représente le nombre d'autonomes au début de la période d'âge $[x^A_i; x^A_{i+1}[$, et $q^A_{x^A_i}$ représente la probabilité qu'un individu d'âge x^A_i décède avant d'atteindre l'âge x^A_{i+1}.

Sous l'hypothèse de distribution de Poisson des nombres de décès, donnant des estimateurs identiques à ceux obtenus par la théorie des modèles de durées (c.f. Section 1.4.3.1),

$$\mathbb{E}[D^A_{x^A_i}] = e^A_{x^A_i} \mu^A_{x^A_i}. \tag{1.41}$$

Remarque : Relation entre mortalité et intensité de décès des autonomes

À partir des résultats de la Section 1.4.1.2, la relation entre l'intensité $\mu^A_{x^A_i}$ et la probabilité de décès $q^A_{x^A_i}$ est donnée par

$$q^A_{x^A_i} = \frac{\mu^A_{x^A_i}}{\mu^A_{x^A_i} + \lambda_{x^A_i}} \left[1 - \exp\left(-\mu^A_{x^A_i}\left(x^A_{i+1} - x^A_i\right)\right)\right].$$

La probabilité de décès dépend ainsi à la fois de l'intensité de décès, mais également de l'intensité de perte d'autonomie au même âge.

Les bases de données disposant de l'information sur l'exposition initiale sont très rares en raison de la présence de phénomènes de censure et de troncature. En revanche, les expositions centrales sont facilement estimables à partir de n'importe quelle base de données. La cohérence entre les estimateurs issus de la théorie des modèles de durée et ceux obtenus sous l'hypothèse de Poisson, ajoutée à la plus grande disponibilité des expositions centrales, justifie le choix de la distribution de Poisson dans l'ensemble des chapitres de cette thèse. En cas de choix de la loi binomiale, les expositions initiales sont parfois approximées par les expositions centrales.

La fonction de lien canonique des GLMs Poisson est la fonction logarithmique. Ainsi, le lien entre la variable à expliquer et le prédicteur linéaire est donné par

$$\log(\mathbb{E}[D^A_{x_i^A}]) = \underbrace{\log(e^A_{x_i^A})}_{\text{offset}} + \underbrace{\log(\mu^A_{x_i^A})}_{X\beta}. \tag{1.42}$$

Dans le cadre de l'estimation des taux bruts, la variable d'âge est considérée comme une variable catégorielle. Un coefficient β_i est estimé par âge, tel que

$$\log(\mu^A_{x_i^A}) = \log(\beta_i). \tag{1.43}$$

Dans le cas où la variable d'intérêt dépend de 2 facteurs, tel que pour la modélisation du nombre de décès en dépendance $D^D_{x_i^D,t_j^D}$, un coefficient de régression $\beta_{i,j}$ est estimé par combinaison possible de l'âge et de la duration.

1.4.3.3 Introduction aux méthodes de lissage paramétriques et semi-paramétriques des intensités

Les taux bruts de mortalité étant estimés indépendamment les uns des autres pour chaque âge, rien ne garantit un aspect lisse de la courbe de mortalité en fonction de l'âge. Cette rugosité vient notamment de la volatilité des estimateurs donnés par les Équations 1.33, 1.34 et 1.35. Les variances sont respectivement

$$\widehat{Var[\mu^A_{x_i^A}]} = \frac{\mu^A_{x_i^A}}{e^A_{x_i^A}}, \tag{1.44}$$

$$\widehat{Var[\lambda_{x_i^A}]} = \frac{\lambda_{x_i^A}}{e^A_{x_i^A}}, \tag{1.45}$$

$$\widehat{Var[\mu^D_{x_i^D,t_j^D}]} = \frac{\mu^D_{x_i^D,t_j^D}}{e^D_{x_i^D,t_j^D}}. \tag{1.46}$$

Ainsi, l'incertitude autour de l'estimation des taux bruts est d'autant plus élevée que l'exposition est faible. Une des solutions pour réduire cette incertitude, et donc l'aspérité de la loi de mortalité en fonction de l'âge, est d'augmenter le nombre d'individus observés. Ceci n'étant pas toujours possible, ou alors à un coût élevé, les actuaires et les démographes font appel à des méthodes de lissage de taux. Celles-ci peuvent être regroupées en 3 familles :

- les modèles paramétriques,
- les méthodes de lissage semi-paramétriques, et
- les méthodes de lissage non paramétriques.

Cette dernière famille n'est pas abordée dans cette introduction de thèse, car elle ne sera pas employée dans les chapitres suivants.

Les modèles paramétriques

Dans le cadre des modèles de lissage paramétriques, une hypothèse sur la forme de la courbe des taux est énoncée a priori. Chacun de ces modèles repose sur une fonction mathématique faisant intervenir des paramètres inconnus. L'objectif consiste alors à identifier les paramètres optimaux permettant de s'approcher au mieux des taux bruts

observés sur la base de données. L'a priori sur la forme de la loi, illustré par le choix de la fonction mathématique, a un impact fort sur la courbe de mortalité lissée.

La régression polynomiale appartient à cette famille. Considérons le cas où la mortalité ne dépend que de l'âge x. Après avoir appliqué une éventuelle transformation $g(\cdot)$ aux taux bruts, la régression polynomiale suppose que $g(\mu_x^A)$ peut s'écrire sous la forme d'un polynôme de degré p, tel que

$$g(\mu_x^A) = \beta_0 + \beta_1 x + \ldots + \beta_p x^p. \tag{1.47}$$

La qualité d'ajustement du polynôme aux taux bruts dépend essentiellement du degré du polynôme. Plus le degré d est élevé, plus le modèle sera complexe et la fonction ajustée sera proche des taux observés. Cependant, un degré trop élevé a pour conséquence un surapprentissage et une perte de l'aspect lisse et régulier que l'on souhaite obtenir pour la courbe de mortalité.

Lorsque l'intensité dépend de plusieurs variables, telles que l'âge et la duration dans un contexte de modélisation de la mortalité des dépendants, un polynôme par variable est considéré. Des effets d'interaction entre âge et duration peuvent également être pris en compte.

Dans le contexte de la modélisation de données de survie, une transformation logarithmique est couramment appliquée aux taux bruts de mortalité (i.e. $g(\cdot) = \log(\cdot)$).

La régression polynomiale, associée aux modèles linéaires généralisés, est utilisée dans le Chapitre 4 de cette thèse. Nous y considérons notamment un effet quadratique de l'âge et un effet cubique de la durée depuis la perte d'autonomie.

Au-delà de la régression polynomiale, certaines fonctions ont été spécialement pensées et introduites afin de refléter les formes usuelles de courbes rencontrées dans le cadre de la modélisation de données de survie. Parmi les modèles paramétriques les plus reconnus dans ce domaine, on retrouve notamment :

- le modèle de Gompertz (1825), et sa généralisation le modèle de Gompertz-Makeham (Makeham, 1860),
- le modèle de Weibull (1951), ou encore
- les modèles logistiques, tels que le modèle de Perks (1932) et celui de Thatcher (1999).

L'a priori sur la forme de la loi de mortalité implique un manque de souplesse limitant la capacité à s'approcher au mieux des taux observés dans la base de données. En revanche, ces méthodes permettent d'extrapoler les lois de mortalité à des tranches d'âges pour lesquelles aucune observation n'est disponible.

Les modèles semi-paramétriques : introduction aux méthodes de lissage par splines

Il est souvent difficile de trouver un modèle paramétrique simple permettant de capturer la variance de l'incidence et de la mortalité observées. Les modèles additifs généralisés (GAM) sont ainsi souvent utilisés par les actuaires pour modéliser les risques biométriques, en supposant une distribution de Poisson des nombres de transitions (entrées en dépendance ou décès) observés. Les GAM, extension des GLMs, permettent de prendre en compte des relations plus complexes entre la variable observée Y et les variables explicatives, telles que

$$G(\mathbb{E}[Y]) = \beta_0 + \sum_j \beta_j s_j(X), \tag{1.48}$$

où $s_j(\cdot)$ est une fonction lisse. Chaque fonction $s_j(\cdot)$ peut modéliser :

- l'effet d'une seule variable explicative continue,
- l'effet d'interaction entre une variable continue et une variable qualitative,
- l'effet d'interaction entre deux (ou plus) variables continues.

Nous présentons ici le cas simple d'un lissage à une seule dimension, lorsqu'une seule variable continue intervient dans le prédicteur linéaire. Il s'agit du cas notamment de la modélisation de l'incidence et de la mortalité des autonomes, où la seule variable continue correspond à l'âge.

Plusieurs méthodes existent pour estimer les fonctions lisses $s_j(\cdot)$. Le lecteur pourra se référer à Simonoff (2012) pour une bonne introduction aux méthodes de lissage les plus courantes. Par souci de simplification, nous présentons dans cette section uniquement la méthode de lissage par P-Splines [5], utilisée dans le Chapitre 2 et le Chapitre 3. Cette méthode, introduite par Eilers and Marx (1996), associe les B-Splines [6] à une pénalité de lissage, ayant pour objectif de limiter les grandes variations entre coefficients de splines adjacentes.

La méthode de lissage par P-Splines, adaptée à la mortalité pour la première fois par Currie et al. (2004) conjointement avec l'hypothèse de Poisson des nombres de décès, se présente comme telle :

1. L'intervalle d'étude de la variable explicative (i.e. l'âge dans notre cas) est subdivisé de telle sorte à répartir de manière équidistante un ensemble de courbes sous forme de cloche appelées splines de base. Un exemple de répartition de splines sur l'intervalle d'âge x est illustré par la Figure 1.8.

5. Penalized Splines
6. Basis Splines

2. Le prédicteur linéaire η s'écrit alors sous la forme

$$\eta = B\theta, \tag{1.49}$$

où B est la matrice représentant la base de splines ; chaque colonne correspondant aux valeurs d'une spline.

3. Un terme de pénalité est alors ajouté à la log-vraisemblance de Poisson, ayant pour objectif d'éviter une trop grande variabilité entre coefficients de splines adjacentes. Cette pénalité s'écrit sous la forme

$$\rho\theta^T P_d \theta, \tag{1.50}$$

où P_d représente une matrice de pénalité dépendant d'un ordre d. Le paramètre ρ matérialise le poids attribué à cette pénalité. Ce dernier est souvent optimisé de telle sorte à minimiser le BIC, comme recommandé par Currie and Durban (2002).

En considérant la fonction de lien logarithmique, l'intensité de décès est fonction de l'âge à travers la relation suivante :

$$\log(\mu_x^A) = \sum_j^J \beta_j B_j(x), \tag{1.51}$$

où $B_j(x)$ représente la j-ième splines, et J correspond au nombre de splines réparties sur l'intervalle.

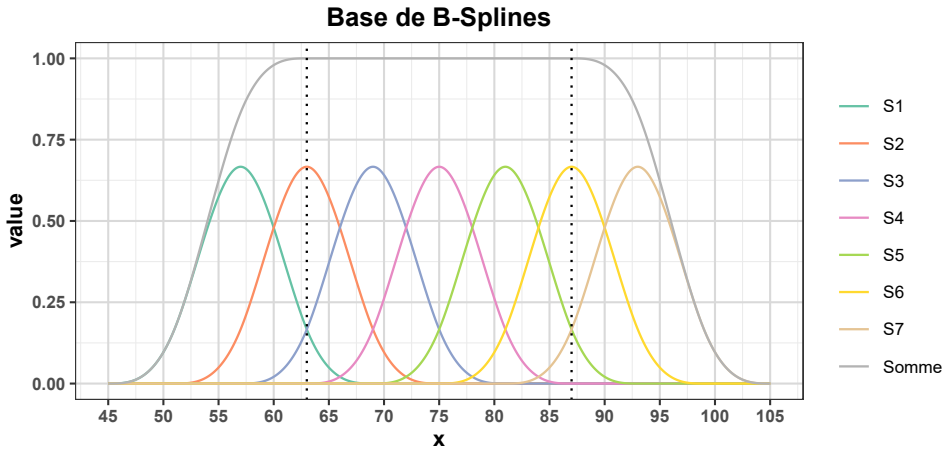


FIGURE 1.8 – Base de B-Splines

Selon Eilers and Marx (2002), le choix des hyperparamètres tels que le degré des splines et leur espacement sur l'intervalle d'étude a peu d'importance sur l'estimation finale. Cet

article propose alors d'utiliser des splines cubiques (degré 3) et de positionner une spline toutes les 4 ou 5 observations en cas d'observations équidistantes sur l'axe étudié.

La méthode de lissage par P-Splines se généralise facilement au cas bidimensionnel en faisant appel au produit de Kronecker. Cette généralisation, décrite en détail dans Currie et al. (2004), permet entre autres le lissage de la mortalité en dépendance, fonction à la fois de l'âge et de la durée depuis la perte d'autonomie.

Pour davantage de détails sur l'application des P-Splines à la modélisation de la mortalité dans le cas uni- et bi-dimensionnel, le lecteur est invité à consulter la thèse de Camarda (2008).

Le lissage P-Splines adapté à la modélisation de données de survie sous l'hypothèse de Poisson est implémenté sous **R** dans la librairie **MortalitySmooth** (Camarda et al., 2012).

1.5 Motivations, contributions et résultats principaux

Le corps de cette thèse se divise en trois chapitres, qui abordent chacun une problématique distincte visant à améliorer la modélisation des produits d'assurance dépendance. Ces chapitres peuvent être lus de manière indépendante. Cependant, le lecteur y trouvera des éléments méthodologiques communs. Cette section a pour objectif de présenter succinctement les motivations ainsi que les contributions et résultats de chacun de ces chapitres.

1.5.1 Chapitre 2 - Coherent extrapolation of mortality rates at old ages applied to Long-Term Care

Ce n'est qu'au début des années 2000 que les produits d'assurance dépendance se sont majoritairement développés. Les assureurs ne disposent donc que d'un historique d'observation d'une vingtaine d'années au plus. Par ailleurs, les contrats d'assurance dépendance en France fixent souvent un âge limite à la souscription, situé entre 75 et 79 ans selon les assureurs. Ainsi, les bases de données ne contiennent que peu ou pas d'observations aux âges avancés, rendant difficile l'estimation de la mortalité. Les assureurs ont alors recours à des méthodes d'extrapolation des lois de mortalité des autonomes et des dépendants, reposant souvent sur une hypothèse quant à la forme de la courbe de mortalité aux âges auxquels aucune observation n'est disponible. Par ailleurs, les courbes de mortalité des autonomes et des dépendants sont souvent extrapolées de manière totalement indépendante, pouvant faire apparaître des incohérences entre elles. On appelle mortalité générale du portefeuille, la loi de mortalité par âge estimée sur l'ensemble du portefeuille indépendamment de l'état de santé des individus. Cette indépendance des extrapolations peut notamment mener

à estimer une mortalité des autonomes et des dépendants plus élevée que la mortalité générale. Cette dernière est par ailleurs plus simple à estimer en raison de l'agrégation de l'ensemble des individus et de la meilleure connaissance de la mortalité en population générale.

Ce chapitre reprend l'article "Coherent extrapolation of mortality rates at old ages applied to Long-Term Care" publié à l'European Actuarial Journal (Le Bastard, 2023a). Un portefeuille d'assurance dépendance est composé exclusivement d'autonomes et de dépendants, représentant à chaque âge deux groupes distincts d'individus. À partir de cette observation, cet article propose une méthode d'extrapolation reposant sur l'idée simple que la mortalité globale du portefeuille, supposée connue, est une moyenne pondérée de la mortalité des deux groupes. Les poids utilisés dans la pondération varient en fonction de l'âge et font intervenir les expositions centrales dans chacun des deux groupes. Il est très courant, dans les bases de données, de n'avoir aucune observation au-delà d'un certain âge avancé. Dans ce cas, les expositions centrales dans chacun des deux groupes sont nulles, posant un problème pour l'estimation des poids. Nous proposons dans ce chapitre un algorithme permettant d'estimer des expositions théoriques pour ces âges avancés, conjointement à l'extrapolation des lois de mortalité. Cette méthode permet une extrapolation simultanée de la mortalité des autonomes et des dépendants, en cohérence avec la loi de mortalité générale. Les intervalles de confiance sont ensuite construits à partir d'un algorithme de simulation inspiré des méthodes de bootstrap. Pour plus d'informations sur la construction d'intervalles de confiance sur des données de mortalité à partir de méthodes de bootstrap, le lecteur intéressé pourra se référer aux travaux de Brouhns et al. (2005); Koissi et al. (2006); Debón et al. (2008).

Afin de pouvoir utiliser cette méthode de manière opérationnelle, une librairie **R** a été développée en parallèle de l'écriture de l'article. Celle-ci, partagée au sein de SCOR, est régulièrement utilisée par l'équipe chargée de l'étude des portefeuilles d'assurance dépendance. Elle a notamment été utilisée dans le mémoire de Le Gal (2021) réalisé au sein de SCOR.

1.5.2 Chapitre 3 - Combining experience data of several Long-Term Care Insurance products with different disability definitions

La définition de la dépendance n'est pas uniforme dans tous les contrats d'assurance dépendance. En effet, nous avons vu dans la Section 1.2.1 qu'il existe en France principalement 2 grilles utilisées pour déterminer le degré de perte d'autonomie d'un individu. Au-delà du choix de cette grille, l'assureur détermine également, dans le contrat, le degré de dépendance minimum déclenchant le versement de la rente jusqu'au décès. Ainsi, certains

contrats prévoient une rente à partir du niveau de dépendance partielle, tandis que d'autres ne couvrent que la dépendance totale. Enfin certains contrats d'assurance dépendance comportent des clauses particulières, comme une période de franchise de fr mois, telle que présentée dans la Section 1.2.2. Dans ce cas, le versement de la rente est conditionnel à la survie au fr premier mois de dépendance. Ainsi, la définition de l'état D du modèle multi-états présenté par la Figure 1.7 varie selon les contrats, rendant complexe l'agrégation de données. Considérons une base de données constituée d'individus ayant des contrats avec des définitions non uniformes. Une des solutions consiste à séparer les individus par type de contrat, puis à procéder à la modélisation du risque de manière indépendante pour chaque définition. Cette approche n'est toutefois pas viable pour les assureurs qui souffrent déjà du manque de données disponibles pour modéliser le risque.

Ce chapitre reprend l'article "Combining experience data of several Long-Term Care Insurance products with different disability definitions" co-écrit avec Loisel Stéphane et Shao Adam (Le Bastard, 2023b). Nous proposons dans cet article deux méthodes permettant d'agréger des données d'individus ayant des contrats dont la définition diffère. Nous considérons la coexistence de deux définitions au sein de la même base de donnée, dont l'une est supposée plus stricte que la seconde. Autrement dit, être dépendant avec la première définition implique la dépendance selon la seconde. La première méthode repose sur les techniques d'optimisation sous-contraintes, tandis que la seconde est inspirée des Penalized Composite Link Model (PCLM) introduits par Eilers (2007). Les deux méthodes sont ensuite appliquées au problème de la gestion de la franchise, en considérant dans un même portefeuille des contrats avec et sans franchise. En comparaison avec la solution consistant à estimer séparément les lois biométriques associées à chaque définition, ces deux méthodes permettent de réduire l'incertitude autour des intensités de transitions grâce à prise en compte de l'ensemble des observations pour leurs estimations. Par ailleurs, la méthode basée sur les PCLM démontre de meilleures performances prédictives en ce qui concerne la mortalité dans l'état de dépendance.

1.5.3 Chapitre 4 - Clustering of pathologies : application to Long-Term Care Insurance

Plusieurs pathologies, souvent liées au vieillissement, peuvent mener à la perte d'autonomie d'un individu. On parle de cause de dépendance lorsque cette pathologie est identifiée comme la source de la dégradation de l'état de santé de l'individu. Parmi les pathologies les plus fréquentes, on retrouve notamment la maladie d'Alzheimer et autres démences, le cancer, les maladies neurologiques ou encore les maladies cardiovasculaires. Cependant, il existe une multitude d'autres causes, moins représentées dans les portefeuilles d'assurance dépendance, telles que les maladies respiratoires, osteoarticulaires ou encore les accidents. Plusieurs études telles que l'article de Biessy (2016a) ont montré l'impact fort de la

pathologie sur la mortalité de l'individu dépendant. Ainsi, les individus dont la cause de dépendance est le cancer ont une probabilité de décès au cours de la première année suivant la perte d'autonomie très élevée comparée à celle des individus déments. Ignorer l'information de la pathologie dans l'estimation de la mortalité conduit à une perte d'information, pouvant entraîner un assureur à une mauvaise estimation des provisions selon la distribution des pathologies représentées dans son portefeuille de rentiers. Cependant, l'information concernant la cause de perte d'autonomie n'est pas disponible dans toutes les bases de données, et certaines pathologies sont trop peu représentées pour pouvoir prétendre estimer indépendamment une table de mortalité spécifique par pathologie. Enfin, par souci de simplicité, les assureurs préfèrent avoir un nombre limité de tables de mortalité, tout en capturant un maximum de la variabilité des observations. Le regroupement de pathologies, similaires en termes de mortalité suivant la perte d'autonomie, semble un bon compromis. Par manque de données ou d'algorithme dédié, les assureurs font souvent appel à l'avis d'un expert pour constituer des groupes de pathologies.

Ce chapitre reprend l'article "Clustering of pathologies : application to Long-Term Care Insurance". Nous y proposons et comparons deux méthodes de clustering, toutes deux s'appuyant sur les modèles linéaires généralisés (GLM). La mortalité en dépendance dépendant à la fois de l'âge atteint et de la durée depuis la perte d'autonomie, l'objectif est de regrouper les pathologies ayant des surfaces de mortalité semblables. La première approche est basée sur les GLM trees, tandis que la seconde est inspirée de la très connue méthode des K-means. Après avoir regroupé les pathologies selon les différentes approches, y compris en demandant l'avis d'un expert, nous montrons l'apport des méthodes de clustering sur les performances des modèles de mortalité associés. La comparaison des BIC [7] de chaque modèle, montre de meilleures performances des modèles utilisant les clusters construits par les méthodes algorithmiques, comparées à celles du modèle basé sur l'avis d'expert.

7. Bayesian Information Criterion

Bibliographie

AG2R (2023). À quel âge souscrire une assurance dépendance ? https ://www.ag2rlamondiale.fr/sante-prevoyance/dependance/conseil-a-quel-age-souscrire-une-assurance-dependance.

AG2R La Mondiale (2019). Assurance Autonomie - Document d'Information sur le Produit d'Assurance.

Asanjarani, A., B. Liquet, and Y. Nazarathy (2022). Estimation of semi-markov multi-state models : a comparison of the sojourn times and transition intensities approaches. *The International Journal of Biostatistics 18*(1), 243–262.

Barndorff, N. (1978). Information and exponential families ; in statistical theory. Technical report.

Biessy, G. (2016a). A semi-Markov model with pathologies for Long-Term Care Insurance. working paper or preprint.

Biessy, G. (2016b). *Modélisation semi-markovienne de la perte d'autonomie chez les personnes âgées : application à l'assurance dépendance*. Theses, Université Paris-Saclay ; Université d'Evry Val d'Essonne.

Biessy, G. (2017). Continuous-time semi-markov inference of biometric laws associated with a long-term care insurance portfolio. *ASTIN Bulletin : The Journal of the IAA 47*(2), 527–561.

Biessy, G. (2019). Smoothing of multidimensional biometric laws in a Long-Term Care Insurance portfolio. working paper or preprint.

Bonnet, C., E. Cambois, C. Cases, and J. Gaymu (2011). La dépendance : aujourd'hui l'affaire des femmes, demain davantage celle des hommes ? *Population Societes 483*(10), 1–4.

Brouhns, N., M. Denuit, and I. Van Keilegom (2005). Bootstrapping the poisson log-bilinear model for mortality forecasting. *Scandinavian Actuarial Journal 2005*(3), 212–224.

Brouste, A., C. Dutang, and T. Rohmer (2020). Closed form Maximum Likelihood Estimator for Generalized Linear Models in the case of categorical explanatory variables : Application to insurance loss modelling. *Computational Statistics*.

Brouste, A., C. Dutang, and T. Rohmer (2022). A closed-form alternative estimator for glm with categorical explanatory variables. *Communications in Statistics-Simulation and Computation*, 1–17.

Camarda, C. G. (2008). *Smoothing methods for the analysis of mortality development*. Ph. D. thesis, Universidad Carlos III de Madrid.

Camarda, C. G. et al. (2012). MortalitySmooth : An R package for smoothing Poisson counts with P-splines. *Journal of Statistical Software 50*(1), 1–24.

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

CNSA (2019). Détail départemental CNSA . https ://www.cnsa.fr/sites/default/files/analyse_des_prix_e

CNSA (consulté le 23 octobre 2023). L'Allocation personnalisée d'autonomie (APA). https ://www.pour-les-personnes-agees.gouv.fr.

Crimmins, E. M., H. Shim, Y. S. Zhang, and J. K. Kim (2019). Differences between men and women in mortality and the health dimensions of the morbidity process. *Clinical chemistry 65*(1), 135–145.

Currie, I. D. and M. Durban (2002). Flexible smoothing with P-splines : A unified approach. *Statistical Modelling 2*(4), 333–349.

Currie, I. D., M. Durban, and P. H. Eilers (2004). Smoothing and forecasting mortality rates. *Statistical modelling 4*(4), 279–298.

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

Dalloz (2015). Code des Assurances : Article R343-3. Décr. no 2015-513 du 7 mai 2015, art. 9-13o, en vigueur le 1er janv. 2016.

Debón, A., F. Montes, and F. Puig (2008). Modelling and forecasting mortality in spain. *European Journal of Operational Research 189*(3), 624–637.

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

DREES (2018). Les français vivent plus longtemps, mais leur espérance de vie en bonne santé reste stable. *Études et résultats*.

DREES (2022). Aides à l'autonomie des personnes âgées : qui paie quoi ? *Les Dossiers de la DREES*.

Dupourqué, E., F. Planchet, and N. Sator (2019). *Actuarial aspects of long term care*. Springer.

Eilers, P. (2007). Ill-posed problems with counts, the composite link model, and penalized likelihood. *Statistical Modelling 7*.

Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing with B -splines and penalties. *Statist. Sci. 11*(2), 89–121.

Eilers, P. H. C. and B. D. Marx (2002). Generalized linear additive smooth structures. *Journal of Computational and Graphical Statistics 11*(4), 758–783.

Folstein, M. F., S. E. Folstein, and P. R. McHugh (1975). "mini-mental state" : A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research 12*(3), 189–198.

Fondation Vaincre Alzheimer (2019). World Population Prospects 2019. https ://population.un.org/wpp2019/Graphs/Probabilistic/PopPerc/65plus/900.

France Alzheimer (consulté le 12 juin 2019). La maladie d'Alzheimer en chiffres. https ://www.francealzheimer.org/maladie-dalzheimer-vos-questions-nos-reponses/maladie-dalzheimer-chiffres/.

Fuino, M. and J. Wagner (2018). Long-term care models and dependence probability tables by acuity level : New empirical evidence from Switzerland. *Insurance : Mathematics and Economics 81*, 51–70.

Gompertz, B. (1825). Xxiv. 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. *Philosophical transactions of the Royal Society of London* (115), 513–583.

Green, P. J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *Journal of the Royal Statistical Society : Series B (Methodological) 46*(2), 149–170.

Guibert, Q. (2015). *Sur l'utilisation des modèles multi-états pour la mesure et la gestion des risques d'un contrat d'assurance*. Theses, Université Claude Bernard - Lyon I.

Hoem, J. M. (1969). Markov chain models in life insurance. *Blätter der DGVFM 9*(2), 91–107.

Huber-Carol, C. (1994). Durées de survie tronquées et censurées. *Journal de la Société de statistique de Paris 135*(4), 3–23.

Hunt, A. and D. Blake (2021). On the structure and classification of mortality models. *North American Actuarial Journal 25*(sup1), S215–S234.

INSEE (2023). Espérance de vie à divers âges : Données annuelles de 1994 à 2022. https ://www.insee.fr/fr/statistiques/2416631#tableau-figure1.

INSEE (2023). *France, portrait social - Édition 2023*. INSEE.

INSEE (2024). Les espérances de vie : outil interactif. https ://www.insee.fr/fr/outil-interactif/6794598.

Kleinbaum, D. G. and M. Klein (1996). *Survival analysis a self-learning text*. Springer.

Koissi, M.-C., A. F. Shapiro, and G. Högnäs (2006). Evaluating and extending the lee–carter model for mortality forecasting : Bootstrap confidence interval. *Insurance : Mathematics and Economics 38*(1), 1–20.

LBPP (2020). Dépendance La Banque Postale. https ://www.my-prevoyance.fr/garanties/dépendance/les-contrats-dépendance/dependance-la-banque-postale.

Le Bastard, L. (2023a). Coherent extrapolation of mortality rates at old ages applied to long term care. *European Actuarial Journal*, 1–30.

Le Bastard, L. (2023b). Combining experience data of several Long-Term Care Insurance products with different disability definitions. preprint.

Le Gal, A. (2021). *Analyse du risque dépendance en France : calibration de lois biométriques et influence des caractéristiques des contrats sur la sinistralité.* Master's thesis, Institut du Risk Management.

Levy, P. (1954). Processus semi-markoviens. In *Proceedings of the International Congress of Mathematicians*, Volume 3, pp. 416–426.

Makeham, W. M. (1860). On the law of mortality and the construction of annuity tables. *Journal of the Institute of Actuaries 8*(6), 301–310.

McCullagh, P. (2019). *Generalized linear models.* Routledge.

Meslé, F. (2004). Écart d'espérance de vie entre les sexes : les raisons du recul de l'avantage féminin. *Revue d'Épidémiologie et de Santé Publique 52*(4), 333–352. "Genres et Santé".

Ministère de la Santé et de la Prévention (2021). Personnes âgées : les chiffres clés. https ://sante.gouv.fr/archives/loi-relative-a-l-adaptation-de-la-societe-au-vieillissement/article/personnes-agees-les-chiffres-cles#Quelques-donnees-cles.

Ministère de l'Economie et des Finances (2012). Arrêté du 18 décembre 2012 relatif à l'égalité entre les hommes et les femmes en assurance. *Journal officiel de la République française*.

Nelder, J. A. and R. W. M. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General) 135*(3), 370–384.

OpinionWay (2021). Les Français et la dépendance.

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

PREDICA (2018). Assurance Autonomie - Document d'Information sur le Produit

d'Assurance. https ://www.ag2rlamondiale.fr/files/live/sites/portail/files/pdf/Sante-Prevoyance/prevoyance/AG2R-LA-MONDIALE-Assurance-Autonomie.pdf.

Renshaw, A. and S. Haberman (1995). On the graduations associated with a multiple state model for permanent health insurance. *Insurance : Mathematics and Economics 17*(1), 1–17.

Seifarth, J. E., C. L. McGowan, and K. J. Milne (2012). Sex and life expectancy. *Gender medicine 9*(6), 390–401.

Serge Braudo (consulté le 02 janvier 2024). Dictionnaire du droit privé. https ://www.dictionnaire-juridique.com/definition/assurance.php.

Service Public (2024). Service-Public.fr, le site officiel de l'administration française. https ://www.service-public.fr/particuliers/vosdroits.

Simonoff, J. S. (2012). *Smoothing methods in statistics*. Springer Science & Business Media.

Soetewey, A., C. Legrand, M. Denuit, and G. Silversmit (2022). Semi-markov modeling for cancer insurance. *European Actuarial Journal 12*.

Thatcher, A. R. (1999). The long-term pattern of adult mortality and the highest attained age. *Journal of the Royal Statistical Society : Series A (Statistics in Society) 162*(1), 5–43.

United Nations (2019). World Population Prospects 2019. https ://population.un.org/wpp2019/Graphs/Probabilistic/PopPerc/65plus/900.

Weibull, W. (1951). A statistical distribution function of wide applicability. *Journal of applied mechanics*.

Xuanyuan, S. and S. Xuanyuan (2023). Application of markov model in long-term care insurance. *Highlights in Science, Engineering and Technology 47*, 9–15.

Chapitre 2

Coherent extrapolation of mortality rates at old ages applied to Long-Term Care

Ce chapitre reprend l'article "Coherent extrapolation of mortality rates at old ages applied to Long Term Care" publié à l'European Actuarial Journal.

Abstract

In an insurance context, Long-Term Care (LTC) products cover the risk of permanent loss of autonomy, which is defined by the impossibility or difficulty of performing alone all or part of the activities of daily living (ADL). From an actuarial point of view, knowledge of risk depends on knowledge of the underlying biometric laws, including the mortality of autonomous insureds and the mortality of disabled insureds. Due to the relatively short history of LTC products and the age limit imposed at underwriting, insurers lack information at advanced ages. This represents a challenge for actuaries, making it difficult to estimate those biometric laws.

In this paper, we propose to complete the missing information at advanced ages on the mortality of autonomous and disabled insured populations using information on the global mortality of the portfolio. In fact, the three previous mortality laws are linked since the portfolio is composed only of autonomous and disabled policyholders. We model the two mortality laws (deaths in autonomy and deaths in LTC) in a Poisson Generalized Linear Model framework, additionally using the P-Splines smoothing method. A constraint is then included to link the mortality laws of the two groups and the global mortality of the portfolio. This new method allows for estimating and extrapolating both mortality laws simultaneously in a consistent manner.

Keywords: Long-Term Care Insurance; Actuarial modelling; Generalized Linear Models; P-Splines; Extrapolation; Penalization.

2.1 Introduction

Long-Term Care (LTC) is linked to the risk that an individual loses their autonomy, resulting in the impossibility or difficulty of performing Activities of Daily Living (ADL), such as washing, eating, moving and dressing. Many causes can lead to a loss of autonomy, but the need for LTC is mostly due to illness occurring at old ages. With the persistent increase in life expectancy and the ageing of the Baby Boom generation, we are entering a period in which the number of people over 80 is likely to continue to grow. Eurostat (2022) estimates that the share of the population in Europe aged 80 years or above is likely to be multiplied by two and a half between 2021 and 2100. Therefore, it is expected that an increasing number of people will need financial support to cover the costs generated by the loss of autonomy. The average age at underwriting of an LTC product is approximately 60, while claims mainly occur after 85. This average underwriting age means for the insurer that only a few observations on the mortality of his portfolio at old ages will be available before the 25th anniversary of the product. This effect combined with the recency of LTC products makes it difficult to estimate the associated risk. Improving knowledge of the LTC risk is then a challenge for actuaries. In contrast, the mortality of the overall population was studied long before the emergence of LTC insurance products.

From an actuarial point of view, the insured's health condition is often represented by an illness-death model composed of three states, namely, "Autonomous", "Disabled" and "Dead". Some insurance contracts cover multiple levels of dependency with different levels of annuity. Actuaries may model these products with multi-state Markov models, with one state for each level of dependency. This choice multiplies the number of laws of transition from one state to another to estimate. The difficulty of calibrating multi-state Markov models comes from the scarcity of data. Most insurers do not observe enough transition from LTC states in their database due to the recency of their product. Therefore, papers modelling LTC products with multiple levels of dependency often make strong assumptions on the intensities of the model, or use big public data as in Biessy (2015) with data from the French LTC public aid called the "Allocation Personnalisée d'Autonomie" (APA). Fleischmann (2015) models an Austrian private health insurance product with 7 levels of severity and assumes that mortality is the same independently from the severity level and that the intensity to reach a severity level is independent of the state of origin and the time spent in that state. In addition to using public data, Biessy (2015) uses parametric laws. Another solution to model products covering multiple states of disability is to consider it as a set of illness-death models. This method is the solution mainly used by insurers. The method developed in this paper can therefore be used to model products covering multiple states of disability.

Since the state "Dead" is an absorbing state, the model is composed of 4 transitions, each

one associated with a biometric law. The first one corresponds to the incidence rates in the disabled state, whereas the second one represents its reverse transition. The two remaining biometric laws correspond to distinct mortality rates for autonomous and disabled lives. In practice it is very hard to reverse loss of autonomy. Recovery probabilities are negligible, especially when the "Disabled" state is associated with a high level of dependency. The definition of LTC in France emphasises the fact that the loss of autonomy must be permanent and irreversible. As in most of LTC product contracts, we consider in this paper that the loss of autonomy is final, which means that no return to autonomy is envisaged. With this hypothesis, which is representative of the real insurance market, only 3 biometric laws need to be estimated. The impact of allowing recovery when taking into account a low level of disability is discussed in Section 2.6.

Since LTC risk is mostly due to ageing pathologies, the estimation of mortality at old ages is of importance for pricing and estimation of risk liabilities. In the context of mortality modelling, a common approach is to fit a parametric model on the crude death rates and assume that information available at younger ages would explain the behaviour at older ages, where we have no or not enough observations. Depending on the selected parametric model, a different underlying assumption on the shape of the mortality curve is made. Some of these models are compared in Hammond (2000) on 13 countries (European, Scandinavian and Japanese) using data from 1960 to 1990. A different approach is used in this paper to extrapolate mortality at old ages, relying on the P-Splines smoothing method introduced by Eilers and Marx (1996). This methodology was adapted to mortality for the first time by Currie et al. (2004).

The incidence rates and mortality rates of autonomous and disabled insureds are usually estimated and extrapolated independently. However, in this way, the consistency between the mortality laws is not guaranteed, and the predicted number of deaths in the whole portfolio might differ from the sum of the predicted numbers of deaths in autonomy and in LTC. Let D_x^G be the number of observed deaths between age x and $x+1$ in group $G \in \{A, D, gen\}$ and $\widehat{D}_x^G$ its predicted value, where A and D represent the groups of autonomous and disabled insureds, respectively, and gen represents the overall portfolio of insureds. Then, $D_x^{gen} = D_x^A + D_x^D$, and in the case of consistent mortality laws, the relation between the expected values must be given by the following equation

$$\widehat{D}_x^{gen} = \widehat{D}_x^A + \widehat{D}_x^D. \tag{2.1}$$

In the literature, the problem of consistency between mortality laws is mostly approached in the context of prospective modelling to ensure that the mortality laws do not diverge indefinitely over time between several groups. This idea of coherent mortality forecasting was first introduced by Li and Lee (2005). Li proposed a method based on the Lee-Carter

model to forecast the mortality of a group of populations by allowing each population to have its own age pattern and level but have a common trend. Later, Zhou et al. (2019) and Li et al. (2017) approached this problem of coherent mortality forecasting with the concept of semicoherence. The idea is to fix a weaker assumption on the coherence between the mortality laws by allowing the mortality trajectories of two populations to diverge, as long as the difference between the two mortality laws does not exceed what they called a tolerance corridor. Noticing that the coherent assumption can be too strong, especially when it is imposed on a large number of populations, Guibert et al. (2020) proposed a new approach based on locally coherent mortality forecasts by assuming that the coherence principle is verified by subgroups of populations.

This paper aims to develop a method that improves the estimation and extrapolation of the mortality laws of autonomous and disabled groups, using knowledge on the mortality of the overall population (union of the 2 groups) while keeping a smooth structure of the mortality laws. To this end, we use the P-Splines smoothing method proposed in Eilers and Marx (1996) and adapted to mortality estimation in Currie et al. (2004) and Macdonald et al. (2018). A constraint based on Equation (2.1) is then included to link the mortality laws of the two groups and the global mortality of the portfolio. The idea of adding constraints to the P-Splines method has already been used in Bollaerts et al. (2006) for research on the cognitive development of children and for mortality modelling in Camarda et al. (2016), Remund et al. (2018) and in Camarda (2019). The method proposed in this paper uses an algorithm converging under certain conditions discussed in Appendix 2.A. The goal of this paper is to provide an algorithm for actuaries in charge of the pricing of LTC products. This paper is not intended to present limit theorems of convergence of estimators because they are difficult to obtain. The simulations support the interest of the method.

We show that this approach leads to a better estimate of the death rates for both autonomous and disabled insureds, providing an estimate of the predicted number of deaths of the overall population close to the sum of the predicted number of deaths in autonomy and LTC.

Section 2.2 of this paper introduces the dataset and the model. In Subsection 2.2.2, we present the continuous multi-state Markov model used in the context of LTC modelling and explain its link with the Poisson model and the Poisson generalized linear model (Poisson-GLM). The P-Splines smoothing method, used to maintain a smooth structure of the mortality laws, is proposed in Subsection 2.2.3. We add a constraint on the consistency between mortality laws based on Equation (2.1) in Subsection 2.2.4.

Section 2.3 focuses on the extrapolation of the mortality laws when no exposures are available at old ages. We propose an extrapolation method with reconstruction of the

exposures using the model proposed in the first part of the paper. Section 2.4 addresses the problem of the choice of the hyper-parameter corresponding to the weight that we give to the consistency constraint. The larger this parameter is, the better the mortality laws estimated by the algorithm satisfy the coherence rule. An application on a real dataset is made in Section 2.5. Recovering from a high level of dependency can easily be assumed as impossible. However, one might wish to assume that recovery is possible when modelling low levels of LTC. Section 2.6 discusses how to use the loopback algorithm to model a product covering several levels of LTC, especially when the lower levels allow recovery. Concluding remarks are provided in Section 2.7.

2.2 Modelling

2.2.1 Data structure

The biometric laws are calibrated to an LTC portfolio observed in a given period. The trajectories of insured individuals, meaning their health status at each time of the period of observation, are observed. For each insured individual, the following information is available:

- date of birth,
- gender,
- underwriting date,
- date of loss of autonomy if it occurred,
- date of death if it occurred, and
- date of exit from the portfolio in case of a contract cancellation.

Since males and females do not have the same mortalities or same probabilities of loss of autonomy at each age, biometric laws are estimated separately by gender. The information of all the insureds is then aggregated to construct two databases per gender. The first one, denoted by DB^A, is used to study the autonomous experience, and the second one, denoted by DB^D, is used to study the experience in LTC. For each integer age x, DB^A and DB^D contain:

- the central exposure to risk between age x and $x+1$,
- the number of observed deaths between ages x and $x+1$, and
- the number of reported losses of autonomy (only for DB^A) between age x and $x+1$.

For a given age x, the central exposure to risk in autonomy (resp. LTC) corresponds to the sum of individual exposures of each insured in autonomous (resp. LTC) state between x and $x+1$. The individual exposure of an insured at age x in a given state (autonomous or LTC) is the fraction of time spent in this state between age x and $x+1$. For example, an insured in autonomous state at his 65th birthday, losing his autonomy 9 months after, and surviving at least until his 66th birthday in LTC has an exposure in the autonomous

state equal to 0.75, which corresponds to 3/4 of a year, and an exposure in LTC equal to 1/4.

2.2.2 Modelling of a Long-Term Care product

We consider in this paper continuous multi-state Markov models, as shown in Figure 2.1, with three states: autonomy (A), LTC/Disability (D) and death. Such models are often used by insurers in practice. As no return to a better state of health is envisaged in this paper, e.g., in Nuttall et al. (1994) and Alegre et al. (2003), only three laws are needed in this model:

- i_x is the incidence intensity at age x,
- λ_x^A is the mortality intensity in state A at age x, and
- λ_x^D is the mortality intensity in state D at age x.

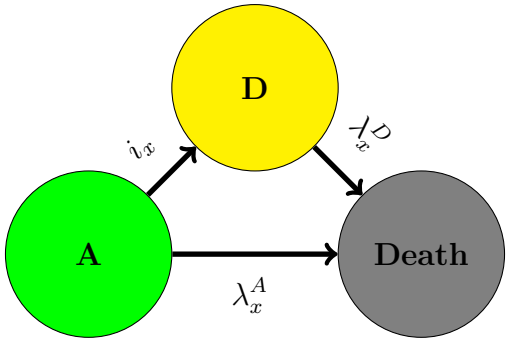


Figure 2.1 – Modelling of an LTC product

Let X_x be the current state of an individual at age $x \in \mathbb{R}^+$. The transition intensities are defined as

$$i_x = \lim_{h \to 0^+} \frac{\mathbb{P}(X_{x+h} = D | X_x = A)}{h} ,$$

$$\lambda_x^A = \lim_{h \to 0^+} \frac{\mathbb{P}(X_{x+h} = \text{Death} | X_x = A)}{h} ,$$

$$\lambda_x^D = \lim_{h \to 0^+} \frac{\mathbb{P}(X_{x+h} = \text{Death} | X_x = D)}{h} .$$

Let x be an integer. The notation is as follows. Consider independent individuals $j = 1, \ldots, n$ observed at least one day in the autonomous state A between x and $x+1$. For each individual j, let $x + {}^j a$ and $x + {}^j b$ be the age at the beginning and end of observation in state A, respectively, for the age band $[x; x+1]$. Let $x + {}^j c$ be the age at the end of observation in the portfolio between x and $x+1$. Then, $0 \leqslant {}^j a \leqslant {}^j b \leqslant {}^j c \leqslant 1$. If j does not enter state D between x and $x+1$, then ${}^j b = {}^j c$. The end of observation in state A of an individual between x and $x+1$ can be due to three reasons:

1. right censoring,

2. loss of autonomy, or
3. death.

Let ${}^{j}d_x$ and ${}^{j}LTC_x$ indicate the cause of the end of observation in state A of j such that:

- ${}^{j}d_x = 1$ if the cause is death and 0 otherwise, and
- ${}^{j}LTC_x = 1$ if the cause is the loss of autonomy and 0 otherwise.

Let ${}^{j}d_x^{LTC} = 1$ if individual j dies in state D between x and $x+1$, 0 otherwise.

The main assumptions that we need in this paper are as follows:

Assumptions

1. the mortality rates remain constant throughout the age interval from x to $x+1$ (where x is an integer),
2. the logarithm of mortality rates may be decomposed on a P-Splines basis (see Section 2.2.3),
3. incidence and general mortality laws are assumed to be known and are not estimated.

Then, using classic tools of survival analysis and methods for modelling competing risks, as explained in Section 2.3 in Porta et al. (2007), and from the definitions of the intensities, the likelihood associated with individual j between x and $x+1$ is given by

$$ {}^{j}L_x = \underbrace{\exp\left(-\int_{{}^{j}a}^{{}^{j}b}(\lambda^A_{x+u}+i_{x+u})\,\mathrm{d}u\right)(i_{x+{}^{j}b})^{{}^{j}LTC_x}(\lambda^A_{x+{}^{j}b})^{{}^{j}d_x}}_{{}^{j}L_x^A}\ \underbrace{\exp\left(-\int_{{}^{j}b}^{{}^{j}c}(\lambda^D_{x+u})\,\mathrm{d}u\right)(\lambda^D_{x+{}^{j}c})^{{}^{j}d_x^{LTC}}}_{{}^{j}L_x^D}. \tag{2.2} $$

The likelihood of individual j in Equation (2.2) can be separated into 2 distinct partial likelihoods. ${}^{j}L_x^A$ corresponds to the experience of j in state A, whereas ${}^{j}L_x^D$ corresponds to its experience in state D. If j is not observed in state D between x and $x+1$, then ${}^{j}L_x^D = 1$. We can then study the experience in states A and D separately.

Under Assumption 1, ${}^{j}L_x^A$ becomes

$$ {}^{j}L_x^A = \exp\left(-(\lambda_x^A+i_x)\,{}^{j}e_x^A\right)(i_x)^{{}^{j}LTC_x}(\lambda_x^A)^{{}^{j}d_x}, \tag{2.3} $$

where ${}^{j}e_x^A = {}^{j}b - {}^{j}a$ is the time of exposure to the risk in the autonomous state of individual j between age x and $x+1$.

Therefore, the likelihood for the age band $[x; x+1]$ for the overall population of individuals

$j = 1, \ldots, n$, being the product of all individuals likelihood is equal to

$$L_x^A = \exp\left(-(\lambda_x^A + i_x)e_x^A\right)(i_x)^{LTC_x}(\lambda_x^A)^{d_x}, \tag{2.4}$$

where e_x^A, called the central exposure to risk is the sum of the time exposed to the risk in autonomy by all individuals in the age band, and LTC_x and d_x are the total number of observed losses of autonomy and deaths, respectively.

Maximizing the likelihood L_x^A with respect to λ_x^A and i_x is equivalent to maximizing separately

- $L_x^{A\to D}(i_x) = \exp\left(-i_x e_x^A\right)(i_x)^{LTC_x}$, and
- $L_x^{A\to \text{Death}}(\lambda_x^A) = \exp\left(-\lambda_x^A e_x^A\right)(\lambda_x^A)^{d_x}$.

We note that $L_x^{A\to D}(i_x)$ and $L_x^{A\to \text{Death}}(\lambda_x^A)$ are proportional to the likelihood of Poisson variables with expectancies equal to $i_x^A e_x^A$ and $\lambda_x^A e_x^A$, respectively. Therefore, the maximum likelihood estimators of i_x and λ_x^A obtained by maximizing Equation (2.4) are equal to the maximum likelihood estimators of the Poisson distributions. The incidence intensity i_x is considered as known in this paper. The likelihood has to be maximized with respect to λ_x^A only. One can therefore assume that the number of observed deaths in autonomy (state A) at age x follows a Poisson distribution with parameter $\lambda_x^A e_x^A$. The same reasoning can be applied to deaths in the LTC state (state D). This result is interesting, allowing us to use the theory of Poisson-GLM to estimate λ_x^A. In the rest of the paper, we assume that the number of deaths at age x in both autonomous and LTC states follows a Poisson distribution with parameter $\lambda_x^G e_x^G$ where $G \in \{A, D, gen\}$.

