duas

Université de Strasbourg

Mémoire d'actuariat

6 novembre 2023

Construction d'une table de mortalité d'expérience avec des processus gaussiens

— Adrien Moyaux —

Tuteur en entreprise : **Patrick GOMIS**

Tuteur académique : **Pierre-Olivier GOFFARD**

Résumé

Mots clés : *table de mortalité, processus gaussiens, régression par processus gaussiens, Hoem, Kaplan-Meier, Whittaker-Henderson, modèle de Cox*

Ce mémoire explore une nouvelle méthode de construction de table de mortalité : les *processus gaussiens*. Plus précisément, la régression par processus gaussiens (*Gaussian Process Regression* ou *GPR*) est utilisée dans le cas de la modélisation de la mortalité.

Dans un premier temps, une table de mortalité est construite selon des méthodes classiques. Les taux bruts sont ainsi calculés avec l'estimateur de Kaplan-Meier, puis lissés avec le lissage de Whittaker-Henderson. Des positionnements sont également effectués avec le modèle de Cox.

Dans un second temps, les taux de mortalité sont déterminés d'une autre manière, en utilisant des processus gaussiens. La comparaison avec les taux issus des méthodes classiques met en lumière un intérêt certain pour l'utilisation de processus gaussiens. En effet, les taux des processus gaussiens sont déjà lissés et proches des taux obtenus par les méthodes classiques. En outre, le modèle ajusté permet également d'extrapoler les taux en des points inconnus, sans avoir besoin de faire d'hypothèse de forme de la courbe de mortalité. De plus, des intervalles de confiance sont disponibles en tout point de prédiction. Ceux-ci semblent plus élevés que ceux des méthodes classiques en des points connus, mais restent relativement petits. Enfin, les processus gaussiens permettent d'effectuer des positionnements sans avoir besoin de recourir à des hypothèses telle que celle des hasards proportionnels du modèle de Cox. Une comparaison des taux de Cox avec ceux des processus gaussiens montre une meilleure adéquation aux taux bruts des taux des processus gaussiens.

Certaines limites des processus gaussiens ont néanmoins été découvertes. La limite la plus importante est qu'aucune croissance des taux par âge n'est supposée par le modèle. Par conséquent, la prédiction en des âges élevés où peu d'exposition est présente et presque aucun décès n'est constaté a mené à des prédictions erronées où la mortalité atteint un pic à un certain âge, puis se met à décroître. Une autre limite découverte est la difficulté de réaliser des positionnements par cohorte, les résultats pouvant être très variables selon les hyperparamètres fournis en entrée des processus gaussiens.

En conclusion, les processus gaussiens semblent être une nouvelle méthode prometteuse de construction de tables de mortalité. Ce procédé pourrait d'ailleurs tout à fait être utilisé pour la construction d'autres types de tables associant des probabilités à des variables d'entrée.

Abstract

Keywords : *graduation of mortality rates, Gaussian processes, Gaussian Process Regression (GPR), Hoem, Kaplan-Meier, Whittaker-Henderson, Cox model*

This thesis explores a novel method for constructing mortality tables : Gaussian processes. Specifically, Gaussian Process Regression (GPR) is used in modeling mortality.

Initially, a mortality table is constructed using traditional methods. Raw rates are calculated using the Kaplan-Meier estimator and then smoothed using Whittaker-Henderson smoothing. Positioning is also performed using the Cox model.

In a second step, the graduation of mortality rates is realised differently, using Gaussian processes. Comparing these rates with those obtained from traditional methods highlights a significant advantage in using Gaussian processes. Indeed, Gaussian process rates are already smoothed and closely resemble rates obtained by traditional methods. Furthermore, the fitted model allows for rate extrapolation at unknown points without the need for assuming a specific mortality curve shape. Additionally, confidence intervals are available at any prediction point. These intervals appear to be higher than those of traditional methods at known points but remain relatively small. Lastly, Gaussian processes enable positioning without the need for assumptions such as the proportional hazards of the Cox model. A comparison between Cox rates and Gaussian process rates shows that Gaussian processes rates align better with raw rates.

However, some limitations of Gaussian processes have been discovered. The most significant limitation is that the model assumes no age-related rate growth. Consequently, predictions at older ages with minimal exposure and almost no observed deaths have led to erroneous predictions where mortality peaks at a certain age and then decreases. Another discovered limitation is the challenge of cohort-based positioning, as results can be highly variable depending on the input hyperparameters provided to the Gaussian processes.

In conclusion, Gaussian processes appear to be a promising new method for constructing mortality tables. This method could potentially be used to construct other types of tables outputting probabilities given input variables.

Remerciements

Je tiens avant tout à exprimer ma profonde gratitude envers Patrick Gomis. Son engagement indéfectible, son expertise inestimable ainsi que sa disponibilité constante ont été d'une importance capitale tout au long de la rédaction de ce mémoire.

J'adresse également mes plus sincères remerciements à mon tuteur pédagogique, Pierre-Olivier Goffard. Ses conseils éclairés, son suivi attentif et sa suggestion de travailler sur les processus gaussiens ont grandement enrichi la qualité de ce travail.