Let x_{min} and x_{max} be integers corresponding to the minimum and maximum observed ages. Let d_x^A and d_x^D be the observed deaths at age x in states A and D, respectively. Since we may not have enough observations at certain ages, we introduce an indicator function w_x^G for each group indicating if the observations in the associated group at age x are reliable and can be included in the log-likelihood.

The total likelihood for all the observations from x_{min} to x_{max} in group G is given by

$$L^G\left(\lambda_{x_{min}}^G, \ldots, \lambda_{x_{max}}^G\right) = \prod_{x=x_{min}}^{x_{max}} \left[\frac{\left(\lambda_x^G e_x^G\right)^{d_x^G}}{d_x^G!}\exp\left(-\lambda_x^G e_x^G\right)\right]^{w_x^G}, \tag{2.5}$$

where $w_x^G = 1$ if the observations in the associated group at age x are reliable and 0 otherwise.

Equation (2.5) is the product of likelihoods of Poisson distributions for each age x between x_{min} and x_{max}.

The maximum likelihood estimators of the intensities $\hat{\lambda_x^G}$ for $x \in \{x_{min}; x_{min}+1; ...; x_{max}\}$

are given by the ratio d_x^G/e_x^G. Each intensity is fully determined by the deaths and exposure at this age, regardless of the observations at neighbouring ages. This can imply a very irregular curve of the mortality law that can be explained by the variance of the estimator, equal to d_x/e_x^2. Smoothing methods can be used to obtain a smoother mortality law, reducing volatility in the results. In this paper, we use the P-Splines smoothing method, which is widely used in the literature to smooth mortality intensities.

2.2.3 A model based on P-Splines

The method of P-Splines is a method of smoothing embedded in the GLM framework described in Eilers and Marx (1996), Marx and Eilers (1998), Eilers and Marx (2002), or Currie and Durban (2002).

Let $n = x_{max} - x_{min}$ be an integer. Let J be an integer representing the number of splines. In this method:

1. Let B be a basis matrix of cubic splines such that $B_{i,j}$ is the value of the cubic spline j at the i^{th} age. For a given group (A or D), the curve of the mortality intensities is considered a linear combination of J cubic splines. Let $\boldsymbol{\theta^G} = \{\theta_1^G, \ldots, \theta_J^G\}$ be the vector of coefficients such that $\log(\lambda^G_{\boldsymbol{\theta^G}, x_i}) = \sum_{j=1}^{J} B_{i,j}^G \theta_j^G$ for each $i \in \{1, \ldots, n+1\}$. Let $\boldsymbol{\Lambda}^G_{\boldsymbol{\theta^G}} = \left(\lambda^G_{\boldsymbol{\theta^G}, x_{min}}, \lambda^G_{\boldsymbol{\theta^G}, x_{min}+1}, \ldots, \lambda^G_{\boldsymbol{\theta^G}, x_{max}}\right)^T$. We can write this linear combination in matrix form as

   $$\log(\boldsymbol{\Lambda}^G_{\boldsymbol{\theta^G}}) = B^G \boldsymbol{\theta^G},$$

   where $B^G \in M_{n+1,J}(\mathbb{R}^+)$ represents the matrix of the J splines at each point $\{x_{min}, \ldots, x_{max}\}$. B_{ij}^G is the value of the j^{th} spline at the i^{th} age of group G. This matrix is called the B-spline basis.
2. Let d be an integer. A penalty term $\rho(D_d^G \boldsymbol{\theta^G})^T (D_d^G \boldsymbol{\theta^G})$ depending on the penalty order d is added to the log-likelihood to avoid complex models with excessively large variability between coefficients of adjacent splines. $\rho \in \mathbb{R}$ is a smoothing parameter giving a weight to the penalty. Let $\Delta\theta_j^G = \theta_j^G - \theta_{j-1}^G$, $\Delta^2\theta_j^G = \Delta(\Delta\theta_j^G) = \theta_j^G - 2\theta_{j-1}^G + \theta_{j-2}^G$, ..., $\Delta^d\theta_j^G = \Delta(\Delta^{d-1}\theta_j^G)$.

   D_d^G is defined as the matrix satisfying $D_d^G \boldsymbol{\theta^G} = \Delta^d \boldsymbol{\theta^G}$.

   Let $P_d^G = \rho(D_d^G)^T D_d^G$; a simpler way to write the penalty term that is used in the rest of the paper is $(\boldsymbol{\theta^G})^T P_d^G \boldsymbol{\theta^G}$.

Therefore, J coefficients θ_j^G must be estimated for each group G instead of one by age. In addition to having a smoother curve, the number of coefficients to estimate is then lower than if no smoothing method were used. The extrapolation mostly depends on the

order of the penalty. The smoothing parameter ρ can be chosen to minimise the BIC as recommended in Currie and Durban (2002). The choice of other parameters, such as the number of nodes or the degree of the splines, may be less critical, as different choices often lead to similar smoothings. Ruppert (2000) and Eilers and Marx (1996) study the choice of the P-Splines parameters. The following rule of thumb is often sufficient:

- In the case of equidistant data, fix 1 node every 4 or 5 observations,
- Use cubic splines (order 3).

Let $(B^G\boldsymbol{\theta^G})_k$ be the k^{th} coefficient of the vector $\log(\boldsymbol{\Lambda}^G_{\boldsymbol{\theta^G}}) = B^G\boldsymbol{\theta^G}$. The penalized log-likelihood for group G is given by

$$\begin{aligned} l^G_{pen}(\boldsymbol{\theta^G}) &= \log(L^G(\boldsymbol{\theta^G})) - \frac{1}{2}(\boldsymbol{\theta^G})^T P^G_d \boldsymbol{\theta^G} \\ &= \sum_{x=x_{min}}^{x_{max}} w^G_x \left[d^G_x \log(\lambda^G_{\boldsymbol{\theta^G},x}) - \lambda^G_{\boldsymbol{\theta^G},x} e^G_x\right] - \frac{1}{2}(\boldsymbol{\theta^G})^T P^G_d \boldsymbol{\theta^G} \\ &= \underbrace{\sum_{k=0}^{x_{max}-x_{min}} w^G_{x_{min}+k}\left[d^G_{x_{min}+k}(B^G\boldsymbol{\theta^G})_{k+1} - \exp((B^G\boldsymbol{\theta^G})_{k+1}) e^G_{x_{min}+k}\right]}_{l^G(\boldsymbol{\theta^G})} - \frac{1}{2}(\boldsymbol{\theta^G})^T P^G_d \boldsymbol{\theta^G}. \end{aligned} \tag{2.6}$$

In the following, the log-likelihood for autonomous and LTC groups is denoted by $l^A(\boldsymbol{\theta^A})$ and $l^D(\boldsymbol{\theta^D})$, respectively.

It is possible to smooth the intensities of the 2 groups simultaneously but independently (i.e. the observations of one group have no influence on the estimate of the mortality of the other). As the respective penalized log-likelihoods are independent, maximizing the sum of these log-likelihoods is equivalent to maximizing both of them. This is then equivalent to maximizing $l^{A/D}_{pen}\left(\boldsymbol{\theta^A}, \boldsymbol{\theta^D}\right)$ given by the following equation

$$\begin{aligned} l^{A/D}_{pen}\left(\boldsymbol{\theta^A}, \boldsymbol{\theta^D}\right) = &\sum_{k=0}^{x_{max}-x_{min}} w^A_{x_{min}+k}\left[d^A_{x_{min}+k}(B^A\boldsymbol{\theta^A})_{k+1} - \exp((B^A\boldsymbol{\theta^A})_{k+1}) e^A_{x_{min}+k}\right] - \frac{1}{2}(\boldsymbol{\theta^A})^T P^A_d \boldsymbol{\theta^A} \\ &+ \sum_{k=0}^{x_{max}-x_{min}} w^D_{x_{min}+k}\left[d^D_{x_{min}+k}(B^D\boldsymbol{\theta^D})_{k+1} - \exp((B^D\boldsymbol{\theta^D})_{k+1}) e^D_{x_{min}+k}\right] - \frac{1}{2}(\boldsymbol{\theta^D})^T P^D_d \boldsymbol{\theta^D}. \end{aligned} \tag{2.7}$$

To this aim, we introduce the basis spline matrix $B = \begin{bmatrix} B^A & 0 \\ 0 & B^D \end{bmatrix} \in M_{2(n+1),2J}(\mathbb{R}^+)$ and the penalty matrix $P = \begin{bmatrix} P^A_d & 0 \\ 0 & P^D_d \end{bmatrix} \in M_{2J,2J}(\mathbb{R})$.

Let $\boldsymbol{\theta}$ be the vector of all the smoothing coefficients, i.e. $\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta^A} \\ \boldsymbol{\theta^D} \end{pmatrix} \in \mathbb{R}^{2J}$, and $\log(\boldsymbol{\Lambda_\theta}) = \begin{pmatrix} \log(\boldsymbol{\Lambda}^A_{\boldsymbol{\theta}^A}) \\ \log(\boldsymbol{\Lambda}^D_{\boldsymbol{\theta}^D}) \end{pmatrix} \in \mathbb{R}^{2(n+1)}$; then:

1. $\log(\boldsymbol{\Lambda_\theta}) = B\boldsymbol{\theta}$,
2. The sum of the penalties of the 2 groups can be written as $\frac{1}{2}\boldsymbol{\theta}^T P \boldsymbol{\theta}$.

We introduce the following vectors:

- $\mathbf{d} = \left(d^A_{x_{min}}, \ldots, d^A_{x_{min}+n}, d^D_{x_{min}}, \ldots, d^D_{x_{min}+n}\right)^T \in \mathbb{N}^{2(n+1)}$,
- $\mathbf{e} = \left(e^A_{x_{min}}, \ldots, e^A_{x_{min}+n}, e^D_{x_{min}}, \ldots, e^D_{x_{min}+n}\right)^T \in \mathbb{R}^{2(n+1)}$, and
- $\mathbf{w} = \left(w^A_{x_{min}}, \ldots, w^A_{x_{min}+n}, w^D_{x_{min}}, \ldots, w^D_{x_{min}+n}\right)^T \in \{0,1\}^{2(n+1)}$,

of length $2(n+1)$. Then, the penalized log-likelihood on all the observations is

$$\begin{aligned} l^{A/D}_{pen}(\boldsymbol{\theta}) &= l^A(\boldsymbol{\theta}) + l^D(\boldsymbol{\theta}) - \underbrace{\frac{1}{2}\boldsymbol{\theta}^T P \boldsymbol{\theta}}_{\text{P-Splines smoothing penalty } Pen^{smoothing}} \\ &= \sum_{k=0}^{n} \mathbf{w}_{k+1}\Big[\mathbf{d}_{k+1}(B\boldsymbol{\theta})_{k+1} - \exp((B\boldsymbol{\theta})_{k+1})\mathbf{e}_{k+1}\Big] \\ &+ \sum_{k=n+1}^{2n+1} \mathbf{w}_{k+1}\Big[\mathbf{d}_{k+1}(B\boldsymbol{\theta})_{k+1} - \exp((B\boldsymbol{\theta})_{k+1})\mathbf{e}_{k+1}\Big] - \frac{1}{2}\boldsymbol{\theta}^T P \boldsymbol{\theta}. \end{aligned}$$

2.2.4 Introduction of a second penalty on the log-likelihood

In this paper, the mortality intensities λ^{gen}_x of the general population are assumed to be known and to be piecewise constant as the mortality laws of groups A and D. The idea of our approach is to minimize the gap between the predicted number of deaths in the general population and the sum of the deaths in autonomous and LTC states (cf. Equation (2.1)). As the portfolio is composed of autonomous and dependent individuals, the total number of deaths at age x in the portfolio is equal to the sum of deaths in autonomy (state A) and LTC state (state D), as shown on Figure 2.2. This figure shows the possible transitions in an LTC product. One observed death is either a death in autonomy or in the LTC state. The exposures in these states are respectively denoted by e^A and e^D. Therefore, the exposure of the portfolio is $e^A + e^D$. In the context of a Poisson distribution for the number of observed deaths, Equation (2.1) can be written as follows

$$\lambda^{gen}_x(e^A_x + e^D_x) = \lambda^A_x e^A_x + \lambda^D_x e^D_x. \tag{2.8}$$

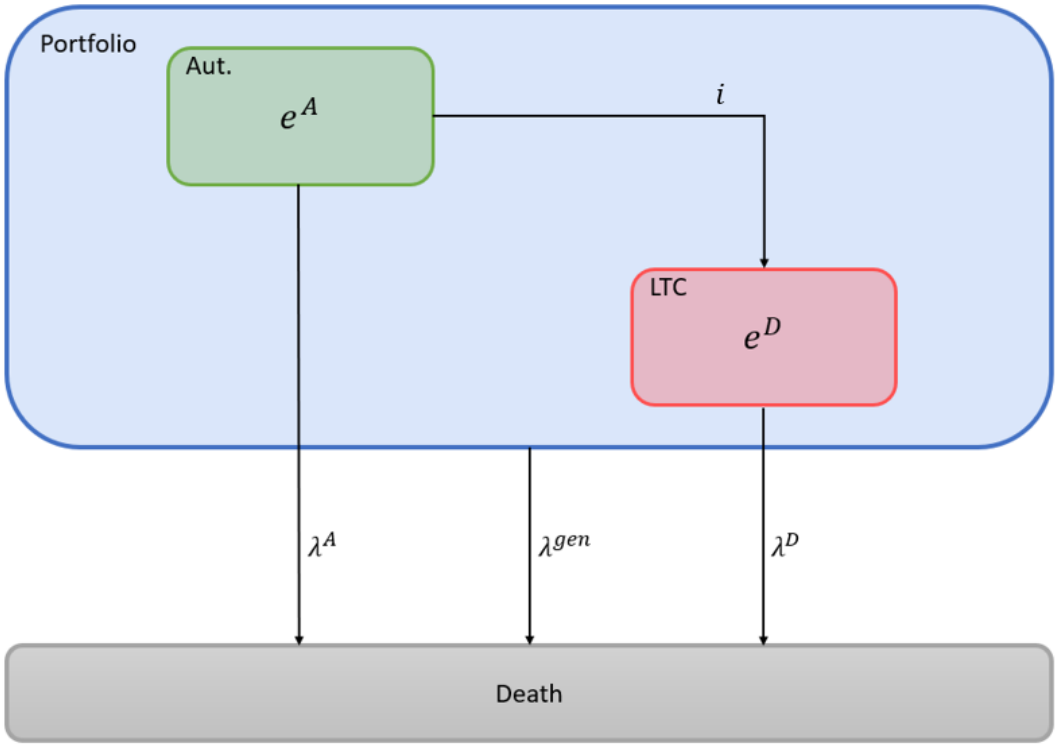


Figure 2.2 – Modelling of an LTC product keeping consistency between mortality laws

To respect the coherence criteria between the 3 mortality laws (autonomous, LTC, and general), a second penalty term given by the following equation

$$Pen^{loopback}(\boldsymbol{\theta}) = \frac{1}{2}K \sum_{x=x_{min}}^{x_{max}} \left(\frac{\lambda_x^{gen}\left(e_x^A + e_x^D\right) - \lambda_{\boldsymbol{\theta},x}^A e_x^A - \lambda_{\boldsymbol{\theta},x}^D e_x^D}{e_x^A + e_x^D} \right)^2 \tag{2.9}$$
$$= \frac{1}{2}K \sum_{k=0}^{n} \left(\frac{\lambda_{x_{min}+k}^{gen}\left(\mathbf{e}_{k+1} + \mathbf{e}_{(n+1)+k+1}\right) - \Lambda_{\boldsymbol{\theta},k+1}\mathbf{e}_{k+1} - \Lambda_{\boldsymbol{\theta},(n+1)+k+1}\mathbf{e}_{(n+1)+k+1}}{\mathbf{e}_{k+1} + \mathbf{e}_{(n+1)+k+1}} \right)^2,$$

is added to the log-likelihood, where K is a parameter corresponding to the weight given to the consistency criteria in the estimation of the mortality laws.

The new penalized log-likelihood now becomes

$$l_{pen}^{loopback}(\boldsymbol{\theta}) = l_{pen}^{A/D}(\boldsymbol{\theta}) - Pen^{loopback}(\boldsymbol{\theta}). \tag{2.10}$$

This penalized log-likelihood is maximized by the Newton-Raphson algorithm. The first and second derivatives with respect to the coefficients θ_i are needed. The matrix forms of the gradient and the Hessian are given by Equations (2.11) and (2.12), respectively.

$$\nabla_{\boldsymbol{\theta}}(l_{pen}) = B^T W(\mathbf{d} - \check{\mathbf{d}}_{\boldsymbol{\theta}}) - P\boldsymbol{\theta} - KB^T \left((\tilde{W}_3^{-1})^2 W_{\boldsymbol{\theta}}^Q \otimes I_2 \right) \check{\mathbf{d}}_{\boldsymbol{\theta}}, \tag{2.11}$$

$$H_{\boldsymbol{\theta}}(l_{pen}) = -B^T W W_{\boldsymbol{\theta}} B - P - K B^T \left[W_{\boldsymbol{\theta}} \left([(\tilde{W}_3^{-1})^2 W_{\boldsymbol{\theta}}^Q] \otimes I_2 \right) \right] B - K \left[\tilde{W}_3^{-1} W_{\boldsymbol{\theta}}^A B_A \quad \tilde{W}_3^{-1} W_{\boldsymbol{\theta}}^D B_D \right]^T \left[\tilde{W}_3^{-1} W_{\boldsymbol{\theta}}^A B_A \quad \tilde{W}_3^{-1} W_{\boldsymbol{\theta}}^D B_D \right], \quad (2.12)$$

where:

- the basis matrices $B \in M_{2(n+1),2J}(\mathbb{R}^+)$, $B_A \in M_{(n+1),J}(\mathbb{R}^+)$, $B_D \in M_{(n+1),J}(\mathbb{R}^+)$, the penalty matrix $P \in M_{2J,2J}(\mathbb{R})$ and the vector of deaths $\mathbf{d} \in \mathbb{N}^{2(n+1)}$ are introduced in Section 2.2.3,
- $\check{\mathbf{d}}_{\boldsymbol{\theta}} = (\Lambda_{\boldsymbol{\theta},i}\mathbf{e}_i)_{i\in\{1,\dots,2(n+1)\}}$ is the expected number of observed deaths with the assumption of a Poisson distribution,
- $W = \text{diag}(\mathbf{w})$,
- I_2 is the 2×2 identity matrix
- $W_{\boldsymbol{\theta}} = \text{diag}(\check{\mathbf{d}}_{\boldsymbol{\theta}}) \in M_{2(n+1),2(n+1)}(\mathbb{R})$,
- $W_{\boldsymbol{\theta}}^G = \text{diag}(\check{\mathbf{d}}_{\boldsymbol{\theta}}^C) \in M_{(n+1),(n+1)}(\mathbb{R})$ where $G \in \{A, D\}$,
- $\tilde{W}_3 = \text{diag}\left((\mathbf{e}_{k+1} + \mathbf{e}_{(n+1)+k+1})_{k\in\{0,\dots,n\}}\right) \in M_{(n+1),(n+1)}(\mathbb{R})$ is the diagonal matrix of the total exposure at each age, and
- $W_{\boldsymbol{\theta}}^Q = \text{diag}\left((\Lambda_{\boldsymbol{\theta},k+1}\mathbf{e}_{k+1} + \Lambda_{\boldsymbol{\theta},(n+1)+k+1}\mathbf{e}_{(n+1)+k+1} - \lambda^{gen}_{x_{min}+k}[\mathbf{e}_{k+1} + \mathbf{e}_{(n+1)+k+1}])_{k\in\{0,\dots,n\}}\right)$ $\in M_{(n+1),(n+1)}(\mathbb{R})$.

The Newton-Raphson algorithm is used to find the optimal coefficients $\boldsymbol{\theta}$. The estimator at step $k+1$ denoted $\hat{\boldsymbol{\theta}}^{(k+1)}$ is given by

$$\widehat{\boldsymbol{\theta}}^{(k+1)} = \widehat{\boldsymbol{\theta}}^{(k)} - H_{\widehat{\boldsymbol{\theta}}^{(k)}}(l_{pen})^{-1} \nabla_{\widehat{\boldsymbol{\theta}}^{(k)}}(l_{pen}). \quad (2.13)$$

The algorithm stops when the maximum relative difference between two coefficients from successive iterations is lower than a previously fixed tolerance ε (i.e $\max_{i\in\{1,\dots,2J\}} \left| \frac{\widehat{\boldsymbol{\theta}}_i^{(k)} - \widehat{\boldsymbol{\theta}}_i^{(k-1)}}{\widehat{\boldsymbol{\theta}}_i^{(k)}} \right| < \varepsilon$). The final estimator of $\boldsymbol{\theta}$ is denoted by $\widehat{\boldsymbol{\theta}}$.

The convergence of the algorithm is discussed in Appendix 2.A.

2.3 Extrapolation of mortality laws: calibration of theoretical exposures at old ages

Extrapolating the mortality laws is necessary when not having enough observations at certain old ages. In this case, the lack of observations does not enable the inclusion of this information at old ages in the likelihood in Equation (2.10). The weights of these ages

are then fixed to 0 ($w_x = 0$). The corresponding intensities are exclusively determined in such a way as to minimize both the P-Splines and consistency penalties ($Pen^{smoothing}$ and $Pen^{loopback}$).

2.3.1 Research problem

The previous methodology is adapted when having non-zero exposures up to the maximum age at which we wish to estimate mortality law. However, the data often contain very few or no observations at old ages. In fact, as age increases, exposures in the portfolio of the insurer tend to decrease because of deaths. As a consequence, the loopback penalty, based on exposures, is of undetermined form. The extrapolation is therefore done entirely with the P-Splines penalty. However, the objective and the interest of the loopback is to extrapolate the mortality laws in a coherent way to better estimate the mortality at old ages, using the information on the general mortality at these ages.

We overcome this point by computing theoretical exposures at these old ages.

2.3.2 Estimation of theoretical exposures and extrapolation

To overcome this issue, a maximum age at which we have enough observations is fixed. All exposures above this age x_M are estimated to compute the distribution of autonomous and LTC people at each age in the general population. The theoretical and estimated exposures are only used in the loopback penalty term $Pen^{loopback}$ of Equation (2.10).

Suppose that the general mortality intensities of the overall portfolio and the incidence are known. Theoretical exposures at each age above x_M and in each group can be obtained by projecting the population observed at the maximum age previously fixed. To this aim, we start the projection with the exposures computed at age x_M. The mortality laws and the incidence intensities are then used to estimate the exposures at older ages.

Let $\mathbf{e}$ and $\mathbf{d}$ be the vectors of observed exposures and deaths in the portfolio, respectively, as defined in Section 2.2.3. Let $\mathbf{e^{est}}$ be the estimated vector of exposures. The k^{th} terms of $\mathbf{e}$ and $\mathbf{e^{est}}$ are equal ($\mathbf{e^{est}}_k = \mathbf{e}_k$) for each k corresponding to an age below x_M. Here, $\mathbf{e^{est}}$ is estimated by projection using mortality laws and incidence intensities, depending on the vector of coefficients of the splines $\boldsymbol{\theta}$. Therefore, $\mathbf{e^{est}}$ is denoted by $\mathbf{e^{est}}(\boldsymbol{\theta})$.

Let $l_{pen}^{loopback}\left(\boldsymbol{\theta}|\mathbf{e^{est}}, \mathbf{d}\right)$ be the penalized log-likelihood from Equation (2.10) given the estimated exposure vector $\mathbf{e^{est}}$ and the vector of death counts $\mathbf{d}$. We want to compute the mortality laws by maximizing this penalized log-likelihood with respect to $\boldsymbol{\theta}$. However, the exposures are needed to compute the mortality laws, and the mortality laws are needed to estimate the exposures by projection.

The problem that we want to solve in this section is then:

$$\max_{\boldsymbol{\theta}} l_{pen}^{loopback}\left(\boldsymbol{\theta}|\mathbf{e}^{\mathbf{est}}(\boldsymbol{\theta}),\mathbf{d}\right).$$

The exposures at old ages are then estimated iteratively with Algorithm 1 by updating the mortality laws and exposures simultaneously. The Newton-Raphson algorithm in Section 2.2.4 is first used with $K = 0$ to compute mortality laws without the consistency constraint. From the resulting mortality laws and the known incidence intensities, exposures are first estimated by projection of the portfolio at age x_M. The following two steps are then repeated several times. First, mortality laws are computed using the Newton-Raphson algorithm from Section 2.2.4 with the estimated exposures and the chosen parameter K to link the estimation of the two mortality laws. The second stage consists of the re-computation of the exposures using the mortality laws from the previous step.

Let:

- $loopback(data, K, expo)$ be the loopback function applied to $data$ with the Newton-Raphson algorithm in Section 2.2.4, with $expo$ the exposures in both autonomous and LTC group (A and D respectively) at each age and K the loopback penalty chosen for the calibration of the model,
- $compute\_expo(expo_{x_M}, incidence, mortality_A, mortality_D)$ be the projection function that estimates the theoretical exposures given the incidence ($incidence$), the mortality laws ($mortality_A$ and $mortality_D$), and the exposures $expo_{x_M}$ at the age chosen for the projection x_M (the last age for which we consider the real exposures). This projection is made by considering the exposure in each group at age x_M as the number of insureds in each group. Under the assumption that all deaths and losses of autonomy occur at the end of the period (i.e., just before the birthday of the insured) and from the transition probabilities at each age, we are then able to estimate the number of insureds in states A and D at age $x \geqslant x_M$.
- $c(mortality_A, mortality_D)$ be the concatenation of the vectors of mortality intensities in states A and D,
- $expo_{data}$ be the real exposures observed at each age.

Algorithm 1 Exposures estimation algorithm

$K \leftarrow 0$
$expo \leftarrow expo_{data}$
$c(mortality_A, mortality_D) \leftarrow loopback(data, K, expo)$
$expo \leftarrow compute\_expo(expo_{x_M}, incidence, mortality_A, mortality_D)$
$K \leftarrow$ penalty parameter chosen for model calibration
for i=1,...,nb.iterations **do**
  $c(mortality_A, mortality_D) \leftarrow loopback(data, K, expo)$
  $expo \leftarrow compute\_expo(expo_{x_M}, incidence, mortality_A, mortality_D)$
end for
return $expo$

The algorithm stops when the maximum over all the ages between the exposures of two successive iterations is lower than a chosen tolerance ε_2. A maximum number of iterations is also fixed. Then, the algorithm returns the vector of the exposures in groups A and D at each age. The values are the real exposures for ages below x_M and theoretical exposures for ages above.

The mortality laws are then obtained by applying the Newton-Raphson algorithm in Section 2.2.4 by maximizing the penalized log-likelihood $l_{pen}^{loopback}$ (cf. Equation (2.10)). The weights w_x^G are fixed to 1 for the ages that are considered in the likelihood and 0 for the others.

The exposures used in the likelihood part of Equation (2.10) are the observed ones, even if $w_x^G = 1$ and the age is above x_M. The vector of theoretical exposure is only used in the loopback penalty.

We have introduced a hyper-parameter K on the penalized log-likelihood. Therefore, an important step for the user of this algorithm is to fix its value.

2.4 Choice of hyper-parameter K, an application on synthetic data

The choice of K is important for mortality law estimation. In fact, K can be considered as the weight given to the coherence criterion. The larger K is, the better the mortality laws estimated by the algorithm satisfy the coherence rule. Let us illustrate this aspect first on synthetic data.

2.4.1 Presentation of the synthetic data

Synthetic mortality laws have been constructed from age 50 to 120. Autonomous and LTC mortality laws have been independently estimated on a real French LTC portfolio covering severe LTC. The general mortality has then been constructed to satisfy the coherence criterion from Equation (2.8) in Section 2.2.4. The obtained laws are plotted from 50 to 120 years old in Figure 2.3. At age 50, the general mortality is equal to the autonomous mortality since we consider a population of 100% autonomous insured at age 50 to construct the general mortality law. As age increases, the proportion of disabled people in the general population changes, as does the general mortality law. We see that in this example of synthetic laws, the LTC mortality law converges to the general mortality since the population is composed of a majority of disabled individuals at old ages.

Let us assume then that the general mortality law is known but that no data are available above age 85 for both A and D groups. This means that exposures and number of deaths

at these ages are null. The loopback algorithm should be able to find the mortality intensities for both autonomous and LTC groups that have been used to construct the general mortality law for ages above 85.

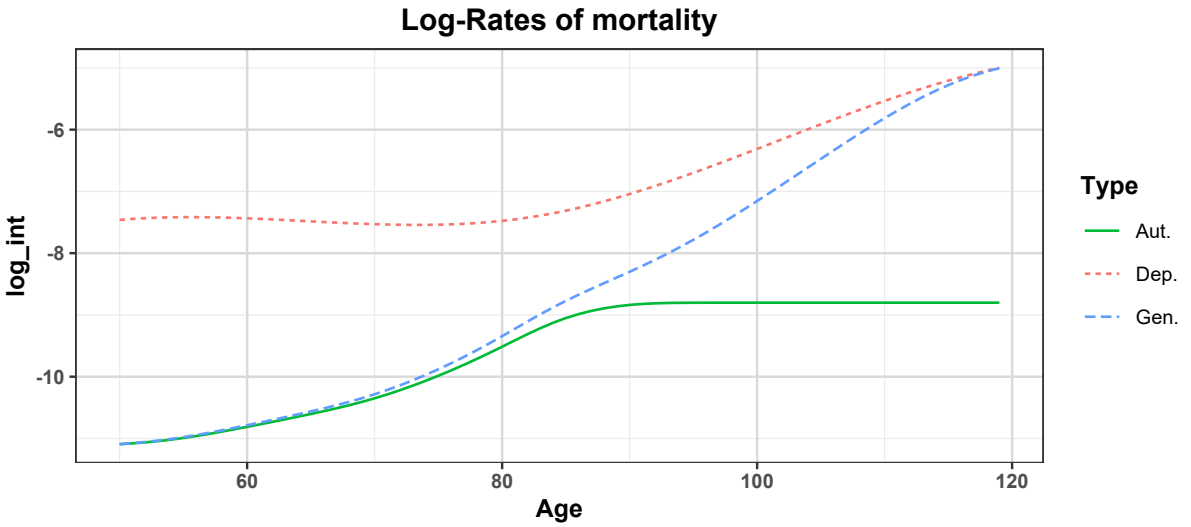


Figure 2.3 – Synthetic mortality laws

2.4.2 Impact of the choice of K on the residual loopback error

The larger K is, the better the mortality laws estimated by the algorithm satisfy the coherence rule, and the lower the residual loopback error given by the formula

$$Error_{loopback} = \sum_{x=x_{min}}^{x_{max}} \left(\frac{\lambda_x^{gen} \left(e_x^A + e_x^D \right) - \lambda_x^A e_x^A - \lambda_x^D e_x^D}{e_x^A + e_x^D} \right)^2 . \tag{2.14}$$

The pattern of the residual loopback error as a function of K is illustrated in Figure 2.4.

2.4.3 Optimization of parameter K

A large value of K leads to a small value of the loopback error. Unfortunately, we cannot choose K as large as possible since it implies problems in the convergence of the algorithm. Indeed, the Hessian becomes non-invertible after a few iterations. We need to find a balance between minimizing the loopback error and having K small enough to converge the algorithm. Figure 2.4 shows that if we accept a residual error smaller than $2e-4$, we have to choose K such that the error is below the red line. This means here that we can choose all K larger than the one at the intersection between the red line and the error curve, which is approximatively equal to 2950.

The idea is to fix a tolerance criterion on the loopback error. We then choose the value of K that leads to a residual loopback error close to this tolerance. The smaller the tolerance

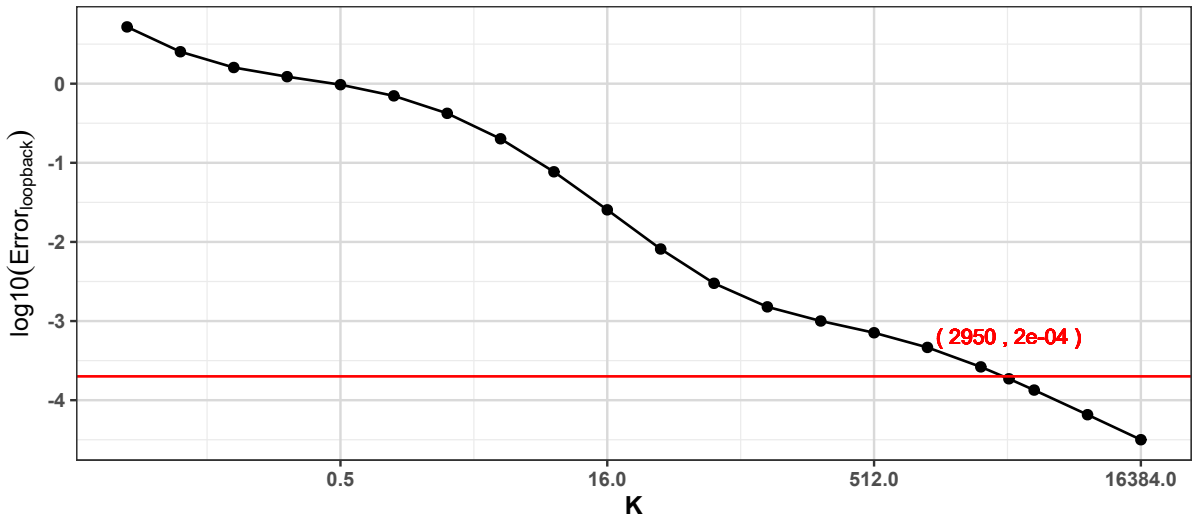


Figure 2.4 – Choice of K given a tolerance on the residual loopback error

is, the larger K. A function has been developed to optimize the choice of parameter K, leading to a tolerance close to the one previously fixed. In this example, we choose a tolerance named ε_3 in the R function equal to $2e-4$ (i.e., $\varepsilon_3 = 2e-4$).

The optimal value for K found by the algorithm is equal to 2741.65, and the residual loopback error is equal to $2e-4$.

2.4.4 Application of the loopback with the optimal K

The loopback algorithm is then used with this optimized parameter K to estimate coherent autonomous and LTC mortality laws. Since no observations are available above 85, Algorithm 1 from Section 2.3.2 is used to estimate theoretical exposures at old ages appearing in the loopback penalty given by Formula 2.9. In this example, the maximum age x_M for which we used the real observed exposures is fixed to 80. All exposures used in the loopback penalty for ages above 80 are computed by projecting the population of age 80 using biometric laws. Proportions of autonomous and disabled individuals are sufficient to compute the penalty. Therefore, those values are computed from the theoretical exposures and plotted in Figure 2.5. Starting with almost only autonomous individuals (97.2%) at age 80, the proportion of disabled individuals increases and reaches 99.4% at age 119.

The resulting mortality laws, plotted in Figure 2.6, are very close to the laws we used to construct the general mortality law. The triangles represent the observations used to fit the laws, and the dots represent the mortality intensities of the synthetic data that are not used in the loopback algorithm. The lines represent the estimated mortality laws with the loopback algorithm. Despite not using this information above age 85, the algorithm successfully manages to return mortality intensities close to those from the original synthetic data with the optimal K. Figure 2.6 shows the added value of the

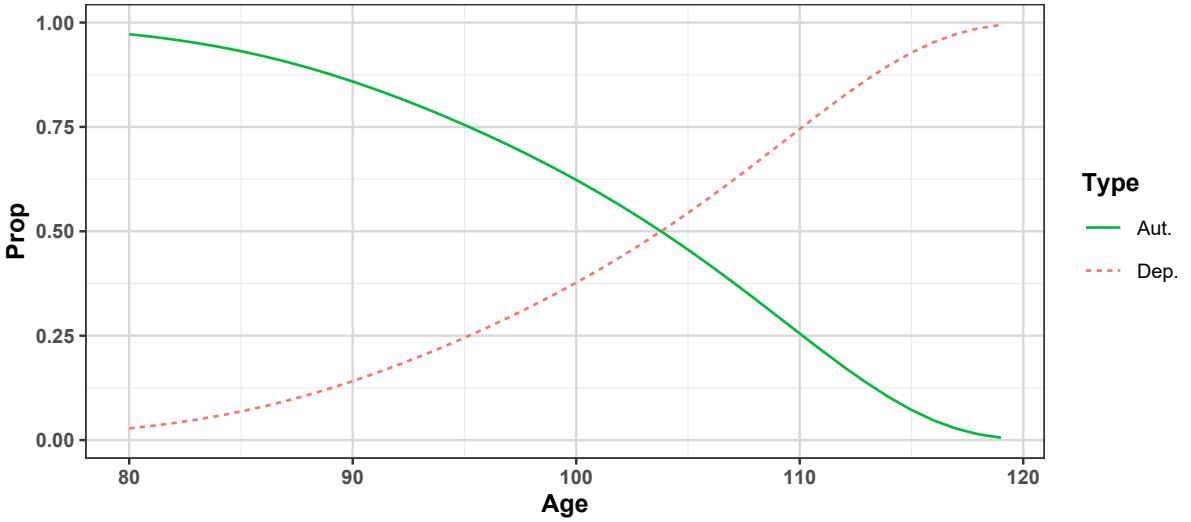


Figure 2.5 – Proportions resulting from calibrated exposures with the optimal parameter K

loopback algorithm. In fact, by not using any coherence penalty, the extrapolation of the mortality laws is driven only by the P-Splines order. Therefore, the extrapolation of the mortality in LTC is quadratic if the order is fixed to 3, as in our example. As a consequence, the mortality law in LTC (group D) diverges from the general mortality of the portfolio, whereas the population is mostly composed of disabled individuals at old ages.

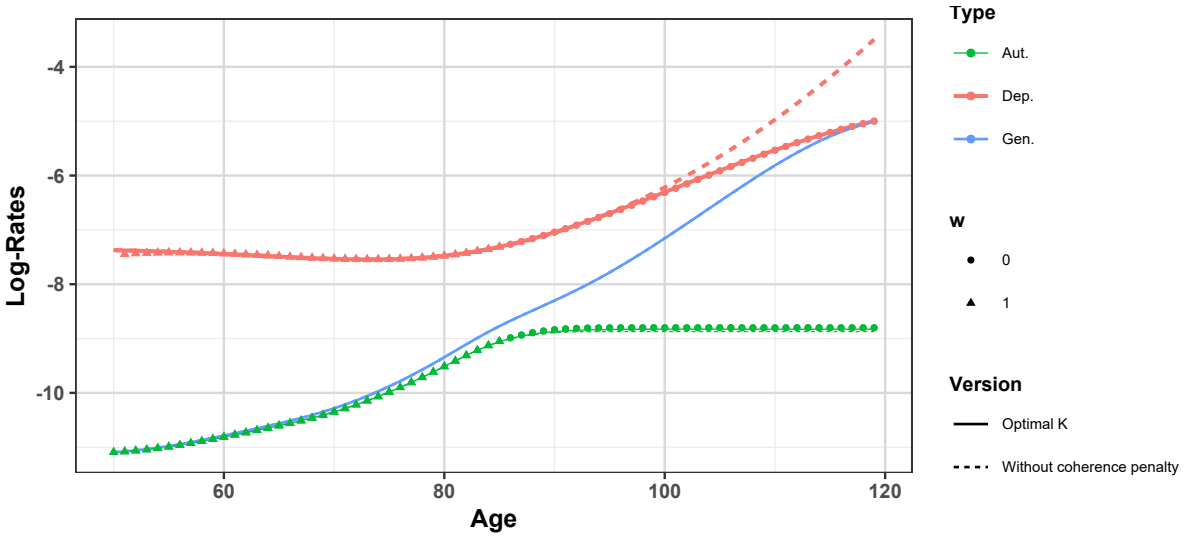


Figure 2.6 – Estimated mortality laws with the optimal parameter K

2.5 A case study on real data

2.5.1 Data

We rely on data coming from 5 medium-to-large French LTC portfolios. The application focuses only on females. The level of the loss of autonomy varies from mild to severe. In this application, we consider only severe LTC, with the GIR12 definition from the AGGIR grid described in Dupourqué (2012), which is used by the French government for the attribution of public aid. From these portfolios, 11,130 deaths are observed in the autonomous state (A), versus 3,681 in LTC (D). To calibrate the mortality laws, two datasets are constructed from the portfolios. The first one, called DB_A represents the dataset of the active contributors, and the second one, called DB_D, represents the dataset of the annuitants who are disabled. The first one is used to calibrate the autonomous mortality and incidence, while the second one is used to calibrate the mortality in LTC.

From these databases, only the observations between age 50 and 91 are used in the likelihood, as we decided to consider only ages with at least 10 observed deaths in our database. Observations at ages with fewer than 10 observed exits are too volatile. In Figures 2.8 and 2.10, representing estimated mortality laws, crude rates are represented as triangles or circles. Triangles represent the data points used in the likelihood, unlike the circles.

The general mortality law used in this section is calibrated on the same portfolios by aggregating the databases DB_A and DB_D and smoothing the crude rates by using the P-Splines smoothing methods. To extrapolate the general mortality law, we assume that the mortality law of the portfolio at old ages is close to the French mortality law, which is well known. The French mortality law used here comes from the « Human Mortality Database (HMD) » with an observation period from 2016 to 2018, available at www.mortality.org (data downloaded in May 2020) thanks to Max Planck Institute for Demographic Research (Germany), University of California, Berkeley (USA), and French Institute for Demographic Studies (France) (2020). The BRASS model, explained in Brass (1971), is used to force the mortality law to converge to the HMD mortality.

The incidence law i used in this section is estimated on the same 5 French LTC portfolios with the P-Splines smoothing method with order $d = 2$. As many LTC products in France exclude recovery, we do not observe any transition from state D to state A.

2.5.2 Application

We first begin with the extrapolation of the mortality laws, excluding any coherence criterion. This is equivalent to using the loopback algorithm with a penalty K equal to 0. We then study the impact of K on the residual loopback error and choose the optimal

hyper-parameter K. The mortality laws are then estimated and extrapolated using the loopback with this optimal penalization parameter.

2.5.2.1 Extrapolation without loopback penalization

When the penalization parameter K is fixed to 0, then the likelihood is equal to the sum of the likelihood of two P-Splines smoothing, one for each group (A and D). Maximizing the sum of these two likelihoods is equivalent to maximizing both of them independently. In this example, the smoothing penalty order is fixed to 1 for state A and 2 for state D. Therefore, as shown on Figure 2.8 with the dotted lines, the mortality law converges to a horizontal line at old ages for the first group and converges to a linear extrapolation for the second group.

2.5.2.2 Selection of parameter K and extrapolation of mortality laws

As seen in Section 2.4, parameter K has a large influence on the residual loopback error, with a larger K leading to a smaller residual loopback error.

For a tolerance ε_3 fixed to $1e-2$, the optimal K defined in Section 2.4 and chosen by the algorithm is equal to 411.56.

The maximum age x_M for which we use the real observed exposures in the loopback penalty is fixed to 90. All exposures used in the loopback penalty for ages above x_M are estimated using Algorithm 1 from Section 2.3.2. As in Section 2.4.4, the estimated proportions of autonomous (group A) and dependent individuals (group D) in the projected population needed for the loopback penalty are plotted in Figure 2.7. At age x_M, 80.6% of exposures are exposures in autonomy. At 99, estimated exposures in autonomy and disability are almost equal. The proportion of dependent individuals converges to 100% as age increases. Therefore, the population at old ages is composed almost only of dependent individuals, and we expect the estimated mortality law in LTC (group D) to converge to the global mortality of the portfolio denoted Gen.

As expected, the associated mortality laws obtained by the loopback algorithm, shown in solid lines in Figure 2.8, present convergence of LTC mortality to general mortality, while the log-intensity of autonomous mortality converges to a constant value. With the incidence law used in this example, the probability of remaining autonomous after 110 years is extremely low. Therefore, as shown in Figure 2.7, almost all the surviving insureds at 110 years are disabled, and the general mortality is equal to the mortality in LTC.

Figure 2.8 shows the impact of both smoothing and coherence penalties on the estimated mortality laws. Mortality laws obtained by maximizing $l^G(\boldsymbol{\theta}^{\boldsymbol{G}})$ given in Equation (2.6) for each group G are represented in dotted lines. Without any smoothing penalty, the resulting mortality laws are very volatile and try to capture all the variance observed in

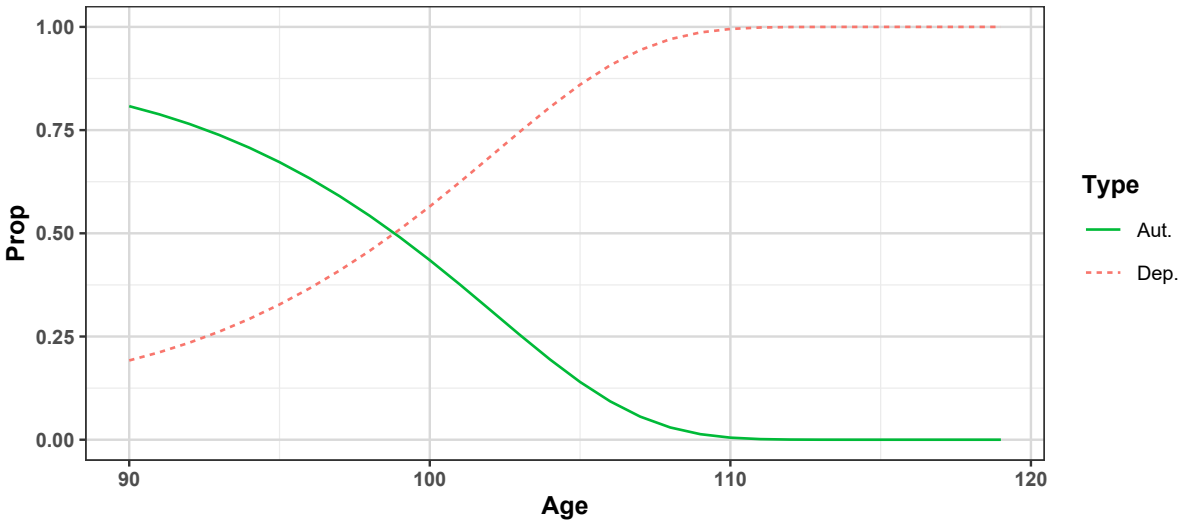


Figure 2.7 – Proportions of autonomous and dependent individuals estimated with the optimal parameter K

the data. As shown in dashed lines, adding a smoothing penalty for each group reduces over-fitting and obtains better extrapolation in the sense that the mortality laws do not explode as age increases. Finally, solid lines represent the mortality laws obtained with optimal parameter K. Adding the coherence penalty allows consistency between the three mortality laws, having the mortality of group D (LTC) converging to the general mortality since the portfolio is composed almost only of dependent individuals at old ages.

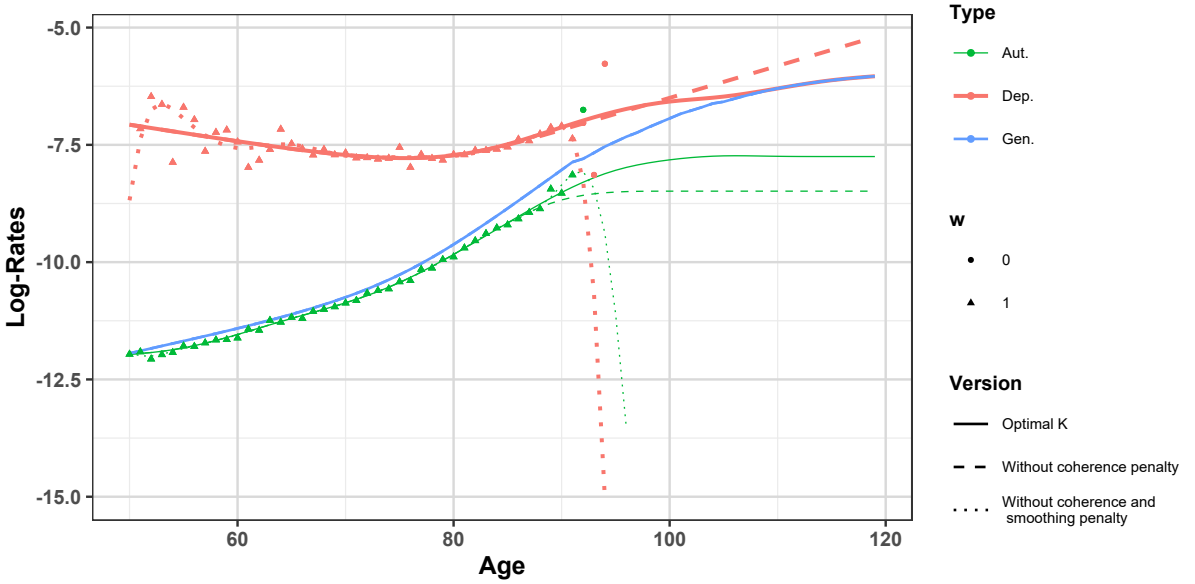


Figure 2.8 – Comparison of the mortality laws for the GIR12 product obtained with the optimal K, without any coherence penalty, and without a smoothing penalty

Given the estimated proportions of autonomous and dependent individuals in Figure 2.7 and the associated calibrated mortality laws in groups A and D plotted in Figure 2.8, the implied mortality law of the porfolio is estimated with Equation 2.8. The proportions of

individuals in states A and D are $\frac{e_x^A}{e_x^A + e_x^D}$ and $\frac{e_x^D}{e_x^A + e_x^D}$, respectively. Figure 2.9 shows how well the algorithm was able to replicate the general mortality law used as a reference. The two mortality laws are really close, except for ages below 70 where the implied mortality law is really close to the autonomous mortality law. This is explained by exposures in state D almost equal to zero at young ages. Therefore, the implied mortality law is almost equal to the autonomous mortality. Moreover, since exposures in state A are high at young ages, the weight of the likelihood of autonomous observations of Equation 2.10 is higher than the weight of the consistency penalty for these ages.

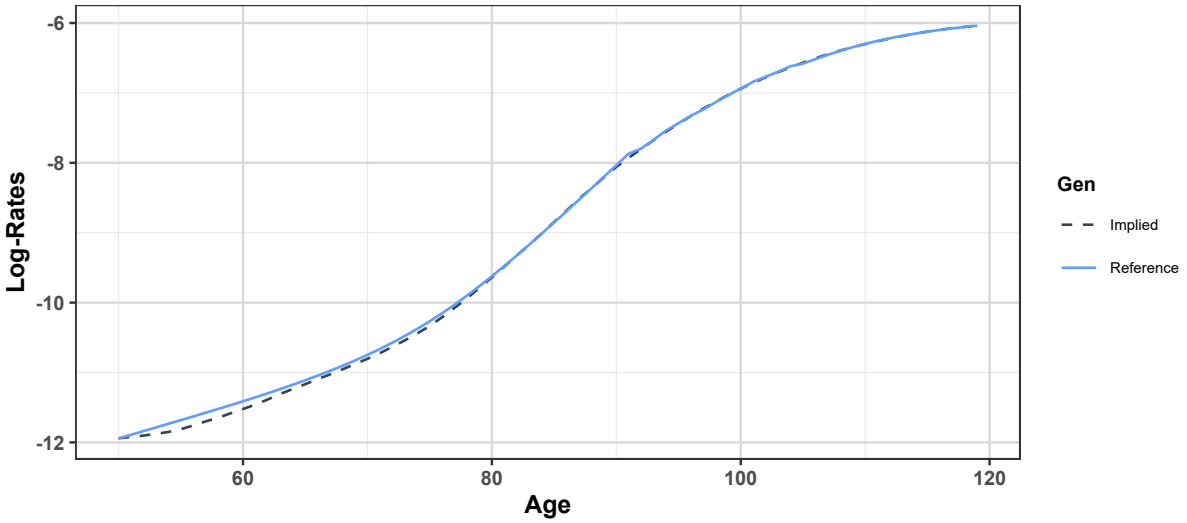


Figure 2.9 – Comparison of the implied mortality of the portfolio to the general mortality law used as a reference

The confidence intervals of the two mortality laws are obtained with a simulation algorithm inspired by the bootstrap method. Using the fitted mortality laws and assuming that the number of deaths is Poisson distributed, new death counts per age and group (A and D) are simulated for each age participating in the log-likelihood term of Equation (2.10). New mortality laws are then fitted using the loopback algorithm on these new simulated data. We must keep in mind that these confidence intervals are computed considering that the general mortality is known. This implies that the uncertainty on the general mortality is not taken into account when computing the confidence intervals of the mortality laws in autonomy and LTC.

The confidence intervals at 95% constructed with 800 simulations are shown in Figure 2.10.

Thanks to the loopback, the autonomous and LTC mortality and the incidence laws represented by λ_x^A, λ_x^D and i_x, respectively, in Figure 2.1, are consistent with the general mortality of the portfolio that is known.

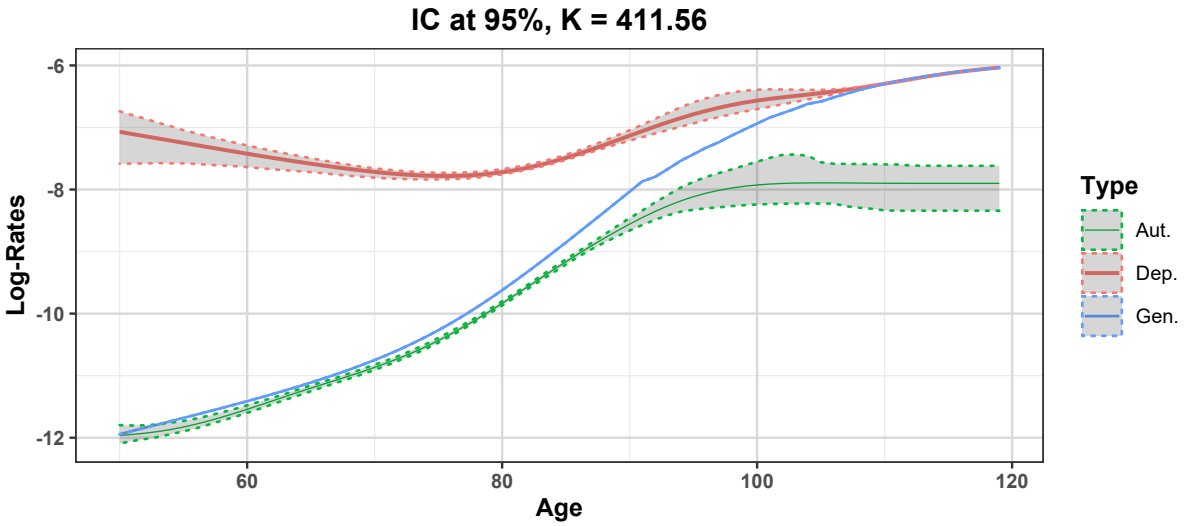


Figure 2.10 – Confidence intervals for the GIR12 product obtained with the optimal K

2.5.2.3 Actuarial application

Using the calibrated laws λ_x^A, λ_x^D and i_x, 50,000 trajectories of the future states of 50-year-old women are simulated. The aim is to estimate the probability for autonomous women to be in the autonomous, disabled or death state at each age above 50. The results are represented in Figure 2.11, where the obtained proportions in each group at each age are plotted. Starting with 100% of autonomous individuals at age 50, the proportion of autonomous individuals decreases with increasing age, since recovery is not considered in the model. Death is an absorbing state, and the proportion can only increase with age. The probability of being in LTC increases until age 93 before decreasing afterwards. Indeed, insureds can both enter and exit the LTC state. Under the calibrated biometric laws, up to age 93, the number of entries into LTC is larger than the number of deaths. This is reversed afterwards, with more deaths expected than loss of autonomy. The last survivor in this simulation dies in LTC at age 118. The last autonomous insured enters LTC at age 110 before dying.

Figure 2.12 represents the proportion of autonomous and disabled people among the survivors. Until age 99, there are still more autonomous than disabled insureds. For an insurer, this means that the proportion of insureds paying their premium is larger than the proportion of disabled insureds receiving an annuity.

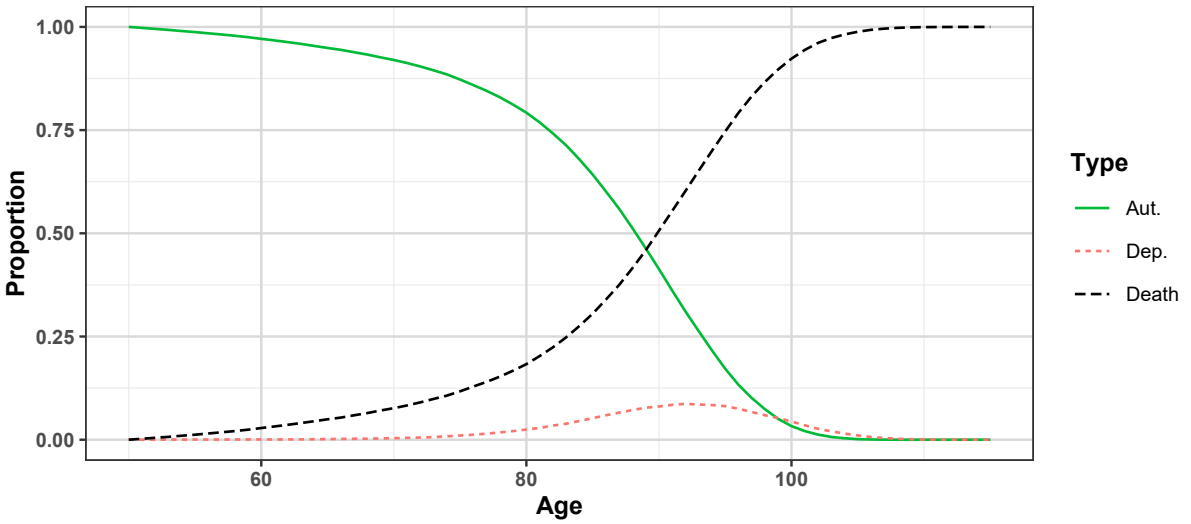


Figure 2.11 – Proportion of insureds in each group considering a 100% autonomous population at age 50

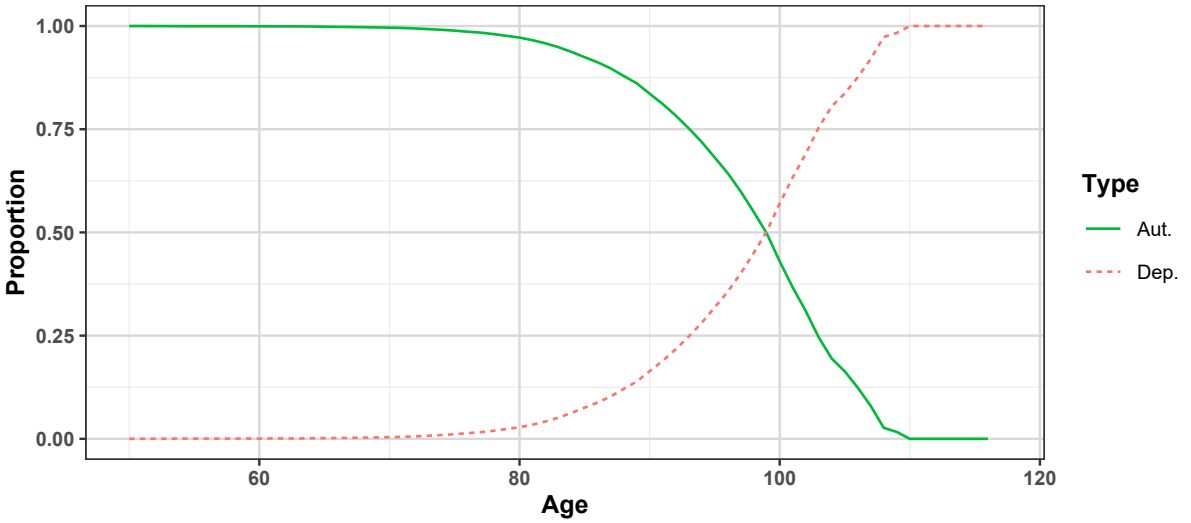


Figure 2.12 – Proportion of autonomous insureds and annuitants considering a 100% autonomous population at age 50

2.6 Modelling products with several levels of dependency and allowing recovery

Let us consider in this section a product covering multiple levels of dependency, with different amounts of annuity depending on the degree of loss of autonomy.

Let us assume 3 levels of dependency:

- Total Dependency (**TD**)
- Partial Dependency (**PD**)
- Light Dependency (**LD**)

Recovering from severe dependency can be assumed to be impossible. However, one could

want to allow recovery from the light level of dependency **LD**.

Section 2.6.1 presents two ways of modelling this product. Subsection 2.6.2 focuses on how to incorporate the recovery in the loopback algorithm.

2.6.1 Two ways of modelling a product covering multiple levels of dependency

In a first step, let consider a product without transition payments. We can model a product covering multiple levels of dependency with a 5-state Markov model (termed **Model 1** in this section), as shown in Figure 2.13. For clarity, the intensity notations are not mentioned in this figure, except for the incidence rates from the autonomous state **A** to the lower level of dependency **LD** denoted i_x and the recovery rates from **LD** to **A** denoted r_x.

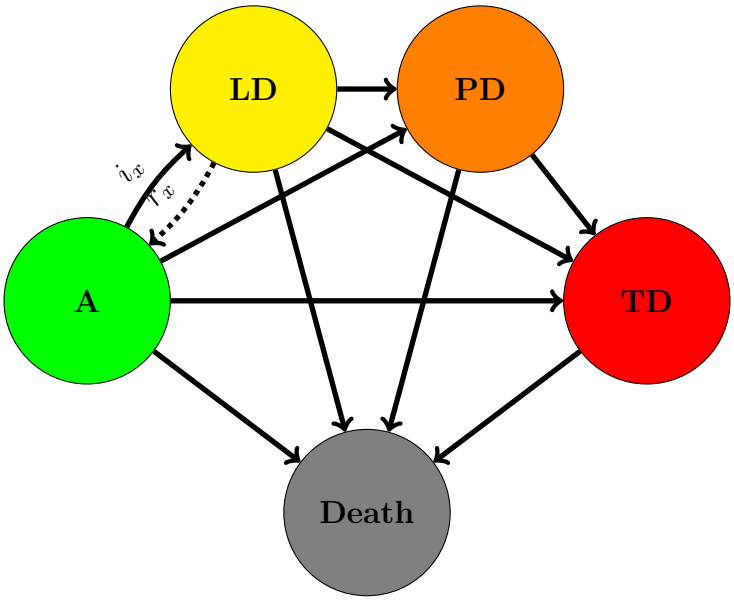
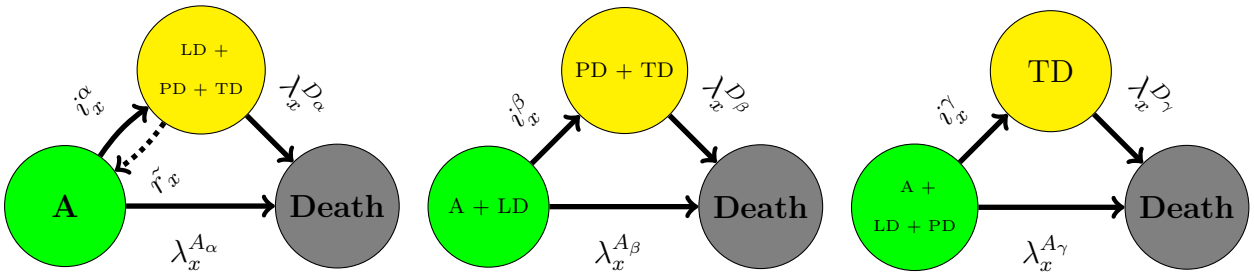


Figure 2.13 – Modelling of an LTC product with multiple degrees of loss of autonomy (**Model 1**)

Most insurers suffer from scarcity of data due to the recency of LTC products covering multiple degrees of loss of autonomy. As a consequence, it is difficult for these insurers to calibrate this type of model without needing to introduce strong assumptions on the intensities, as in Fleischmann (2015), where, for example, the intensity to reach a specific level of dependency is assumed to be independent of the state of origin. Another way of modelling this product is to consider it as a set of 3 products. Underwriting to this product covering 3 levels of severity of loss of autonomy, with an annuity depending on this severity, is equivalent to underwriting to 3 LTC contracts denoted α, β and γ, as represented in Figure 2.14 and described as follows:

- Product α covers all degrees of LTC (light, partial and total dependency) with the same annuity amount. The conditions of the contract are as follows:
  - The insured pays the premium P_α as long as the insured is autonomous.
  - The insurer pays an annuity R_α as long as the insured is in light, partial or total dependency.
  - The insured can recover from dependency. In this case, the insurer stops paying the annuity, and the insured is considered autonomous.
- Product β covers partial and total dependency such that:
  - The insured pays the premium P_β as long as the insured is autonomous or in light dependency.
  - The insurer pays an annuity R_β as long as the insured is in partial or total dependency.
  - Recovery is not possible.
- Product γ covers only total dependency such that:
  - The insured pays the premium P_γ as long as the insured is alive and not in total dependency.
  - The insurer pays an annuity R_γ as long as the insured is in total dependency.
  - Recovery is not possible.



(a) Product α (b) Product β (c) Product γ

Figure 2.14 – Modelling of an LTC product covering multiple degrees of loss of autonomy with a set of 3 LTC products (**Model 2**)

We note that $\tilde{r_x} < r_x$ because only insureds in light dependency (**LD**) can recover.

Let:

- P be the premium of the product covering the multiple levels of LTC.
- $\tilde{R}_{level}$ be the annuity paid to a dependent insured with the level of severity $level \in \{LD, PD, TD\}$.
- P_j and R_j be the premium and annuity amounts for Product $j \in \{\alpha, \beta, \gamma\}$, respectively.

Let us compare both models by analysing the cash-flows of the insured depending on its health status. To model the multi-level product with a set of LTC products, the cash flows given in Table 2.1, of both models have to be equal. In this case, underwriting to

a contract covering the 3 degrees of dependency is equivalent to underwriting to the 3 products described in Figure 2.14.

	Cash flows	
Health status	**Model 1**	**Model 2**
Autonomous	$-P$	$-P_\alpha - P_\beta - P_C$
LD	$\tilde{R}_{LD}$	$R_\alpha - P_\beta - P_\gamma$
PD	$\tilde{R}_{PD}$	$R_\alpha + R_\beta - P_\gamma$
TD	$\tilde{R}_{TD}$	$R_\alpha + R_\beta + R_\gamma$

Table 2.1 – Comparison of cash flows of Model 1 and Model 2

A product offering transition payments, can also be considered as a set of 3 products as in Figure 2.14 if going through intermediate LTC states does not change the total amount received by the insured. Here is an example of transition payments of such a product:

- In case of entry in LD from A, the insured receives a capital $\tilde{K_{LD}}$ at the time of entry in LD,
- In case of entry in PD from A, the insured receives a capital $\tilde{K_{PD}}$ at the time of entry in PD,
- In case of entry in TD from A, the insured receives a capital $\tilde{K_{TD}}$ at the time of entry in TD,
- If the insured enters state PD from state LD, then he receives $\tilde{K_{PD}} - \tilde{K_{LD}}$ at the time of entry in PD,
- If the insured enters state TD from state LD, then he receives $\tilde{K_{TD}} - \tilde{K_{LD}}$ at the time of entry in TD,
- If the insured enters state TD from state PD, then he receives $\tilde{K_{TD}} - \tilde{K_{PD}}$ at the time of entry in TD,
- An insured entering state $S \in \{LD, PD, TD\}$ from state A after recovering from LD receives $\tilde{K_S} - \tilde{K_{LD}}$ at the time of entry in S.

Model 2 is a good way to model a multi-level product when not having a large database without making strong assumptions on the shape of the rates. This model is often used by insurers.

2.6.2 Taking into account the possibility to recover

Product α represented in Figure 2.14a allows recovery. Allowing this transition has only a slight impact on the algorithm presented in this paper. In fact, allowing recovery has an impact only if one needs to estimate exposures, as in Section 2.3. In this case, Algorithm 1 becomes:

Algorithm 2 Exposures estimation algorithm in the case of recovery

$K \leftarrow 0$
$expo \leftarrow expo_{data}$
$c(mortality_A, mortality_D) \leftarrow loopback(data, K, expo)$
$expo \leftarrow compute\_expo(expo_{x_M}, incidence, \mathbf{recovery}, mortality_A, mortality_D)$
$K \leftarrow$ penalty parameter chosen for model calibration
for i=1,...,nb.iterations **do**
  $c(mortality_A, mortality_D) \leftarrow loopback(data, K, expo)$
  $expo \leftarrow compute\_expo(expo_{x_M}, incidence, \mathbf{recovery}, mortality_A, mortality_D)$
end for
return *expo*

where *recovery* represents the transition rates from the LTC to the autonomous state.

In this case, the function $compute\_expo(expo_{x_M}, incidence, \mathbf{recovery}, mortality_A, mortality_D)$ has to take into account the *recovery* law.

Let us assume that recoveries occur at the end of the period as the deaths and losses of autonomy in Section 2.3. An insured recovering at age x is in state A at age $x+1$ and cannot enter state D a second time or die before age $x+1$. At the i^{th} iteration,

$$expo^A_{x+1}(i) = expo^A_x(i-1)\exp\left(-(\lambda^A_x + i_x)\right) + \mathbf{expo^D_x(i-1)}\left[\mathbf{1} - \exp\left(-(\lambda^{\mathbf{D}}_{\mathbf{x}} + \tilde{\mathbf{r_x}})\right)\right]\frac{\tilde{\mathbf{r_x}}}{\lambda^{\mathbf{D}}_{\mathbf{x}} + \tilde{\mathbf{r_x}}},$$
$$expo^D_{x+1}(i) = expo^D_x(i-1)\exp\left(-(\lambda^D_x + \tilde{\mathbf{r_x}})\right) + expo^A_x(i-1)\left[1 - \exp\left(-(\lambda^A_x + i_x)\right)\right]\frac{i_x}{\lambda^A_x + i_x},$$

where $expo^G_{x+1}(i)$ denotes the exposure in group G at age x at the i_{th} iteration, and $\tilde{r_x}$ denotes the intensity rate of recovery. The terms added by allowing the recovery are highlighted in boldface.

2.7 Discussion

In this paper, we introduce an approach to simultaneously estimate the mortality laws of two subgroups A and D (where A and D represent the autonomous and disabled insured groups, respectively), knowing the mortality of the overall group ($A \cup D$). To do so, we rely on the P-Splines smoothing method combined with Poisson-GLM, to which we add a consistency constraint. The aim of this constraint is to link the mortality of the overall group, named general mortality in this paper, to both mortality laws in groups A and D. This constraint is based on the idea that each death in the overall group is a death in either subgroup A or subgroup D. Therefore, the sum of deaths in A and D is equal to the number of deaths in the overall group (gen). If D^G_x denotes the random variable of the death counts at age x in group $G \in \{A, D, gen\}$, then $D^{gen}_x = D^A_x + D^D_x$. As in the

Poisson-GLM part of the model, we assume that the count of deaths in each group G at each age x exhibits a Poisson distribution of parameters proportional to the central exposure and the mortality intensities. This allows us to link the mortality rates of the three groups. This constraint is added in the form of a penalty in the likelihood. The mortality intensities are then estimated by maximizing the penalized log-likelihood.

We then address the problem of extrapolation of mortality laws in the case where no or not enough observations are available at old ages. This is often the case in an insurance context, particularly when estimating the risk associated with LTC products. In fact, the recency of these products combined with the fact that they are sold to individuals on average 60 years old are responsible for the data paucity beyond 85 years old. Extrapolation of mortality laws is therefore necessary for actuaries to assess the risk. We introduce an iterative approach to estimate the missing exposures at old ages. To do so, we successively estimate the mortality laws with the method described in Section 2.2 and then the exposures using the probability of transition from group A to D and the mortality laws from the previous step. We then re-estimate the mortality laws using the estimated exposures. These new mortality laws then lead to new estimations of the exposures. The algorithm stops when tolerance criteria are reached between successive estimations of exposures.

In the first step, the algorithm developed in this paper is tested on synthetic data. Mortality laws are known until 119 years old, but we hide the observations above 85 to the algorithm and see how the algorithm is able to reproduce these mortality intensities between ages 86 and 119.

We introduce methods to fix the hyper-parameter and to construct confidence intervals and perform testing on the synthetic dataset. We show that our approach improves the extrapolation of the mortality laws. In fact, the extrapolation with a consistency penalty is much closer to the actual mortality intensities from 86 to 119 than the extrapolation without a penalty. In a second step, the approach presented in this paper is used on real data from five medium-to-large LTC portfolios. As with the synthetic data, we compare the results of estimations and extrapolations with and without a consistency penalty. Compared to the estimation without penalty, adding the consistency criteria results in lower estimated mortality rates in LTC (group D) at old ages and higher mortality rates for the autonomous group (group A). An insurer not using consistency criteria would overestimate the mortality of the annuitants, leading to underestimation of the provisions.

As the loopback algorithm is based on P-Splines, orders of splines penalties d introduced in Section 2.2.3 for each group (A and D) are considered as hyper-parameters. As seen in Section 2.2.3, the choice of d is crucial since it drives the age extrapolation results. In particular, without a loopback penalty, the age extrapolation is linear on the log-scale for $d = 2$ and constant for $d = 1$. Adding consistency penalty decreases the impact of this

choice. The extrapolation is no longer driven only by this order but also by the consistency criteria. Nevertheless, a careful choice must be made whether one assumes that the mortality intensity continues to grow log-linearly with age even at old ages, as in Gavrilov and Gavrilova (2019), or if mortality stops growing at old ages, as in Barbi et al. (2018). In this paper, the order is fixed to 2 for disabled mortality. The order 2 allows a linear extrapolation and gives more degrees of freedom. At old ages, the probability of being autonomous is very low. Most insureds are either disabled or dead. Therefore, autonomous mortality at old ages has a relatively low impact on product pricing and reserving. In our application, fixing $d = 2$ for autonomous leads to intersecting mortality curves. In fact, in the first step of Algorithm 1, exposures are estimated with the mortality laws without a consistency constraint. With $d = 2$ for both autonomous and disabled groups, these extrapolated mortality laws at first step intersect, and the autonomous mortality is truly high at old ages (higher than the general mortality), leading to estimated exposures equal to 0 in autonomy. Therefore, for the second and next steps of the algorithm, the autonomous mortality in the constraint has a negligible or even zero weight. Hence, the constraint has only an impact on the extrapolation of the mortality in LTC, and the autonomous mortality remains higher than both the general and the disabled mortality.

In the context of modelling LTC products, many insurers use two-dimensional mortality rates for the LTC group, using semi-Markov models. In fact, mortality in LTC may depend on attained age but also on time spent in disability. Future research should implement this algorithm with a one-dimensional mortality law for group A depending on age and a two-dimensional mortality for group D depending on age and duration.

Appendices

Appendix 2.A Convergence of the Newton Raphson algorithm

To be the maximum penalized likelihood estimator of $\boldsymbol{\theta}$, the Hessian matrix $H_{\widehat{\boldsymbol{\theta}}}$ at the final step of the algorithm has to be negative semi-definite.

Let us analyse the Hessian matrix.

- The first term $-B^T W W_{\boldsymbol{\theta}} B$ is negative semi-definite for all $\boldsymbol{\theta}$. Indeed, recalling that $W_{\boldsymbol{\theta}}$ is diagonal with only non-negative terms,
$$h^T B^T W_{\boldsymbol{\theta}} B h = (Bh)^T W_{\boldsymbol{\theta}} (Bh) \geqslant 0 \quad \forall h \in \mathbb{R}^{2M}.$$
- The second term $-P$, which does not depend on $\boldsymbol{\theta}$, is also negative semi-definite. Indeed, from 2.2.3, we know that $P_d = D_d^T D_d$. Therefore, $h^T P_d h \geqslant 0 \quad \forall h \in \mathbb{R}^{2M}$.
- The third term $-KB^T \left[W_{\boldsymbol{\theta}} \left([(\tilde{W}_3^{-1})^2 W_{\boldsymbol{\theta}}^Q] \otimes I_2 \right) \right] B$ is not necessarily negative semi-definite for all $\boldsymbol{\theta}$. In fact, the weight matrix $\left[W_{\boldsymbol{\theta}} \left([(\tilde{W}_3^{-1})^2 W_{\boldsymbol{\theta}}^Q] \otimes I_2 \right) \right]$ is diagonal, but not all coefficients are greater than 0 for some $\boldsymbol{\theta}$. The terms of the diagonal matrix are non-positive if some terms of $W_{\boldsymbol{\theta}}^Q$ are non-negative. This is the case when
$$\lambda_{\boldsymbol{\theta},x}^A e_x^A + \lambda_{\boldsymbol{\theta},x}^D e_x^D \leqslant \lambda_x^{gen} [e_x^A + e_x^D], \text{ for some } x_{min} \leqslant x \leqslant x_{max}.$$
- The fourth term $-K \left[\tilde{W}_3^{-1} W_{\boldsymbol{\theta}}^A B_A \quad \tilde{W}_3^{-1} W_{\boldsymbol{\theta}}^D B_D \right]^T \left[\tilde{W}_3^{-1} W_{\boldsymbol{\theta}}^A B_A \quad \tilde{W}_3^{-1} W_{\boldsymbol{\theta}}^D B_D \right]$ is negative semi-definite for all $\boldsymbol{\theta}$. In fact,

$$h^T \left[\tilde{W}_3^{-1} W_{\boldsymbol{\theta}}^A B_A \quad \tilde{W}_3^{-1} W_{\boldsymbol{\theta}}^D B_D \right]^T \left[\tilde{W}_3^{-1} W_{\boldsymbol{\theta}}^A B_A \quad \tilde{W}_3^{-1} W_{\boldsymbol{\theta}}^D B_D \right] h = \left\| \left[\tilde{W}_3^{-1} W_{\boldsymbol{\theta}}^A B_A \quad \tilde{W}_3^{-1} W_{\boldsymbol{\theta}}^D B_D \right] h \right\|_2^2 \geqslant 0.$$

Then, a sufficient condition for $H_{\boldsymbol{\theta}}(l_{pen})$ to be negative semi-definite and therefore for $\widehat{\boldsymbol{\theta}}$ to be the optimal parameter is that the third term is negative semi-definite. The condition is given by

$$\lambda_{\boldsymbol{\theta},x}^A e_x^A + \lambda_{\boldsymbol{\theta},x}^D e_x^D \leqslant \lambda_x^{gen} (e_x^A + e_x^D), \quad \forall x_{min} \leqslant x \leqslant x_{max}. \tag{2.15}$$

This means that the sum of the predicted number of deaths in states A and D has to be lower than or equal to the predicted number of deaths of the overall population.

Bibliography

Alegre, A., E. Pociello, A. Pons, J. Varea, and A. Vicente (2003). Actuarial valuation of long-term care annuities. *Insurance Mathematics and Economics Volume 32*.

Barbi, E., F. Lagona, M. Marsili, J. W. Vaupel, and K. W. Wachter (2018). The plateau of human mortality: Demography of longevity pioneers. *Science 360*(6396), 1459–1461.

Biessy, G. (2015). 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.

Bollaerts, K., P. Eilers, and I. Mechelen (2006). Simple and multiple P-splines regression with shape constraints. *The British journal of mathematical and statistical psychology 59*, 451–69.

Brass, W. (1971). *Biological Aspects of Demography*. Taylor and Francis.

Camarda, C. (2019). Smooth constrained mortality forecasting. *Demographic Research 41*, 1091–1130.

Camarda, C. G., P. H. Eilers, and J. Gampe (2016). Sums of smooth exponentials to decompose complex series of counts. *Statistical Modelling 16*(4), 279–296.

Currie, I. D. and M. Durban (2002). Flexible smoothing with P-splines: A unified approach. *Statistical Modelling 2*(4), 333–349.

Currie, I. D., M. Durban, and P. H. Eilers (2004). Smoothing and forecasting mortality rates. *Statistical modelling 4*(4), 279–298.

Dupourqué, E. (2012). AGGIR, the work of grids. *Long-Term Care News 32*.

Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing with B -splines and penalties. *Statistical Science 11*(2), 89–121.

Eilers, P. H. C. and B. D. Marx (2002). Generalized linear additive smooth structures. *Journal of Computational and Graphical Statistics 11*(4), 758–783.

Eurostat (2022). Population structure and ageing. pp. 575–94.

Fleischmann, A. (2015). Calibrating intensities for long-term care multiple-state Markov insurance model. *European Actuarial Journal 5*, 327–354.

Gavrilov, L. and N. Gavrilova (2019). New trend in old-age mortality: Gompertzialization of mortality trajectory. *Gerontology 65*, 1–7.

Guibert, Q., S. Loisel, O. Lopez, and P. Piette (2020). Bridging the Lee-Carter's gap: a

locally coherent mortality forecast approach. https://hal.archives-ouvertes.fr/hal-02472777.

Hammond, M. (2000). The forces of mortality at ages 80 to 120. *International Journal of Epidemiology 29*(2), 384–384.

Li, J. S.-H., W.-S. Chan, and R. Zhou (2017). Semicoherent multipopulation mortality modeling: The impact on longevity risk securitization. *Journal of Risk and Insurance 84*(3), 1025–1065.

Li, N. and R. Lee (2005). Coherent mortality forecasts for a group of population: An extension of the Lee–Carter method. *Demography 42*, 575–94.

Macdonald, A. S., S. J. Richards, and I. D. Currie (2018). *Modelling Mortality With Actuarial Applications*. Cambridge University Press.

Marx, B. D. and P. H. Eilers (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis 28*(2), 193 – 209.

Max Planck Institute for Demographic Research (Germany), University of California, Berkeley (USA), and French Institute for Demographic Studies (France) (2020). Human mortality database. www.mortality.org. (data downloaded on 2020-10-15).

Nuttall, S. R., R. J. L. Blackwood, B. M. H. Bussell, J. P. Cliff, M. J. Cornall, A. Cowley, P. L. Gatenby, and J. M. Webber (1994). Financing long-term care in Great Britain. *Journal of the Institute of Actuaries 121*(1), 1–68.

Porta, N., G. Gomez, M. Calle, and N. r. Malats (2007). Competing risks methods. https://upcommons.upc.edu/bitstream/handle/2117/2201/TR_CR.pdf.

Remund, A., T. Riffe, and C. Camarda (2018). A cause-of-death decomposition of the young adult mortality hump. *Demography 55*, 957–978.

Ruppert, D. (2000). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics Volume 11*, 735–757.

Zhou, R., G. Xing, and M. Ji (2019). Changes of relation in multi-population mortality dependence: An application of threshold VECM. *Risks 7*(1).

Chapitre 3

Combining experience data of several Long-Term Care Insurance products with different disability definitions

Ce chapitre reprend l'article "Combining experience data of several Long-Term Care Insurance products with different disability definitions".

Abstract

Long-term care (LTC) products cover the risk of permanent loss of autonomy. While the global definition of the loss of autonomy is the impossibility or difficulty of performing activities of daily living (ADL) alone, in the LTC insurance market, the exact definition of the health state leading to a claim varies across different markets and even within the same market. A difference in the disability definition implies a difference in the mortality rates of the autonomous and disabled policyholders. Insurers or reinsurers often have experience data coming from several long-term care products with differing definitions of risk. One solution is to separate the data to estimate mortality rates for each definition independently.

In this paper, we propose two methods to aggregate the experience data of two portfolios with different disability definitions to improve the estimations of the mortality. The mortality laws of the two products are modelled in a Poisson Generalized Linear Model framework. The first method uses a constrained optimization model and is solved by sequential quadratic programming. The second method uses the Penalized Composite Link Model (PCLM). These methods allow better and simultaneous estimation of mortality for both products by combining all available data.

Keywords: Long-Term Care Insurance; Actuarial modelling; Multiple definitions; Penalized Composite Link Model; Constrained optimization.

3.1 Introduction

Long-term care (LTC) costs are among the major risks faced by individuals in retirement. For example, Chapman (2012) reports that approximately two-thirds of people aged 65+ need LTC. By using the U.S. Health and Retirement Study data, Johnson (2019) estimates a large gap between the risk of needing vs. receiving LTC. As reported in Johnson (2019), 70% of individuals past age 65 need LTC, whereas less than 50% receive some form of LTC services. Similar levels of risk of needing LTC have been reported in many other studies (Shao et al., 2017; Kemper et al., 2005; Hurd et al., 2017).

LTC costs have been steadily increasing over the past decades in many countries, and this trend is projected to continue (Colombo et al., 2011; Shi and Zhang, 2013; Jin et al., 2023). A recent report by OECD (OECD, 2020) shows that on average, LTC costs were approximately USD 760 per capita in OECD countries in 2018, accounting for 1.5% of GDP. In some countries, such as Denmark, Norway and Sweden, LTC spending was as high as 3.5% of their GDP. In France, LTC spending was 2.5% of its GDP.

The fundamental source of funding for LTC costs in most countries is public programs. For example, in the U.S. LTC expenses are mainly funded by its public health program Medicaid (Colombo et al., 2008; Kaye et al., 2010); in France, LTC expenses are mainly funded by its national allowance program called Allocation Personalisée d'Autonomie (APA) with a cap (Or and Penneau, 2021). Coupled with ageing populations, the increasing trend of LTC costs may place a large burden on public health programs in many counties. It has become very important in many countries to develop or to enhance the private insurance market to help fund LTC costs (Shao et al., 2019; Colombo et al., 2011; Productivity Commission of Australia, 2013).

A typical private LTC insurance policy entitles the policyholder to regular payments, such as cash benefits or on a reimbursement basis, when the policyholder loses autonomy according a certain definition (Haberman and Pitacco, 1999; Shao et al., 2017). The definitions of "losing autonomy" or "becoming disabled" vary across different markets, and these definitions are not uniform even within the same market.

In the U.S. market, the disability definition in a typical LTC insurance policy is the loss of independence in performing two or more activities of daily living (ADLs) and/or having cognitive impairment (Pritchard, 2006). In the French market, the disability definition is mostly based on the GIR (Groupe Iso-Ressources) assessment rules. Within the French LTC insurance market, different disability definitions can be used in different insurance policies. For example, some LTC insurance policies include a deferred period (such as 3 months, 6 months, or 9 months), and some start making payments at the date of the loss of autonomy.

Differing disability definitions may result in very different transition rates between health states (including disability rate and mortality rate) in the pricing model and can therefore have substantially different financial impacts. An insurer or reinsurer may have experience data for many different portfolios of LTC insurance policies with differing disability definitions, where they typically analyse experience for each portfolio of LTC insurance policies. This separate modelling approach can result in information loss from a parameter estimation perspective. For example, two portfolios with different definitions, from the same insurer with similar claims management team and system, should have shared experience that can provide better estimations. This combined modelling approach can also help experience analysis for portfolios with limited experience or areas with limited experience data.

Let us consider an insured population such that the health state of each policyholder is observed with only one of the two definitions. Instead of separating the policyholders in two portfolios depending on the definition with which they are observed, we jointly estimate the mortality rates of each product to prevent information loss.

In this paper, we develop two methods that make the best use of data available by combining experience data of the two portfolios with different disability definitions. These two methods are the constrained optimization method and the Penalized Composite Link Model (PCLM) method. A common approach to estimate the parameters of a model is the maximum likelihood estimation. The first method proposed in this paper estimates parameters by maximizing the likelihood subject to some constraints linking biometric laws of the two portfolios. The second approach proposed in this paper uses the Penalized Composite Link Model introduced by Eilers (2007) as an extension of the Composite Link Model proposed by Thompson and Baker (1981). The Penalized Composite Link Model has already been used in the context of mortality modelling in Remund et al. (2018) for cause-of-death decomposition and in Camarda et al. (2016) by considering that the mortality curve is a sum of 3 smooth components. In the latter paper, the first component represents infant mortality, the second component captures mortality due to ageing, and the third component models the accident hump for early adult ages. The constrained optimization model and the Penalized Composite Link Model (PCLM) developed in this paper are embedded in the GLM framework. Finally, the two models are also compared with the separate modelling approach to show the benefits gained by combining experience data.

The two models developed in this paper use the P-splines smoothing method to avoid overfitting. This smoothing method introduced by Eilers and Marx (1996) is based on the idea of penalizing models with large variability between coefficients of adjacent splines.

The remainder of this paper is arranged as follows. Section 3.2 introduces the research

question and modelling framework. In Section 3.3, we provide technical details on estimation methods, including P-splines smoothing, constrained optimization model, and the Penalized Composite Link Model (PCLM). These models are applied to a real-life problem in Section 3.4; in particular, the application is for the estimation results of two portfolios with and without the deferred period. Section 3.5 concludes.

3.2 Modelling of the multidefinition problem

In this paper, long-term care products are modelled by using the semi-Markov framework, where mortality in disability states depends on both attained age and time already spent in the disability state.

Let us consider two disability definitions, and let T_k denote the LTC state associated with definition $k \in \{1, 2\}$. Let $X_x^{(k)}$ denote the health status of a policyholder observed with definition k. The three possible health statuses in a long-term care insurance product are healthy, disabled and dead.

Let Ω_k denote the set of possible health statuses of a policyholder observed with disability definition k, then

$$\Omega_k = \{H_k, T_k, Death\},$$

where H_k denotes the healthy state of definition k.

Let F_α and F_β denote two levels of disability such that $F_\alpha \cap F_\beta = \varnothing$.

The two disability definitions are defined as follows:

- $T_1 = F_\alpha \cup F_\beta$, and
- $T_2 = F_\beta$,

such that T_2 is included in T_1. Being disabled with definition 2 implies disability with definition 1.

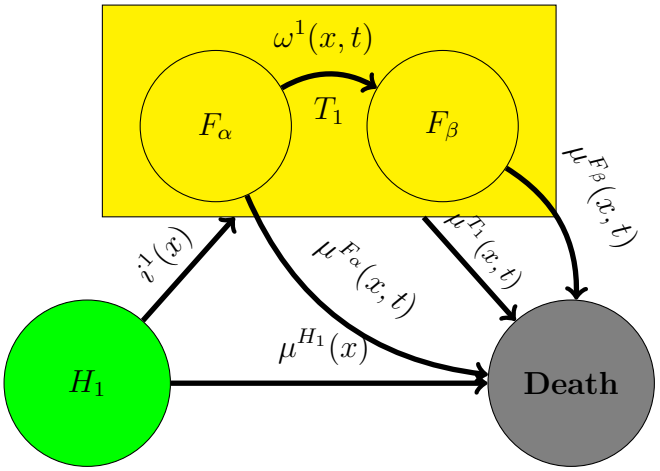
By construction, the autonomous state of the second definition is $H_2 = H_1 \cup F_\alpha$. A policyholder being disabled with definition 1 is either in health state F_α or F_β. In this paper, we consider that for each disabled policyholder observed with definition 1 (i.e. $X_x^{(1)} = T_1$), its refined LTC state F_α or F_β is known.

Let us assume the following:

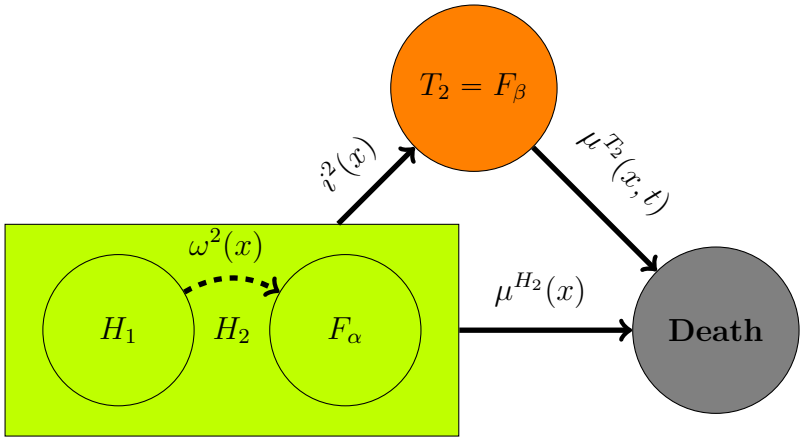
Assumption. 3.2.1

No return to a better health state is envisaged, i.e., recovery is assumed to be impossible.

We then have transition diagrams for each of the two types of LTC insurance, as represented in Figure 3.2.1, where $i^k(x)$, $\mu^{H_k}(x)$ with $k \in \{1,2\}$ and $\mu^{F_g}(x,t)$ with $g \in \{\alpha, \beta\}$ are the transition intensities.



(a) Definition 1



(b) Definition 2

Figure 3.2.1 – Modelling of the long-term care product with two disability definitions

The transition intensities are given by

- $i^k(x) = \lim_{h\to 0} \dfrac{\mathbb{P}(X^{(k)}_{x+h} = T_k | X^{(k)}_x = H_k)}{h}$,
- $\omega^2(x) = \lim_{h\to 0} \dfrac{\mathbb{P}(X^{(2)}_{x+h} = F_\alpha | X^{(2)}_x = H_1)}{h}$,
- $\mu^{H_k}(x) = \lim_{h\to 0} \dfrac{\mathbb{P}(X^{(k)}_{x+h} = \text{Death} | X^{(k)}_x = H_k)}{h}$,
- $\mu^{F_g}(x,t) = \lim_{h\to 0} \dfrac{\mathbb{P}(X^{(1)}_{x+h} = \text{Death} | X^{(1)}_x = F_g, X^{(1)}_{(x-t)} = F_g, X^{(1)}_{(x-t)^-} \neq F_g)}{h}$, and
- $\mu^{T_k}(x,t) = \lim_{h\to 0} \dfrac{\mathbb{P}(X^{(k)}_{x+h} = \text{Death} | X^{(k)}_x = T_k, X^{(k)}_{x-t} = T_k, X^{(k)}_{(x-t)^-} = H_k)}{h}$.

The dashed arrow represents a transition that we cannot observe with disability definition 2. Transitions from F_α to F_β are assumed to be observed for individuals who are disabled according to disability definition 1.

Therefore, $\mu^{F_g}(x,t)$ denotes the mortality of a disabled of age x, who is in state F_g since t years, and $\mu^{T_k}(x,t)$ denotes the mortality of an individual of age x, disabled since t years according to the definition k.

Each disability definition can be associated with a different long-term care product. In the remainder of the paper, the terms "product" and "disability definition" are equivalent.

We consider the following assumptions:

Assumption. 3.2.2

The policyholders are homogeneous and independent under the two types of insurance.

Assumption. 3.2.3

The interval of ages can be divided into subintervals such that mortality and incidence rates from healthy states are constant in each interval. The split points are denoted $\{x_1^H, x_2^H, ..., x_{M^H+1}^H\}$, where M^H denotes the number of intervals of ages for healthy states.

Assumption. 3.2.4

The interval of ages and durations can be divided into subintervals such that mortality rates in states $F_g, g \in \{\alpha, \beta\}$ are constant in each interval. The split points of the age and duration dimensions are denoted $\{x_1^F, x_2^F, ..., x_{M_x^F+1}^F\}$ and $\{t_1^g = 0, t_2^g, ..., t_{M_t^g+1}^g\}$, respectively.

$$\mu^{F_g}(x,t) = \mu^{F_g}(x_p^F, t_q^g), \forall t | t_q^g \leqslant t < t_{q+1}^g, \forall x | x_p^F \leqslant x < x_{p+1}^F.$$

M_x^F and M_t^g denote the number of subdivisions of the intervals of ages and durations for the disabled state F_g, respectively.

The subdivision of ages can be different for healthy and disabled states. However, it is assumed in the following that the set of ages $\{x_1^F, x_2^F, ..., x_{M_x^F+1}^F\}$ is included in $\{x_1^H, x_2^H, ..., x_{M^H+1}^H\}$. Therefore, the subdivision of ages in the healthy state is a more refined subdivision of the interval of ages.

Let us note that the subdivision of the interval of ages is the same for the disabled states F_α and F_β, whereas the subdivision of the duration interval can differ.

As intensities $\mu^{H_k}()$ and $i^k(), k \in \{1,2\}$ and $\mu^{F_g}(,), g \in \{\alpha, \beta\}$ are piecewise constants (Assumption 3.2.3 and Assumption 3.2.4), we denote

- $\mu_p^{H_k} = \mu^{H_k}(x_p^H), \forall p \in \{1, \ldots, M^H\}$,

- $i_p^k = i^k(x_p^H), \forall p \in \{1, \ldots, M^H\}$,
- $\mu_{p,q}^{F_g} = \mu^{F_g}(x_p^F, t_q^g), \forall p \in \{1, \ldots, M_x^F\}, q \in \{1, \ldots, M_t^g\}$.