Un grand merci s'étend aussi à tous les membres du DAPP (*Département Actuariat Produits Prévoyance*) et de l'ADE (*Assurance des Emprunteurs*) de BPCE Assurances. Leur accueil chaleureux a rendu cette expérience de recherche extrêmement enrichissante.

Pour conclure, je souhaite exprimer ma sincère gratitude envers Marie Le Pennec. Grâce à elle, j'ai eu l'opportunité d'intégrer BPCE Assurances et de réaliser mon mémoire au sein de cette entreprise. Sa présence au sein de l'équipe a été une source d'inspiration constante tout au long de nos alternances et de la rédaction de nos mémoires respectifs, m'incitant à donner le meilleur de moi-même.

Ces remerciements reflètent ma profonde reconnaissance envers ceux qui ont rendu possible la réalisation de ce mémoire, une étape cruciale de mon parcours académique et professionnel. Votre soutien, vos inspirations, et votre générosité d'esprit ont été des éléments déterminants de cette réussite, et je suis sincèrement reconnaissant pour cela.

Sommaire

Note

Dans la version PDF de ce mémoire, **tous les titres sont cliquables**, tant ceux dans le sommaire que ceux dans le corps du texte. Ainsi, le clic sur un titre dans le corps du texte renvoie à sa position dans le sommaire. De plus, en haut de chaque page se trouve le nom du chapitre et de la section en cours, là encore sous forme de liens, permettant en un clic de retourner au début de l'endroit voulu. Enfin, en bas de chaque page, le numéro de page renvoie au début du sommaire.

Chapitre I

Contexte

L'objectif de ce mémoire est de construire une table de mortalité d'expérience en utilisant une méthode non encore utilisée par les praticiens : les processus gaussiens. L'idée que les processus gaussiens puissent être utilisés pour la construction d'une table de mortalité est récente, avec la publication en 2018 d'un article de Ludkovski et al.[Lud18] Dans celui-ci, les auteurs proposent un modèle non paramétrique unique pour la graduation des taux et la prévision de l'amélioration de la mortalité.

Ce mémoire consiste par conséquent en la construction d'une table de mortalité de deux manières différentes : par des méthodes classiques (calcul des taux bruts, puis lissage) ainsi que par l'utilisation de processus gaussiens.

Dans le chapitre II traitant des méthodes classiques, la table est construite en estimant les taux bruts grâce à l'estimateur de Kaplan-Meier. Elle est comparée à celle obtenue par l'estimateur de Hoem pour s'assurer de la similarité des résultats. Les taux bruts de Kaplan-Meier sont ensuite lissés pour obtenir la table de mortalité finale. Enfin, des positionnements sont également effectués sur le sexe et le réseau (BP ou CE) à l'aide du modèle de Cox.

Dans le chapitre III sur les processus gaussiens, une première section se charge d'expliquer la théorie sous-jacente au modèle. Ensuite, le cas particulier de la mortalité est abordé dans une seconde section. Enfin, l'ajustement du modèle à une base d'exposition décès et les résultats obtenus sont présentés et commentés dans les deux dernières sections de ce chapitre. En particulier, les taux de mortalité des processus gaussiens sont comparés à ceux déterminés par les méthodes classiques.

Pour finir, une application pratique est réalisée dans le chapitre IV en utilisant chacune de ces deux méthodes pour construire une table de mortalité après refus. Celles-ci sont ensuite comparées, avec en particulier une analyse de leur impact pour le calcul du *Best Estimate*.

I.1) La prévoyance chez BPCE Assurances

I.1.1) BPCE Assurances

BPCE Assurances, anciennement Natixis Assurances, est une entité du groupe BPCE. Le groupe BPCE est issu de la fusion en 2009 des réseaux Banque Populaire et Caisse d'Epargne. Il est le deuxième groupe bancaire en France ainsi que la quatrième plus grosse banque en terme de produit net bancaire.[Cor22] BPCE est aussi le sixième plus gros assureur français en assurance de personnes (chiffres 2020).[Ass21] De plus, en 2022, il est le quatrième bancassureur en France (classement par rapport au chiffre d'affaire assurance).[Dan22]