The total likelihood for all individuals from products 1 and 2 is given by

$$\begin{aligned} L_{tot} \propto & \prod_{p=1}^{M^H} \exp\left(-\left(i_p^1 + \mu_p^{H_1}\right) e^{H_1}(p)\right) \left(i_p^1\right)^{N^1(p)} \left(\mu_p^{H_1}\right)^{D^{H_1}(p)} \\ & \prod_{p=1}^{M_x^F} \prod_{q=1}^{M_t^\alpha} \exp\left(-\mu_{p,q}^{F_\alpha} e^\alpha(p,q)\right) \left(\mu_{p,q}^{F_\alpha}\right)^{D^\alpha(p,q)} \times \\ & \prod_{p=1}^{M^H} \exp\left(-\left(i_p^2 + \mu_p^{H_2}\right) e^{H_2}(p)\right) \left(i_p^2\right)^{N^2(p)} \left(\mu_p^{H_2}\right)^{D^{H_2}(p)} \\ & \prod_{p=1}^{M_x^F} \prod_{q=1}^{M_t^\beta} \exp\left(-\mu_{p,q}^{F_\beta} e^\beta(p,q)\right) \left(\mu_{p,q}^{F_\beta}\right)^{D^\beta(p,q)}, \end{aligned} \tag{3.1}$$

where i^1, μ^{H_1}, i^2, μ^{H_2}, μ^{F_α}, and μ^{F_β} are unknown, while the following are observed and are denoted such that:

- $e^{H_k}(p)$ is the sum of the central exposures in state H_k of all policyholders from product k between integer ages x_p^H and x_{p+1}^H,
- $e^\alpha(p,q)$ is the sum of the central exposures of all disabled policyholders from product 2 being in state F_α between integer ages x_p^T and x_{p+1}^T and durations t_q^g and t_{q+1}^g,
- $e^\beta(p,q)$ is the sum of the central exposures of all disabled policyholders in state F_β from both products between integer ages x_p^T and x_{p+1}^T and durations t_q^g and t_{q+1}^g,
- $D^{H_1}(p)$ denotes the number of deaths from the healthy state between x_p^H and x_{p+1}^H in the first type of insurance,
- $D^{H_2}(p)$ denotes the number of deaths from the healthy state between x_p^H and x_{p+1}^H in the second type of insurance,
- $D^\alpha(p,q)$ denotes the number of deaths from the F_α disabled state with age at death between x_p^T and x_{p+1}^T and duration between t_q^α and t_{q+1}^α,
- $D^\beta(p,q)$ is the total number of deaths of disabled policyholders with age at death between x_p^T and x_{p+1}^T and duration between t_q^β and t_{q+1}^β from both products.
- $N^1(p)$ denotes the number of transitions from H_1 to the disabled state T_1 between x_p^H and x_{p+1}^H, observed in product 1,
- $N^2(p)$ denotes the number of transitions from H_2 to the disabled state T_2 between x_p^H and x_{p+1}^H, observed in product 2.

The detailed proof of Equation 3.1 is given in Appendix 3.A.

The last term of Equation 3.1 represents the likelihood of all observations from products 1 and 2 in state F_β.

The log likelihood function of the combined observations from products 1 and 2 is given by

$$\begin{aligned} l_{tot} = & \sum_{p=1}^{M^H} \left(- \left(i_p^1 + \mu_p^{H_1} \right) e^{H_1}(p) \right) + N^1(p) \log \left(i_p^1 \right) + D^{H_1}(p) \log \left(\mu_p^{H_1} \right) + \\ & \sum_{p=1}^{M_x^F} \sum_{q=1}^{M_t^\alpha} \left(-\mu_{p,q}^{F_\alpha} e^\alpha(p,q) \right) + D^\alpha(p,q) \log \left(\mu_{p,q}^{F_\alpha} \right) + \\ & \sum_{p=1}^{M^H} \left(- \left(i_p^2 + \mu_p^{H_2} \right) e^{H_2}(p) \right) + N^2(p) \log \left(i_p^2 \right) + D^{H_2}(p) \log \left(\mu_p^{H_2} \right) + \\ & \sum_{p=1}^{M_x^F} \sum_{q=1}^{M_t^\beta} \left(-\mu_{p,q}^{F_\beta} e^\beta(p,q) \right) + D^\beta(p,q) \log \left(\mu_{p,q}^{F_\beta} \right) + cst. \end{aligned} \tag{3.2}$$

where *cst* denotes a constant.

Equation 3.2 is equivalent to a sum of log-likelihoods of the Poisson distribution, where:

- $N^1(p) \sim \text{Poisson}\left(i_p^1 e^{H_1}(p)\right)$,
- $N^2(p) \sim \text{Poisson}\left(i_p^2 e^{H_2}(p)\right)$,
- $D^{H_1}(p) \sim \text{Poisson}\left(\mu_p^{H_1} e^{H_1}(p)\right)$,
- $D^{H_2}(p) \sim \text{Poisson}\left(\mu_p^{H_2} e^{H_2}(p)\right)$,
- $D^\alpha(p,q) \sim \text{Poisson}\left(\mu_{p,q}^{F_\alpha} e^\alpha(p,q)\right)$, and
- $D^\beta(p,q) \sim \text{Poisson}\left(\mu_{p,q}^{F_\beta} e^\beta(p,q)\right)$.

Therefore, we assume in the following that the counts of deaths and the counts of losses of autonomy have a Poisson distribution.

Moreover, Equation 3.2 shows that each transition rate can be estimated separately and independently. As we focus this research on mortality rates, incidence rates are considered constants in Equation 3.2. The problem is therefore simplified to

$$\max_{\mu_{\cdot}^{H_1}, \mu_{\cdot}^{H_2}, \mu_{\cdot,\cdot}^{F_\alpha}, \mu_{\cdot,\cdot}^{F_\beta}} l_{tot},$$

where l_{tot} is given by

$$\begin{aligned} l_{tot} = & \underbrace{\sum_{p=1}^{M^H} \left(-\mu_p^{H_1} e^{H_1}(p) \right) + D^{H_1}(p) \log \left(\mu_p^{H_1} \right)}_{l^{H_1}} + \underbrace{\sum_{p=1}^{M_x^F} \sum_{q=1}^{M_t^\alpha} \left(-\mu_{p,q}^{F_\alpha} e^\alpha(p,q) \right) + D^\alpha(p,q) \log \left(\mu_{p,q}^{F_\alpha} \right)}_{l^{F_\alpha}} + \\ & \underbrace{\sum_{p=1}^{M^H} \left(-\mu_p^{H_2} e^{H_2}(p) \right) + D^{H_2}(p) \log \left(\mu_p^{H_2} \right)}_{l^{F_\beta}} + \underbrace{\sum_{p=1}^{M_x^F} \sum_{q=1}^{M_t^\beta} \left(-\mu_{p,q}^{F_\beta} e^\beta(p,q) \right) + D^\beta(p,q) \log \left(\mu_{p,q}^{F_\beta} \right)}_{l^{H_2}}. \end{aligned} \tag{3.3}$$

3.3 Methods

In this section, we introduce the two proposed methods to make better use of all the available information from observations with different disability definitions. In the first part of this section, we focus on the P-splines smoothing framework that is used in the two methods to prevent overfitting and to produce smooth estimated mortality laws. For the second part, we present the first method that uses the constrained optimization algorithm. The third and final part of this section is devoted to the presentation of the Penalized Composite Link Model, corresponding to the second method proposed in this paper. While the first method assumes that the mortality intensities in each state H_1, H_2, F_α and F_β can be expressed as the exponential of a combination of basis-splines, the second method uses the same assumption only on the three following states H_1, F_α and F_β. No assumption on the shape of the mortality in H_2 is made in the second model, leading to fewer coefficients to estimate.

3.3.1 P-splines smoothing framework

Common parametric smoothing methods are introduced by Gompertz and Gompertz-Makeham, assuming that mortality increases exponentially with age, or as in Perks (1932), Beard (1959) allowing a deceleration of the mortality increase at old ages. Unlike these models, the P-splines smoothing method is a nonparametric smoothing technique. One of the strengths of these techniques is that they do not assume any particular shape of the mortality function.

A spline is a piecewise polynomial function that is continuous and has continuous derivatives up to a certain order. This type of function is commonly used for smoothing problems. The P-splines smoothing method, introduced in Eilers and Marx (1996), uses a B-spline basis with penalties to prevent overfitting. This smoothing method, also described in Marx and Eilers (1998), is applied to mortality estimation by using the Poisson-GLM framework in Currie and Durban (2002) and Macdonald et al. (2018).

First, let us consider the case of a one-dimensional mortality law, as the mortality in H_1 depends only on age. J^H splines are uniformly positioned over the entire interval of ages. Splines are therefore equidistant. An important property of the spline basis is that the sum of the B-splines equals 1 at any point on the support. Let $B^{H_1} \in \mathcal{M}_{M^H, J^{H_1}}$ denote the matrix of the spline basis for estimation of mortality in H_1, where each column corresponds to a spline, and each row corresponds to an observation age ($x_p^H, p \in \{1, \dots, M^H\}$ in the case of mortality in healthy states). In the context of mortality estimation, the P-splines smoothing method requires the following assumption:

Assumption. 3.3.1

Mortality rates can be expressed as the exponential of a combination of basis-splines.

Therefore,

$$\mu^{H_1} = \exp\left(B^{H_1}\theta^{H_1}\right), \text{ where } \theta^{H_1} \in \mathbb{R}^{J^{H_1}}. \tag{3.4}$$

A term $\frac{1}{2}\rho\|\Delta^{H_1}\theta^{H_1}\|_2^2$ is then added to the log-likelihood, where Δ^{H_1} is a matrix penalizing complex models with large variability between coefficients of adjacent splines. For simplicity, the penalty term can be written as

$$\frac{1}{2}\left(\theta^{H_1}\right)^T P^{H_1}\theta^{H_1}, \tag{3.5}$$

where $P^{H_1} \in \mathcal{M}_{J^{H_1},J^{H_1}}$ is given by $P^{H_1} = \rho\left(\Delta^{H_1}\right)^T\Delta^{H_1}$.

The problem is therefore

$$\max_{\theta^{H_1}} l^{H_1}\left(\theta^{H_1}\right) - \frac{1}{2}\left(\theta^{H_1}\right)^T P^{H_1}\theta^{H_1}. \tag{3.6}$$

The output of this smoothing method depends on multiple hyperparameters listed below:

1. The number of splines distributed in the interval,
2. The degree of the splines,
3. The order of the penalty, denoted d,
4. The smoothing parameter ρ is the weight of the penalty in the penalized log-likelihood .

As the extrapolation of the fitted mortality is mostly driven by the form of the penalty matrix, the choice of the order d and the smoothing weight ρ have a significant impact. In contrast, the number of splines and their degree are less critical, and have a rather limited impact on the fitted mortality laws. Ruppert (2002) and Eilers (2007) suggest that the following choice is often sufficient for these two parameters:

- Use cubic splines (degree 3),
- Fix a knot every 4 or 5 observations.

As recommended in Currie and Durban (2002), parameters ρ and d are often chosen to minimize the Bayesian Information Criterion (BIC), as defined in Schwarz (1978), given by

$$BIC = -2l^{H_1}\left(\theta^{H_1}\right) + \log(M^H) \times df,$$

where df is the degree of freedom of the model.

The one-dimensional P-splines smoothing method can be generalized to two-dimensional smoothing problems to consider the duration in the context of estimating mortality in disabled states. To this aim, two matrices of basis splines are needed. Let $g \in \{\alpha, \beta\}$. For each state F_g, let $B_x^g \in \mathcal{M}_{M_x^F, J_x^g}$ and $B_t^g \in \mathcal{M}_{M_t^g, J_t^g}$ denote the matrices of splines for the age and duration dimensions, respectively. J_t^g and J_x^g denote the number of splines for each dimension.

Notation: Let A be a matrix of dimensions $r \times c$, $A_{vec} = vec(A) = (A_{.1}^T, \ldots, A_{.c}^T)^T$, where $A_{.k} \in \mathbb{R}^r$ is the k^{th} column of matrix A.

Then,

$$\mu_{vec}^{F_g} = \exp\left(B^g \theta^g\right), \tag{3.7}$$

with $B^g = B_t^g \otimes B_x^g \in \mathcal{M}_{M_x^F \times M_t^g, J_x^g \times J_t^g}$, where $\otimes$ represents the Kronecker product.

The penalty matrix for two-dimensional P-splines smoothing is the sum of a penalty term on the age dimension and a penalty term on the duration dimension. The overall penalty matrix is given by the following equation:

$$P^g = (I_{J_t^g} \otimes P_x^g) + (P_t^g \otimes I_{J_x^g}) \in \mathcal{M}_{J_x^g \times J_t^g, J_x^g \times J_t^g}, \tag{3.8}$$

where P_x^g and P_t^g are the penalty matrix as described for the one-dimensional case and $I_{J_t^g}$ and $I_{J_t^g}$ are the identity matrices of dimensions J_x^g and J_t^g, respectively.

3.3.2 Optimization with constraint

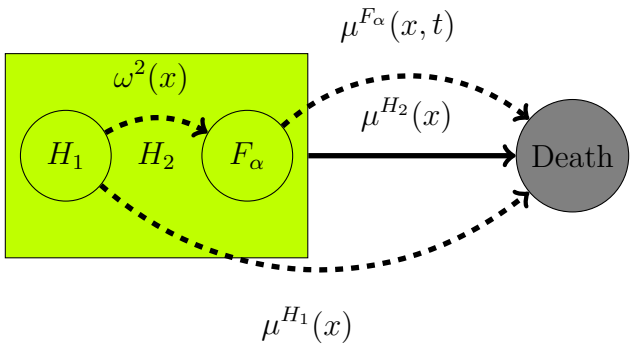


Figure 3.3.1 – Zoom on transitions from state H_2 to Death

Figure 3.3.1 shows that $H_2 = H_1 \cup F_\alpha$. Therefore, at each age x, the mortality in state H_2 is a weighted average of the mortality in H_1 and the mortality in F_α. The 3 mortality laws μ^{H_2}, μ^{H_1} and μ^{F_α}, are linked by the following Equation (3.9)

$$\begin{aligned}\mu_p^{H_2} &= \lim_{h\to 0} \frac{\mathbb{P}(X^{(2)}_{x_p^H+h} = \text{Death} | X^{(2)}_{x_p^H} = H_2)}{h} \\ &= \frac{\mathbb{P}(X_x^{(1)} = H_1)}{\mathbb{P}(X_x^{(1)} = H_1 \cup X_x^{(1)} = F_\alpha)} \mu_p^{H_1} + \sum_{q=1}^{M_t^\alpha} \frac{\mathbb{P}(X_x^{(1)} = F_\alpha, X^{(1)}_{(x-t_q^\alpha)} = F_\alpha, X^{(1)}_{(x-t_{q+1}^\alpha)} = H_1)}{\mathbb{P}(X_x^{(1)} = H_1 \cup X_x^{(1)} = F_\alpha)} \mu_{p,q}^{F_\alpha}.\end{aligned} \tag{3.9}$$

A detailed proof of Equation (3.9) is given in Appendix 3.B.

By using the P-splines smoothing framework for each of the four mortality laws, the optimal coefficients $\theta^{H_1}, \theta^{H_2}, \theta^\alpha$ and θ^β are obtained by solving the following constrained optimization problem

$$\begin{aligned}\max_{\theta^{H_1},\theta^{H_2},\theta^\alpha,\theta^\beta} & l^{H_1}(\theta^{H_1}) - \frac{1}{2}\left(\theta^{H_1}\right)^T P^{H_1}\theta^{H_1} + l^{H_2}(\theta^{H_2}) - \frac{1}{2}\left(\theta^{H_2}\right)^T P^{H_2}\theta^{H_2} + \\ & l^\alpha(\theta^\alpha) - \frac{1}{2}\left(\theta^\alpha\right)^T P^\alpha\theta^\alpha + l^\beta(\theta^\beta) - \frac{1}{2}\left(\theta^\beta\right)^T P^\beta\theta^\beta,\end{aligned} \tag{3.10}$$

subject to

$$\begin{aligned}\mu_p^{H_2}\left(\theta^{H_2}\right) = & \frac{\mathbb{P}(X_x^{(1)} = H_1)}{\mathbb{P}(X_x^{(1)} = H_1 \cup X_x^{(1)} = F_\alpha)} \mu_p^{H_1}\left(\theta^{H_1}\right) + \\ & \sum_{q=1}^{M_t^\alpha} \frac{\mathbb{P}(X_x^{(1)} = F_\alpha, X^{(1)}_{(x-t_q^\alpha)} = F_\alpha, X^{(1)}_{(x-t_{q+1}^\alpha)} = H_1)}{\mathbb{P}(X_x^{(1)} = H_1 \cup X_x^{(1)} = F_\alpha)} \mu_{p,q}^{F_\alpha}\left(\theta^\alpha\right).\end{aligned} \tag{3.11}$$

The optimization problem is solved by using Sequential Quadratic Programming (SQP) as described in Kraft (1988) and Boggs and Tolle (1995). The SQP method is implemented on the statistical programming software **R** in the **slsqp()** function from the **nloptr** package. The SQP method replaces the original problem of optimization with a sequence of quadratic problems where the objectives are second-order approximations of the Lagrangian and the constraints are first-order approximations of the original constraints. One of the major advantages of the SQP methods is that the initial point does not need to satisfy all the constraints of the original problem. The documentation of the package can be found in Jelmer Ypma (2022).

3.3.3 Penalized Composite Link Model

The Penalized Composite Link Model was first proposed by Eilers (2007). This method is an extension of the Composite Link Model introduced by Thompson and Baker (1981), that uses the Generalized Linear Model framework (GLM) (Nelder and Wedderburn, 1972;

McCullagh, 2019).

By assuming that the policyholders are homogeneous and independent under the two types of insurance, policyholders in state F_β from both products 1 and 2 can be aggregated.

As in Section 3.2, $D^{H_k} \in \mathbb{N}^{M^H}$ denotes the vector of counts of deaths in the healthy state $H_k, k \in \{1,2\}$. $D^g \in \mathcal{M}_{M_x^F, M_t^g}(\mathbb{N})$ is the matrix of counts of deaths in state F_g. Since policyholders cannot be observed in state F_α with product 2, D^α is only composed of deaths of policyholders from product 1. However, since state F_β is observed for both products, $D^\beta \in \mathcal{M}_{M_x^F, M_t^\beta}(\mathbb{N})$ is the sum of the matrices of counts of deaths in F_β of the two products. D_{ij}^β denotes the number of deaths in F_β at age $x_i^F < x < x_{i+1}^F$ and duration $t_j^\beta < t < t_{j+1}^\beta$. Let D be the vector of counts of deaths such that $D = (D^{H_1}, D_{vec}^\alpha, D_{vec}^\beta, D^{H_2})^T \in \mathbb{N}^{2.M^H + M_x^F.(M_t^\alpha + M_t^\beta)}$. Similar to the vector of counts of deaths, $e = (e^{H_1}, e_{vec}^\alpha, e_{vec}^\beta, e^{H_2})^T \in \mathbb{N}^{2.M^H + M_x^F.(M_t^\alpha + M_t^\beta)}$ is the concatenation of the vectors of central exposures in each state H_1, F_α, F_β, and H_2.

Let us assume that the mortality rates in states H_1 and $F_g, g \in \{\alpha, \beta\}$ can be expressed as the exponential of a combination of basis splines as in Assumption 3.3.1. Let

- $\mu^{H_1} = \left(\mu^{H_1}(x_1^H), \dots, \mu^{H_1}(x_{M^H}^H)\right)^T \in \mathbb{R}^{M^H}$, and
- $\mu_{vec}^{F_g} = vec(\mu^{F_g}) \in \mathbb{R}^{M_x^F.M_t^g}$, where $\mu^{F_g} \in \mathcal{M}_{M_x^F, M_t^g}(\mathbb{R})$ is the matrix of the mortality intensities in state F_g such that

$$\mu_{p,q}^{F_g} = \mu^{F_g}\left(x_p^F, t_q^g\right) \forall p \in \{1, \dots, M_x^F\}, q \in \{1, \dots, M_t^g\}.$$

Then,

$$\mu^{H_1} = \exp(B^{H_1}\theta^{H_1}), \tag{3.12}$$

$$\mu_{vec}^{F_g} = \exp(B^g\theta^g), g \in \{\alpha, \beta\}, \tag{3.13}$$

where $B^{H_1} \in \mathcal{M}_{M^H, J^{H_1}}(\mathbb{R})$ and $B^g \in \mathcal{M}_{M_x^F \times M_t^g, J^g}(\mathbb{R})$ are the splines basis for states H_1 and F_g, respectively. J^{H_1} and $J^g, g \in \{\alpha, \beta\}$ denote the number of coefficients of splines needed to estimate the mortality laws in the healthy state and disabled states g, respectively. Using notations from Section 3.3.1, $J^g = J_x^g \times J_t^g$. Therefore, $\theta^{H_1} \in \mathbb{R}^{J^{H_1}}$ and $\theta^g \in \mathbb{R}^{J^g}$.

A difference with Section 3.3.2 is that no assumption on the mortality in state H_2 is made.

Based on Equation 3.2, it is assumed that the counts of deaths have a Poisson distribution with parameter $e \times \mu$, where e and μ denote the vectors of central exposures and transition intensities, respectively. This assumption is commonly used with Penalized Composite Link Models, as in Eilers (2007), Remund et al. (2018) or Rizzi et al. (2015). Therefore,

$$\hat{D^{H_k}} = \mathbb{E}[D^{H_k}] = e^{H_k} \times \mu^{H_k}, k \in \{1,2\}, \tag{3.14}$$

$$\hat{D^g} = \mathbb{E}[D^g] = e^g \times \mu^{F_g}, g \in \{\alpha, \beta\}. \tag{3.15}$$

With observations from product 1, we can create a fictitious state H_2 composed of policyholders in H_1 and F_α from product 1. We denote H_2^{prod1} this fictitious state. Given Assumption 3.2.2, mortality in H_2^{prod1} is the same as in H_2. Therefore,

$$\hat{D}^{H_2^{prod1}} = e^{H_2^{prod1}} \times \mu^{H_2}, \tag{3.16}$$

where $D^{H_2^{prod1}}$ and $e^{H_2^{prod1}}$ denote the vectors of counts of deaths and central exposures in the state $H_2^{prod1} = H_1 \cup F_\alpha$, respectively. The subdivision of the age interval of this state is the same as for states H_1 and H_2.
Let $p \in \{1, \cdots, M^H\}$. From Assumption 3.2.4, we know that the subdivision of the interval of ages in healthy states is either the same or more refined than is the subdivision of ages for the disabled states. Moreover,

$$\forall p \in \{1, \cdots, M^H\}, \exists \tilde{p} \in \{1, \cdots, M_x^F\}, x_{\tilde{p}}^F \leqslant x_p^H < x_{p+1}^H \leqslant x_{\tilde{p}+1}^F.$$

Therefore, if not directly calculable from the database, the central exposure in state F_α between ages x_p^H and x_{p+1}^H (subdivision of the interval of ages used in states H_k) for duration between t_q^α and t_{q+1}^α can be estimated by

$$e_{sub_H}^\alpha(p,q) = e^\alpha(\tilde{p},q)\frac{x_{p+1}^H - x_p^H}{x_{\tilde{p}+1}^F - x_{\tilde{p}}^F},$$

as illustrated in Figure 3.3.2.

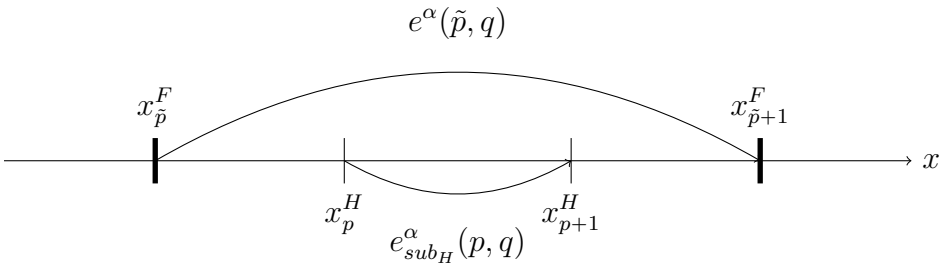


Figure 3.3.2 – Subdivision of the age interval

A death in state H_2^{prod1} is either a death in H_1 or F_α. With product 1, we are able to decompose this count of deaths as a sum of deaths in H_1 and F_α. Therefore,

$$\hat{D}^{H_2^{prod1}}(p) = e^{H_1}(p) \times \mu_p^{H_1} + \sum_{q=1}^{M_t^\alpha} e_{sub_H}^\alpha(p,q) \times \mu^{F_\alpha}(x_p^H, t_q^\alpha). \tag{3.17}$$

Since $x_{\tilde{p}}^F \leqslant x_p^H < x_{\tilde{p}+1}^F$,

$$\mu^{F_\alpha}(x_p^H, t_q^\alpha) = \mu^{F_\alpha}(x_{\tilde{p}}^F, t_q^\alpha) = \mu_{\tilde{p},q}^{F_\alpha}. \tag{3.18}$$

Therefore, Equation 3.17 becomes

$$\hat{D}^{H_2^{prod1}}(p) = e^{H_1}(p) \times \mu_p^{H_1} + \sum_{q=1}^{M_t^\alpha} e_{sub_H}^{\alpha}(p,q) \times \mu_{\tilde{p},q}^{F_\alpha}. \tag{3.19}$$

Moreover, Equation 3.14 and Equation 3.16 lead to

$$\hat{D^{H_2}}(p) = e^{H_2}(p) \times \frac{\hat{D}^{H_2^{prod1}}(p)}{e^{H_2^{prod1}}(p)}. \tag{3.20}$$

Then,

$$\hat{D^{H_2}}(p) = e^{H_2}(p)\frac{e^{H_1}(p)}{e^{H_2^{prod1}}(p)} \times \mu_p^{H_1} + \sum_{q=1}^{M_t^\alpha} e^{H_2}(p)\frac{e_{sub_H}^{\alpha}(p,q)}{e^{H_2^{prod1}}(p)} \times \mu_{\tilde{p},q}^{F_\alpha}. \tag{3.21}$$

Therefore, all the expected values of the Poisson variables $D^{H_k}(p), k \in \{1,2\}$ and $D^g(p,q), g \in \{\alpha,\beta\}$ are linear combinations of the components of the vector $\Lambda = (\mu^{H_1}, \mu_{vec}^{F_\alpha}, \mu_{vec}^{F_\beta})$.

We then write $D \sim Poisson(C\Lambda)$, where:

- $C \in \mathcal{M}_{(M^H+M^H+(M_t^\alpha+M_t^\beta).M_x^F),(M^H+(M_t^\alpha+M_t^\beta).M_x^F)}(\mathbb{R})$, and
- $\Lambda(\Theta) = \begin{bmatrix} \mu^{H_1}(\theta^{H_1}) \\ \mu_{vec}^{F_\alpha}(\theta^\alpha) \\ \mu_{vec}^{F_\beta}(\theta^\beta) \end{bmatrix} \in \mathbb{R}^{(M^H+(M_t^\alpha+M_t^\beta).M_x^F)}$, with $\Theta = (\theta^{H_1}, \theta^\alpha, \theta^\beta)^T \in \mathbb{R}^{J^{H_1}+J^\alpha+J^\beta}$.

Details for the structure of matrix C are given in Appendix 3.C.

Therefore,

$$\hat{D} = C\Lambda(\Theta), \tag{3.22}$$

$$\Lambda(\Theta) = \exp(B\Theta), \tag{3.23}$$

where

$$B = \begin{bmatrix} B^{H_1} & 0 & 0 \\ 0 & B^\alpha & 0 \\ 0 & 0 & B^\beta \end{bmatrix} \in \mathcal{M}_{(M^H+(M_t^\alpha+M_t^\beta).M_x^F),(J^{H_1}+J^\alpha+J^\beta)}(\mathbb{R}). \tag{3.24}$$

To have smooth mortality rates, a penalty inspired from the P-splines smoothing method is added to the log-likelihood such that

$$l^{pen}(\Theta) = l(\Theta) - \frac{1}{2}\Theta^T P \Theta, \tag{3.25}$$

where:

- $l(\Theta)$ is the log-likelihood associated with the random vector $D = (D^{H_1}, D^{\alpha}_{vec}, D^{\beta}_{vec}, D^{H_2})^T$,
- $$P = \begin{bmatrix} P^{H_1} & 0 & 0 \\ 0 & P^{\alpha} & 0 \\ 0 & 0 & P^{\beta} \end{bmatrix} + \left[\begin{array}{c|cc} 0 & 0 & 0 \\ \hline 0 & \multicolumn{2}{c}{\multirow{2}{*}{$P^{\alpha/\beta}$}} \\ 0 & & \end{array}\right] \in \mathcal{M}_{J^{H_1}+J^{\alpha}+J^{\beta}}(\mathbb{R}), \tag{3.26}$$
  with $P^G = \rho^G (\Delta^G)^T \Delta^G, G \in \{H_1, \alpha, \beta, \alpha/\beta\}$, and
- ρ^G represents the weight given to the penalty, as in Section 3.3.1.

Penalty matrices $P^{H_1} \in \mathcal{M}_{J^{H_1}}(\mathbb{R})$, $P^{\alpha} \in \mathcal{M}_{J^{\alpha}}(\mathbb{R})$, and $P^{\beta} \in \mathcal{M}_{J^{\beta}}(\mathbb{R})$ are the typical penalty matrices used in the P-splines smoothing method. Each state H_1, F_α, and F_β has its own matrix, penalizing coefficients of adjacent splines to ensure that the fitted mortality laws are smooth.

One might want to add a penalty between coefficients of the states F_α and F_β. This is the aim of the penalty matrix $P^{\alpha/\beta} \in \mathcal{M}_{J^{\alpha}+J^{\beta}}(\mathbb{R})$. If not, $P^{\alpha/\beta} = 0$. Using vector notations, we have:

- $\mu^{H_1}_p = \Lambda_p, p \in \{1, \dots, M^H\}$,
- $\mu^{F_\alpha}_{p,q} = (\mu^{\alpha}_{vec})_{(q-1).M^F_x+p} = \Lambda_{M^H+(q-1).M^F_x+p}, p \in \{1, \dots, M^F_x\}, q \in \{1, \dots, M^{\alpha}_t\}$,
- $\mu^{F_\beta}_{p,q} = (\mu^{\beta}_{vec})_{(q-1).M^F_x+p} = \Lambda_{M^H+M^F_x.M^{\alpha}_t+(q-1).M^F_x+p}, p \in \{1, \dots, M^F_x\}, q \in \{1, \dots, M^{\beta}_t\}$, and
- $\mu^{H_2}_p = \Lambda_{M^H+M^F_x.(M^{\alpha}_t+M^{\beta}_t)+p}, p \in \{1, \dots, M^H\}$.

Therefore,

$$\begin{aligned} l(\Theta) = & \sum_{p=1}^{M^H} -(\Lambda(\Theta) \times e)_p + D_p \log\left((\Lambda(\Theta) \times e)_p\right) + \\ & \sum_{p=1}^{M^F_x} \sum_{q=1}^{M^{\alpha}_t} -(\Lambda(\Theta) \times e)_{M^H+(q-1).M^F_x+p} + D_{M^H+(q-1).M^F_x+p} \log\left((\Lambda(\Theta) \times e)_{M^H+(q-1).M^F_x+p}\right) + \\ & \sum_{p=1}^{M^F_x} \sum_{q=1}^{M^{\beta}_t} -(\Lambda(\Theta) \times e)_{M^H+M^F_x.M^{\alpha}_t+(q-1).M^F_x+p} + \\ & \qquad D_{M^H+M^F_x.M^{\alpha}_t+(q-1).M^F_x+p} \log\left((\Lambda(\Theta) \times e)_{M^H+M^F_x.M^{\alpha}_t+(q-1).M^F_x+p}\right) + \\ & \sum_{p=1}^{M^H} -(\Lambda(\Theta) \times e)_{M^H+M^F_x.(M^{\alpha}_t+M^{\beta}_t)+p} + \\ & \qquad D_{M^H+M^F_x.(M^{\alpha}_t+M^{\beta}_t)+p} \log\left((\Lambda(\Theta) \times e)_{M^H+M^F_x.(M^{\alpha}_t+M^{\beta}_t)+p}\right). \end{aligned} \tag{3.27}$$

The optimal coefficient $\hat{\Theta}$ is obtained by maximizing the penalized log-likelihood given by Equation 3.25, i.e.,

$$\hat{\Theta} = \arg\max_{\Theta} \; l^{pen}(\Theta). \tag{3.28}$$

We are then able to write the model as a Composite Link Model. This allows us to simultaneously estimate three smooth mortality laws while having a constraint allowing us to include observations from a portfolio with a different disability definition. Then, from Eilers (2007) and Remund et al. (2018), the coefficient $\hat{\Theta}$ is estimated by repeatedly solving the following system:

$$(\breve{X}^T\tilde{W}\breve{X} + P)\tilde{\Theta} = \breve{X}^T(D - \tilde{D}) + \breve{X}^T\tilde{W}\breve{X}\tilde{\Theta}, \tag{3.29}$$

where $\breve{X} = \tilde{W}^{-1}C\tilde{\Gamma}B$; $\tilde{W} = \text{diag}(\tilde{D})$; $\tilde{\Gamma} = \text{diag}(\Lambda)$; $\tilde{D} = C\Lambda(\tilde{\Theta})$.

A tilde indicates the current approximation at each iteration.

Starting values of the parameter $\tilde{\Theta}$ are needed. A convenient way is to start with the coefficient obtained by separately estimating the mortality rates in both healthy and disabled states (H_1, F_α, and F_β) by using the P-splines smoothing method with observations from product 1 only.

3.4 Application to the problem of the deferred period

3.4.1 Introduction to the deferred period and the problem with the data

The deferred period is the minimum period that the loss of autonomy must last before the benefit begins. The length of the deferred period has a significant impact on the probability of the insurer paying the benefit to the policyholder. The premium is therefore lower as the length of the deferred period increases. The usual length in French long-term care contracts is 3 months. One of the advantages of the deferred period is that it reduces the number of short claims and, therefore, the management costs linked to the payment of the annuities.

Some long-term care products have a deferred period, while some do not. For contracts with a deferred period, as no annuity is paid to policyholders during the deferred period, the loss of autonomy is not reported in the database if the policyholder dies during this period, i.e., before the first annuity, as shown in Figure 3.4.1. This death is considered a death in autonomy. Only policyholders surviving until the end of the deferred period, at least until the payment of the first annuity, have a date of loss of autonomy reported in the database. Therefore, policyholders are considered healthy as long as they have not received any annuity. Therefore, the healthy group is composed of autonomous and newly disabled individuals, as shown in Figure 3.4.1.

Let us assume that the real date of loss of autonomy is available for all disabled policyholders

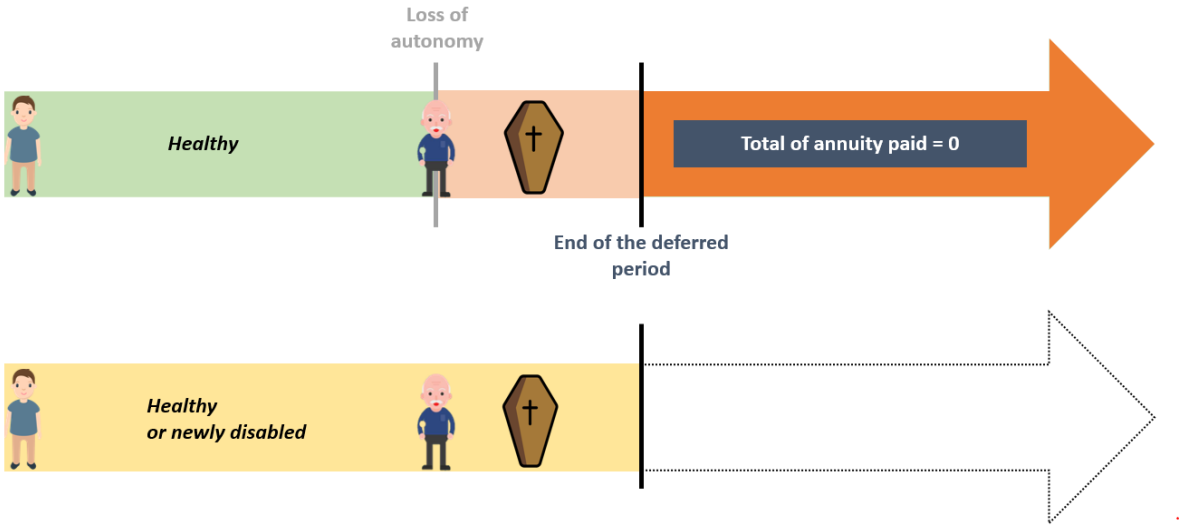


Figure 3.4.1 – Illustration of the consequences of a deferred period on the data observed in the portfolio

who survived the deferred period. If not, the date of loss of autonomy can easily be estimated by subtracting fr months from the date of the first annuity payment.

As the mortality during the first few months of disability is very high and the deaths occurring during the deferred period are reported as deaths in the healthy state, estimating mortality laws on such a database without considering this information can have multiple consequences:

- The mortality in the healthy state is overestimated.
- The mortality of the disabled policyholders during the deferred period is underestimated (only disabled individuals surviving the deferred period are considered exposed, but no deaths are reported during this period).
- The incidence is underestimated since the loss of autonomy is reported only if the newly disabled policyholder survives until the first annuity.

Let us assume that we have a database with some policyholders having a deferred period and some policyholders covered from the first day of the loss of autonomy (without any deferred period). This situation can also occur in the case of having several databases from different insurers. In this situation, one cannot simply aggregate the data basis without accounting for the deferred period of one product.

This situation corresponds to a problem of multidefinition as described in Section 3.2, with one definition included in the other one.

3.4.2 Modelling of the product

In this case, product 1 corresponds to the product without a deferred period, and product 2 corresponds to the product with a deferred period. Being disabled for product 2 implies

being disabled for more than fr months. Therefore, a policyholder disabled with definition 2 (product 2) is disabled with definition 1 (product 1). Thus, $T_2 \subset T_1$.

These products can be represented as in Figure 3.2.1, where:

- H_1 denotes the state of healthy policyholders,
- F_α denotes the state of newly disabled policyholders. The loss of autonomy occurred less than fr months ago,
- F_β denotes the state of disabled policyholders who have lost their autonomy more than fr months ago, and
- H_2 denotes the state of healthy and newly disabled policyholders who have not yet received any annuity.

In this case,

$$\omega^1(x,t) = \begin{cases} 0, & \text{if } t \neq fr/12 \\ +\infty, & \text{if } t = fr/12. \end{cases} \tag{3.30}$$

The transition from F_α to F_β, if observed, systematically occurs at the end of the deferred period. It is natural to think that there is no jump in the mortality function before and after the deferred period. $D^\alpha(p,q), p \in \{1,\dots,M_x^F\}, q \in \{1,\dots,M_t^\alpha\}$ denotes the number of deaths occurring during the deferred period (only observed with product 1). As the maximum duration in F_α is $t = fr/12$, then $t^\alpha_{M_t^\alpha+1} = fr/12$.

$D^\beta(p,q), p \in \{1,\dots,M_x^F\}, q \in \{1,\dots,M_t^\alpha\}$ denotes the sum of counts of deaths occurring after the deferred period for products 1 and 2. As disabled policyholders enter state F_β at the end of the deferred period, the mortality before the end of the deferred period $\mu^{F_\alpha}_{p,M_t^\alpha}$ should be close to the mortality $\mu^{F_\beta}_{p,1}$ during the first subinterval of duration in F_β. It is therefore interesting to consider the mortality in F_α and F_β as only one smooth mortality law $\mu^F(x,t) = \mu^{T_1}(x,t)$, such that $F = \{F_\alpha \cup F_\beta\}$.

As introduced in Section 3.2,

$$\mu^F(x,t) = \lim_{h\to 0} \frac{\mathbb{P}(X_{x+h} = \text{Death} | X_x \in F, X_{x-t} \in F, X_{(x-t)^-} = H_1)}{h}. \tag{3.31}$$

A policyholder who entered state F less than fr months ago is necessarily in state F_α. In contrast, a disabled policyholder who lost autonomy t years ago, with $t > fr/12$, is necessarily in state F_β for $t - fr/12$ years. Therefore,

$$\mu^F(x,t) = \begin{cases} \mu^{F_\alpha}(x,t), & \text{if } t < fr/12 \\ \mu^{F_\beta}\left(x, t - \dfrac{fr}{12}\right), & \text{if } t \geqslant fr/12 \end{cases}. \tag{3.32}$$

A more detailed proof of this equation (Equation 3.32) is available in Appendix 3.D.

Note: The second parameter of $\mu^{F_\beta}(,)$ denotes the time since entry into state F_β. This corresponds to the time since the end of the deferred period. In contrast, the second parameter of $\mu^F(x,t)$ denotes the time since the loss of autonomy, corresponding to the time since entry in F_α.

Let $D^F = [D^\alpha : D^\beta] \in \mathcal{M}_{M_x^F, M_t^\alpha + M_t^\beta}$ be the augmented matrix of the counts of deaths. $D^F_{.,M_t^\alpha}$ denotes the vector of counts of deaths in LTC during the last period before time $fr/12$. $D^F_{.,M_t^\alpha+1} = D^\beta_{.,1}$ denotes the vector of counts of deaths in LTC occurring during the first period after the deferred period.

Therefore,

$$D^F_{vec} = (D^\alpha_{vec}, D^\beta_{vec})^T. \tag{3.33}$$

As the subdivision of the interval of ages is the same for F_α and F_β, the subdivision of the interval of ages of F is $\{x_1^F, \ldots, x_{M_x^F}^F\}$. By using the subdivision of the duration interval from F_α and F_β, the subdivision of the interval for the overall LTC state F is

$$\left\{t_1^F, \ldots, t_{M_t^F+1}^F\right\} = \left\{t_1^\alpha, \ldots, t_{M_t^\alpha+1}^\alpha = t_1^\beta + \frac{fr}{12}, \ldots, t_{M_t^\beta+1}^\beta + \frac{fr}{12}\right\}, \tag{3.34}$$

where $M_t^F = M_t^\alpha + M_t^\beta$.

A common basis of B-splines is used for the age dimension for the 2 states. The matrix of splines for ages in disabled states is denoted $B_x^F, \in \mathcal{M}_{M_x^F, J_x^F}$. Splines are positioned over the entire interval of duration $\left[t_1^F; t_{M_t^F+1}^F = t_{M_t^\beta+1}^\beta + \frac{fr}{12}\right]$. Some splines are common to F_α and F_β, as shown in Figure 3.4.2, which represents the splines basis on the overall duration interval, $J_t^F \leqslant J_t^\alpha + J_t^\beta$. In the example of Figure 3.4.2, 4 splines are common to F_α and F_β. Therefore, $J_t^F = J_t^\alpha + J_t^\beta - 4$.

The matrix of splines for the duration dimension is denoted $B_t^F \in \mathcal{M}_{M_t^\alpha + M_t^\beta, J_t^F}$.

To ensure the same coefficients associated with the splines shared by F_α and F_β, $B^F = B_t^F \otimes B_x^F$ is of a slightly different structure than $\left[\begin{array}{c|c} B^\alpha & 0 \\ \hline 0 & B^\beta \end{array}\right]$.

More details about the structure of the matrices B_t^F and B^F are given in Appendix 3.E.

3.4.2.1 Constrained optimization method applied to the deferred period

The method explained in Section 3.3.2 is applied to the problem of the deferred period. Having one single spline basis for state $F = F_\alpha \cup F_\beta$ has only a slight impact on the form

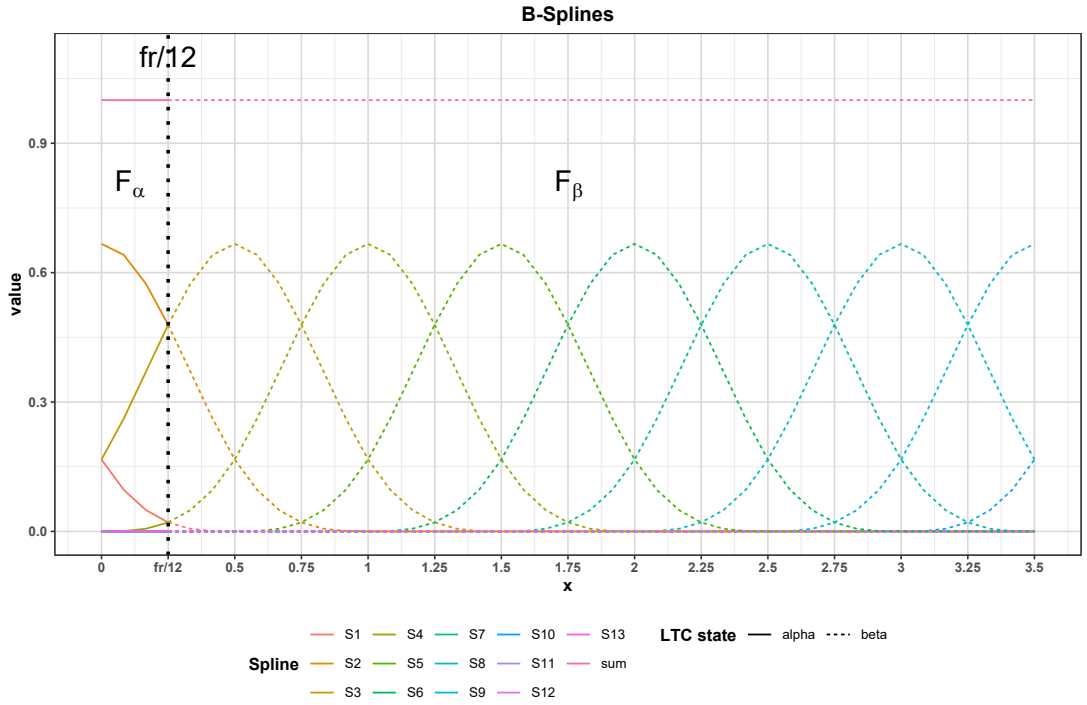


Figure 3.4.2 – Basis of splines for the duration for states F_α and F_β

of the objective function, which becomes

$$\max_{\theta^{H_1},\theta^{H_2},\theta^F} l^{H_1}(\theta^{H_1}) - \frac{1}{2}\left(\theta^{H_1}\right)^T P^{H_1}\theta^{H_1} + l^{H_2}(\theta^{H_2}) - \frac{1}{2}\left(\theta^{H_2}\right)^T P^{H_2}\theta^{H_2} + \boldsymbol{l^{\alpha}(\theta^{F})} + \boldsymbol{l^{\beta}(\theta^{F})} - \frac{\mathbf{1}}{\mathbf{2}}\left(\boldsymbol{\theta^{F}}\right)^{\boldsymbol{T}} \boldsymbol{P^{F}\theta^{F}}, \tag{3.35}$$

subject to

$$\mu_p^{H_2}\left(\theta^{H_2}\right) = \frac{\mathbb{P}(X_x = H_1)}{\mathbb{P}(X_x = H_1 \cup X_x = F_\alpha)}\mu_p^{H_1}\left(\theta^{H_1}\right) + \sum_{q=1}^{M_t^\alpha} \frac{\mathbb{P}(X_x = F_\alpha, X_{(x-t_q^\alpha)} = F_\alpha, X_{(x-t_{q+1}^\alpha)} = H_1)}{\mathbb{P}(X_x = H_1 \cup X_x = F_\alpha)}\mu_{p,q}^{F_\alpha}\left(\boldsymbol{\theta^{F}}\right). \tag{3.36}$$

3.4.2.2 PCLM method applied to the deferred period

The submatrix $\begin{bmatrix} B^\alpha & 0 \\ 0 & B^\beta \end{bmatrix}$ from matrix B in Equation 3.24 is replaced by a single matrix B^F. This slight modification does not affect the remaining formulas or the method for the estimation of the optimal coefficients.

With the same reasoning, the submatrix $\begin{bmatrix} P^\alpha & 0 \\ 0 & P^\beta \end{bmatrix}$ from matrix P in Equation 3.26 is replaced by a single matrix P^F. This allows us to add constraints between splines before

and after the deferred period to ensure smoothness of the mortality law on the axis of the duration.

Therefore, $B = \left[\begin{array}{c|cc} B^{H_1} & 0 & 0 \\ \hline 0 & \multicolumn{2}{c}{\multirow{2}{*}{B^F}} \\ 0 & & \end{array}\right]$, and $P = \left[\begin{array}{c|cc} P^{H_1} & 0 & 0 \\ \hline 0 & \multicolumn{2}{c}{\multirow{2}{*}{P^F}} \\ 0 & & \end{array}\right]$.

3.4.3 Data: Application to a single portfolio by recreating a fictitious deferred period

3.4.3.1 Presentation of the data set

We rely on data from a large French LTC portfolio. This portfolio does not have any deferred period. As the incidence and mortality laws greatly differ for males and females, biometric functions have to be estimated separately for each gender. We focus the application only on females. In this application, we consider mild and severe LTC, with the GIR1234 definition from the AGGIR grid used by the French government for the attribution of public aid and described in Dupourqué (2012). In this portfolio, 1,388 deaths are observed in the autonomous state (H_1), versus 832 in LTC ($T_1 = F_\alpha \cup F_\beta$).

To ensure that Assumption 3.2.2 of homogeneity and independence is met, this portfolio is divided into two data sets. The first one, denoted DB_1, corresponds to the definition 1, without a deferred period. This portfolio is only a subset of the initial portfolio. Policyholders not represented in DB_1 are selected in the second portfolio DB_2. This second database is then modified to fictitiously create a deferred period by postponing the eventual date of loss of autonomy by $fr = 3$ months.

Mortality rates are piecewise constant for the duration. However, since the mortality is very high at the date of occurrence of the loss of autonomy and decreases substantially during the first year, it is common in long-term care modelling to fix smaller steps on the subdivision of the interval of duration during the first year. In the following application, we assume a constant mortality rate by month during the first year of loss of autonomy. Starting from the second year, mortality rates are assumed to be yearly constants.

For the age dimension, mortality rates are assumed to be constant between two integer ages.

After cleaning the original data at the individual granularity, observations are aggregated to obtain exposures and counts of deaths for each subdivision of ages and durations on which mortality rates are assumed to be constants.

Observations from DB_1 are:

- the vector of central exposures in H_1: e^{H_1},
- the vector of counts of deaths in H_1: D^{H_1},

- the matrix of central exposures in F_α: e^α,
- the matrix of counts of deaths in F_α: D^α,
- the matrix of central exposures in F_β (only individuals in DB_1): $e^{\beta,1}$, and
- the matrix of counts of deaths in F_β (only individuals in DB_1): $D^{\beta,1}$.

Observations from DB_2 are:

- the vector of central exposures in H_2: e^{H_2},
- the vector of counts of deaths in H_2: D^{H_2},
- the matrix of central exposures in F_β (only individuals in DB_2): $e^{\beta,2}$, and
- the matrix of counts of deaths in F_β (only individuals in DB_2): $D^{\beta,2}$.

3.4.3.2 Results

In the case of the deferred period, insurers are mostly interested in mortality in H_1 and $F = F_\alpha \cup F_\beta$. Observations from H_2 are used only to improve the estimations of the two other mortality laws. The mortality for H_2 has a rather limited interest for insurers.

In this section, we compare the results of estimations of the 2 mortality laws μ^{H_1} and μ^F with 3 different methods:

1. Independent estimations of each mortality law with the P-splines smoothing method (Without constr.),
2. Constrained optimization (Constr.) from Section 3.3.2,
3. Penalized Composite Link Model (PCLM) from Section 3.3.3.

The smoothing parameter ρ in each of these methods is selected to minimize the BIC.

Estimated mortality intensities in the healthy state and their associated confidence intervals are plotted in Figure 3.4.3. While estimated intensities are similar at young ages where data contain many observations, the mortality curve estimated with the PCLM method diverges from the two others after 85 years old. The estimated mortality function at old ages is higher with the PCLM method compared to that of the mortality obtained by only using information from the portfolio without a deferred period and the mortality estimated with the constrained optimization model. In fact, the optimal smoothing parameter ρ associated with mortality in H_1 minimizing the BIC is smaller with the PCLM method. Therefore, the linear smoothing penalty is weaker, and the fitted mortality function is more flexible than that with the two other methods. It gives more weight to the observations and less to the linear constraint, enabling the capture of more variance at old ages. The size of the confidence interval is smaller with the constrained optimization method than with the two other methods. Confidence intervals are constructed using a bootstrap approach with a residual resampling method as explained in Sartori et al. (2011). A narrower interval indicates that the estimation is more robust to a slight modification of the observed counts of death.

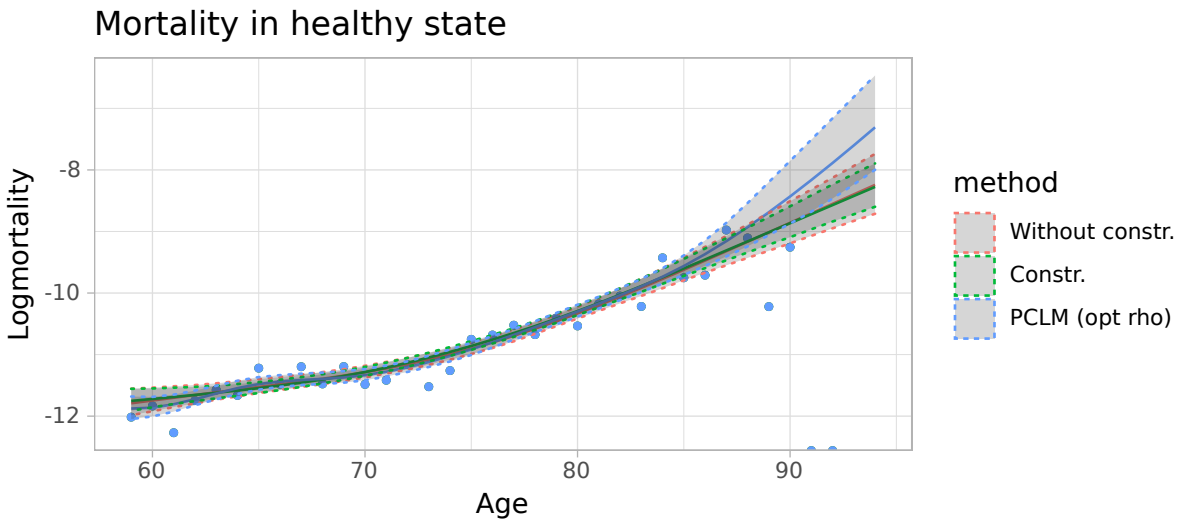


Figure 3.4.3 – Mortality in the healthy state H_1

Mortality in LTC depends on 2 inputs, the attained age and the duration since the loss of autonomy. Therefore, the log-mortality is represented as a surface instead of a curve for mortality in the healthy state. The surfaces of the log-mortality rates estimated with the three methods are plotted in Figure 3.4.4. The resulting mortality rates obtained either with the independent estimation or with the constrained optimization methods are similar. Adding the constraints to the maximization of the likelihood does not significantly change the optimum. However, the PCLM method is more different in the formulation of the problem since no coefficients are associated with mortality in state H_2.

As a consequence, the maximum likelihood estimators of this method result in different mortality rates in LTC compared to those produced by the other two methods. As shown in Figure 3.4.4, the mortality function estimated with the PCLM method does not seem to depend greatly on the duration. Moreover, the pattern of this estimated mortality law shows that the mortality of disabled policyholders has a smile shape on the age dimension. For a fixed duration, the mortality is high at young and old ages and seems to reach the minimum at approximately 80 years old. This phenomenon can be explained by the prevalence of the pathologies affecting the disabled policyholders. This prevalence depends greatly on the attained age. Pathologies affecting young disabled policyholders are mostly diseases affecting mortality, such as cancer. In contrast, the most represented pathologies at approximately 80 years old are Alzheimer's disease and dementia. These pathologies are known to have a limited impact on mortality. Therefore, as explained in Biessy (2016), cancer has a high contribution to mortality in LTC at young ages, especially for low durations.

Figure 3.4.5 compares the estimated mortality laws of disabled policyholders at 85 years old as a function of duration. For all methods, the order of the penalty is fixed to 2 on

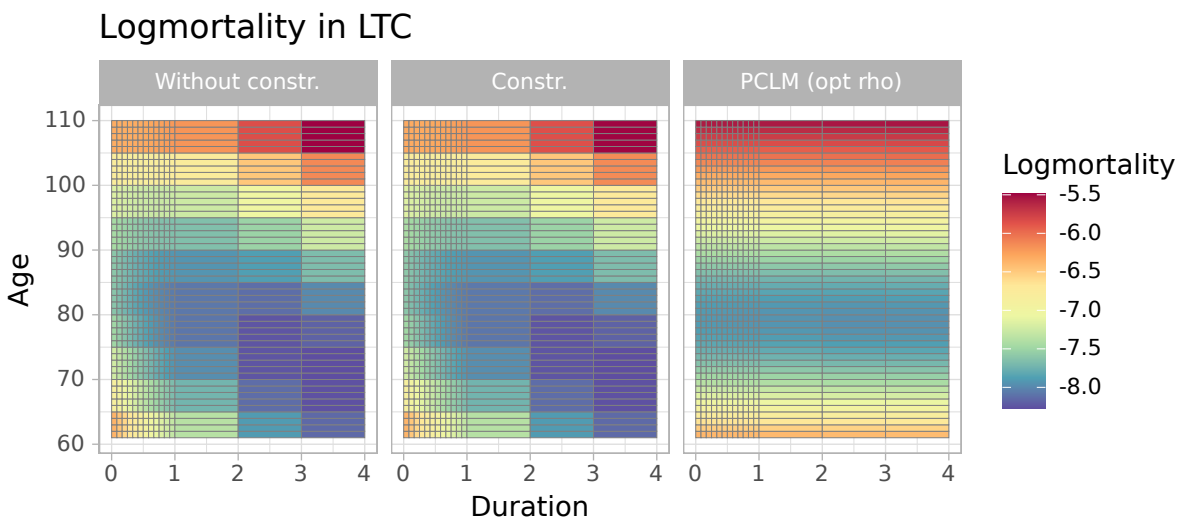


Figure 3.4.4 – Mortality in Long-Term Care (state F)

the age dimension and 1 on the duration dimension to prevent unreasonable divergence of the mortality with increasing duration. For a fixed attained age, the mortality with and without constraints has a smile shape, with mortality decreasing at low durations and increasing after the first year. In contrast, the mortality of disabled policyholders aged 85 years old, estimated with the PCLM method, is essentially the same for all durations. The PCLM method gives more weight to the smoothing penalty ρ on the duration shape compared to that of the 2 other methods. Since this penalty tries to minimize the variations in the duration dimension, the mortality is more likely to be constant in this dimension. The confidence interval obtained with the PCLM method is narrower than those estimated with the two other methods.

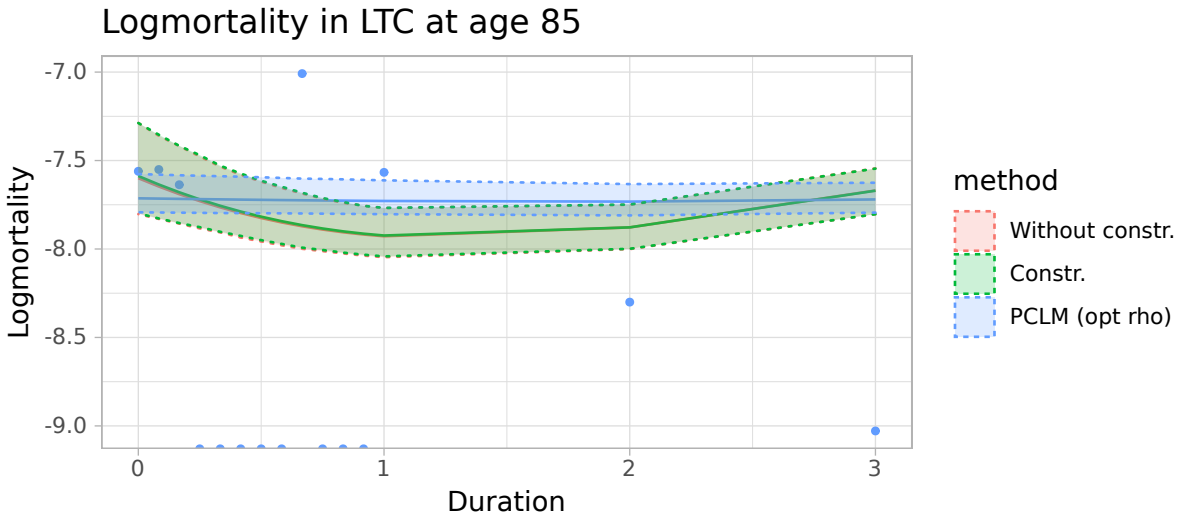


Figure 3.4.5 – Mortality in Long-Term Care (state F) at age 85

One might want to evaluate the evolution of mortality with duration for a given entry

age. Figure 3.4.6 shows the mortality rates of disabled policyholders losing their autonomy at 70 years old. For all methods, mortality decreased with time during the 3 first years. However, the decrease rate is smaller with the PCLM method. While the PCLM method gives the lowest mortality rates during the first three months, the estimated mortality for higher durations is higher than those with the two other methods.

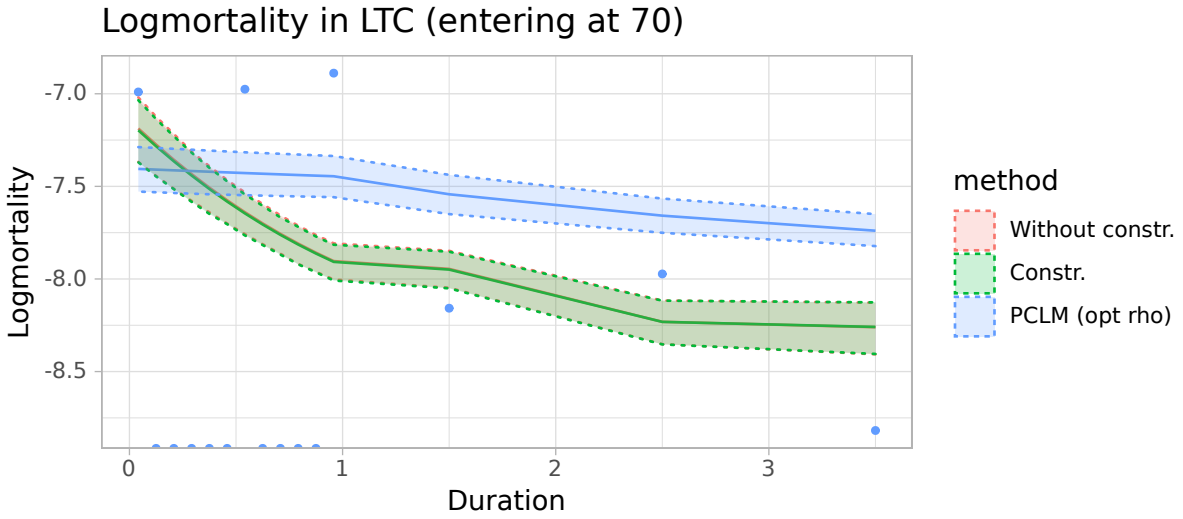


Figure 3.4.6 – Mortality of a policyholder losing autonomy at age 70

Figure 3.4.7 shows the mortality intensities during the first month following the loss of autonomy. As anticipated with Figure 3.4.4, the mortality function has a smile shape on the age dimension. The confidence interval obtained with the PCLM method is narrower than those associated with the two other methods.

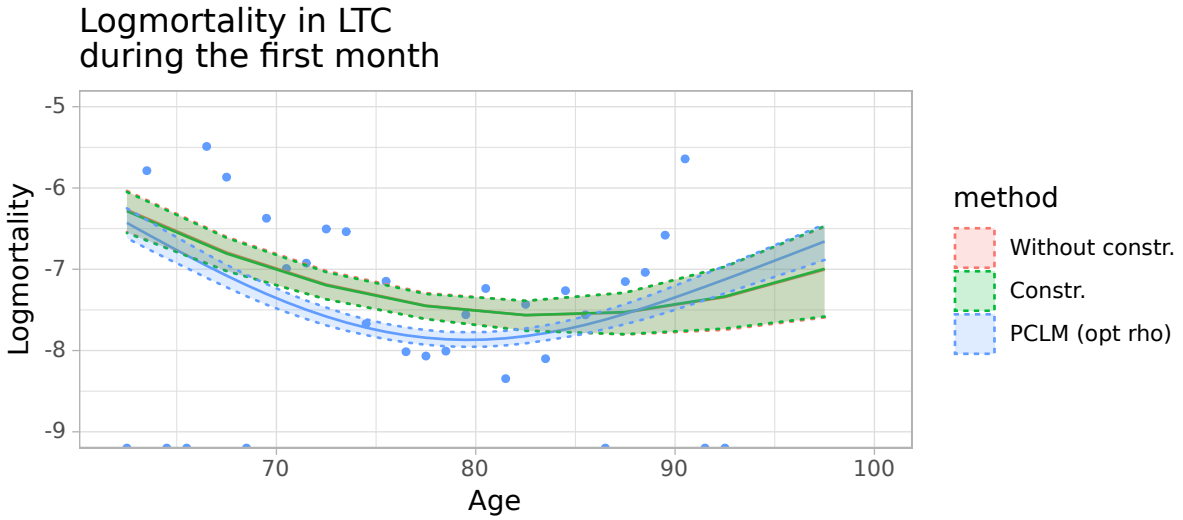


Figure 3.4.7 – Mortality of disabled policyholders during the first month following the loss of autonomy

The deviance residuals in the healthy (H_1) and disabled (F) states are plotted in Figures 3.4.8 and 3.4.9, respectively. We use these graphics to analyse if the residuals are within the

interval $[-2; 2]$ and if there exists a trend in the residuals (lower residuals at low ages and higher residuals at old ages, for example). A residual within the interval $[-2; 2]$ indicates that the model does not significantly underestimate or overestimate the associated count of death. The existence of a trend may indicate that the model does not well capture the effect of one of the dimensions (age and/or duration) on the mortality rates. In the healthy state, all models seem to perform well. In the disabled state, models with and without constraints seem to underestimate mortality after the first year. In contrast, the PCLM model overestimates mortality after the first year. The residuals during the first year seem homogeneous and quite similar among the 3 models.

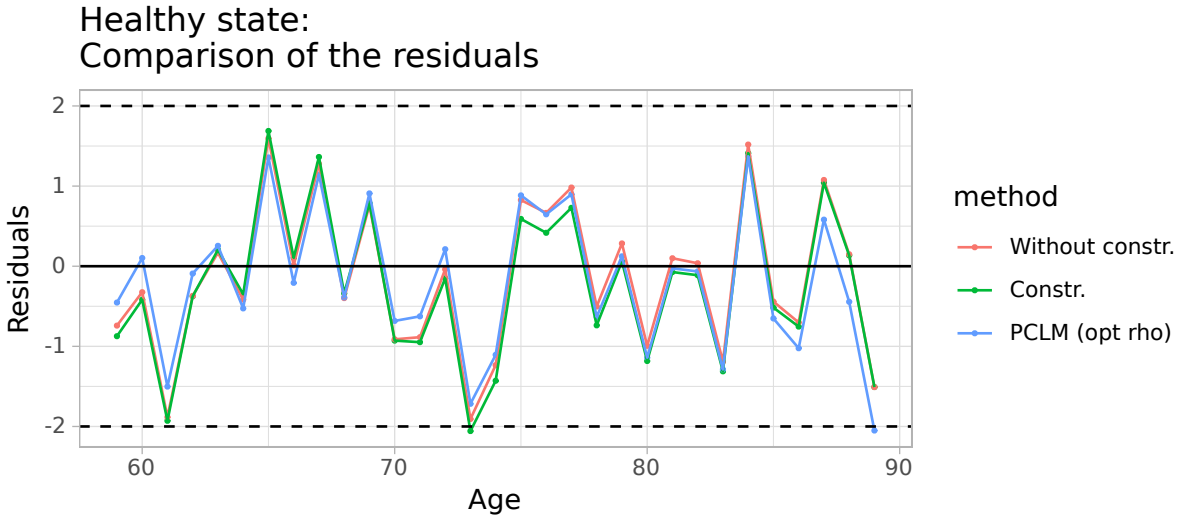


Figure 3.4.8 – Deviance residuals in the healthy state H_1

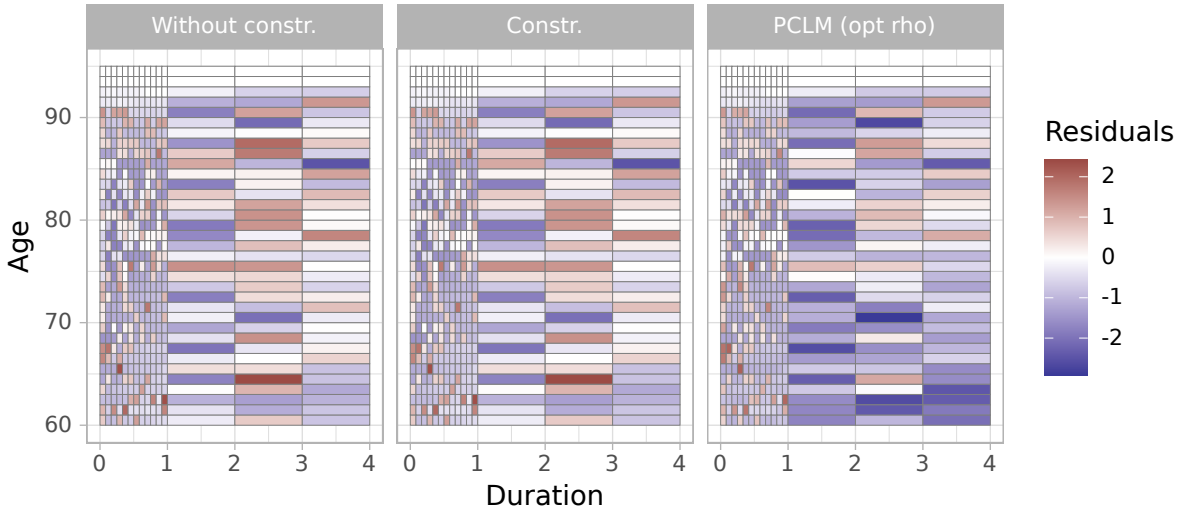


Figure 3.4.9 – Deviance residuals in the disabled state F

Figure 3.4.10 displays the first month of disability. The deviance residuals are within the specified interval $[-2; 2]$ and seem quite homogeneous among all ages.

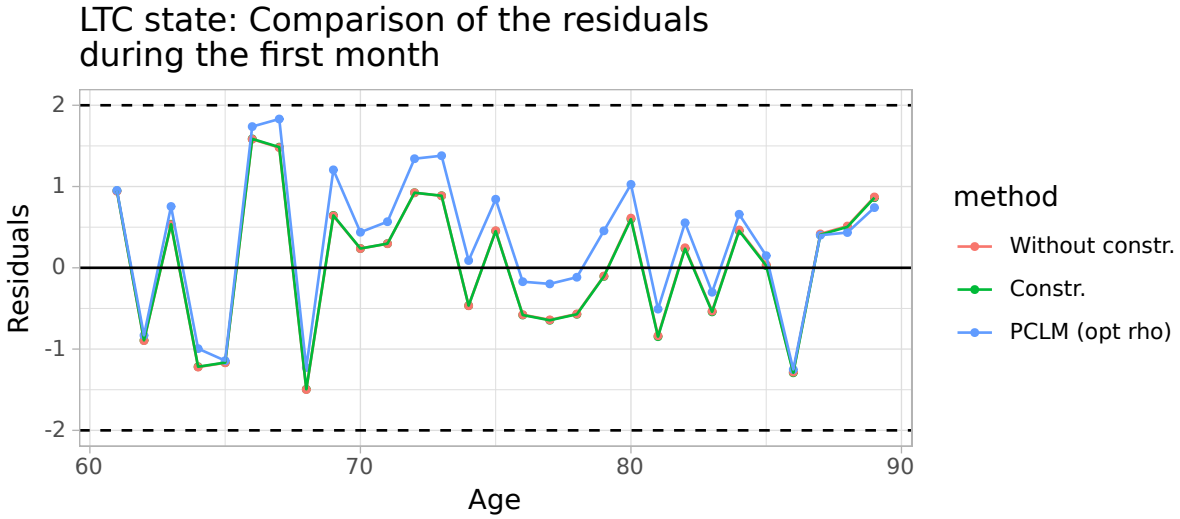


Figure 3.4.10 – Deviance residuals in the disabled state F during the first month following the loss of autonomy

3.4.3.3 Model Performance

The predictive performances of the three methods are evaluated with a 5-Fold cross-validation. Other resample methods to compare mortality models can be found in Atance et al. (2020). The following metrics are analysed:

- the root mean square error (RMSE)

$$\sqrt{\frac{1}{N^G}\sum_{k=1}^{N^G}(D_k^G - \hat{D}_k^G)^2}, G \in \{H_1, F\},$$

  where $N^{H_1} = M^H$, $N^F = M_x^F(M_t^\alpha + M_t^\beta)$, and $\hat{D}_k^G$ denotes the k^{th} term of the vector of expected counts of deaths in group G,
- the root mean percentage error (RMPE)

$$\sqrt{\frac{1}{N^G}\sum_{k=1}^{N^G}\left(\frac{D_k^G - \hat{D}_k^G}{\hat{D}_k^G}\right)^2}, G \in \{H_1, F\},$$

- the χ^2 statistic

$$\frac{1}{N^G}\sum_{k=1}^{N^G}\left(\frac{D_k^G - \hat{D}_k^G}{\sqrt{\hat{D}_k^G}}\right)^2, G \in \{H_1, F\},$$

- the deviance

$$\sum_{k=1}^{N^G} 2\left(D_k^G \log\left(\frac{D_k^G}{\hat{D}_k^G}\right) - (D_k^G - \hat{D}_k^G)\right).$$

Model performance on the healthy state is evaluated on the age range $[60, 90]$, and model performance on the disabled state is evaluated on the first year of disability for ages between 65 and 90 years old. When the metrics of the two methods presented in this paper are better than those of the method without constraints (independent estimation of each biometric law), then the metric is displayed in bold. The best value for each metric is written in red. The results are shown in Table 3.4.1 and Table 3.4.2.

Table 3.4.1 shows that adding a constraint to the P-splines smoothing method improves the Root Mean Square Error (RMSE), the Root Mean Percentage Error (RMPE), the deviance and the χ^2 statistics of mortality in the healthy state. This method also improves the metrics for the mortality of disabled policyholders, as shown in Table 3.4.2.

The PCLM method mostly improves the predictive performance of the mortality of disabled policyholders as shown in Table 3.4.2. The root mean percentage error and χ^2 are the lowest for this method compared to those of the two others. This method also improves the Root Mean Percentage Error (RMPE) on the prediction of the mortality of autonomous policyholders (Table 3.4.1).

Method	RMSE	RMPE	χ^2	Deviance
Constr.	**2.19**	**0.60**	**173.32**	**204.82**
Without Constr.	2.19	0.63	180.95	205.96
PCLM	2.21	**0.57**	**176.70**	215.38

Table 3.4.1 – Cross validation: mortality in the healthy state

Method	RMSE	RMPE	χ^2	Deviance
Constr.	0.30	**5.76**	**1,338.40**	**509.40**
Without Constr.	0.30	5.80	1,344.35	509.52
PCLM	0.30	**4.69**	**1,226.24**	514.33

Table 3.4.2 – Cross validation: mortality in disabled state

Additional plots of confidence intervals and analyses of Poisson residuals for each state H_1, H_2 and F are available in Appendix 3.F.

3.5 Discussion

In this paper, we present two methods to estimate transition rates by using combined experience data of two long-term care portfolios with differing disability definitions. In this paper, we focus on the case when one of the definitions is included in the other. In this situation, the healthy state (denoted H_2) of one of the products/definitions is a mixture of

the healthy state (H_1) and the disabled state (T_1) of the other product, as described in Section 3.2. The disabled state T_1 is divided in two distinct states, namely F_α and F_β, such that $F_\alpha = T_1 \backslash T_2$. F_β is the state containing policyholders disabled according to both definitions.

We focus on the estimation of mortality in healthy and disabled states by using combined experiences from both portfolios. Mortality in the healthy state is considered a function of age only; whereas, mortality in disabled states is assumed to be a function of age and duration since the loss of autonomy. In fact, mortality in the disabled state often depends greatly on the duration, with a higher mortality in the first months following entry into the disabled state.

Assuming that the population of both portfolios is homogeneous, the mortality of this healthy state H_2 is a mixture of the mortality in H_1 and F_α. The three mortality laws are then linked. To avoid overfitting, the two methods introduced in this paper rely on the P-splines smoothing method embedded in the Poisson Generalized Linear Model framework as presented in Section 3.3.1. With the Poisson assumption justified in Section 3.2, counts of deaths are assumed to have a Poisson distribution of parameters proportional to the central exposure and the mortality intensities.

The presented methods are based on the idea of maximizing the likelihood. The first approach described in Section 3.3.2 consists of maximizing the Poisson likelihood subject to multiple constraints linking the biometric functions of the two portfolios. In this method, the mortality laws (mortality in H_1, H_2, F_α and F_β) are assumed to be expressed as exponential of the combination of basis-splines. A set of spline coefficients must be estimated for each state. The second approach presented in Section 3.3.3 uses the Penalized Composite Link Model introduced by Eilers (2007). This method uses the fact that the expectancy of the counts of deaths in each state (H_1, H_2, F_α and F_β) can be expressed as a linear combination of the mortality in H_1, F_α and F_β. Therefore, no coefficients are associated with mortality in H_2. This method allows us to reduce the number of estimated coefficients compared to that of the first method.

The presented methods are then applied to solve the problem when having two portfolios with and without a deferred period. In this case, being disabled for the product with a deferred period implies having the disabled status for the product without a deferred period. This situation is then a special case of the problem solved in this paper, since one of the disability definitions is included in the other. To meet the assumption of homogeneity between the two portfolios, a single French long-term care portfolio is randomly divided into two. One of the data sets is then modified to include a fictitious deferred period. The two methods are then evaluated on these portfolios and compared with the separate modelling approach to show the benefits of combining information of the two portfolios

despite the difference in the disability definitions.

We show that combining information improves the predictive performance. First, adding a constraint to the maximization of the P-splines likelihood helps to reduce the size of the confidence interval of the mortality in the healthy state, without having a large impact on the mortality estimated independently on one single portfolio. However, this first method has almost no impact on the estimated mortality in the disabled state and its associated confidence interval. The results of the cross-validation show that this method has better predictive performance than that of the separate modelling approach. The PCLM method helps reduce the confidence interval of the estimated mortality in the disabled state, but it has an impact on the estimated mortality compared to that of the estimation without combining information of the portfolios. The cross-validation shows that this method gives better predictive performance on the RMPE (Root Mean Percentage Error) and χ^2 metrics. However, this method slightly increases the residual deviance of the model compared to that in the separate modelling approach.

The methods introduced in this paper can be extended to other cases where portfolios have different disability definitions. For example, some long-term care products offer the possibility for the policyholder to be covered for several levels of loss of autonomy, namely "Mild" and "Severe" disability. Policyholders only covered for severe disability are considered "autonomous" if they are only mildly disabled. The long-term care portfolio of the insurer is composed of policyholders with different disability definitions. This portfolio can therefore be divided into two different portfolios, one for each definition. The first covers both mild and severe disability, whereas the second covers only severe disability. In this situation, being disabled with the second definition implies being disabled for the first definition. It is therefore a special case of the problem of multiple definitions studied in this paper. The two methods that we introduced can be used to estimate simultaneously the biometric laws of both definitions by combining information of all policyholders, instead of separating policyholders covered for the mild disability from the other policyholders.

The methods introduced in this paper assume that the counts of deaths have a Poisson distribution. One of the properties of the Poisson distribution is that its mean equals variance. However, overdispersion is often observed in the context of mortality modelling. Accordingly, the Poisson assumption may be too restrictive. In future studies, a dispersion parameter can be introduced to improve the model. The models assume homogeneity of policyholders in the two portfolios. This assumption can be difficult to satisfy because of antiselection. In the example of the deferred period, the population of policyholders choosing a long-term care contract without a deferred period is likely different from the population of policyholders underwriting a contract with a deferred period. The metrics used in this paper to compare the models do not take into account the economic impact of an error in the prediction of the count of death. In fact, an error of estimation of

mortality rates does not have the same economic impact depending on the age. It would be interesting to explore metrics considering weights depending on the financial impacts. Finally, in future research the methods developed in this paper may be generalized to parametric forms of mortality.

Appendices

Appendix 3.A Proof of the total likelihood of the combined observations

Let j denote an individual. Each individual is observed with only one product (1 or 2). We introduce:

- $\overline{x^S}(j)$, the minimum age at which we observe individual j (age at underwriting, or age at the beginning of the observation period),
- $\overline{x^H}(j)$, the age of end of observation in healthy state (equal to the age of death, age of loss of autonomy, or age at censoring),
- $\overline{x^T}(j)$, the age of end of observation in disabled state T_k where k denotes the definition with which we observe individual j (age of death, or age at censoring if individual j is in state T_k at the end of observation period),
- $\overline{x^{F_g}}(j)$, the age of end of observation in disabled state F_g where $g \in \{\alpha, \beta\}$ (age of death, age of transition from F_α to F_β or age at censoring if individual j is in state F_g at the end of observation period),
- $c^H(j)$, the cause of exit of healthy state for individual j such that:
  - $c^H(j) = 0$ if individual j is still in the healthy state at the end of the observation (right censoring),
  - $c^H(j) = 1$ if individual j dies in healthy state,
  - $c^H(j) = 2$ if the cause of exit of the healthy state is the loss of autonomy (entry in state T_k if the health status of individual j is observed with definition k),
- $c^T(j)$, the cause of exit of disabled state ($T_k, k \in \{1, 2\}$) for individual j such that:
  - $c^T(j) = 0$ if individual j is in state T_k at the end of the observation (right censoring),
  - $c^T(j) = 1$ if individual j dies in the disabled state (T_1 or T_2 depending on the definition with which we observe individual j)
- $c^{F_g}(j)$, the cause of exit of disabled state ($F_g, g \in \{\alpha, \beta\}$) for individual j such that:
  - $c^{F_g}(j) = 1$ if individual j dies in disabled state F_g,
  - $c^{F_\alpha}(j) = 2$ if an individual in state F_α goes in state F_β,
  - $c^{F_g}(j) = 0$ otherwise.

If individual j does not lose autonomy during the observation period, then $\overline{x^T}(j) = \overline{x^H}(j)$. If individual j is observed with definition 2, then $\overline{x^{F_\alpha}}(j) = \overline{x^H}(j)$. If individual j dies in state F_α then $\overline{x^{F_\beta}}(j) = \overline{x^{F_\alpha}}(j) = \overline{x^T}(j)$.

The likelihood associated with one individual observed with definition k is given by

$$\begin{aligned}
L_j = &\exp\left(-\int_{\overline{x^S}(j)}^{\overline{x^H}(j)}\left(i^k(x)+\mu^{H_k}(x)\right)dx\right)\left(i^k\left(\overline{x^H}(j)\right)\right)^{\mathbb{1}_{\{c^H(j)=2\}}}\left(\mu^{H_k}\left(\overline{x^H}(j)\right)\right)^{\mathbb{1}_{\{c^H(j)=1\}}}\\
&\exp\left(-\int_0^{\overline{x^{F_\alpha}}(j)-\overline{x^H}(j)}\left(\mu^{F_\alpha}\left(\overline{x^H}(j)+t,t\right)+\omega^1\left(\overline{x^H}(j)+t,t\right)\right)dt\right)\\
&\left(\mu^{F_\alpha}\left(\overline{x^{F_\alpha}}(j),\overline{x^{F_\alpha}}(j)-\overline{x^H}(j)\right)\right)^{\mathbb{1}_{\{c^{F_\alpha}(j)=1\}}}\left(\omega^1\left(\overline{x^{F_\alpha}}(j),\overline{x^{F_\alpha}}(j)-\overline{x^H}(j)\right)\right)^{\mathbb{1}_{\{c^{F_\alpha}(j)=2\}}\mathbb{1}_{\{k=1\}}}\\
&\exp\left(-\int_0^{\overline{x^{F_\beta}}(j)-\overline{x^{F_\alpha}}(j)}\mu^{F_\beta}\left(\overline{x^{F_\alpha}}(j)+t,t\right)dt\right)\left(\mu^{F_\beta}\left(\overline{x^{F_\beta}}(j),\overline{x^{F_\beta}}(j)-\overline{x^{F_\alpha}}(j)\right)\right)^{\mathbb{1}_{\{c^{F_\beta}(j)=1\}}}.
\end{aligned} \tag{3.37}$$

As the transition intensities from F_α to F_β are known $\left(\omega^1\left(\overline{x^{F_\alpha}}(j),\overline{x^{F_\alpha}}(j)-\overline{x^H}(j)\right)\right)^{\mathbb{1}_{\{c^{F_\alpha}(j)=2\}}\mathbb{1}_{\{k=1\}}}$ and $\omega^1\left(\overline{x^H}(j)+t,t\right)$ are considered as constants for the maximization of the likelihood given by Equation 3.37.

Therefore, the total likelihood for all individuals from products 1 and 2 is given by

$$\begin{aligned}
L_{tot}\propto &\prod_{j=1}^{n_1}\exp\left(-\int_{\overline{x^S}(j)}^{\overline{x^H}(j)}\left(i^1(x)+\mu^{H_1}(x)\right)dx\right)\left(i^1\left(\overline{x^H}(j)\right)\right)^{\mathbb{1}_{\{c^H(j)=2\}}}\left(\mu^{H_1}\left(\overline{x^H}(j)\right)\right)^{\mathbb{1}_{\{c^H(j)=1\}}}\\
&\exp\left(-\int_0^{\overline{x^{F_\alpha}}(j)-\overline{x^H}(j)}\mu^{F_\alpha}\left(\overline{x^H}(j)+t,t\right)dt\right)\left(\mu^{F_\alpha}\left(\overline{x^{F_\alpha}}(j),\overline{x^{F_\alpha}}(j)-\overline{x^H}(j)\right)\right)^{\mathbb{1}_{\{c^{F_\alpha}(j)=1\}}}\\
&\exp\left(-\int_0^{\overline{x^{F_\beta}}(j)-\overline{x^{F_\alpha}}(j)}\mu^{F_\beta}\left(\overline{x^{F_\alpha}}(j)+t,t\right)dt\right)\left(\mu^{F_\beta}\left(\overline{x^{F_\beta}}(j),\overline{x^{F_\beta}}(j)-\overline{x^{F_\alpha}}(j)\right)\right)^{\mathbb{1}_{\{c^{F_\beta}(j)=1\}}}\times\\
&\prod_{j=1}^{n_2}\exp\left(-\int_{\overline{x^S}(j)}^{\overline{x^H}(j)}\left(i^2(x)+\mu^{H_2}(x)\right)dx\right)\left(i^2\left(\overline{x^H}(j)\right)\right)^{\mathbb{1}_{\{c^H(j)=2\}}}\left(\mu^{H_2}\left(\overline{x^H}(j)\right)\right)^{\mathbb{1}_{\{c^H(j)=1\}}}\\
&\exp\left(-\int_0^{\overline{x^{F_\beta}}(j)-\overline{x^H}(j)}\mu^{F_\beta}\left(\overline{x^H}(j)+t,t\right)dt\right)\left(\mu^{F_\beta}\left(\overline{x^{F_\beta}}(j),\overline{x^{F_\beta}}(j)-\overline{x^H}(j)\right)\right)^{\mathbb{1}_{\{c^{F_\beta}(j)=1\}}},
\end{aligned} \tag{3.38}$$

where $n_k, k\in\{1,2\}$ denotes the number of observed policyholders from product k.

Let $e_j^{H_k}(p)$ denote the central exposure to risk of individual j in state H_k between integer ages x_p^H and x_{p+1}^H. Let $e_j^g(p,q)$ denote the exposure to risk of individual j in state F_g between integer ages x_p^T and x_{p+1}^T and durations t_q^g and t_{q+1}^g. Let $e^{H_k}(p)=\sum_{j=1}^{n_k}e_j^{H_k}(p)$ be the sum of the central exposures in state H_k of all policyholders from product k. Let $e^\alpha(p,q)=\sum_{j=1}^{n_1}e_j^\alpha(p,q)$. F_β is the only state observed in both products. Therefore, let $e^{\beta,k}(p,q)=\sum_{j=1}^{n_k}e_j^\beta(p,q)$ denote the sum of the central exposures in state F_β for product

k. We denote $e^{\beta}(p,q) = e^{\beta,1}(p,q) + e^{\beta,2}(p,q)$ as the sum of the central exposures of both products.

Let $N^1(p)$ denote the number of transitions from H_1 to T_1 between x_p^H and x_{p+1}^H. Let $N^2(p)$ denote the number of transitions from H_2 to T_2 between x_p^H and x_{p+1}^H.

Let $D^{H_1}(p)$ denote the number of deaths from the healthy state between x_p^H and x_{p+1}^H in the first type of insurance. Let $D^{H_2}(p)$ denote the number of deaths from the healthy state between x_p^H and x_{p+1}^H in the second type of insurance. Let $D^{\alpha}(p,q)$ denote the number of deaths from the F_α disabled state with age at death between x_p^T and x_{p+1}^T and duration between t_q^α and t_{q+1}^α. For the central exposure, let $D^{\beta,k}(p,q)$ denote the number of deaths from the F_β disabled state with age at death between x_p^T and x_{p+1}^T and duration between t_q^β and t_{q+1}^β for product k. We denote $D^{\beta}(p,q) = D^{\beta,1}(p,q) + D^{\beta,2}(p,q)$.

As intensities $\mu^{H_k}()$ and $i^k(), k \in \{1,2\}$ and $\mu^{F_g}(,), g \in \{\alpha,\beta\}$ are piecewise constants (Assumption 3.2.3 and Assumption 3.2.4), we denote

- $\mu_p^{H_k} = \mu^{H_k}(x_p^H), \forall p \in \{1,\dots,M^H\}$,
- $i_p^k = i^k(x_p^H), \forall p \in \{1,\dots,M^H\}$, and
- $\mu_{p,q}^{F_g} = \mu^{F_g}(x_p^F, t_q^g), \forall p \in \{1,\dots,M_x^F\}, q \in \{1,\dots,M_t^g\}$,

and Equation 3.38 becomes

$$\begin{aligned} L_{tot} \propto & \prod_{p=1}^{M^H} \exp\left(-\left(i_p^1 + \mu_p^{H_1}\right) e^{H_1}(p)\right) \left(i_p^1\right)^{N^1(p)} \left(\mu_p^{H_1}\right)^{D^{H_1}(p)} \\ & \prod_{p=1}^{M_x^F} \prod_{q=1}^{M_t^\alpha} \exp\left(-\mu_{p,q}^{F_\alpha} e^{\alpha}(p,q)\right) \left(\mu_{p,q}^{F_\alpha}\right)^{D^{\alpha}(p,q)} \times \\ & \prod_{p=1}^{M^H} \exp\left(-\left(i_p^2 + \mu_p^{H_2}\right) e^{H_2}(p)\right) \left(i_p^1\right)^{N^2(p)} \left(\mu_p^{H_2}\right)^{D^{H_2}(p)} \\ & \prod_{p=1}^{M_x^F} \prod_{q=1}^{M_t^\beta} \exp\left(-\mu_{p,q}^{F_\beta} e^{\beta}(p,q)\right) \left(\mu_{p,q}^{F_\beta}\right)^{D^{\beta}(p,q)}. \end{aligned} \tag{3.39}$$

The last term of Equation 3.39 represents the likelihood of all observations from products 1 and 2 in state F_β.

The log likelihood function of the combined observations from products 1 and 2 is given by

$$
\begin{aligned}
l_{tot} = & \sum_{p=1}^{M^H} \left(- \left(i_p^1 + \mu_p^{H_1} \right) e^{H_1}(p) \right) + N^1(p) \log \left(i_p^1 \right) + D^{H_1}(p) \log \left(\mu_p^{H_1} \right) + \\
& \sum_{p=1}^{M_x^F} \sum_{q=1}^{M_t^\alpha} \left(-\mu_{p,q}^{F_\alpha} e^\alpha(p,q) \right) + D^\alpha(p,q) \log \left(\mu_{p,q}^{F_\alpha} \right) + \\
& \sum_{p=1}^{M^H} \left(- \left(i_p^2 + \mu_p^{H_2} \right) e^{H_2}(p) \right) + N^2(p) \log \left(i_p^1 \right) + D^{H_2}(p) \log \left(\mu_p^{H_2} \right) + \\
& \sum_{p=1}^{M_x^F} \sum_{q=1}^{M_t^\beta} \left(-\mu_{p,q}^{F_\beta} e^\beta(p,q) \right) + D^\beta(p,q) \log \left(\mu_{p,q}^{F_\beta} \right) + cst. \qquad (3.40)
\end{aligned}
$$

where *cst* denotes a constant.

Appendix 3.B Proof of the constraint

Let us prove the constraint given by Equation (3.9) in Section 3.3.2.

$$\mu_p^{H_2} = \mu^{H_2}(x_p^H) = \lim_{h\to 0} \frac{\mathbb{P}(X^{(2)}_{x_p^H+h} = \text{Death}|X^{(2)}_{x_p^H} = H_2)}{h}. \tag{3.41}$$

$$\begin{aligned}
\mathbb{P}(X^{(2)}_{x+h} = \text{Death}|X^{(2)}_x = H_2) =& \mathbb{P}(X^{(1)}_{x+h} = \text{Death}|X^{(1)}_x = H_1 \cup X^{(1)}_x = F_\alpha)\\
=& \frac{\mathbb{P}(X^{(1)}_{x+h} = \text{Death}, \{X^{(1)}_x = H_1 \cup X^{(1)}_x = F_\alpha\})}{\mathbb{P}(X^{(1)}_x = H_1 \cup X^{(1)}_x = F_\alpha)}\\
=& \frac{\mathbb{P}(X^{(1)}_{x+h} = \text{Death}, X^{(1)}_x = H_1) + \mathbb{P}(X^{(1)}_{x+h} = \text{Death}, X^{(1)}_x = F_\alpha)}{\mathbb{P}(X^{(1)}_x = H_1 \cup X^{(1)}_x = F_\alpha)}\\
=& \frac{\mathbb{P}(X^{(1)}_{x+h} = \text{Death}|X^{(1)}_x = H_1)\mathbb{P}(X^{(1)}_x = H_1)}{\mathbb{P}(X^{(1)}_x = H_1 \cup X^{(1)}_x = F_\alpha)} +\\
& \sum_{q=1}^{M_t^\alpha} \frac{\mathbb{P}(X^{(1)}_{x+h} = \text{Death}|\{X^{(1)}_x = F_\alpha, X^{(1)}_{(x-t_q^\alpha)} = F_\alpha, X^{(1)}_{(x-t_{q+1}^\alpha)} = H_1\})}{\mathbb{P}(X^{(1)}_x = H_1 \cup X^{(1)}_x = F_\alpha)} \times\\
& \mathbb{P}(X^{(1)}_x = F_\alpha, X^{(1)}_{(x-t_q^\alpha)} = F_\alpha, X^{(1)}_{(x-t_{q+1}^\alpha)} = H_1)\\
=& \frac{\mathbb{P}(X^{(1)}_{x+h} = \text{Death}|X^{(1)}_x = H_1)\mathbb{P}(X^{(1)}_x = H_1)}{\mathbb{P}(X^{(1)}_x = H_1 \cup X^{(1)}_x = F_\alpha)} +\\
& \sum_{q=1}^{M_t^\alpha} \frac{\mathbb{P}(X^{(1)}_{x+h} = \text{Death}|\{X^{(1)}_x = F_\alpha, X^{(1)}_{(x-t_q^\alpha)} = F_\alpha, X^{(1)}_{(x-t_q^\alpha)^-} = H_1\})}{\mathbb{P}(X^{(1)}_x = H_1 \cup X^{(1)}_x = F_\alpha)} \times\\
& \mathbb{P}(X^{(1)}_x = F_\alpha, X^{(1)}_{(x-t_q^\alpha)} = F_\alpha, X^{(1)}_{(x-t_{q+1}^\alpha)} = H_1).
\end{aligned}$$

Therefore, Equation (3.41) becomes

$$\begin{aligned}
\mu_p^{H_2} =& \lim_{h\to 0} \frac{\mathbb{P}(X^{(2)}_{x_p^H+h} = \text{Death}|X^{(2)}_{x_p^H} = H_2)}{h}\\
=& \frac{\mathbb{P}(X^{(1)}_x = H_1)}{\mathbb{P}(X^{(1)}_x = H_1 \cup X^{(1)}_x = F_\alpha)} \lim_{h\to 0} \frac{\mathbb{P}(X^{(1)}_{x_p^H+h} = \text{Death}|X^{(1)}_{x_p^H} = H_1)}{h} +\\
& \sum_{q=1}^{M_t^\alpha} \frac{\mathbb{P}(X^{(1)}_x = F_\alpha, X^{(1)}_{(x-t_q^\alpha)} = F_\alpha, X^{(1)}_{(x-t_{q+1}^\alpha)} = H_1)}{\mathbb{P}(X^{(1)}_x = H_1 \cup X^{(1)}_x = F_\alpha)} \times\\
& \lim_{h\to 0} \frac{\mathbb{P}(X^{(1)}_{x+h} = \text{Death}|\{X^{(1)}_x = F_\alpha, X^{(1)}_{(x-t_q^\alpha)} = F_\alpha, X^{(1)}_{(x-t_q^\alpha)^-} = H_1\})}{h}\\
=& \frac{\mathbb{P}(X^{(1)}_x = H_1)}{\mathbb{P}(X^{(1)}_x = H_1 \cup X^{(1)}_x = F_\alpha)} \mu_p^{H_1} + \sum_{q=1}^{M_t^\alpha} \frac{\mathbb{P}(X^{(1)}_x = F_\alpha, X^{(1)}_{(x-t_q^\alpha)} = F_\alpha, X^{(1)}_{(x-t_{q+1}^\alpha)} = H_1)}{\mathbb{P}(X^{(1)}_x = H_1 \cup X^{(1)}_x = F_\alpha)} \mu_{p,q}^{F_\alpha}.
\end{aligned} \tag{3.42}$$

Appendix 3.C Structure of matrix C

Let E^{H_1} and $E^{H_2} \in \mathcal{M}_{M^H}(\mathbb{R})$ be the diagonal matrices of the central exposures in states H_1 and H_2, respectively. Let $E^{\alpha} \in \mathcal{M}_{M_x^F, M_t^{\alpha}}(\mathbb{R})$ and $E^{\beta} \in \mathcal{M}_{M_x^F, M_t^{\beta}}(\mathbb{R})$ be the matrix of the sum of the central exposures in F_α and F_β, respectively. We denote ${}^{v}E^{g}_{.k} = \text{diag}(E^{g}_{.,k}) \in \mathcal{M}_{M_x^F}(\mathbb{R}), g \in \{\alpha, \beta\}$ as the diagonal matrix of the central exposures in state F_g for the k^{th} duration.

- $C = \begin{bmatrix} C_0 \\ C_1 \end{bmatrix} \in \mathcal{M}_{(M^H + M^H + (M_t^{\alpha} + M_t^{\beta}).M_x^F),(M^H + (M_t^{\alpha} + M_t^{\beta}).M_x^F)}(\mathbb{R})$,

- $C_0 = \begin{bmatrix} E^{H_1} & 0 & \dots & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & {}^{v}E^{\alpha}_{.1} & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & 0 & \ddots & 0 & \dots & \dots & \dots & \dots & 0 \\ 0 & \dots & 0 & \ddots & 0 & \dots & \dots & \dots & 0 \\ 0 & \dots & \dots & 0 & {}^{v}E^{\alpha}_{.M_t^{\alpha}} & 0 & \dots & \dots & 0 \\ 0 & \dots & \dots & \dots & 0 & {}^{v}E^{\beta}_{.1} & 0 & \dots & 0 \\ 0 & \dots & \dots & \dots & \dots & 0 & \ddots & 0 & 0 \\ 0 & \dots & \dots & \dots & \dots & \dots & \dots & \ddots & 0 \\ 0 & \dots & \dots & \dots & \dots & \dots & \dots & 0 & {}^{v}E^{\beta}_{.M_t^{\beta}} \end{bmatrix} \in \mathcal{M}_{(M^H + (M_t^{\alpha} + M_t^{\beta}).M_x^F)}(\mathbb{R})$,

- $C_1 = \begin{bmatrix} \dfrac{E^{H_2}.E^{H_1}}{E^{H_1} + \sum_{q=1}^{M_t^{\alpha}} {}^{v}E^{\alpha}_{.,q}} & \dfrac{E^{H_2}.{}^{v}E^{\alpha}_{.,1}}{E^{H_1} + \sum_{q=1}^{M_t^{\alpha}} {}^{v}E^{\alpha}_{.,q}} & \dots & \dfrac{E^{H_2}.{}^{v}E^{\alpha}_{.,M_t^{\alpha}}}{E^{H_1} + \sum_{q=1}^{M_t^{\alpha}} {}^{v}E^{\alpha}_{.,q}} & 0 & \dots & 0 \end{bmatrix}$

$$C_1 = (E^{H_1} + \sum_{q=1}^{M_t^{\alpha}} {}^{v}E^{\alpha}_{.,q})^{-1}.E^{H_2}.\begin{bmatrix} E^{H_1} & {}^{v}E^{\alpha}_{.,1} & \dots & {}^{v}E^{\alpha}_{.,M_t^{\alpha}} & 0 & \dots & 0 \end{bmatrix}.$$

The matrix $C_1 \in \mathcal{M}_{M^H,(M^H + (M_t^{\alpha} + M_t^{\beta}).M_x^F)}(\mathbb{R})$ is composed of "concatenation" of $M_t^{\beta} + 1$ diagonal matrices of dimension M^H, and the matrix is completed with columns of 0.

Appendix 3.D Proof of the mortality rates in state F in the application to the problem of the deferred period

Let $X_t^{(1)} \in \{H_1, F_\alpha, F_\beta, Death\}$ denote the health state at time t observed with LTC product 1, as modelled with the diagram in Figure 3.2.1a. Let $F = \{F_\alpha \cup F_\beta\}$
Let $X_t^{(2)} \in \{H_2, F_\beta, Death\}$ denote the health state at time t observed with LTC product 1, as modelled with the diagram in Figure 3.2.1a.

Let $\tau_k = \min\{u > \tau_{k-1} | X_u \neq X_{\tau_{k-1}}\}$ denote the sequence of jump times, with $\tau_0 = 0$. Let $Z_k = X_{\tau_k}$ be the sequence of visiting states.

$$
\begin{aligned}
\mu^F(x,t) &= \lim_{h\to 0} \frac{1}{h}\mathbb{P}\left(X_{x+h}^{(1)} = Death | X_x^{(1)} \in F, X_{x-t}^{(1)} \in F, X_{(x-t)^-}^{(1)} = H_1\right) \\
&= \lim_{h\to 0} \mathbb{P}\left(X_{x+h}^{(1)} = Death | X_x^{(1)} \in F, X_{\tau_1}^{(1)} \in F, \tau_1 = (x-t), X_0 = H_1\right) \\
&= \lim_{h\to 0} \frac{\mathbb{P}\left(X_{x+h}^{(1)} = Death, X_x^{(1)} \in F | X_{\tau_1}^{(1)} \in F, \tau_1 = (x-t), X_0 = H_1\right)}{\mathbb{P}\left(X_x^{(1)} \in F | X_{\tau_1}^{(1)} \in F, \tau_1 = (x-t), X_0 = H_1\right)} \\
&= \lim_{h\to 0} \frac{\mathbb{P}\left(X_{x+h}^{(1)} = Death, X_x^{(1)} \in F | X_{\tau_1}^{(1)} \in F, \tau_1 = (x-t), X_0 = H_1\right)}{\mathbb{P}\left(X_x^{(1)} \in F | X_{\tau_1}^{(1)} \in F, \tau_1 = (x-t), X_0 = H_1\right)} \\
&= \lim_{h\to 0} \frac{\mathbb{P}\left(X_{x+h}^{(1)} = Death, X_x^{(1)} = F_\alpha | X_{\tau_1}^{(1)} \in F, \tau_1 = (x-t), X_0 = H_1\right)}{\mathbb{P}\left(X_x^{(1)} = F_\alpha | X_{\tau_1}^{(1)} \in F, \tau_1 = (x-t), X_0 = H_1\right) + \mathbb{P}\left(X_x^{(1)} = F_\beta | X_{\tau_1}^{(1)} \in F, \tau_1 = (x-t), X_0 = H_1\right)} + \\
&\quad \lim_{h\to 0} \frac{\mathbb{P}\left(X_{x+h}^{(1)} = Death, X_x^{(1)} = F_\beta | X_{\tau_1}^{(1)} \in F, \tau_1 = (x-t), X_0 = H_1\right)}{\mathbb{P}\left(X_x^{(1)} = F_\alpha | X_{\tau_1}^{(1)} \in F, \tau_1 = (x-t), X_0 = H_1\right) + \mathbb{P}\left(X_x^{(1)} = F_\beta | X_{\tau_1}^{(1)} \in F, \tau_1 = (x-t), X_0 = H_1\right)}.
\end{aligned}
\tag{3.43}
$$

In the case of the deferred period, the transition rate from H_1 to F_β is equal to 0. Therefore,

$$\{X_{\tau_1}^{(1)} \in F, \tau_1 = (x-t), X_0 = H_1\} = \{X_{\tau_1}^{(1)} = F_\alpha, \tau_1 = (x-t), X_0 = H_1\}.$$

Additionally, as no return to a better health state is possible,

$$\mathbb{P}\left(X_x^{(1)} = F_\alpha | X_{\tau_1}^{(1)} = F_\alpha, \tau_1 = (x-t), X_0 = H_1\right) = \mathbb{P}\left(\tau_2 > x | X_{\tau_1}^{(1)} = F_\alpha, \tau_1 = (x-t), X_0 = H_1\right).$$

Moreover, from Equation 3.30

$$\omega^1(x,t) = \begin{cases} 0, & \text{if } t \neq fr/12 \\ +\infty, & \text{if } t = fr/12. \end{cases}$$

Therefore,

$$\begin{aligned} \mathbb{P}\left(\tau_2 > x | X^{(1)}_{\tau_1} = F_\alpha, \tau_1 = (x-t), X_0 = H_1\right) &= \exp\left(-\int_{x-t}^{x} \left(\mu^{F_\alpha}\left(u, u-(x-t)\right) + \omega^1\left(u, u-(x-t)\right)\right) \mathrm{d}u\right) \\ &= \exp\left(-\int_0^t \left(\mu^{F_\alpha}\left(x-t+y, y\right) + \omega^1\left(x-t+y, y\right)\right) \mathrm{d}y\right) \\ &= \begin{cases} 0 & \text{if } t \geqslant fr/12 \\ \exp\left(-\int_0^t \mu^{F_\alpha}\left(x-t+y, y\right) \mathrm{d}u\right) & \text{if } t < fr/12. \end{cases} \end{aligned} \tag{3.44}$$

Another consequence of the no return to a better health state is that

$$\begin{aligned} \mathbb{P}\left(X^{(1)}_x = F_\beta | X^{(1)}_{\tau_1} = F_\alpha, \tau_1 = (x-t), X_0 = H_1\right) &\leqslant \mathbb{P}\left(\tau_2 \leqslant x, X^{(1)}_{\tau_2} = F_\beta | X^{(1)}_{\tau_1} = F_\alpha, \tau_1 = (x-t), X_0 = H_1\right) \\ &= \int_0^t \mathbb{P}\left(\tau_2 \geqslant x-t+y | X^{(1)}_{\tau_1} = F_\alpha, \tau_1 = (x-t), X_0 = H_1\right) \times \\ &\qquad \delta_{fr/12}(y)\ \mathrm{d}y, \end{aligned}$$

$$\text{where } \delta_{fr/12}(y) = \begin{cases} +\infty & \text{if } t = fr/12 \\ 0 & \text{otherwise.} \end{cases}$$

Therefore,

$$\mathbb{P}\left(X^{(1)}_x = F_\beta | X^{(1)}_{\tau_1} = F_\alpha, \tau_1 = (x-t), X_0 = H_1\right) = 0, \forall t < fr/12. \tag{3.45}$$

Then,

- if $t < fr/12$, from Equation 3.43 and Equation 3.45

$$\begin{aligned} \mu^F(x,t) &= \lim_{h\to 0} \frac{1}{h} \frac{\mathbb{P}\left(X^{(1)}_{x+h} = Death, X^{(1)}_x = F_\alpha | X^{(1)}_{\tau_1} = F_\alpha, \tau_1 = (x-t), X_0 = H_1\right)}{\mathbb{P}\left(X^{(1)}_x = F_\alpha | X^{(1)}_{\tau_1} = F_\alpha, \tau_1 = (x-t), X_0 = H_1\right)} \\ &= \lim_{h\to 0} \frac{1}{h} \mathbb{P}\left(X^{(1)}_{x+h} = Death | X^{(1)}_x = F_\alpha, X^{(1)}_{\tau_1} = F_\alpha, \tau_1 = (x-t), X_0 = H_1\right) \\ &= \mu^{F_\alpha}(x,t), \end{aligned}$$

- if $t \geqslant fr/12$, from Equation 3.43 and Equation 3.44

$$\begin{aligned}\mu^F(x,t) &= \lim_{h\to 0}\frac{1}{h}\frac{\mathbb{P}\left(X^{(1)}_{x+h} = Death, X^{(1)}_x = F_\beta | X^{(1)}_{\tau_1} = F_\alpha, \tau_1 = (x-t), X_0 = H_1\right)}{\mathbb{P}\left(X^{(1)}_x = F_\beta | X^{(1)}_{\tau_1} = F_\alpha, \tau_1 = (x-t), X_0 = H_1\right)} \\ &= \lim_{h\to 0}\frac{1}{h}\mathbb{P}\left(X^{(1)}_{x+h} = Death | X^{(1)}_x = F_\beta, X^{(1)}_{\tau_1} = F_\alpha, \tau_1 = (x-t), X_0 = H_1\right).\end{aligned}$$

As the transition from F_α to F_β can occur only fr months after entry into F_α, and no return to a better health state is possible, then

$$\begin{aligned}&\mathbb{P}\left(X^{(1)}_{x+h} = Death | X^{(1)}_x = F_\beta, X^{(1)}_{\tau_1} = F_\alpha, \tau_1 = (x-t), X_0 = H_1\right) = \\ &\mathbb{P}\left(X^{(1)}_{x+h} = Death | X^{(1)}_x = F_\beta, X^{(1)}_{\tau_2} = F_\beta, \tau_2 = (x-t+fr/12), X^{(1)}_{\tau_1} = F_\alpha, \tau_1 = (x-t), X_0 = H_1\right).\end{aligned}$$

Therefore,

$$\begin{aligned}\mu^F(x,t) &= \lim_{h\to 0}\frac{1}{h}\mathbb{P}\left(X^{(1)}_{x+h} = Death | X^{(1)}_x = F_\beta, X^{(1)}_{\tau_2} = F_\beta, \tau_2 = (x-t+fr/12)\right) \\ &= \mu^{F_\beta}(x, t-fr/12).\end{aligned}$$

Appendix 3.E Details on the structure of the matrix B^F for the application to the deferred period

Let $B_t^\alpha \in \mathcal{M}_{M_t^\alpha, J_t^\alpha}$ be the matrix of the values of the splines at each subdivision point before the deferred period ($\left\{t_1^\alpha, \ldots, t_{M_t^\alpha+1}\right\}$). J_t^α denotes the number of splines from B_t^F having a non-null value on the interval $\left[t_1^\alpha; t_{M_t^\alpha}^\alpha\right]$.

Let $B_t^\beta \in \mathcal{M}_{M_t^\beta, J_t^\beta}$ be the matrix of the values of the splines at each subdivision point after the deferred period ($\left\{t_1^\beta + \frac{fr}{12}, \ldots, t_{M_t^\beta+1} + \frac{fr}{12}\right\}$). J_t^β denotes the number of splines from B_t^F having a non-null value on the interval $\left[t_1^\beta + \frac{fr}{12}; t_{M_t^\beta+1} + \frac{fr}{12}\right]$.

Let s denote the number of splines shared by states F_α and F_β ($s = 4$ in Figure 3.4.2). The first s columns of B_t^β are the continuous extension of the last s columns of B_t^α. Let $B_t^g(s), g \in \{\alpha, \beta\}$ denote the submatrix of B_t^g composed of only splines shared by F_α and F_β. Let $B_t^g(-s), g \in \{\alpha, \beta\}$ denote the submatrix of B_t^g composed of only splines not shared by F_α and F_β.

Then, $B_t^F = \left[\begin{array}{c|c|c} B_t^\alpha(-s) & B_t^\alpha(s) & 0 \\ 0 & B_t^\beta(s) & B_t^\beta(-s) \end{array}\right]$.

Therefore, from Section 3.3.1, the matrix B^F for the estimation of the mortality in two dimensions in LTC is given by

$$B^F = B_t^F \otimes B_x^F \tag{3.46}$$

$$\left[\begin{array}{c|c|c} B_t^\alpha(-s) \otimes B_x^F & B_t^\alpha(s) \otimes B_x^F & 0 \\ 0 & B_t^\beta(s) \otimes B_x^F & B_t^\beta(-s) \otimes B_x^F \end{array}\right]. \tag{3.47}$$

Appendix 3.F Additional plots

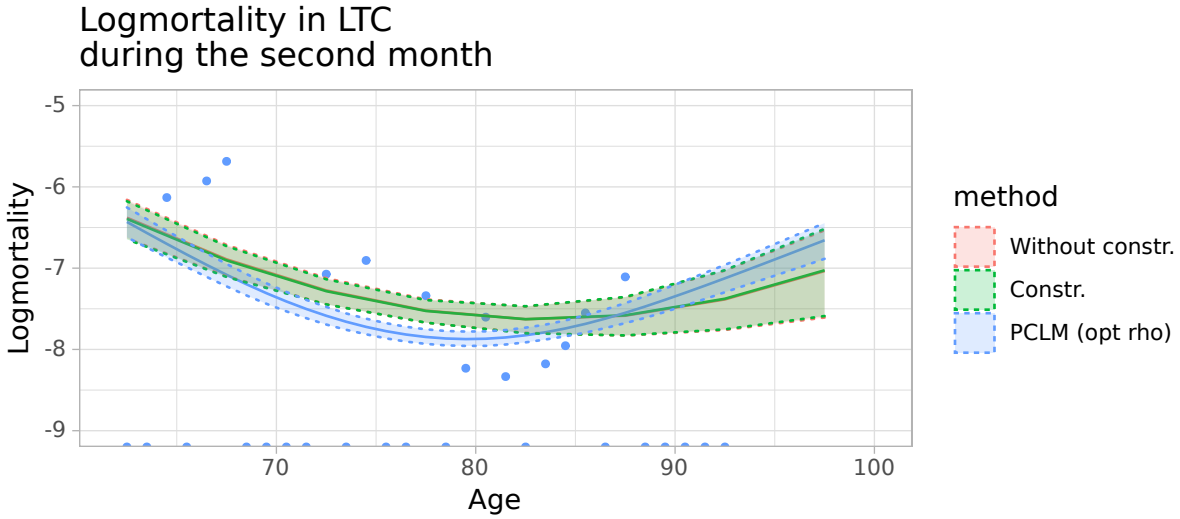


Figure 3.F.1 – Mortality of disabled policyholders during the second month following the loss of autonomy

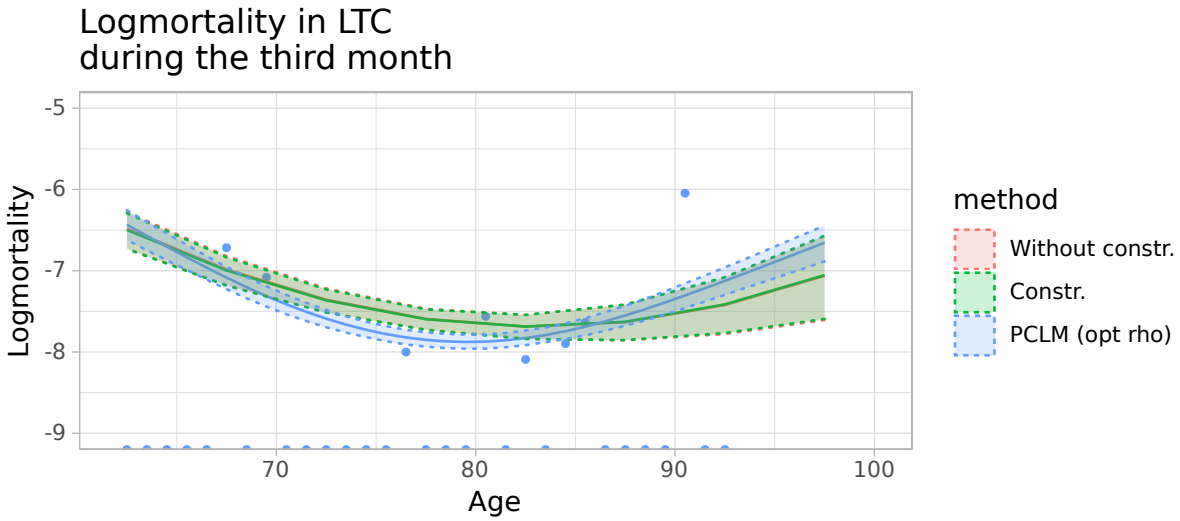


Figure 3.F.2 – Mortality of disabled policyholders during the third month following the loss of autonomy

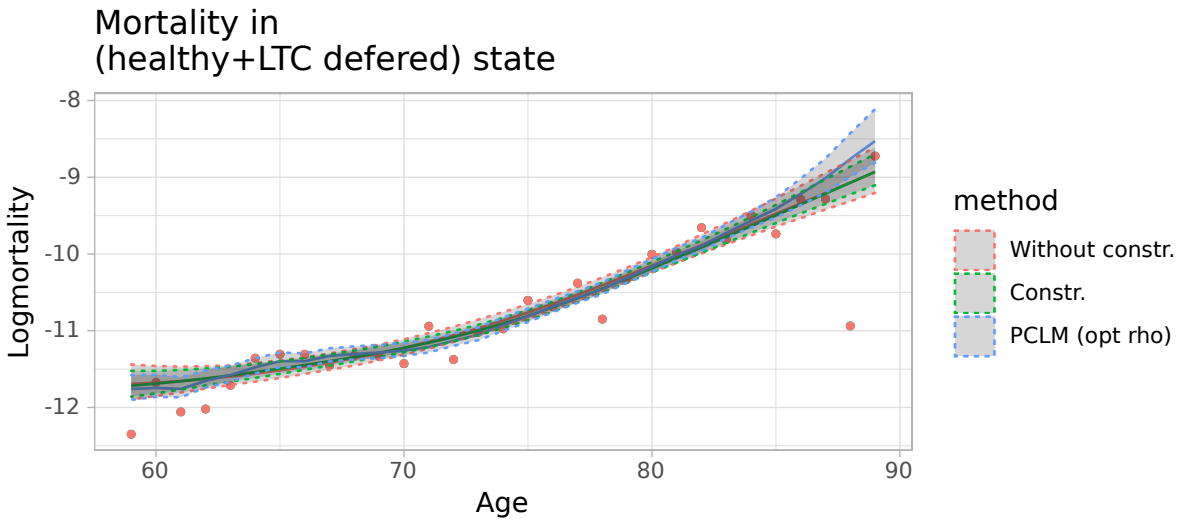


Figure 3.F.3 – Mortality in state H_2

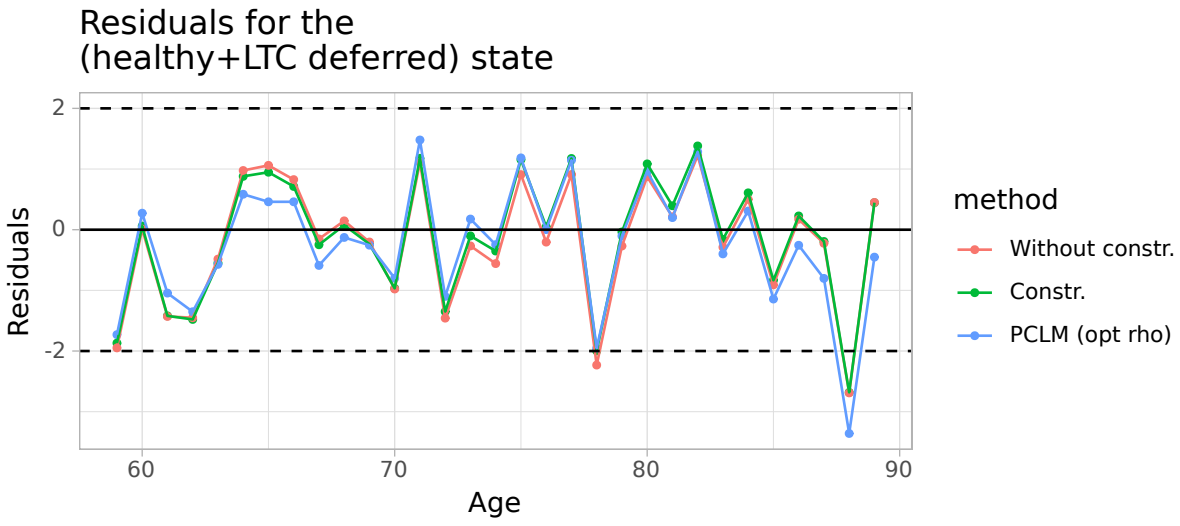


Figure 3.F.4 – Deviance residuals in the state H_2

Bibliography

Atance, D., A. Debón, and E. Navarro (2020). A comparison of forecasting mortality models using resampling methods. *Mathematics 8*(9), 1550.

Beard, R. E. (1959). *Appendix: Note on Some Mathematical Mortality Models*, pp. 302–311. John Wiley & Sons, Ltd.

Biessy, G. (2016). A semi-Markov model with pathologies for Long-Term Care Insurance. working paper or preprint.

Boggs, P. T. and J. W. Tolle (1995). Sequential quadratic programming. *Acta numerica 4*, 1–51.

Camarda, C. G., P. H. C. Eilers, and J. Gampe (2016). Sums of smooth exponentials to decompose complex series of counts. *Statistical Modelling 16*, 279 – 296.

Chapman, C. (2012). A conversation on responsibility. *The Actuary, Society of Actuaries 8*(6), 8–10.

Colombo, F., A. Llena-Nozal, J. Mercier, and F. Tjadens (2008). Medicaid eligibility issues for long-term care insurance partnership programs. *Center for Health Care Strategies Long-term Care Partnership Expansion Project March*.

Colombo, F., A. Llena-Nozal, J. Mercier, and F. Tjadens (2011). Help wanted? Providing and paying for long-term care. *OECD Health Policy Studies. OECD Publishing, Paris, France*.

Currie, I. D. and M. Durban (2002). Flexible smoothing with P-splines: A unified approach. *Statistical Modelling 2*(4), 333–349.

Dupourqué, E. (2012). AGGIR, the work of grids. *Long-Term Care News 32*.

Eilers, P. (2007). Ill-posed problems with counts, the composite link model, and penalized likelihood. *Statistical Modelling 7*.

Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing with B-splines and penalties. *Statistical Science 11*(2), 89 – 121.

Haberman, S. and E. Pitacco (1999). *Actuarial models for disability insurance*. FL: Chapmanand Hall/CRC Press.

Hurd, M. D., P. C. Michaud, and S. Rohwedder (2017). Distribution of lifetime nursing home use and of out-of-pocket spending. *Proceedings of the National Academy of Sciences of the United States of America 114*(37), 9838–9842.

Jelmer Ypma, S. G. J. (2022). R interface to nlopt. *CRAN*.

Jin, H., Y. Su, Y. Ping, S. Pickersgill, X. Chen, X. Liu, D. Watkins, Y. Li, H. Liu, and C. Wu (2023). Projecting long-term care costs for home and community-based services in china from 2005 to 2050. *Journal of the American Medical Directors Association 24*(2), 228–234.

Johnson, R. W. (2019). What is the lifetime risk of needing and receiving long-term services and supports? *Assistant Secretary for Planning and Evaluation Reports April*.

Kaye, H. S., C. Harrington, and M. P. LaPlante (2010). Long-term care: who gets it, who provides it, who pays, and how much? *Health affairs (Project Hope) 29*(1), 11–21.

Kemper, P., H. L. Komisar, and L. Alecxih (2005). Long-term care over an uncertain future: what can current retirees expect? *Inquiry: a journal of medical care organization, provision and financing 42*(4), 335–350.

Kraft, D. (1988). *A Software Package for Sequential Quadratic Programming*. Deutsche Forschungs- und Versuchsanstalt für Luft- und Raumfahrt Köln: Forschungsbericht. Wiss. Berichtswesen d. DFVLR.

Macdonald, A. S., S. J. Richards, and I. D. Currie (2018). *Modelling Mortality With Actuarial Applications*. Cambridge University Press.

Marx, B. D. and P. H. Eilers (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis 28*(2), 193 – 209.

McCullagh, P. (2019). *Generalized linear models*. Routledge.

Nelder, J. A. and R. W. M. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General) 135*(3), 370–384.

OECD (2020). Spending on long-term care. *OECD Report Nov*.

Or, Z. and A. Penneau (2021). *Pricing long-term care for older persons*, Chapter Case Study - France. WHO Centre for Health Development.

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

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

Productivity Commission of Australia (2013). An ageing australia: Preparing for the future. *Commission Research Paper*.

Remund, A., C. Camarda, and T. Riffe (2018). A cause-of-death decomposition of young adult excess mortality. *Demography 55*(3), 957–978.

Rizzi, S., J. Gampe, and P. H. C. Eilers (2015). Efficient Estimation of Smooth Distributions From Coarsely Grouped Data. *American Journal of Epidemiology 182*(2), 138–147.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics 11*(4), 735–757.

Sartori, S. et al. (2011). *Penalized regression: Bootstrap confidence intervals and variable selection for high-dimensional data sets*. Ph. D. thesis, Università degli Studi di Milano.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 461–464.

Shao, A. W., H. Chen, and M. Sherris (2019). To borrow or insure? long term care costs and the impact of housing. *Insurance: Mathematics and Economics 85*, 15–34.

Shao, A. W., M. Sherris, and J. H. Fong (2017). Product pricing and solvency capital requirements for long-term care insurance. *Scandinavian Actuarial Journal 2017*(2), 175–208.

Shi, P. and W. Zhang (2013). Managed care and health care utilization: Specification of bivariate models using copulas. *North American Actuarial Journal 17*(4), 306–324.

Thompson, R. and R. J. Baker (1981). Composite link functions in generalized linear models. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 30*(2), 125–131.

Chapitre 4

Clustering of pathologies : application to Long-Term Care Insurance

Abstract

Long-term care products cover the risk of permanent loss of autonomy. In the event of a loss of autonomy, the insurer must pay an annuity until the death or recovery of the insured. As recovery is very unlikely, long-term care product risk modellers in many countries assume that the only cause of ending annuity payments is death. Therefore, estimating the mortality of disabled insured individuals is crucial for insurers, as it significantly impacts the pricing and reserving of long-term care products. Multiple pathologies can lead to the loss of autonomy of an individual. Experience data show that the pathology responsible for the long-term care needs of an individual has a significant impact on its mortality. Therefore, accounting for the pathology is important, especially for reserving. However, the small number of data observations does not enable insurers to estimate a single mortality table for each pathology.

In this paper, we present two clustering approaches to create groups of pathologies with similar mortality rates. This allows us to aggregate the data of different pathologies and reduce the number of different mortality tables without losing too much specific information on each pathology. The first method relies on GLM trees, while the second method is a generalized K-means approach. We then show that accounting for pathologies using the clusters obtained from the proposed methods improves the predictive performance of mortality. Finally, estimating different mortality rates according to pathology allows insurers to improve reserving and to study the impact of an increase or decrease in the incidence of a specific pathology on the mortality of disabled insured individuals.

Keywords: Clustering; Pathologies; Long-Term Care Insurance; Actuarial modelling; Generalized Linear Models; GLM Trees; K-Means.

4.1 Introduction

Most long-term care insurance products cover the risk of loss autonomy by providing an annuity until death. Therefore, the estimation of mortality after a loss of autonomy is crucial for insurers. An overestimation of mortality would lead to an underestimation of the premium, while an underestimation of mortality could prompt individuals to underwrite the contract with another insurer due to an overestimation of the premium.

Multiple pathologies can lead to the loss of autonomy and the need for long-term care (LTC). The principal causes of such autonomy loss are dementia, cancer, neurological disease, and cardiovascular disease. Less common causes include respiratory and osteoarticular diseases, as well as accidents. As shown in Biessy (2016), pathologies have a significant impact on mortality in LTC. In this paper, the author uses a multi-state model with the states of autonomy, death and four states of LTC, combined with a continuous semi-Markov framework. The four states of LTC correspond to four different groups of pathologies: "cancer", "dementia", "neurological diseases" and "other causes". Therefore, neglecting pathology in the estimation of mortality leads to a loss of information for the insurer.

Due to the recent development of long-term care products, insurers often lack data related to the mortality of disabled insured individuals. Furthermore, information about pathologies is not always available. This scarcity of claims makes it difficult or impossible for them to estimate a specific mortality table for each pathology. Moreover, for operational reasons, insurers often prefer the simplicity of having a lower number of tables while capturing most of the variance. Clustering pathologies with similar mortality rates seems a good compromise to reduce the number of different mortality tables without losing too much specific information on each pathology.

As in Biessy (2016), we consider that mortality in LTC depends on age and duration. We are thus in a context of two-dimensional mortality laws. Therefore, this paper aims to study surface clustering methods.

The problem of curve clustering was previously addressed by Abraham et al. (2003). As the problem consists of clustering functions with an infinite dimension, Abraham et al. (2003) proposed a two-step approach. In the initial step, the dimension of each element is reduced by approximating the curve with a finite number of parameters. In this paper, the authors rely on B-splines with the following condition: the splines have to be identically distributed on the interval for each curve. In the second step, the K-means method is applied to the B-spline coefficients. Theoretically, this approach could be extended to surface clustering for pathologies. In this context, one might have to independently estimate B-spline coefficients for each element. Pathologies would then be clustered using the K-means algorithm on the basis spline coefficients. However, this method may not be suitable for our problem. Indeed, the scarcity of observations for some pathologies makes

it challenging or even impossible to fit two-dimensional P-splines or B-splines for certain elements.

Clustering methods have already been used in the context of mortality modelling. Using fuzzy clustering implemented in the R package e1071, Debón et al. (2017) clustered mortality surfaces of EU countries. In this method, each country can be associated with more than one cluster. A coefficient indicates the strength of the association between a country and each of the clusters. In another study, Carracedo et al. (2018) proposed another method to detect mortality clusters of European countries, using Local Moran's Index. More recently, relying on data from the Human Mortality Database, Léger and Mazzuco (2021) used three different functional clustering methods to group countries with similar mortality trends.

Two methods, both relying on the generalized linear model (GLM) framework (Nelder and Wedderburn, 1972; McCullagh, 2019), are developed in this paper to cluster pathologies into homogeneous groups in terms of mortality. The first one, presented in Section 4.4.1, uses GLM trees. The second approach, presented in Section 4.4.2, is inspired by K-means, the most famous nonhierarchical clustering technique.

We start by describing the dataset used in this paper in Section 4.2 and analysing the heterogeneity induced by the plurality of pathologies. Then, Section 4.3 focuses on the basics of mortality modelling in the context of long-term care insurance and how generalized linear models are used for this purpose in the remainder of the paper. Clustering methods are then presented and applied to our dataset in Section 4.4. The performances of the clustering methods are then compared in Section 4.5. Using the clusters resulting from the best method according to the previous section, the benefits of accounting for pathology when modelling the mortality of disabled policyholders are discussed in Section 4.6. Section 4.7 concludes this paper by summarising the results and providing suggestions for future research.

4.2 Data

This section begins by introducing the dataset used for this study. We then analyse the heterogeneity of mortality between pathologies using descriptive statistics, highlighting the importance of considering this covariate in the estimation of mortality for disabled policyholders.

4.2.1 Presentation of the data

This study relies on data from a non-French health fund. Given that this paper focuses on mortality in LTC, only disabled policyholders are retained in the database. A total

of 11,115 disabled individuals were observed from 2008 to 2016. Unlike French LTC insurance products, this health fund considers recovery from disability as being possible. The definition used for disability differs from the usual definitions in the French market, leading to the presence of transitions from disability to the autonomous state in the database. Given that this paper focuses only on mortality, recovery is treated as right censoring. Additionally, this health fund imposes a maximum lifetime benefit of five years. As a consequence, each disabled policyholder exits the disabled state after a maximum of 5 years, and health status is not observed beyond this period. This maximum lifetime benefit implies right censoring, and no observations are available for durations exceeding 5 years. Information regarding the pathology responsible for the loss of autonomy of each disabled policyholder is available. The detailed diagnosis is provided by the health fund, along with a broader pathology group at a more macro level. Examples of pathology groups include "Respiratory diseases", "Cardiovascular diseases" and "Dementia". The health fund distinguishes 14 groups of pathologies. Rare pathologies are grouped as "Other". A total of 5,000 deaths of disabled insured individuals were observed during the observation period.

Table 4.2.1 summarises the number of claims and observed deaths per pathology. In this health fund, many disabled insured individuals have multiple pathologies identified as the cause of the loss of autonomy. The most prevalent causes of claims include dementia, cancer, cardiovascular diseases, and accidents.

Pathology	**Censoring**	**Death**	**Total**
Accident	740	295	1,035
Cancer	504	1,501	2,005
Cardiovascular disease	752	494	1,246
Dementia	1,235	907	2,142
Endocrine disease	28	28	56
Gastrointestinal disease	26	16	42
Infectious disease	31	15	46
Neurological disease	427	174	601
Osteoarticular	136	63	199
Other	360	88	448
Psychiatric disease & depression	198	132	330
Respiratory disease	55	67	122
Several	1,601	1,159	2,760
Urological or kydney disorder	32	51	83

Table 4.2.1 – Number of observed insured individuals and causes of exit for each pathology

Although the original data are at the individual level, the models presented in the paper use aggregated data. The observations are grouped by attained age, duration, gender,

and pathology by defining a discretization grid for age and duration. Age and duration intervals are split on an annual basis. However, due to significant variations in mortality in the early onset of disability, the first year following the loss of autonomy is commonly partitioned on a monthly basis. As a result, the database used in the remainder of the paper contains one line for each combination of these four variables. For each of these lines, the central exposure and count of deaths are calculated using the original individual database.

Notation: x_i denotes the i^{th} split point of the discretization grid on the age interval, and t_j denotes the j^{th} split point on the duration interval.

The central exposure associated with the combination of age x_i, duration t_j, gender g and pathology p is the sum of the individual central exposures of disabled policyholders with pathology p and gender g, for an attained age between x_i and x_{i+1} and a duration between t_j and t_{j+1}.

4.2.2 Heterogeneity of mortality of disabled insured individuals between pathologies

Biessy (2016) shows the significant impact of pathologies on mortality within the French LTC insurance market. Our data exhibit the same phenomenon. The impact of pathologies on mortality can be illustrated by calculating the ratios of the actual counts of deaths over the expected counts under the null hypothesis H_0: "Given the gender, age and duration since the loss of autonomy, the mortality rates are equal for all pathologies".

Assuming that mortality rates are independent of the pathology responsible for the loss of autonomy, crude mortality rates are estimated on the entire database of disabled policyholders for each gender, without accounting for the pathology. The expected counts of deaths for each pathology, attained age, and duration are then computed using the central exposures and crude mortality rates of disabled policyholders.

The ratios of actual over expected counts of deaths for each pathology and gender are plotted in Figure 4.2.1. Under the assumption that pathology has no impact on mortality in LTC, the observed counts of deaths should be close to the expected counts. Therefore, the actual-to-expected ratios should be close to 1. With a significance level of $\alpha = 5\%$ and under the Poisson assumption of the number of deaths, the tolerance intervals of the ratios of actual over expected values are constructed. The purpose of these intervals is to account for the variance in the random variable of the observed counts of deaths. Considering pathology p, if the ratio is above the tolerance interval, we can conclude that the mortality observed with pathology p is significantly underestimated by the common mortality law aggregating all pathologies. If the ratio is below this interval, the mortality

is significantly overestimated. Figure 4.2.1 shows that using common mortality rates for all pathologies leads to a significant underestimation of the mortality of disabled insured individuals with cancer for both males and females. Moreover, using a unique mortality law for all pathologies results in an overestimation of mortality of disabled policyholders with cardiovascular disease, dementia, neurological disease, psychiatric disease, as well as those disabled because of an accident. Further details about statistical hypothesis testing using the metric of actual over expected values are provided in Section 4.5.2.1.

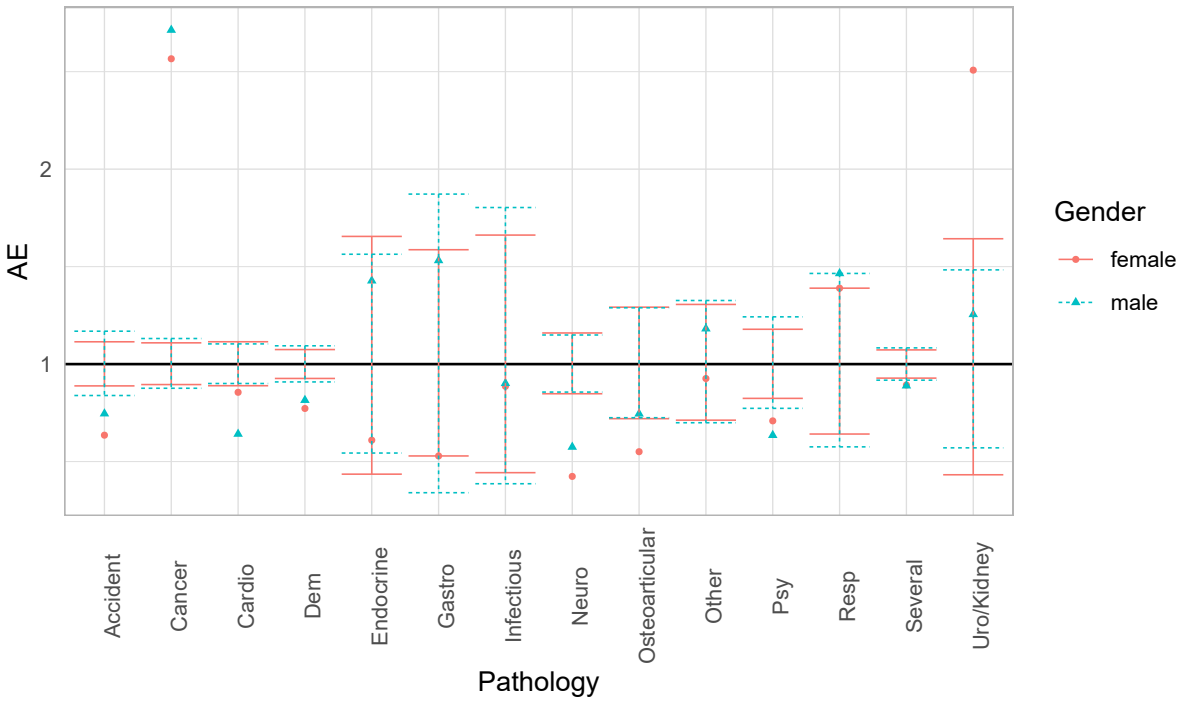


Figure 4.2.1 – Actual-to-expected ratios obtained by considering a common mortality law for all pathologies

This finding highlights the importance of considering pathology when estimating the mortality of disabled policyholders in the context of modelling LTC insurance products. However, datasets containing information about the pathology leading to the need for long-term care are scarce, and some pathologies are also poorly represented in portfolios. Table 4.2.1 shows that fewer than 100 individuals were observed by the health fund during the study period for the following pathologies: endocrine diseases, gastrointestinal diseases, infectious diseases, and urological or kidney disorders. Therefore, while accounting for pathologies is crucial, the lack of data for some of them makes estimating a specific mortality table for each pathology challenging. Moreover, for the sake of simplicity, insurers prefer to maintain a reasonable number of distinct mortality tables for disabled policyholders. Therefore, clustering pathologies with similar mortality rates is an effective means of enhancing mortality modelling without introducing excessive complexity.

4.3 The fundamentals of mortality modelling using the GLMs

This section focuses on the common tools used to model and estimate mortality of disabled policyholders in long-term care insurance products. We first describe the tools for survival analysis and the common assumptions used for such products in Section 4.3.1. Two different methods for clustering pathologies are presented and tested in this paper. Since both approaches rely on generalized linear models (GLMs), we describe the fundamentals of GLMs as applied to mortality modelling in Section 4.3.2.

4.3.1 Modelling the mortality of disabled policyholders

In the survival model theory, mortality is often described by the force of mortality μ, also called mortality intensity. Numerous papers on LTC, such as Czado and Rudolph (2002), Biessy (2017) and Biessy (2019), show that the force of mortality of disabled policyholders depends on the time that has elapsed since the loss of autonomy. Semi-Markov models, initially introduced in the context of disability modelling by Janssen (1966), followed by Hoem (1972), assume that the intensities of the process depend not only on the current state but also on the time spent in this state. As in more recent papers, such as Pitacco (2014) for disability modelling and Soetewey et al. (2022), Fuino and Wagner (2018), and Xuanyuan and Xuanyuan (2023) for long-term care insurance modelling, we assume the following:

Assumption. 4.3.1

Mortality rates depend on attained age and time spent since the loss of autonomy.

Let $\mathcal{X}_x$ denote the health status of a disabled policyholder at age x. Let A, LTC and $Death$ denote the autonomous, disabled and death states, respectively. The force of mortality $\mu_{x,t}$ is given by

$$\mu_{x,t} = \lim_{h \to 0} \frac{\mathbb{P}\left(\mathcal{X}_{x+h} = D | \mathcal{X}_x = LTC, \mathcal{X}_{(x-t)} = LTC, \mathcal{X}_{(x-t)^-} = A\right)}{h} \tag{4.1}$$

The force of mortality $\mu_{x,t}$ is commonly assumed to be piecewise constant by age and duration, as in Soetewey et al. (2022). This assumption is reasonable for appropriate steps of age and duration. In the remainder of the paper, we assume the following:

Assumption. 4.3.2

Mortality rates are piecewise constant by age and duration on a yearly basis, except for the first year following the loss of autonomy, in which mortality rates are piecewise constant on a monthly basis on the duration axis.

The split points of the age and duration intervals are denoted as $(x_i)_{i\in\{1,\dots,M_x\}}$ and $(t_j)_{j\in\{1,\dots,M_t\}}$, respectively.

Mortality can also be described by q_{x_i,t_j}, denoting the probability of an individual of integer age x_i who has been disabled for t_j years dying before reaching duration t_{j+1} or age x_{i+1}. μ_{x_i,t_j} and q_{x_i,t_j} are related by the following equation:

$$\begin{aligned} q_{x_i,t_j} &= 1 - \exp\left(-\int_0^{\min(t_{k+1}-t_k,\,x_{i+1}-x_i)} \mu_{x_i+u,t_j+u}\mathrm{d}u\right) \\ &= 1 - \exp\left(-\mu_{x_i,t_j} \times \min\left(t_{k+1}-t_k, x_{i+1}-x_i\right)\right). \end{aligned} \tag{4.2}$$

As men and women are not equal when it comes to the risk of mortality and loss of autonomy, the mortality rates of disabled policyholders depend on gender g.

We denote by $\mu^{g,p}_{x_i,t_j}$ and $q^{g,p}_{x_i,t_j}$ the force of mortality and the probability of death associated with gender g and pathology p for age $x, x_i \leqslant x < x_{i+1}$ and duration $t, t_j \leqslant t < t_{j+1}$, respectively.

4.3.2 Basics of GLMs

Both methods used in this paper to cluster pathologies rely on generalized linear models (GLMs) (Nelder and Wedderburn, 1972; McCullagh, 2019). GLMs are a generalization of linear models that allow the probability distribution of the response variable Y to be chosen from the exponential family introduced in Barndorff (1978). Examples of distributions from this family are Normal, Exponential, Log-Normal, Gamma, Binomial, Poisson, and Inverse Gaussian. Let X denote the covariates. The conditional mean of the response variable given X, is expressed as a function of a linear combination of X through the following expression:

$$\mathbb{E}[Y|X] = G^{-1}(X\beta), \tag{4.3}$$

where β is a coefficient vector, $G(\cdot)$ is the link function, and $X\beta$ is the linear predictor.

The GLM depends on three elements:

- the probability distribution,
- the link function $G(\cdot)$,

- the linear predictor $\eta = X\beta$.

The optimal coefficient vector β is obtained by maximizing the log-likelihood.

The cross effects of a qualitative covariate with another covariate (qualitative or quantitative) can be included in the linear predictor to allow group-specific coefficients in the model. In this case, some elements of the vector of coefficients β may differ depending on the value of the qualitative covariate involved in the cross effect.

In some situations, it may also be useful to add an additional covariate to the linear predictor, with a fixed coefficient that does not have to be estimated. This term is called an offset. In this case, Equation 4.3 becomes

$$G(\mathbb{E}[Y|X]) = X\beta + \text{offset}. \tag{4.4}$$

In the context of mortality modelling, the response variable is the number of deaths. The common probability distributions used in this context are Poisson and Binomial distributions (Hunt and Blake, 2021).

Let $D^{g,p}_{x_i,t_j}$ denote the number of deaths occurring at age x between x_i and x_{i+1} with a duration in claim t between t_j and t_{j+1} for gender g and pathology p.

Assuming a binomial distribution of the count of deaths, $\mathbb{E}[D^{g,p}_{x_i,t_j}] = {}^0e^{g,p}_{x_i,t_j} q^{g,p}_{x_i,t_j}$, where ${}^0e^{g,p}_{x_i,t_j}$ denotes the initial number of disabled live policyholders with pathology p at age x_i with a duration of claim t_j. Assuming a Poisson distribution of the count of deaths, $\mathbb{E}[D^{g,p}_{x_i,t_j}] = e^{g,p}_{x_i,t_j}\mu^{g,p}_{x_i,t_j}$, where $e^{g,p}_{x_i,t_j}$ denotes the central exposure to risk (average number of disabled policyholders with pathology p and gender g with attained age between x_i and x_{i+1}, for a claim duration between t_j and t_{j+1}).

While central exposures are widely available and easily estimated from a database, initial exposures are more complex to estimate because of censoring and truncating. However, initial exposures can be approximated from central exposures. Both distributions were tested in this study, but we retained the Poisson distribution due to its better predictive performance in terms of the log-likelihood values.

In GLMs, the link function is chosen by the user and determines the link between the regressor and the estimated value of the response. However, as explained in Myers and Montgomery (1997), each distribution has a special link function called the canonical link function, which has the nice mathematical properties described in Nelder and Wedderburn (1972). The canonical function of the Poisson GLM is the log function. In this paper, we work with the canonical link function because of its useful properties.

With the number of deaths as the response variable, and using the Poisson assumption combined with the log link function, we have the following equation:

$$\begin{aligned}\log(\mathbb{E}[D_{x,t}^{g,p}] &= \log(e_{x,t}^{g,p}\mu_{x,t}^{g,p}) \\ &= \underbrace{\log(e_{x,t}^{g,p})}_{\text{offset}} + \log(\mu_{x,t}^{g,p}).\end{aligned} \tag{4.5}$$

Therefore, using the Poisson regression to model the counts of deaths, an offset accounting for the central exposure is added to the linear predictor.

The covariates considered in this paper are the gender, age and duration. In what follows, we use a formula assuming a quadratic impact of age, and a cubic impact of duration. As mortality in LTC greatly differs for males and females, a cross effect of gender with both age and duration is considered in the GLM formula.

4.4 Clustering methods and initial results

The first clustering approach, which rely on GLM Trees, is described in Subsection 4.4.1. The second method proposed in this paper is inspired by the well known K-means algorithm. This method will be further explained in Subsection 4.4.2.

4.4.1 First method: GLM trees

As can be seen from its name, this method introduced by Achim Zeileis and Hornik (2008) combines generalized linear models (GLMs) with tree algorithms. The basic unsupervised learning tree algorithm builds a tree of clusters. The root cluster contains all the observations, and each node contains a subset of the data of its parent. The tree successively splits each node according to a specific metric. If all the observations in the node are considered to belong to the same group according to the metric, then the node is not split further. The final leaves indicate the final label of each observation.

The GLM tree method is as follows. As in Section 4.3.2, let Y denote the variable to be explained, and let X denote the covariates included in the linear predictor to explain Y. Let Z denote the covariates considered for splitting the tree and called the split variables. A variable can be used both as a split variable and as a covariate in the linear predictor in a GLM, at the same time. As with the basic tree algorithm, the root node contains all the observations. At each step, for each node of the current level of the tree, the algorithm starts by estimating the GLM parameters using all the observations from that node. Thereafter, the stability of the parameters is tested over the split variables. The node is then split according to the split variable with the highest parameter instability. These steps are repeated until the parameter stability threshold is met or the tree has reached a maximum depth chosen in advance by the user. According to Achim Zeileis and

Hornik (2008), one of the main benefits of this method is the use of the same objective function for both parameter estimation and partitioning. While the formula of the GLM fitted in each node implies linear relationships between Y and the covariates X, recursive partitioning allows for a nonlinear relationship by allowing cuts of quantitative variables at some identified optimal split points.

The formula of the GLM tree is written as follows:

$$Y \sim f(X)|Z,$$

where $f(X)$ describes the form of the linear predictor.

In this paper, we focus on the clustering of pathologies. For the sake of simplicity, insurers prefer to have identical clusters of pathologies for both genders. Therefore, the only variable considered as a split variable is pathology. The covariates considered as explanatory variables in the GLM are gender, attained age of the insured, and duration since the loss of autonomy. GLM trees can be fitted using a recursive partitioning algorithm implemented in **R** (R Core Team, 2020) within the **partykit** library. A detailed description of the library is provided in Hothorn and Zeileis (2015).

The formula used for our application is:

$$\begin{aligned}\text{Formula: } Count \sim Age \times Gender + Age^2 \times Gender + Duration \times Gender + \\ Duration^2 \times Gender + Duration^3 \times Gender|Pathology.\end{aligned}$$

Assuming a Poisson distribution of the counts of deaths and using the log link function, an offset accounting for the central exposure in each cluster is added to the GLM, as presented in Equation 4.4. Let $\kappa(p)$ denote the assigned cluster of pathology p; the offset for cluster k is given by

$$\text{Offset}_{x,t}^{g,k} = \log\left(\sum_p \mathbb{1}\{\kappa(p) = k\} e_{x,t}^{g,p}\right)$$

Figure 4.4.1 shows the results of the GLM tree clustering. The analysis of the actual-to-expected ratios in Section 4.2.2 shows that the mortality of disabled policyholders who lose their autonomy because of cancer greatly differs from the mortality associated to other pathologies. This difference in mortality is detected by the GLM tree, which sorts cancer into its own cluster. Let us analyse the recursive partitioning in parallel with the results of the actual-to-expected ratios plotted in Figure 4.2.1. We use "global disabled mortality" to denote the mortality of the overall portfolio of disabled policyholders without accounting for the pathology. The GLM tree seems to first separate the pathologies whose observed mortalities are underestimated by the global disabled mortality, from the other pathologies. These include cancer; respiratory disease; urological and kidney disorders; and the "other" group, corresponding to all pathologies not listed among the 14 studied pathologies.

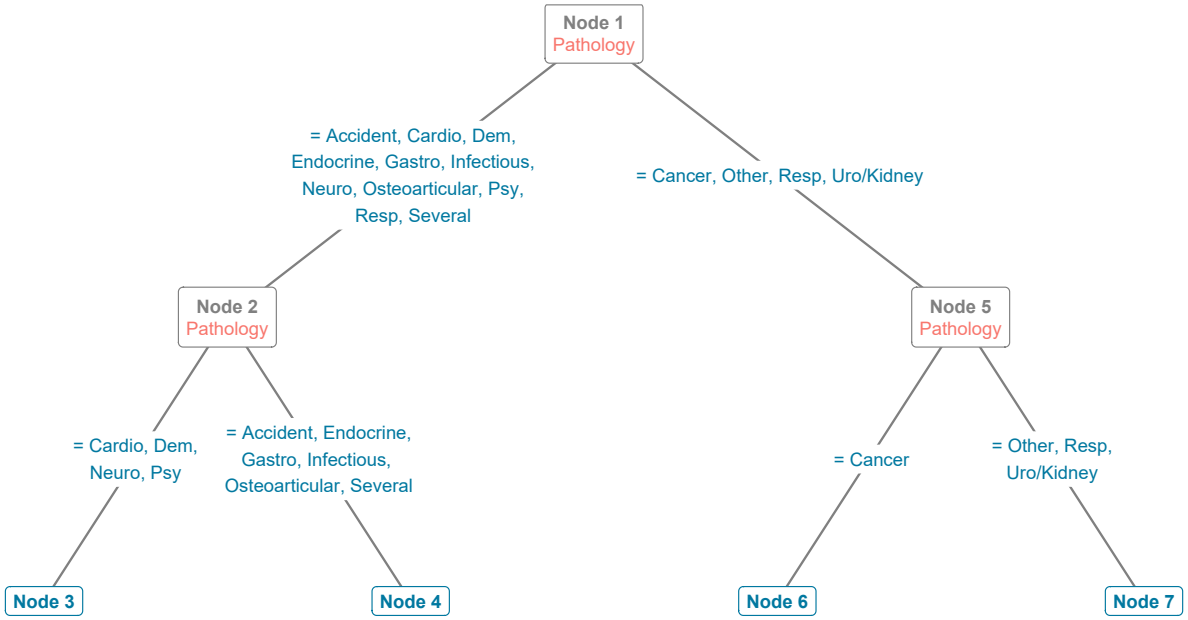


Figure 4.4.1 – Resulting tree obtained by the GLM tree method

4.4.2 Second method: Generalized K-means

K-means is one of the most widely used unsupervised clustering algorithms. In this section, we propose a K-means approach to cluster pathologies with similar mortality rates. Section 4.4.2.1 starts by introducing the basics of the K-means algorithm and the chosen features required to apply it to our problem of surface clustering. The first step of the K-means approach depends on a random distribution of the pathologies among the K groups. Section 4.4.2.2 studies the impact of this random initialization on the composition of the final clusters. Since this impact seems rather significant in our application, a similarity matrix is built by running the K-means algorithm many times. Two approaches presented in Section 4.4.2.3 are then subsequently tested to cluster pathologies based on this similarity matrix.

4.4.2.1 Generalized K-means for pathology clustering

After fixing the desired number of final clusters K, the goal of clustering is to make the elements within the clusters as similar as possible while maximizing the dissimilarity between elements of different groups. K-means clustering requires the choice of a distance or similarity measure between single elements and clusters. Let a be an element and C denote a cluster. If the elements are points from an Euclidian space $\mathbb{R}^n$, then the common distance measure $dist(a, C)$ is the Euclidian distance,

$$dist(a, C) = \sum_{i=1}^{n} (a_i - C_i)^2.$$

Any other similarity or dissimilarity measure can be used to adapt the K-means algorithm depending on the context and the problem. In this study, each pathology p is represented by an element a^p containing information about the observed counts of deaths and central exposures by attained age, duration and gender. In this work, mortality associated with each cluster is estimated using GLMs. Therefore, our method combines GLMs with K-means. The use of K-means in a GLM framework has already been addressed in several papers as in Zhang and Lin (2021) and Abraham et al. (2003). The latter proposed a two-step approach for unsupervised curve clustering. After fitting B-splines to each element, the similarity between curves is assessed by measuring the distance between the vectors of the spline coefficients. This approach is based on the idea that curves with similar B-spline coefficients are close. However, this approach cannot be used in our work because of the rather limited number of observations for certain pathologies, preventing a robust estimation of their associated spline coefficients.

Let $C^k, k \in \{1, \dots, K\}$ denote K distinct clusters. Each cluster can be represented by its mortality surface, which represents the common mortality of all the pathologies within this cluster. The similarity $S(a^p, C^l)$ between an element a^p and a cluster C^k is assessed by the Poisson log-likelihood of the observed counts of deaths from pathology p assuming the mortality rates associated with cluster C^k. The distance measure is given by the following equation:

$$\begin{aligned} dist(a^p, C^k) &= -S(a^p, C^k) \\ &= -\left(\sum_{g=1}^{2} \sum_{i=1}^{M_x} \sum_{j=1}^{M_t} -\mu_{x_i,t_j}^{g,k} e_{x_i,t_j}^{g,p} + d_{x_i,t_j}^{g,p} \log\left(\mu_{x_i,t_j}^{g,k} e_{x_i,t_j}^{g,p} \right) \right), \end{aligned} \tag{4.6}$$

where:

- M_x is the number of subdivisions in the age interval,
- M_t is the number of subdivisions in the duration interval,
- $\mu_{x_i,t_j}^{g,k}$ is the common mortality intensity of pathologies in cluster C^k, and
- $e_{x_i,t_j}^{g,p}$ is the central exposure of disabled policyholders of gender g with pathology p.

Remark: As in Section 4.4.1, we assume similar clusters of pathologies for males and females.

The steps of the adapted K-means clustering algorithm are as follows:

Algorithm. 4.4.1: K-means clustering algorithm

1. Randomly assign a cluster from 1 to K to each pathology in the dataset, corresponding to the initial cluster of the pathology,
2. Iterate the following until the distribution of pathologies among the clusters stops changing:
   (a) Fit the GLM for each cluster without accounting for the pathology. The central exposures and counts of deaths from pathologies within the same cluster are added.
   (b) Assign each pathology to the cluster with the closest mortality surface, using the distance measure $dist(\cdot, \cdot)$ from Equation 4.6.

Since K-means requires fixing the number of clusters in advance, our method is combined with the generalized information criterion (GIC) to select the best hyperparameter K when this value is unknown. As recommended in Zhang and Lin (2021), K is chosen to minimize the Bayesian information criterion (BIC). The choice of K is further discussed in Section 4.5.1.

4.4.2.2 Impact of the random initialization and construction of a distance matrix

The first step of the adapted K-means algorithm is to randomly assign a cluster to each of the M_P pathologies. Let us analyse the impact of this random initialization on the composition of the final groups of diseases for a chosen number of final clusters $K = 4$. We then propose a method to construct a matrix of distances between pathologies based on their assignments to the K groups for a large number of random initializations.

The K-means algorithm is run successively with 1,000 random initializations. The distribution of the assigned cluster for each pathology is plotted in Figure 4.4.2.

After building the 4 groups and fitting a GLM to estimate the mortality surfaces associated with them, the groups were ranked by the level of mortality at age 70 for the first duration. This arbitrary choice is used to label the clusters in each iteration. As shown in Figure 4.4.2, cancer seems to always be assigned to the cluster 4, which has the highest mortality rate at 70 years old immediately after the loss of autonomy. In contrast, other pathologies are rarely assigned to this group. This finding shows that cancer differs from other pathologies in particular by its high mortality rate for low durations. Interpretation is more difficult for other pathologies since random initialization seems to have a greater impact on their final assigned clusters.

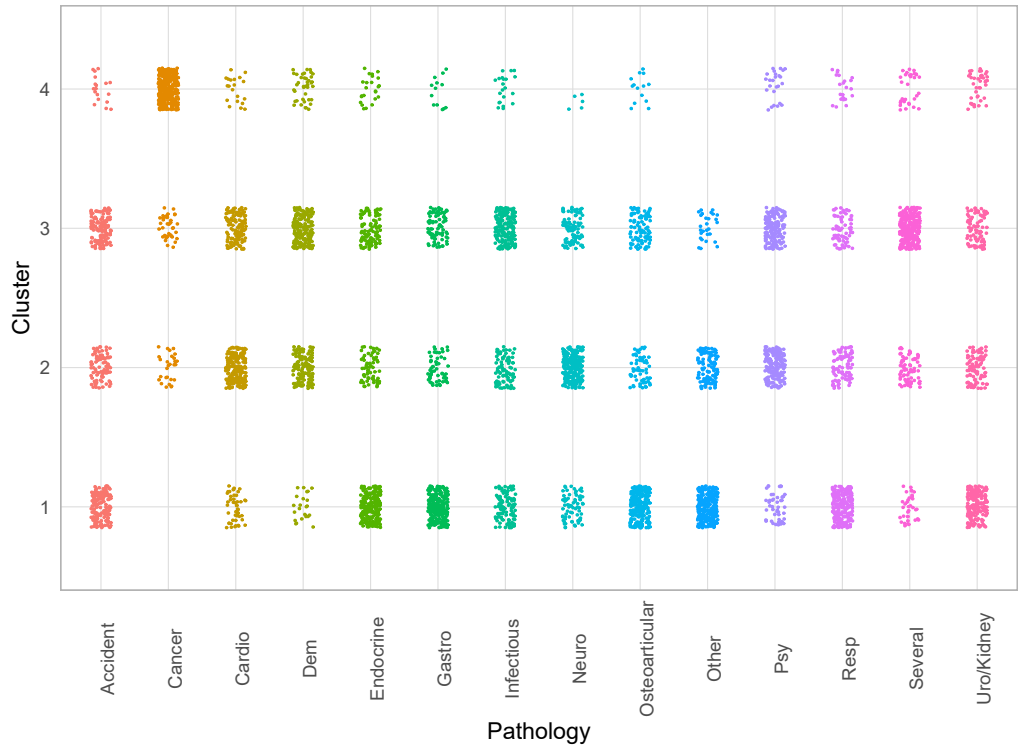


Figure 4.4.2 – Impact of the random initialization on the cluster assigned to each pathology

We therefore need a metric to assess the similarity between pathologies. This can be estimated by answering the following question: how often are two pathologies assigned to the same cluster? A similarity matrix $\mathcal{S} \in \mathcal{M}_P(\mathbb{R})$ is built, where $\mathcal{S}_{p,\tilde{p}} \in [0;1]$ denotes the similarity between pathologies p and $\tilde{p}$.

Let R denote the number of random initializations. Let $\kappa^r(p) \in \{1,\dots,K\}$ denote the assigned cluster of pathology p for the r^{th} random initialization; then $\mathcal{S}_{p,\tilde{p}}$ is estimated by

$$\mathcal{S}_{p,\tilde{p}} = \frac{\sum_{r=1}^{R} \mathbb{1}\{\kappa^r(p) = \kappa^r(\tilde{p})\}}{R}. \tag{4.7}$$

Note: $\mathcal{S}$ is symmetric by construction, and $\mathcal{S}_{p,p} = 1$.

The similarity matrix $\mathcal{S}$ associated with the $R = 1{,}000$ random initializations plotted in Figure 4.4.2 is given by Table 4.4.1. Values above 0.6 in the upper diagonal are highlighted in yellow for values between 0.6 and 0.7, orange for values between 0.7 and 0.8 and red for values above 0.8. The two pathologies the most often assigned to the same group are dementia and psychiatric diseases. Then, urological and kidney diseases are assigned to the same cluster as respiratory diseases in 83% of the 1,000 initializations. With all similarity values between 0.6 and 0.8, we can see in yellow and orange that endocrine, gastrointestinal, osteoarticular, infectious and respiratory diseases seem similar in terms of mortality when they are responsible for the loss of autonomy. Other similarity values

are more difficult to interpret. Therefore, statistical methods are needed to construct optimal clusters based on this similarity matrix. Two approaches are presented and tested in Section 4.4.2.3 to address this problem.

	Accident	Cancer	Cardio	Dem	Endocrine	Gastro	Infectious	Neuro	Osteo-articular	Other	Psy	Resp	Several	Uro Kidney
Accident	1.00	0.00	0.22	0.34	0.55	0.49	0.62	0.38	0.57	0.38	0.32	0.37	0.40	0.24
Cancer	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.00	0.00	0.00	0.02
Cardio	0.22	0.00	1.00	0.55	0.10	0.01	0.12	0.57	0.08	0.20	0.46	0.18	0.41	0.18
Dem	0.34	0.00	0.55	1.00	0.10	0.07	0.19	0.62	0.10	0.10	0.84	0.06	0.39	0.03
Endocrine	0.55	0.00	0.10	0.10	1.00	0.78	0.65	0.04	0.78	0.55	0.06	0.71	0.46	0.59
Gastro	0.49	0.00	0.01	0.07	0.78	1.00	0.62	0.06	0.79	0.45	0.04	0.64	0.32	0.53
Infectious	0.62	0.00	0.12	0.19	0.65	0.62	1.00	0.13	0.63	0.30	0.14	0.52	0.60	0.49
Neuro	0.38	0.00	0.57	0.62	0.04	0.06	0.13	1.00	0.11	0.09	0.64	0.04	0.22	0.02
Osteo-articular	0.57	0.00	0.08	0.10	0.78	0.79	0.63	0.11	1.00	0.46	0.10	0.60	0.45	0.47
Other	0.38	0.13	0.20	0.10	0.55	0.45	0.30	0.09	0.46	1.00	0.07	0.69	0.20	0.56
Psy	0.32	0.00	0.46	0.84	0.06	0.04	0.14	0.64	0.10	0.07	1.00	0.01	0.28	0.00
Resp	0.37	0.00	0.18	0.06	0.71	0.64	0.52	0.04	0.60	0.69	0.01	1.00	0.40	0.83
Several	0.40	0.00	0.41	0.39	0.46	0.32	0.60	0.22	0.45	0.20	0.28	0.40	1.00	0.40
Uro Kidney	0.24	0.02	0.18	0.03	0.59	0.53	0.49	0.02	0.47	0.56	0.00	0.83	0.40	1.00

Table 4.4.1 – Similarity matrix for $K = 4$

4.4.2.3 Clustering based on the similarity matrix

Using the similarity matrix $\mathcal{S}$ built with R iterations of generalized K-means for a fixed number of final clusters K, the goal of this section is to construct K clusters of pathologies. This can be seen as an optimal permutation problem (OPP), introduced in Morone (2022). In this paper, the authors proposes a theoretical framework to find an optimal permutation of the rows and columns of the matrix to obtain a new matrix as close as possible to a desired clustered form. Indeed, if there exists a permutation matrix P such that $P^t\mathcal{S}P$ is a block matrix composed of K blocks, then each block defines a cluster. In this hypothetical situation, the similarity between pathologies of seperate clusters is zero. A clear structure of clusters where all similarity values between two pathologies of seperate groups are zero is an extreme situation. In most cases, the goal is to maximize values within the blocks while minimizing values outside the blocks. One disadvantage of this method is the need to set the number of elements within each cluster in advance in addition to the number of clusters K, leading to less flexibility in clustering. Therefore, we propose two other approaches to cluster pathologies based on the similarity matrix $\mathcal{S}$ without setting the number of elements within each group in advance. While both methods are bottom-up hierarchical clustering algorithms, as described in Algorithm 4.4.2, they differ in how they assess the distances between clusters.

Algorithm. 4.4.2: Bottom-Up hierarchical clustering algorithm

1. Assign each object to its own cluster.
2. Iterate the following until there is a single cluster:
   (a) Aggregate the two clusters with the highest similarity or, equivalently, the smallest distance;
   (b) Update the distances or similarity values between all clusters using a chosen formula.

Starting with M_P elements, the number of clusters after the l^{th} iteration is $M_P - l$. The clusters constructed at each step are called "temporary clusters" in what follows. Let $(c^k_{(l)})_{k\in\{1,\dots,M_P-l\}}$ denote the set of $M_P - l$ temporary clusters at the end of the l^{th} iteration. The difference between the two approaches lies in step 2b of Algorithm 4.4.2.

First method for updating the similarity between clusters

In the first approach, the computation of the distance between two temporary clusters relies on conditional probability theory. At step l, the similarity between $c^m_{(l)}$ and $c^{\tilde{m}}_{(l)}$ is assessed by the probability that these two temporary clusters are subsets of the same final cluster obtained with K-means, conditional on the following event:

"For all temporary clusters $c^m_{(l)}$ of step l, there exists a final cluster $C^k, k \in \{1, \dots, K\}$ such

that all pathologies in $c_{(i)}^m$ are in C^k". This event is mathematically written as

$$\forall m \in \{1, \ldots, M_P - l\}, \exists k \in \{1, \ldots, K\}, c_{(l)}^m \subset C^k.$$

Therefore, the similarity between $c_{(l)}^m$ and $c_{(l)}^{\tilde{m}}$ after l iterations is

$$\mathbb{P}\left(\exists k \in \{1, \ldots, K\}, \left\{c_{(l)}^m \cup c_{(l)}^{\tilde{m}}\right\} \subset C^k \middle| \left\{\forall m \in \{1, \ldots, M_P - l\}, \exists k \in \{1, \ldots, K\}, c_{(l)}^m \subset C^k\right\}\right) \tag{4.8}$$

Figure 4.4.3 shows the construction of the groups using this clustering approach. Each pathology has its own cluster at the beginning, as illustrated on the left. As $K = 4$ and $M_P = 14$, we do not need to go further than the 10^{th} iteration of Algorithm 4.4.2. The numbers in the intermediate nodes represent the iterations of the algorithm at which this aggregation occurred.

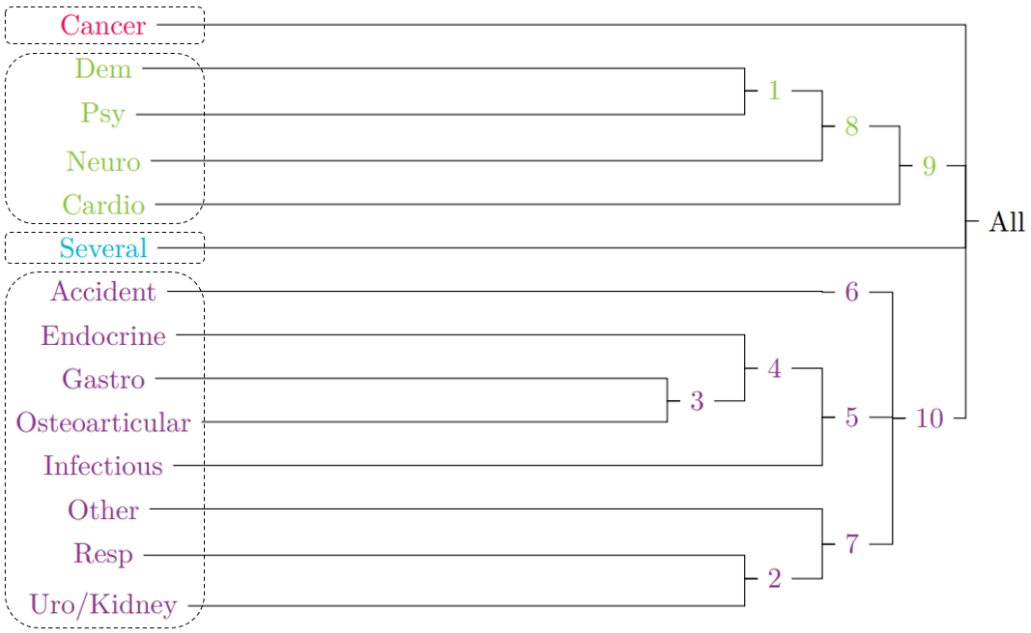


Figure 4.4.3 – Dendrogram obtained with the conditional probability approach ($K = 4$, Poisson GLM, Formula $= x * Gender + I(x^2) * Gender + t * Gender + I(t^2) * Gender + I(t^3) * Gender$")

Despite having the same number of final clusters as the GLM tree approach, this method yields a different distribution of the pathologies within the groups. As anticipated in Section 4.4.2.3, the algorithm first groups dementia and psychiatric diseases. In the second iteration, respiratory, urological and kidney diseases are aggregated. As shown by analysis of the similarity matrix given in Table 4.4.1, cancer seems to greatly differ from other pathologies and therefore has its own cluster. The final clusters are circled using different colours in Figure 4.4.3. Despite being associated with infectious diseases in 60% of the K-means iterations, the group "Several", corresponding to the case where the disabled policyholder has multiple diseases, has its own final cluster.

Since the resulting groups are different from those obtained with the GLM tree method, the predictive performances of these two approaches will be further studied in Section 4.5.2.

Second method of updating the similarity between clusters

In the second approach, the distance between two temporary clusters $c_{(l)}^{m}$ and $c_{(l)}^{\tilde{m}}$ is assessed by the maximum distance between an element of $c_{(l)}^{m}$ and an element of $c_{(l)}^{\tilde{m}}$. This method is implemented in **R** within the **stats** library. The function **hclust()** implements hierarchical clustering algorithms based on a set of dissimilarities. The similarity matrix constructed in Section 4.4.2.2 is transformed into a dissimilarity or distance matrix $\mathcal{D} = 1 - \mathcal{S}$. While several agglomeration methods are provided in this function, corresponding to different ways of recomputing the distances at each step, the complete linkage method assesses the distances between clusters according to the maximum distance between two elements of these clusters, as is wanted here. More details about the **hclust()** function are provided in RDocumentation (2023). The resulting dendrogram, obtained with $K = 4$, is plotted in Figure 4.4.4. As anticipated in Section 4.4.2.2 by analysing the values of the similarity matrix $\mathcal{S}$, cancer has its own cluster because of the specificity of its mortality surface. Moreover, we are not surprised to see that psychiatric diseases and dementia are in the same group, and that urological, kidney and respiratory diseases are all included in the same cluster. While endocrine, gastrointestinal, osteoarticular, and infectious diseases are grouped in the same cluster, respiratory diseases that also look similar to endocrine diseases are grouped with urological and kidney disorders instead. This can be explained by a greater similarity with urological and kidney disorders and low similarity between respiratory diseases and all other pathologies associated with endocrine diseases.

This approach for updating the similarity measure at each step of the Algorithm 4.4.2 yields the same final groups as does the GLM tree approach.

4.5 Choice of the number of clusters in the generalized K-means methods and comparison between all clustering approaches

In this section, we compare the performances of the three clustering methods. We start by discussing the choice of the number of clusters K in the two generalized K-means approaches. Many insurance and reinsurance companies rely on expert judgments to cluster pathologies instead of using clustering algorithms. How would an expert aggregate the 14 pathologies examined in this paper to create 4 homogeneous groups in terms of mortality? We asked an epidemiologist working for the reinsurer SCOR to construct clusters of pathologies with similar mortality rates, based on her experience and expert

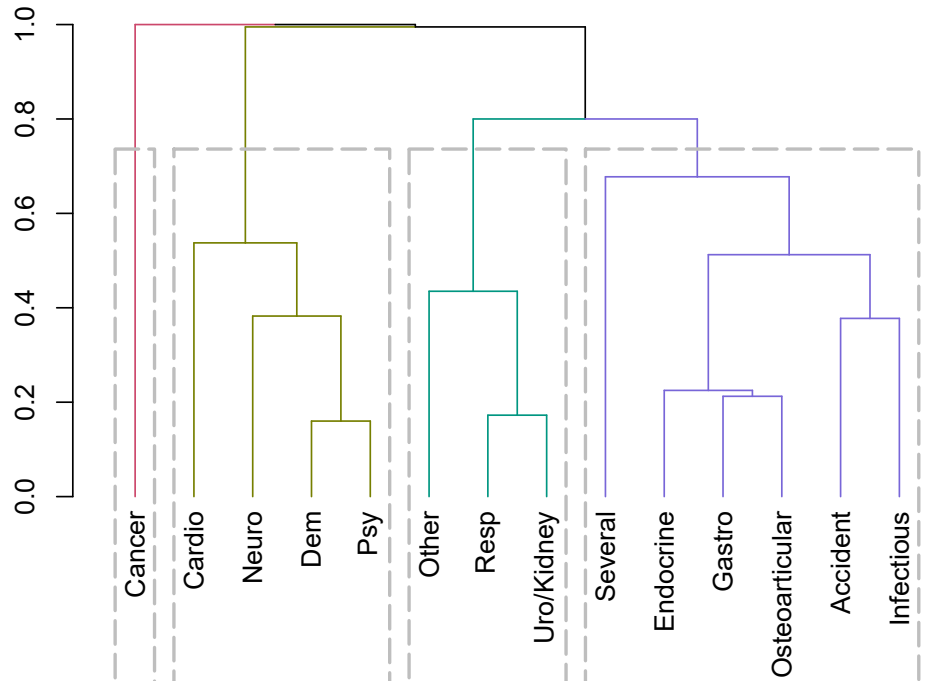


Figure 4.4.4 – Dendrogram obtained with the complete linkage method from **hclust()** ($K = 4$, Poisson GLM, Formula $= x * Gender + I(x^2) * Gender + t * Gender + I(t^2) * Gender + I(t^3) * Gender$")

judgment. We want to compare the performance of the model using clusters produced by expert judgment to that of the models using groups from the three clustering methods presented in this paper. After selecting the optimal hyperparameter K, the performances of all the clustering methods including expert judgment clustering, are compared. For each clustering method, mortality rates are estimated using the resulting groups of pathologies. The goodness of fit of each mortality model is then assessed by the actual-to-expected ratios as in Section 4.2.2, and the Bayesian information criterion. The fitted mortality laws of the best model according to the metrics used for the performance comparison, are then analysed in Section 4.5.3.

4.5.1 Choice of the number of clusters

The generalized K-means algorithm requires to set the number of final clusters K in advance. This choice therefore has an impact on the final groups and on the estimation of mortality for each pathology. Several methods can be used to select the optimal hyperparameter K. As in Zhang and Lin (2021), our optimization relies on the minimization of the Bayesian information criterion (BIC) proposed by Schwarz (1978), which is given by

$$BIC = -2\log(L) + \log(N)\Lambda, \tag{4.9}$$

where:

- L denotes the associated maximum likelihood of the model,
- N denotes the number of individuals in the database,
- Λ denotes the number of parameters of the model.

Since the same GLM formula is used for all clusters, Λ is proportional to the number of clusters K in our case. Increasing the number of final clusters leads to an increase in the number of parameters, resulting in an increase in the log-likelihood due to greater flexibility. However, adding complexity can lead to overfitting. The goal of generalized information criterions as the BIC is to find a trade-off between the simplicity and the goodness of fit of the model by adding a penalty that increases with the number of estimated parameters Λ.

Figure 4.5.1 shows the evolution of the BIC with K for the two aggregation methods based on generalized K-means. As we can see, allowing for two mortality tables instead of a single common mortality table for all pathologies leads to a significant decrease in the BIC. After decreasing with K, the BIC starts to increase. BIC reaches a minimum for $K = 3$ for the two distance update formulas. Considering only this statistic would lead us to choose $K = 3$. However, the difference in the BIC between $K = 3$ and $K = 4$ is very small. Moreover, for comparison purposes, we prefer to have the same number of final clusters K for all methods, including the GLM tree approach. Therefore, $K = 4$ in the remainder of the paper.

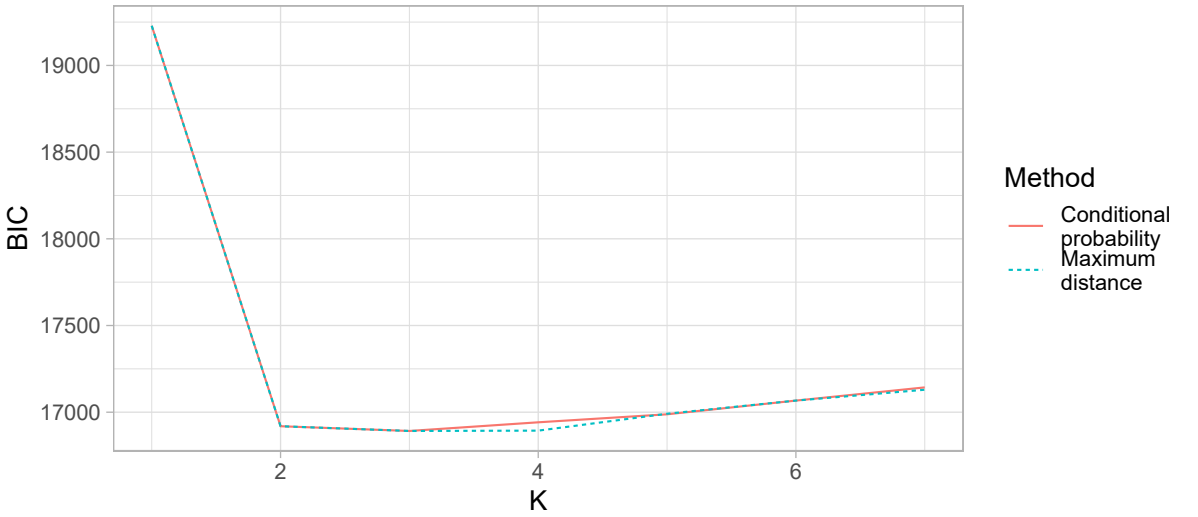


Figure 4.5.1 – Variation in the Bayesian information criterion with the number of final clusters K

4.5.2 Comparison of the methods: Goodness of fit

To assess the goodness of fit of each model fitted with the resulting clusters of each clustering method, we first analyse the ratios of actual over expected counts of deaths by gender and pathology. We then compare the Bayesian information criterion computed for each clustering method. Since the GLM tree yields the exact same final groups as the generalized K-means algorithms combined with the complete linkage method for the distance update, only three models are compared:

1. the GLM tree and generalized K-means with complete linkage method,
2. generalized K-means combined with the conditional probability method to update the distances, and
3. the expert judgement clustering.

4.5.2.1 Comparison of the ratios of actual over expected counts of deaths

After estimating mortality rates by using the groups of pathologies resulting from each clustering methods, the performance of each method is assessed by comparing the expected counts of deaths in the portfolio to the observed counts. The ratios of actual over expected counts of deaths by pathology and gender are plotted in Figure 4.5.2 for each clustering method, and can be compared to those obtained in the model that does not account for pathology plotted in Figure 4.2.1. A model has good predictive performance if the observed number of deaths is close to the expected number of deaths, that is, if the actual-to-expected ratio is close to 1 for each pathology.

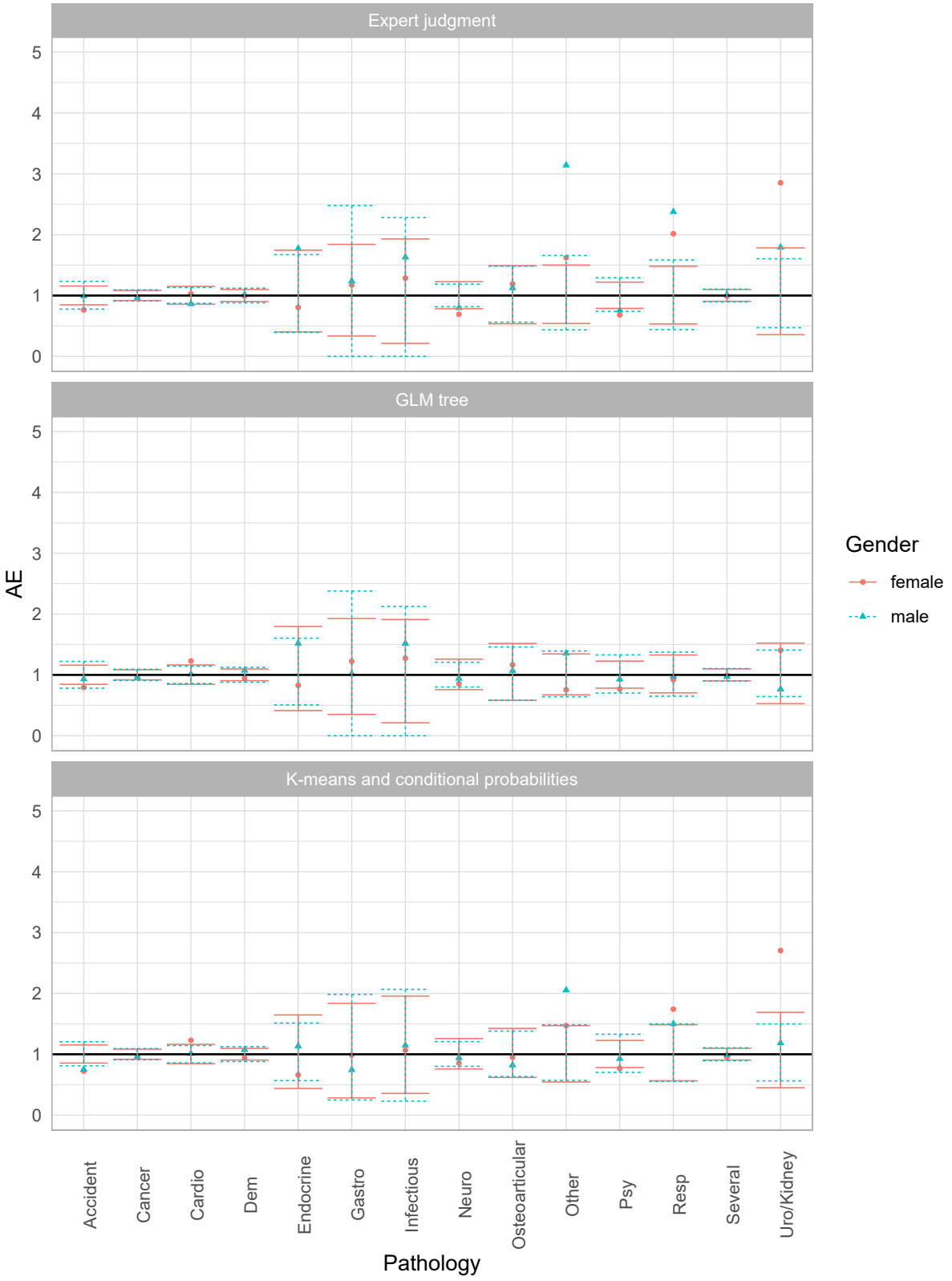


Figure 4.5.2 – Comparison of the ratios of actual over expected counts of deaths for all clustering methods

Figure 4.5.2 shows that the 3 models accounting for pathology, regardless of the clustering method, seem to improve the predictive performance. Let D^p denote a Poisson random variable representing the total number of deaths for pathology p in the database. We want to test the hypothesis

$$\mathcal{H}_0 : D^p \sim Poisson(E^p), \text{ where } E^p \text{ denotes the total expected count of deaths}$$

for each pathology p, with a significance level $\alpha = 5\%$.

Given that only one observation of D^p is available for each pathology p, the well-known chi-squared test, commonly used for Poisson distribution testing, is not suitable. Therefore, another statistical test based on the ratios of actual over expected counts of deaths is constructed to decide whether to reject the null hypothesis $\mathcal{H}_0$ for each pathology.

Let $q_\gamma(E)$ denote the quantile of a Poisson distribution of mean E. $q_\gamma(E)$ is the smallest integer d such that $\mathbb{P}(D \leqslant d) \geqslant \gamma$.

Under $\mathcal{H}_0$,

$$\mathbb{P}\left(\frac{q_{\alpha/2}(E^p)}{E^p} \leqslant \frac{D^p}{E^p} \leqslant \frac{q_{1-\alpha/2}(E^p)}{E^p}\right) \geqslant 1-\alpha,$$

for each pathology p.

A detailed proof of this inequality is given in Appendix 4.A.

Figure 4.5.2 shows the bounds of this interval for each pathology p. $\mathcal{H}_0$ is rejected for pathology p, with the signifance level $\alpha = 5\%$, if the ratio of actual over expected counts of deaths is not included in the interval.

Considering the 14 pathologies and two genders, a total of 28 hypotheses were tested for each model. A summary of the statistical hypothesis tests is given in Table 4.5.1. For each method, this table summarises the number of rejected hypotheses. Hypothesis $\mathcal{H}_0$ is rejected in 64.3% of the combinations of gender and pathology for the model not accounting for the pathology information. Moreover, the test rejects the hypothesis for highly represented pathologies such as cancer and dementia. The model leading to the fewest rejected hypotheses is the one using the GLM tree clustering method. Only 10.7% of the tests lead to a rejection of $\mathcal{H}_0$. Moreover, $\mathcal{H}_0$ is rejected for females for accident, cardiovascular diseases and psychiatric diseases, representing 6.8%, 5.6% and 2% of the observed claims, respectively. The two other clustering methods also lead to rejection of $\mathcal{H}_0$ for females for accidents and cardiovascular and psychiatric diseases, as well as other combinations of gender and pathology.

Method	Rejected	Not rejected	Percent rejected (%)
Without pathology	18	10	64.3
GLM tree	3	25	**10.7**
K-means combined with conditional probability	9	19	32.1
Expert judgment	12	16	42.9

Table 4.5.1 – Summary of the statistical hypothesis testing

Therefore, based on these performance metrics, the GLM tree clustering approach seems to be the most suitable method for clustering pathologies in the context of modelling the mortality of disabled policyholders.

4.5.2.2 Comparison of the Bayesian information criterion

A common metric for comparing models is the Bayesian information criterion (BIC), which was previously used in this paper to choose the optimal number of clusters in Section 4.5.1. The BIC was computed for each clustering method, as well as for the model not accounting for pathology in the estimation of mortality. The results are summarised in Table 4.5.2. The lower BIC is, the better the fitting performance of the model. All models accounting for pathology through clusters have lower BIC values than the model fitting a single mortality law, regardless of the pathology.

Based on this performance metric, the best model is the one using the clusters resulting from the GLM tree method presented in Section 4.4.1 or from the generalized K-means with the complete linkage method presented in Section 4.4.2.

Method	BIC
Without pathology	19,228.78
GLM tree	**16,893.62**
K-means combined with conditional probability	16,941.71
Expert judgment	17,052.63

Table 4.5.2 – Comparison of the Bayesian information criterion before and after clustering for each method

Therefore, the model using clusters obtained with the GLM tree method and the generalized K-means combined with the complete linkage method is considered the best model in the remainder of the paper.

4.5.3 Fitted mortality laws resulting from the best model

The fitted mortality surfaces resulting from the best model are plotted in Figure 4.5.3. We recall that although the clusters are identical for males and females, mortality differs with gender. With 4 clusters, we therefore have 8 fitted mortality surfaces depending on attained age and duration, organised as follows: each row represents a cluster, and each column represents a gender. As we can see, the force of mortality associated with cancer is especially high for short durations compared to that associated with other pathologies. Two years after the loss of autonomy, the excess mortality of disabled policyholders with cancer seems to decrease. Dementia and cardiovascular, psychiatric and neurological diseases seem to be the pathologies associated with the lowest mortality after the loss of autonomy for both genders. Disabled policyholders affected by respiratory, urological and kidney diseases are more likely to experience mortality than are those who are disabled because of accidents; endocrine, gastrointestinal, infectious and osteoarticular diseases; or several diseases. The mortality rates resulting from the other clustering approaches are plotted in Appendix 4.B.

4.6 Actuarial application

Using the best model chosen in Section 4.5.2, we can derive some actuarial consequences of accounting for pathology when estimating the mortality of disabled policyholders. We start by estimating the impact of the pathology information on reserving. Then, by accounting for the heterogeneity of mortality between pathologies by deriving 4 different mortality tables associated with the 4 clusters, we can estimate the impact of a change in the incidence rates of a given pathology p on the global mortality in LTC.

4.6.1 Reserving

Let us consider a LTC insurance product that pays a monthly annuity of $R = 1000€$ in the event of loss of autonomy. The loss of autonomy is assumed to be permanent. Therefore, the only cause of ending annuity payments is the death of the disabled policyholder. Let $i = 3\%$ denote the discount rate. Let $\mu^g_{x,t}$ denote the mortality intensity of a disabled insured individual at age x, knowing his or her gender g and the time since the loss of autonomy t. Let $\mu^{g,p}_{x,t}$ denote the mortality intensity knowing pathology p in addition to the other covariates x, t and g. The actuarial valuation of a lifetime annuity paid to a newly disabled policyholder of age x and gender g is given by the following formula:

$$\Pi^g_x = \sum_{t=1}^{+\infty} \frac{1}{(1+i)^{\frac{t}{12}}} \exp\left(-\frac{1}{12}\sum_{k=0}^{t-1} \mu^g_{x+\frac{k}{12},\frac{k}{12}}\right) R. \tag{4.10}$$

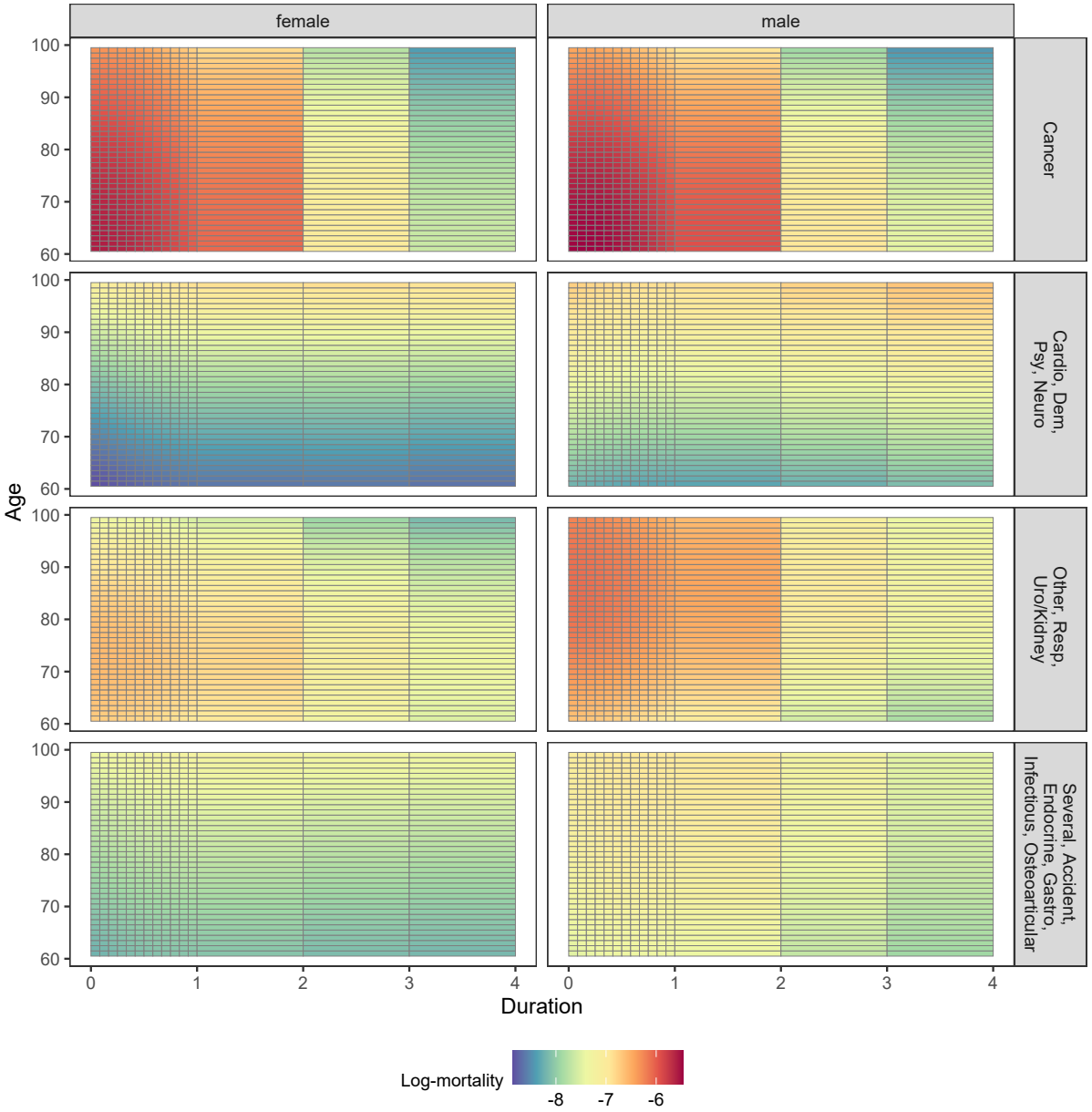


Figure 4.5.3 – The fitted mortality rates associated to each cluster resulting from the GLM tree or generalized K-means combined with the complete linkage method for updating similarity between clusters

Knowing the pathology p responsible for his or her loss of autonomy, the valuation of the lifetime annuity paid to the newly disabled policyholder of age x and gender g is given by the following formula:

$$\Pi_x^{g,p} = \sum_{t=1}^{+\infty} \frac{1}{(1+i)^{\frac{t}{12}}} \exp\left(-\frac{1}{12}\sum_{k=0}^{t-1} \mu^{g,p}_{x+\frac{k}{12},\frac{k}{12}}\right) R. \tag{4.11}$$

Ignoring information about pathology when estimating the mortality of disabled policyholders, an insurer would consider same valuations of the lifetime annuities of all newly disabled policyholders of same age and gender, regardless of the pathologies that caused their loss of autonomy. Let us consider four women who lost their autonomy at age $x = 75$ for distinct causes. The four considered causes of disability are cancer, dementia, respiratory disease and osteoarticular disease. According to the resulting clusters of the GLM tree method, these four pathologies have different mortality rates.

Without accounting for the pathology, the valuations of the lifetime annuities for these four disabled policyholders, computed with Equation 4.10, are equal to 35,121.68 €.

Accounting for the pathology, the valuations of the annuities of these four disabled policyholders given the pathology are shown in Table 4.6.1.

Pathology	$\mathbf{\Pi_x^{g,p}}$	$\mathbf{\frac{\Pi_x^{g,p} - \Pi_x^g}{\Pi_x^g}}$
Cancer	12,009.41€	−65.8%
Dementia	47,130.45€	+34.2%
Resp	23,473.14€	−33.1%
Osteoarticular	40,609.34€	+15.6%

Table 4.6.1 – Valuation of a lifetime annuity of a disabled woman knowing the pathology

Table 4.6.1 shows that the reserves required for newly disabled policyholders of the same age at loss of autonomy strongly depend on the pathology causing their disability. An insurer not accounting for the pathology when estimating reserves would strongly underestimate the resources needed to meet the costs of the loss of autonomy of an insured policyholder with dementia or osteoarticular diseases, as well as any pathology belonging to the same cluster (i.e., accidents; cardiovascular, neurological, endocrine, gastrointestinal, infectious and psychiatric diseases; and several diseases). In contrast, cancer is an aggressive disease that leads to high mortality, especially in the first year following the loss of autonomy, as shown in Figure 4.5.3. Therefore, an insurer ignoring pathology information would strongly overestimate the resources needed to meet the costs of the annuity of a disabled policyholder for whom cancer is identified as the cause of loss of autonomy. Considering the pathology, the reserves needed for a policyholder disabled because of cancer are 65.8%

lower than the reserves estimated without the information of the pathology.

4.6.2 Application of a shock

Assuming a unique mortality table for all disabled insured individuals can be sufficient for an insurer if the distribution of the pathologies in the portfolio of disabled policyholders remains the same. The unique mortality table of the disabled individuals can be seen as a weighted average of the mortality of each pathology. The weights are given by the distribution of each pathology at each age and duration. However, due to the heterogeneity of mortality among the different pathologies, a change in the distribution of the pathologies has an impact on the resulting global mortality in LTC. Estimating a unique mortality table regardless of the pathology does not enable an insurer to estimate the impact of a change in the incidence rates associated with a specific pathology. A reduction or increase in the mortality of a specific pathology would also change the distribution of the pathologies in the portfolio, resulting in a modification of the global mortality in LTC.

In this section, we focus on the impact of a reduction or increase in the incidence of a certain pathology. This may occur in the event of preventive actions proposed by the insurer to reduce the incidence of certain pathologies.

Figure 4.6.1 shows the impact of an increase or decrease of 10% in the incidence of cancer at each age on mortality during the second year. Reducing the incidence of cancer implies a decrease in the prevalence of this pathology among the disabled policyholders. Since these insured individuals have a much greater mortality than other disabled policyholders, this change in the incidence implies a decrease in the global mortality in LTC observed in the portfolio. In contrast, increasing the incidence of cancer leads to an increase in the mortality intensity estimated for the second year of disability. As the difference in mortality between clusters is more pronounced at a young age than at greater ages, a variation in the incidence of cancer has a greater impact on global mortality in LTC, all pathologies combined.

The impact of a change in the incidence of dementia at each age on mortality during the second year is plotted in Figure 4.6.2. Unlike cancer, dementia is one of the pathologies associated with the highest life expectancy following the loss of autonomy. Therefore, increasing the incidence of dementia increases its prevalence among disabled policyholders, leading to a decrease in the global mortality in LTC.

The smile shape of the mortality curve observed in Figure 4.6.1 and Figure 4.6.2 is often observed when estimating the mortality of disabled policyholders for a fixed duration t, especially for low durations. This phenomenon has already been observed in French LTC portfolios, as in Le Bastard et al. (2023). This is mostly due to the evolution of the distribution of pathologies with age among disabled policyholders. The resulting mortality

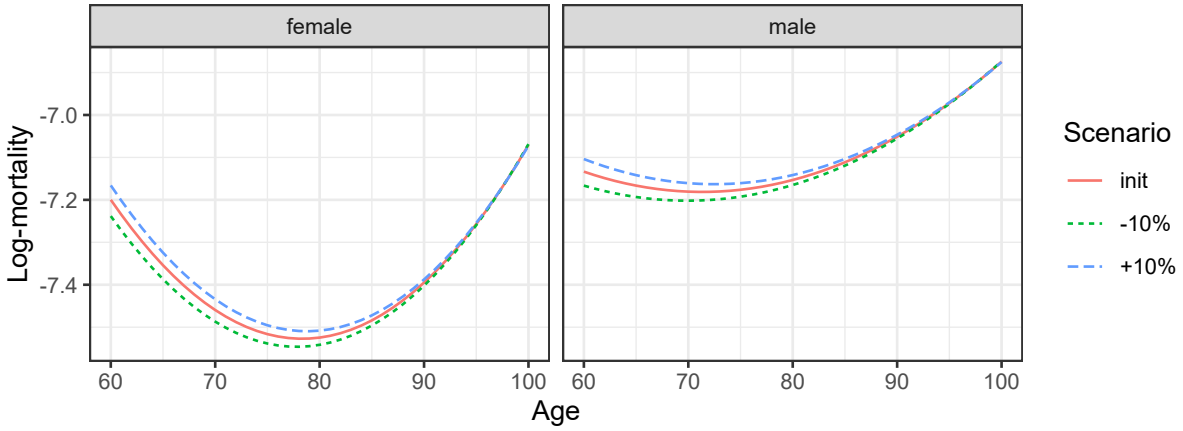


Figure 4.6.1 – Shocks applied to the incidence of cancer for females: impact on the global mortality in LTC during the second year following the loss of autonomy ($\mu^f_{x,1}$)

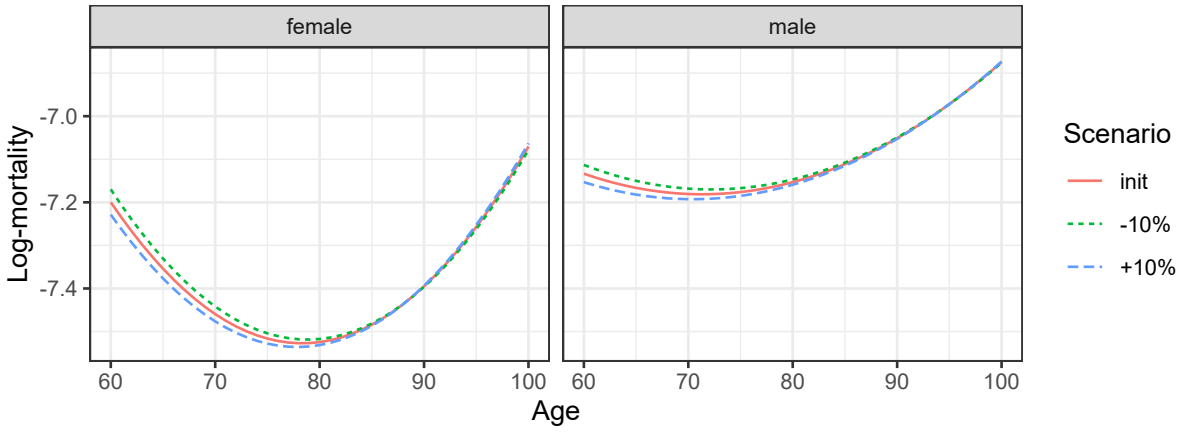


Figure 4.6.2 – Shocks applied to the incidence of dementia for females: impact on the global mortality in LTC during the second year following the loss of autonomy ($\mu^f_{x,1}$)

in LTC during the second year of disability, if the loss of autonomy due to cancer was not covered by the insurer, is plotted in Figure 4.6.3. The smile shape disappears by setting the incidence of cancer to 0, showing that this specific pattern comes from the prevalence of cancer among the disabled policyholders at young ages.

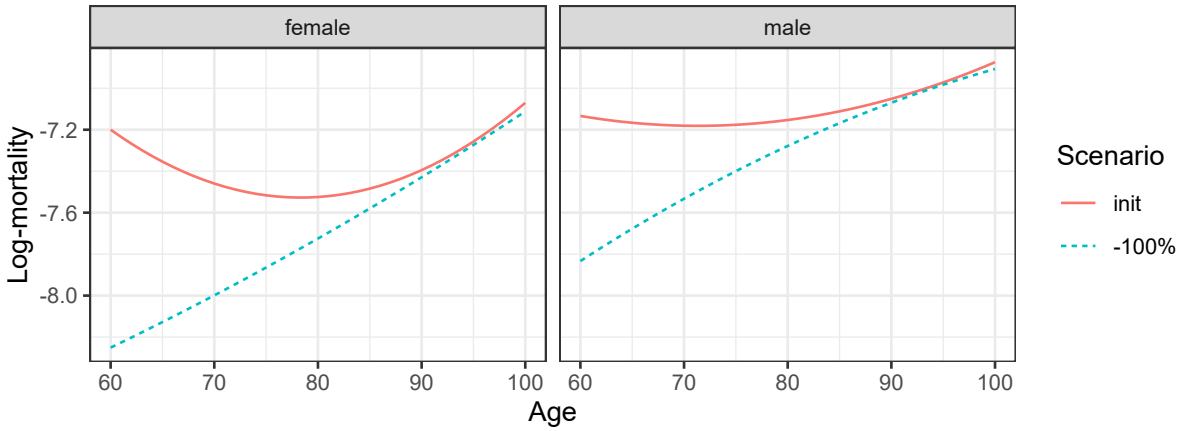


Figure 4.6.3 – Mortality in LTC during the second year following the loss of autonomy if cancer is excluded from the possible causes of disability

4.7 Discussion

In the context of studying long-term care insurance products, multiple pathologies can be the cause of an individual's loss of autonomy. Since LTC insurance datasets containing pathology information are rare, literature about the impact of pathology on the mortality of disabled policyholders is still scarce. In this paper, we rely on data from a health fund. After demonstrating the heterogeneity of the force of mortality induced by the plurality of causes of disability, this paper studies two clustering methods to determine which pathologies lead to similar mortality rates after the loss of autonomy. In fact, although accounting for pathology when estimating the mortality of disabled policyholders seems important for capturing the variance, the lack of data does not enable us to independently estimate specific mortality tables depending on gender, attained age and duration of the claim for each pathology. It is therefore necessary to construct groups of homogeneous pathologies. For the sake of simplicity and to reduce the complexity of the models, groups of pathologies are assumed to be identical for both males and females. Despite having the same clusters, the force of mortality is still estimated separately for each gender to account for the heterogeneity between males and females.

Both methods presented in this paper rely on generalized linear models. For each gender and cluster, we assume a quadratic effect of the attained age and a cubic effect of the duration of the claim. The first method uses GLM trees implemented in the **partykit** package in the statistical software **R**. The second method developed in this paper is

a generalized K-means approach, resulting in the construction of a similarity measure between pairs of pathologies. Optimal clusters are subsequently constructed using bottom-up hierarchical algorithms. Starting with one cluster for each pathology, each step leads to the merging of the two closest clusters based on a distance measure. The distance measures between each pair of clusters are then updated after each step. Two formulas for updating the distances are tested in this paper. The first relies on the theory of conditional probabilities. The second one, called the complete linkage method, is implemented with the **stats** package in **R**.

As of today, because of the lack of literature on clustering pathologies for long-term care insurance products, most insurers rely on expert judgment to group them. The three clustering methods proposed in this paper, as well as the expert judgment approach, are compared in terms of the goodness of fit of the resulting mortality models. To do this, actual and estimated counts of deaths are compared for each pathology for each clustering approach. The performances of the resulting models are also compared using the Bayesian information criterion. We show that the three clustering methods proposed in this paper yield better fitting performance than does the expert judgment approach. Indeed, experts perform reasoning based on their preconceived view regarding the distribution of the severity of the pathologies in each subgroup of pathologies. An interesting example is that all types of cancer do not have the same severity. Therefore, the clustering is biased by their preconceived view. Using clustering methods based on the portfolio's experience enables the creation of homogeneous groups of pathologies with similar mortality rates, while avoiding the potentially biased expert judgment.

Finally, using the best model based on the two previous performance metrics, we show through actuarial applications the benefits for an insurer of accounting for pathology when estimating the mortality of disabled insureds.

Tables containing the coefficients of the Poisson GLMs fitted using the resulting groups from each clustering method can be found in Appendix 4.C. The p-values presented in the fifth column show the significance of the coefficients related to the variable indicating the cluster, as well as the coefficients associated with interactions of variables involving the cluster. This information again highlights the importance of accounting for pathology as a covariate for predicting mortality.

In the context of modelling LTC products, P-splines are used to fit mortality in LTC in several papers such as Le Bastard et al. (2023), which enables to consider more complex structure of the surface that describes mortality. As mortality rates vary greatly between different pathologies, the distribution of pathologies among disabled insureds in the portfolio significantly impacts the global mortality in LTC. Although the shape of the mortality surface in LTC is complex, part of this complexity comes from the variation in

the distribution of pathologies across age and duration. Therefore, a simpler impact of age and duration given the pathology is assumed in this paper, by considering an additive impact of age and duration. A cross effect of age and duration has been considered, but did not improve the performance of the models. Access to data containing more observations would allow us to consider using P-splines.

We remind the reader that this work relies on data coming from a foreign market having a specific definition of the LTC insurance. The resulting groups of pathologies may not be transposable to other countries for two main reasons: the heterogeneity of the populations and potential differences in the definition of the loss of autonomy. However, most current data in France do not include detailed information on the pathology. While our work provides an initial insight on groups of similar pathologies, future research should study the sensitivity of the constructed groups of pathologies with respect to the country.

Future research should also consider different groups of pathologies for males and females. Another idea would be to consider different groupings depending on attained age and duration. To this end, we suggest using the survival trees described in Bou-Hamad et al. (2011). In this method, all covariates can be partitioned recursively to create groups of homogeneous observations. Additionally, future research could also consider a joint mortality model to consider the dependence between mortality laws of different pathologies, as proposed in Debón et al. (2011) to account for the interdependence between mortality tables in several Spanish regions.

Appendices

Appendix 4.A Proof of the statistical hypothesis test for the actual-to-expected ratios

We want to test the hypothesis

$$\mathcal{H}_0 : D \sim Poisson(E) \text{ where } E \text{ denote the expected count of deaths,}$$

with a significance level α.

Let $\gamma \in [0; 1]$. Let $q_\gamma(E)$ denotes the quantile of a Poisson distribution with parameter E. $q_\gamma(E)$ is the smallest integer d such that

$$\mathbb{P}\left(D \leqslant d\right) \geqslant \gamma.$$

Therefore we have

$$\mathbb{P}\left(D < q_{\tilde{\gamma}}(E)\right) < \tilde{\gamma}.$$

Let γ and $\tilde{\gamma}$ denote two probability values; then,

$$\begin{aligned}\mathbb{P}\left(q_{\tilde{\gamma}}(E) \leqslant D \leqslant q_\gamma(E)\right) &= \mathbb{P}\left(D \leqslant q_\gamma(E)\right) - \mathbb{P}\left(D < q_{\tilde{\gamma}}(E)\right) \\ &\geqslant \gamma - \tilde{\gamma}.\end{aligned}$$

Let $\gamma = 1 - \alpha/2$ and $\tilde{\gamma} = \alpha/2$; then,

$$\mathbb{P}\left(q_{\alpha/2}(E) \leqslant D \leqslant q_{1-\alpha/2}(E)\right) \geqslant 1 - \alpha,$$

and

$$\mathbb{P}\left(\frac{q_{\alpha/2}(E)}{E} \leqslant D \leqslant \frac{q_{1-\alpha/2}(E)}{E}\right) \geqslant 1 - \alpha. \tag{4.12}$$

Therefore, the null hypothesis $\mathcal{H}_0$ is rejected if 1 is not included in the interval $\left[\frac{q_{\alpha/2}(E)}{E}; \frac{q_{1-\alpha/2}(E)}{E}\right]$, with a significance level α.

Appendix 4.B Plots of mortality rates resulting from the nonselected clustering approaches

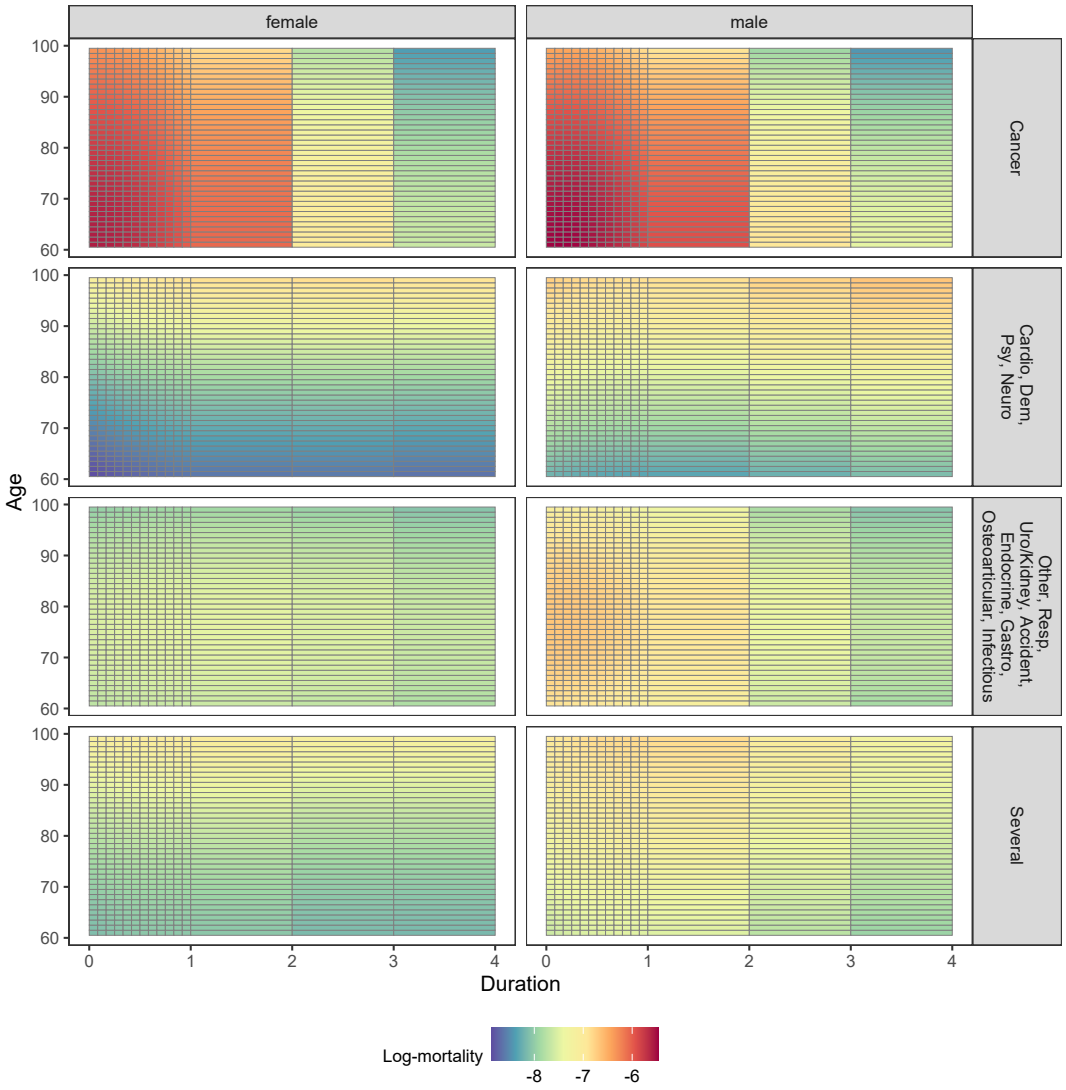


Figure 4.B.1 – The fitted mortality rates associated to each cluster resulting from the generalized K-means algorithm with distance measure using conditional probabilities

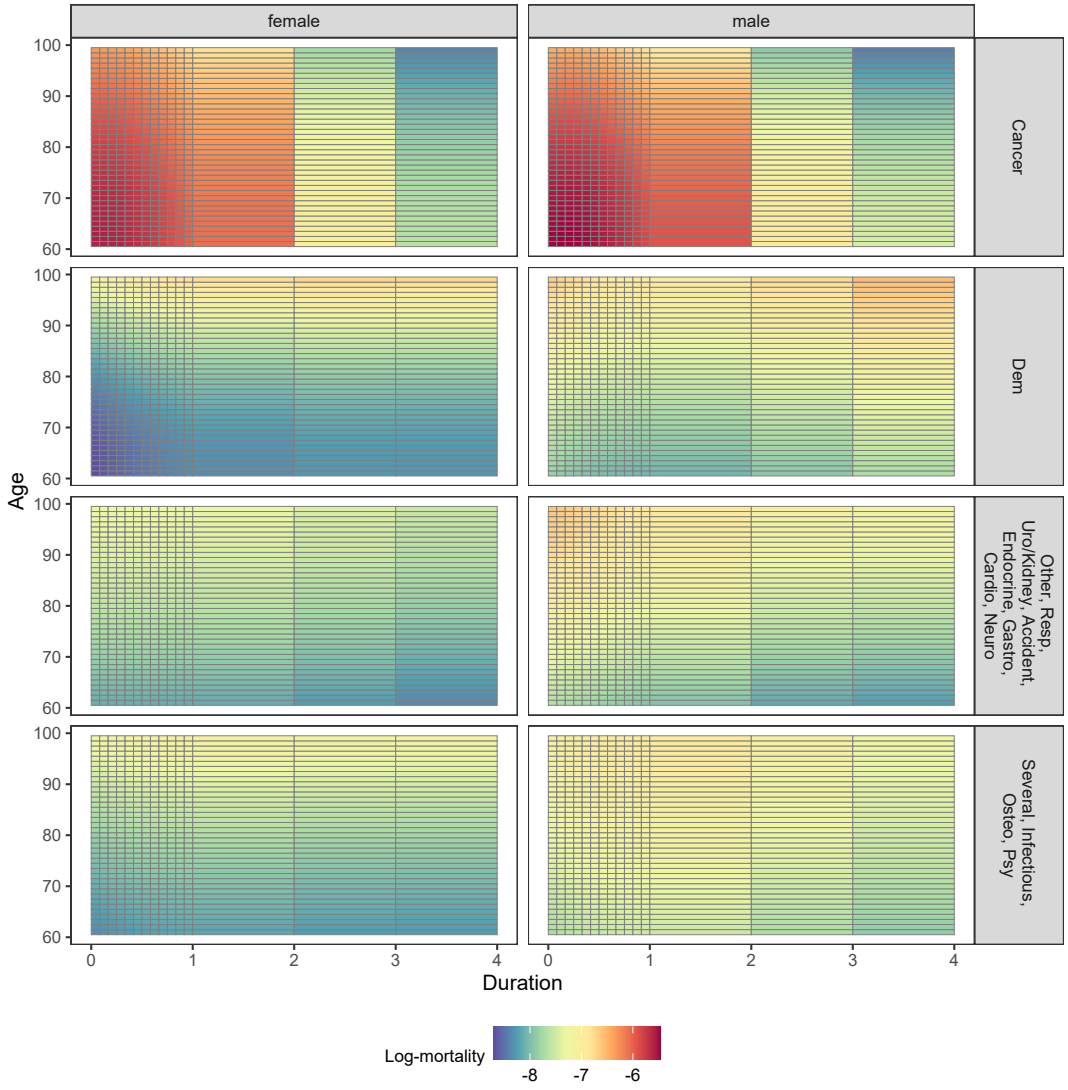


Figure 4.B.2 – The fitted mortality rates associated to each cluster resulting from the expert judgment approach

Appendix 4.C Coefficients of the Poisson GLM using pathology groups from each clustering methods

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-9.60	0.96	-10.05	0.00	***
x	-0.01	0.02	-0.35	0.73	
Gendermale	-2.54	1.67	-1.52	0.13	
cluster2	0.48	1.25	0.38	0.70	
cluster3	2.40	1.03	2.32	0.02	*
cluster4	-2.72	1.39	-1.95	0.05	·
I(x^2)	0.00	0.00	2.11	0.03	*
t	0.43	0.23	1.90	0.06	·
I(t^2)	-0.21	0.13	-1.63	0.10	
I(t^3)	0.03	0.02	1.48	0.14	
x:Gendermale	0.09	0.04	2.08	0.04	*
x:cluster2	0.02	0.03	0.60	0.55	
x:cluster3	0.06	0.03	2.30	0.02	*
x:cluster4	0.17	0.04	4.18	0.00	***
Gendermale:cluster2	2.15	1.96	1.10	0.27	
Gendermale:cluster3	2.09	1.76	1.19	0.23	
Gendermale:cluster4	2.23	2.16	1.03	0.30	
Gendermale:I(x^2)	0.00	0.00	-2.18	0.03	*
cluster2:I(x^2)	0.00	0.00	-1.29	0.20	
cluster3:I(x^2)	0.00	0.00	-4.24	0.00	***
cluster4:I(x^2)	0.00	0.00	-4.78	0.00	***
Gendermale:t	-0.74	0.33	-2.21	0.03	*
cluster2:t	-0.11	0.31	-0.34	0.73	
cluster3:t	-0.43	0.31	-1.36	0.17	
cluster4:t	-0.52	0.65	-0.80	0.42	
Gendermale:I(t^2)	0.43	0.20	2.18	0.03	*
cluster2:I(t^2)	-0.01	0.18	-0.05	0.96	
cluster3:I(t^2)	-0.41	0.22	-1.85	0.06	·
cluster4:I(t^2)	0.03	0.44	0.07	0.94	
Gendermale:I(t^3)	-0.07	0.03	-2.06	0.04	*
cluster2:I(t^3)	0.01	0.03	0.27	0.79	
cluster3:I(t^3)	0.10	0.04	2.51	0.01	*
cluster4:I(t^3)	0.01	0.08	0.12	0.91	
x:Gendermale:cluster2	-0.05	0.05	-1.05	0.29	
x:Gendermale:cluster3	-0.07	0.05	-1.53	0.13	
x:Gendermale:cluster4	-0.10	0.06	-1.60	0.11	
Gendermale:cluster2:I(x^2)	0.00	0.00	0.95	0.34	
Gendermale:cluster3:I(x^2)	0.00	0.00	1.45	0.15	
Gendermale:cluster4:I(x^2)	0.00	0.00	1.87	0.06	·
Gendermale:cluster2:t	0.82	0.46	1.78	0.08	·
Gendermale:cluster3:t	0.91	0.49	1.86	0.06	·
Gendermale:cluster4:t	0.76	0.89	0.85	0.39	
Gendermale:cluster2:I(t^2)	-0.61	0.28	-2.19	0.03	*
Gendermale:cluster3:I(t^2)	-0.60	0.37	-1.62	0.11	
Gendermale:cluster4:I(t^2)	-0.63	0.60	-1.05	0.29	
Gendermale:cluster2:I(t^3)	0.10	0.05	2.19	0.03	*
Gendermale:cluster3:I(t^3)	0.11	0.07	1.55	0.12	
Gendermale:cluster4:I(t^3)	0.12	0.10	1.12	0.26	

Table 4.C.1 – Coefficients of the Poisson GLM using clusters from the GLM trees

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-12.21	0.82	-14.86	0.00	***
x	0.12	0.02	5.29	0.00	***
Gendermale	0.31	1.02	0.30	0.76	
cluster2	5.00	0.91	5.49	0.00	***
cluster3	2.61	1.26	2.07	0.04	*
cluster4	4.65	1.27	3.66	0.00	***
I(x^2)	0.00	0.00	-4.79	0.00	***
t	0.18	0.35	0.50	0.61	
I(t^2)	-0.11	0.22	-0.49	0.62	
I(t^3)	0.01	0.04	0.34	0.73	
x:Gendermale	0.01	0.03	0.43	0.67	
x:cluster2	-0.07	0.03	-2.57	0.01	*
x:cluster3	-0.13	0.03	-3.89	0.00	***
x:cluster4	-0.15	0.03	-4.58	0.00	***
Gendermale:cluster2	-0.76	1.16	-0.66	0.51	
Gendermale:cluster3	-2.85	1.96	-1.45	0.15	
Gendermale:cluster4	-2.10	1.87	-1.12	0.26	
Gendermale:I(x^2)	0.00	0.00	-0.32	0.75	
cluster2:I(x^2)	0.00	0.00	1.75	0.08	·
cluster3:I(x^2)	0.00	0.00	4.93	0.00	***
cluster4:I(x^2)	0.00	0.00	5.06	0.00	***
Gendermale:t	-0.12	0.49	-0.25	0.80	
cluster2:t	-0.17	0.41	-0.42	0.68	
cluster3:t	0.26	0.42	0.61	0.54	
cluster4:t	0.08	0.43	0.18	0.86	
Gendermale:I(t^2)	-0.26	0.31	-0.85	0.39	
cluster2:I(t^2)	-0.52	0.28	-1.81	0.07	·
cluster3:I(t^2)	-0.10	0.26	-0.41	0.68	
cluster4:I(t^2)	-0.11	0.26	-0.42	0.67	
Gendermale:I(t^3)	0.07	0.05	1.33	0.18	
cluster2:I(t^3)	0.12	0.05	2.33	0.02	*
cluster3:I(t^3)	0.02	0.04	0.42	0.67	
cluster4:I(t^3)	0.03	0.04	0.69	0.49	
x:Gendermale:cluster2	0.01	0.03	0.17	0.87	
x:Gendermale:cluster3	0.08	0.05	1.47	0.14	
x:Gendermale:cluster4	0.05	0.05	1.07	0.28	
Gendermale:cluster2:I(x^2)	0.00	0.00	-0.31	0.75	
Gendermale:cluster3:I(x^2)	0.00	0.00	-1.51	0.13	
Gendermale:cluster4:I(x^2)	0.00	0.00	-1.16	0.25	
Gendermale:cluster2:t	0.30	0.61	0.49	0.62	
Gendermale:cluster3:t	-0.61	0.59	-1.03	0.30	
Gendermale:cluster4:t	0.29	0.62	0.48	0.63	
Gendermale:cluster2:I(t^2)	0.09	0.44	0.21	0.83	
Gendermale:cluster3:I(t^2)	0.69	0.37	1.89	0.06	·
Gendermale:cluster4:I(t^2)	0.12	0.39	0.31	0.76	
Gendermale:cluster2:I(t^3)	-0.03	0.08	-0.36	0.72	
Gendermale:cluster3:I(t^3)	-0.14	0.06	-2.21	0.03	*
Gendermale:cluster4:I(t^3)	-0.05	0.06	-0.77	0.44	

Table 4.C.2 – Coefficients of the Poisson GLM using clusters from the K-means algorithm with the distance measure using conditional probabilities

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-7.20	0.40	-18.23	0.00	***
x	0.06	0.01	4.36	0.00	***
Gendermale	-0.45	0.54	-0.84	0.40	
cluster2	-4.07	0.79	-5.15	0.00	***
cluster3	2.35	3.11	0.75	0.45	
cluster4	-1.84	0.95	-1.93	0.05	·
I(x^2)	0.00	0.00	-4.61	0.00	***
t	0.00	0.22	0.02	0.98	
I(t^2)	-0.62	0.18	-3.47	0.00	***
I(t^3)	0.13	0.03	3.87	0.00	***
x:Gendermale	0.02	0.02	1.06	0.29	
x:cluster2	0.02	0.02	0.67	0.50	
x:cluster3	-0.18	0.08	-2.37	0.02	*
x:cluster4	-0.05	0.03	-2.12	0.03	*
Gendermale:cluster2	0.21	1.06	0.19	0.85	
Gendermale:cluster3	-3.81	4.24	-0.90	0.37	
Gendermale:cluster4	-0.54	1.28	-0.43	0.67	
Gendermale:I(x^2)	0.00	0.00	-1.07	0.28	
cluster2:I(x^2)	0.00	0.00	0.97	0.33	
cluster3:I(x^2)	0.00	0.00	3.18	0.00	**
cluster4:I(x^2)	0.00	0.00	3.61	0.00	***
Gendermale:t	0.17	0.36	0.49	0.63	
cluster2:t	0.04	0.34	0.11	0.92	
cluster3:t	0.59	0.38	1.56	0.12	
cluster4:t	0.33	0.32	1.02	0.31	
Gendermale:I(t^2)	-0.17	0.31	-0.55	0.59	
cluster2:I(t^2)	0.55	0.24	2.28	0.02	*
cluster3:I(t^2)	0.35	0.25	1.40	0.16	
cluster4:I(t^2)	0.40	0.22	1.78	0.08	·
Gendermale:I(t^3)	0.04	0.06	0.67	0.50	
cluster2:I(t^3)	-0.12	0.04	-2.75	0.01	**
cluster3:I(t^3)	-0.09	0.04	-2.07	0.04	*
cluster4:I(t^3)	-0.09	0.04	-2.25	0.02	*
x:Gendermale:cluster2	0.00	0.03	-0.12	0.90	
x:Gendermale:cluster3	0.12	0.11	1.17	0.24	
x:Gendermale:cluster4	0.03	0.03	0.94	0.35	
Gendermale:cluster2:I(x^2)	0.00	0.00	0.44	0.66	
Gendermale:cluster3:I(x^2)	0.00	0.00	-1.22	0.22	
Gendermale:cluster4:I(x^2)	0.00	0.00	-1.01	0.31	
Gendermale:cluster2:t	-0.54	0.51	-1.05	0.29	
Gendermale:cluster3:t	-1.28	0.58	-2.19	0.03	*
Gendermale:cluster4:t	-0.03	0.50	-0.06	0.95	
Gendermale:cluster2:I(t^2)	0.27	0.39	0.70	0.48	
Gendermale:cluster3:I(t^2)	0.78	0.42	1.87	0.06	·
Gendermale:cluster4:I(t^2)	0.04	0.38	0.10	0.92	
Gendermale:cluster2:I(t^3)	-0.05	0.07	-0.64	0.52	
Gendermale:cluster3:I(t^3)	-0.13	0.08	-1.73	0.08	·
Gendermale:cluster4:I(t^3)	-0.02	0.07	-0.30	0.76	

Table 4.C.3 – Coefficients of the Poisson GLM using clusters from expert judgment

Bibliography

Abraham, C., P. A. Cornillon, E. Matzner-Løber, and N. Molinari (2003). Unsupervised curve clustering using B-splines. *Scandinavian Journal of Statistics 30*(3), 581–595.

Achim Zeileis, T. H. and K. Hornik (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics 17*(2), 492–514.

Barndorff, N. (1978). Information and exponential families; in statistical theory. Technical report.

Biessy, G. (2016). A semi-Markov model with pathologies for Long-Term Care Insurance. preprint.

Biessy, G. (2017). Continuous-time semi-markov inference of biometric laws associated with a long-term care insurance portfolio. *ASTIN Bulletin: The Journal of the IAA 47*(2), 527–561.

Biessy, G. (2019). Smoothing of multidimensional biometric laws in a Long-Term Care Insurance portfolio. working paper or preprint.

Bou-Hamad, I., D. Larocque, and H. Ben-Ameur (2011). A review of survival trees. *Statistics Surveys 5*(none), 44 – 71.

Carracedo, P., A. Debón, A. Iftimi, and F. Montes (2018). Detecting spatio-temporal mortality clusters of european countries by sex and age. *International journal for equity in health 17*, 1–19.

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

Debón, A., F. Montes, and F. Martínez-Ruiz (2011). Statistical methods to compare mortality for a group with non-divergent populations: an application to spanish regions. *European Actuarial Journal 1*, 291–308.

Debón, A., L. Chaves, S. Haberman, and F. Villa (2017). Characterization of between-group inequality of longevity in European Union countries. *Insurance: Mathematics and Economics 75*, 151–165.

Fuino, M. and J. Wagner (2018). Long-term care models and dependence probability tables by acuity level: New empirical evidence from switzerland. *Insurance: Mathematics and Economics 81*, 51–70.

Hoem, J. M. (1972). Inhomogeneous semi-Markov processes, select actuarial tables, and duration-dependence in demography. In T. Greville (Ed.), *Population Dynamics*, pp. 251–296. Academic Press.

Hothorn, T. and A. Zeileis (2015). partykit: A modular toolkit for recursive partytioning in R. *The Journal of Machine Learning Research 16*(1), 3905–3909.

Hunt, A. and D. Blake (2021). On the structure and classification of mortality models. *North American Actuarial Journal 25*(sup1), S215–S234.

Janssen, J. (1966). Application des processus semi-markoviens à un problème d'invalidité. *Bulletin de l'Association Royale des Actuaries Belges 63*, 35–52.

Le Bastard, L., S. Loisel, and A. W. Shao (2023). Combining experience data of several Long-Term Care Insurance products with different disability definitions. working paper or preprint.

Léger, A.-E. and S. Mazzuco (2021). What can we learn from the functional clustering of mortality data? An application to the Human Mortality Database. *European Journal of Population 37*.

McCullagh, P. (2019). *Generalized linear models*. Routledge.

Morone, F. (2022). Clustering matrices through optimal permutations. *Journal of Physics: Complexity 3*(3), 035007.

Myers, R. H. and D. C. Montgomery (1997). A tutorial on generalized linear models. *Journal of Quality Technology 29*(3), 274–291.

Nelder, J. A. and R. W. M. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General) 135*(3), 370–384.

Pitacco, E. (2014). Health insurance. *Basic Actuarial Models, Cham, Switzerland: Springer Verlag*.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

RDocumentation (2023). hclust: Hierarchical clustering. https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/hclust.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics 6*(2), 461 – 464.

Soetewey, A., C. Legrand, M. Denuit, and G. Silversmit (2022). Semi-Markov modeling for cancer insurance. *European Actuarial Journal 12*.

Xuanyuan, S. and S. Xuanyuan (2023). Application of Markov model in long-term care insurance. *Highlights in Science, Engineering and Technology 47*, 9–15.

Zhang, T. and G. Lin (2021). Generalized k-means in GLMs with applications to the out-

break of COVID-19 in the United States. *Computational Statistics & Data Analysis 159*, 107217.

Conclusion et perspectives de recherche

La gestion du risque de dépendance est devenue, ces dernières années, l'un des enjeux majeurs pour l'ensemble des pays développés en raison du vieillissement de leurs populations. Ces derniers sont confrontés à une augmentation du pourcentage de la population nécessitant l'aide d'une tierce personne pour les activités basiques de la vie quotidienne. La perte d'autonomie des personnes âgées représente, pour ces pays, une problématique majeure en raison du coût financier qu'elle engendre. C'est dans ce contexte que se sont développés les produits d'assurance dépendance, permettant de couvrir une partie des frais liés à la perte d'autonomie, en complément des aides publiques telles que l'APA [1] versée par la Sécurité sociale. En France, la majorité des contrats prévoient le versement d'une rente jusqu'au décès en cas de perte d'autonomie. En contrepartie, l'assuré verse une cotisation tant qu'il est autonome. En raison du peu d'historique des produits d'assurance dépendance, s'étant essentiellement développés au début des années 2000, les bases de données permettant d'estimer le risque de dépendance se font rares, rendant complexe l'estimation du risque. Cette thèse détaille plusieurs contributions à la modélisation du risque de dépendance en s'attaquant à 3 problématiques distinctes.

En raison de la présence d'un âge limite à la souscription dans les contrats, le manque de données est surtout marqué aux grands âges. L'estimation de la mortalité aux âges avancés nécessite alors l'utilisation de méthodes d'extrapolation. La population peut être séparée en deux groupes, représentant d'un côté les autonomes et de l'autre les dépendants. La mortalité générale estimée sur un portefeuille d'assurance est par conséquent liée à celle des deux groupes. Une extrapolation indépendante des taux de mortalité des deux groupes peut faire apparaître des incohérences entre les trois courbes de mortalité. Le Chapitre 2 propose une méthode permettant d'extrapoler simultanément et de manière cohérente la mortalité des autonomes et des dépendants, en utilisant la connaissance sur la mortalité globale du portefeuille. Pour cela, un terme pénalisant l'incohérence entre les mortalités des trois groupes est ajouté à la vraisemblance totale. L'algorithme développé

1. Allocation personnalisée d'autonomie

dans cet article permet l'obtention de courbes de mortalité à la fois lisses et cohérentes entre elles. L'étude de la performance de l'algorithme sur des données synthétiques montre que l'ajout d'une pénalité de cohérence entre les trois lois de mortalité permet une meilleure estimation de la mortalité aux grands âges. Une librairie **R** implémentant cet algorithme a été développée au cours de cette thèse, permettant à SCOR d'utiliser régulièrement cette nouvelle approche à des fins de suivi d'expérience de portefeuilles d'assurance dépendance.

Afin de disposer de plus d'observations, il peut paraître intéressant d'agréger des observations provenant de diverses sources. Cependant, il n'existe pas de consensus commun entre tous les assureurs sur la définition exacte de l'état de dépendance et sur les conditions nécessaires au déclenchement du versement de la rente viagère. La pluralité des définitions de la dépendance entre les différents contrats proposés par les assureurs rend difficile l'agrégation de leurs données. Le Chapitre 3 s'intéresse au problème de la non-uniformité de la définition utilisée pour définir l'état de santé de l'assuré au sein d'une même base de données. Nous y avons proposé deux méthodes permettant de traiter le problème en présence de deux définitions, dont l'une est supposée plus stricte que l'autre. Ces méthodes permettent d'utiliser l'ensemble des observations de la base de données afin d'estimer conjointement les lois biométriques associées à chaque définition, sans souffrir de la perte d'information qui serait engendrée par une modélisation indépendante du risque pour chacune d'entre elles. La présence ou non d'une période de franchise dans certains contrats est l'une des sources d'hétérogénéité de la définition au sein d'une base de données d'assurance dépendance. Comparées aux performances de l'approche par modélisation indépendante des lois avec et sans franchise, les méthodes présentées dans ce chapitre ont permis de réduire l'incertitude autour de l'estimation de la mortalité pour chacune des deux définitions et d'améliorer les performances prédictives des modèles.

Enfin, il existe une multitude de pathologies pouvant mener à la perte d'autonomie d'un individu. Celles-ci sont très différentes en termes de sévérité, mais également en termes d'impact sur la mortalité de l'individu. En effet, nous avons montré le caractère significatif de la pathologie pour la prédiction de la mortalité d'un assuré dépendant. Cependant, la rareté des données d'individus en état de dépendance ne permet pas aux assureurs d'estimer une table de mortalité distincte par pathologie, en particulier pour celles les moins représentées dans le portefeuille. Le Chapitre 4 présente deux méthodes de clustering permettant de créer des groupes de pathologies pour lesquelles la mortalité est homogène. La mortalité en dépendance est fonction à la fois de l'âge atteint et de la durée depuis la perte d'autonomie. Elle peut donc être représentée comme une surface. Le problème de clustering de courbes de mortalité, en fonction de modalités d'une variable explicative, a été traité dans de précédents articles. Ce chapitre s'intéresse quant à lui au clustering de fonctions bidimensionnelles. Les méthodes présentées dans ce chapitre permettent de détecter les pathologies ayant des surfaces de mortalité semblables. En réduisant le nombre

de modalités de la variable de pathologie, et en agrégeant les observations d'un même cluster, ces méthodes permettent de prendre en compte l'information de la pathologie dans l'estimation des probabilités de décès. Pour un assureur disposant d'une base de données suffisamment fournie pour modéliser la mortalité pour chacune des pathologies indépendamment, ces méthodes permettent de réduire la complexité de son modèle en diminuant le nombre final de tables de mortalité, tout en capturant une grande part de la variabilité observée entre les pathologies.

Dans le Chapitre 2, nous faisons l'hypothèse d'un modèle Markovien en considérant que les intensités de décès pour chaque sexe ne dépendent que de l'âge atteint. Une extension pourrait consister à généraliser l'algorithme développé dans ce chapitre au cas semi-Markovien, en supposant des intensités de mortalité des dépendants bidimensionnelles, fonctions à la fois de l'âge atteint et du temps écoulé depuis la perte d'autonomie. Dans ce cas, pour chaque âge x, la pénalité de cohérence lie l'intensité de décès des autonomes et l'ensemble des intensités de décès en dépendance associées à l'âge atteint x jusqu'à une durée maximale de survie en dépendance à laquelle l'ensemble des individus est supposé être décédé. Par ailleurs, la loi de mortalité générale du portefeuille est supposée connue. Cela nécessite donc une estimation préalable de celle-ci. Il pourrait être intéressant de l'estimer simultanément avec les deux autres lois. Pour cela, on propose de s'appuyer sur les Penalized Composite Link Models, utilisés dans le Chapitre 3. Enfin, il serait intéressant de comparer les lois de mortalité obtenues à partir de la méthode développée dans ce chapitre avec celles extrapolées avec les approches plus traditionnelles. Il serait notamment intéressant d'évaluer l'impact du choix de la méthode d'extrapolation sur la tarification et le provisionnement.

Dans le Chapitre 3, nous supposons une distribution de Poisson pour les variables aléatoires associées aux nombres de décès. Une des propriétés fortes de cette distribution de probabilité est l'égalité de la moyenne et de la variance. En pratique, on observe très souvent un phénomène de surdispersion : la variance est supérieure à la moyenne. L'hypothèse de Poisson peut alors être trop restrictive. Il pourrait être intéressant de prendre en compte un paramètre de surdispersion afin d'améliorer la modélisation. Par ailleurs, nous faisons l'hypothèse d'homogénéité des assurés entre les deux types de contrats ayant des définitions différentes. Il s'agit là encore d'une hypothèse forte en raison du phénomène d'antisélection. Les assurés les plus à risque ayant tendance à choisir le contrat le moins restrictif, les deux populations ne sont pas égales face au risque de perte d'autonomie. Dans la pratique, en présence de bases de données provenant de sources différentes, il serait pertinent de procéder à une étude préalable de la mortalité des populations de chaque base. À défaut de s'assurer de l'homogénéité face au risque de perte d'autonomie, difficile à estimer en raison de la différence de définition, cette étude préalable permettrait de juger de l'homogénéité des deux populations face au risque de décès. Un paramètre pourrait alors être introduit

pour prendre en compte le sur-risque d'une des deux populations.

Dans le Chapitre 4, nous utilisons un modèle de lissage paramétrique. Il serait intéressant, sous condition de disposer de plus de données, d'appliquer les méthodes de clustering développées dans celui-ci, combinées avec la méthode de lissage par P-Splines utilisée dans les Chapitres 2 et 3. Par ailleurs, la définition de la dépendance peut grandement varier entre pays. Ce chapitre repose sur des données étrangères, pour lesquelles la définition de la dépendance permet le retour à l'autonomie, contrairement aux produits usuels d'assurance dépendance en France. De plus, il existe une hétérogénéité du risque entre les populations de différents pays, provenant en partie des modes de vie. Les groupes de pathologies et résultats obtenus dans ce chapitre ne sont donc pas directement transposables à des produits d'assurance dépendance proposés en France. Il serait intéressant d'effectuer cette même étude sur des données de plusieurs pays et de constater les similarités et différences des regroupements de pathologies produits par le clustering. Enfin, par mesure de simplicité pour l'assureur, nous avons constitué des groupes de pathologies identiques pour les hommes et les femmes. Dans le cas où l'on souhaiterait autoriser des regroupements distincts pour chaque sexe, âge et durée depuis la perte d'autonomie, nous proposons d'utiliser la méthode des Survival Trees (Bou-Hamad et al., 2011).

Enfin, l'ensemble des chapitres de cette thèse fait abstraction de l'évolution temporelle du risque. Les éventuelles améliorations des taux d'incidence et de mortalité dues aux progrès de la médecine, ou les détériorations engendrées par l'apparition de nouvelles maladies, ne sont donc pas modélisées. Ceci se justifie par l'historique court des produits d'assurance dépendance, responsable d'un manque de données suffisantes. Ainsi, la tendance est rarement prise en compte par les assureurs dans l'estimation des lois biométriques, et donc dans la tarification et le provisionnement. L'engagement à long terme pris par les assureurs lors de l'établissement d'un contrat, s'expliquant par l'écart relativement long entre l'âge de souscription et l'âge moyen de survenue de la perte d'autonomie, constitue un risque pour les assureurs. En effet, une tendance à l'augmentation de l'incidence et/ou à la diminution de la mortalité des dépendants, non prévue par l'assureur, peut être responsable de pertes financières importantes dans le futur pour ce dernier. Une modélisation prospective serait intéressante afin d'anticiper au mieux l'évolution du risque.

Bibliographie

Bou-Hamad, I., D. Larocque, and H. Ben-Ameur (2011). A review of survival trees. *Statistics Surveys 5*(none), 44 – 71.