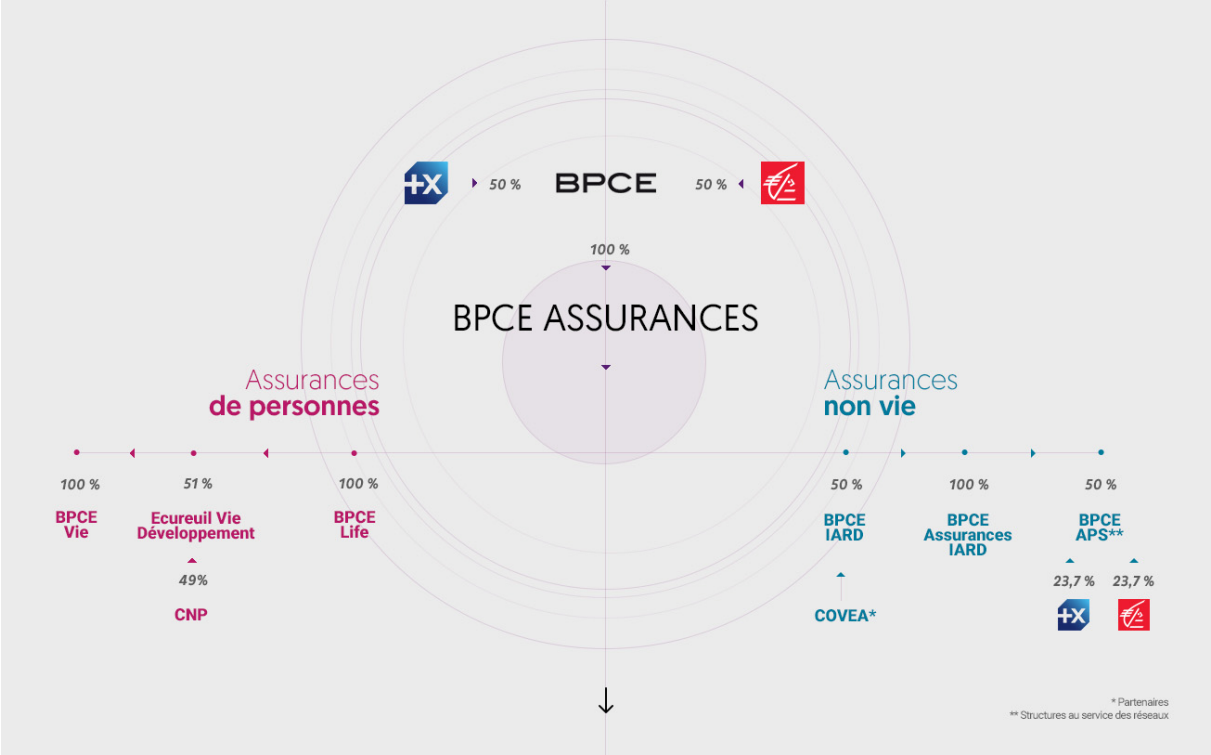

Figure I.1 — *Structure de BPCE Assurances*

BPCE Assurances conçoit, souscrit et gère les contrats d'assurance distribués par les Banques Populaires (BP) et les Caisses d'Epargne (CE). Les clients sont à la fois des particuliers et des professionnels. Elle est organisée en deux métiers, comme l'illustre la figure I.1 :

- Assurances de personnes : assurance vie, épargne, transmission de patrimoine, retraite, assurance décès, assurance dépendance, assurance des emprunteurs...
- Assurances non vie : assurance automobile, multirisque habitation, complémentaire santé, garantie des accidents de la vie, assurance des équipements multimédia, pro-

tection juridique, assurance parabancaire, télésurveillance, assurances des professionnels...

I.1.2) Département Actuariat Produits Prévoyance

La département chargé de la prévoyance se situe dans le métier Assurances de personnes de BPCE Assurances. La prévoyance regroupe « les opérations ayant pour objet la prévention et la couverture du risque décès, des risques portant atteinte à l'intégrité physique de la personne ou liés à la maternité, des risques d'incapacité de travail ou d'invalidité ou du risque chômage ».[Loi89]

Le Département Actuariat Produits Prévoyance, abrégé DA2P, fait partie de la Direction Offre et Pilotage Commercial du pôle Assurances de Personnes des Assurances du Groupe BPCE. Il traite des problématiques d'actuariat Prévoyance Individuelle (tarification – rentabilité) ainsi que de la modélisation de la prévoyance plus largement (prévoyance individuelle et ADE). En revanche, les problématiques concernant l'inventaire ne sont pas de son ressort, un département y étant spécialisé.

Plus spécifiquement, les missions réalisées par le Département Actuariat Produits Prévoyance sont :

Données

- Validation des bases mensuelles
- DWH PI (maintenance, évolution, validation)

Solvabilité II

- Données et hypothèses (production des *model points*, cartographie des hypothèses, qualification des données)
- QRT / ENS

Suivi du risque

- Risque de souscription (décès, arrêt de travail, dépendance, obsèques, résiliations...)
- Études actuarielles / produit
- Politique de souscription (rédaction et suivi)

Loi d'expérience — Modélisation actuarielle

- Mortalité

- Incidence et maintien en arrêt de travail
- Refus
- Attrition (résiliations)

Rentabilité

- Définition des normes de rentabilité ex-ante et ex-post
- Outils de rentabilité
- Production d'études de rentabilité (produits / activités)

Produits

- Cartographie des produits
- Tarification : prime pure, paramétrage de l'infocentre de gestion prévoyance, enveloppes commerciales
- Relecture de conditions générales

Veille

- Actuarielle
- Assurantielle
- Outils (Data science, langages informatiques adaptés à l'activité, etc.)

I.2) Mesure du risque de mortalité chez un assureur

En assurance de personnes, la mortalité est le risque prépondérant. La connaissance de celle-ci permet de tarifer les contrats d'assurance emprunteur, d'épargne, ou encore, comme dans le cas de ce mémoire, de prévoyance. Le risque décès est effectivement la brique de base des contrats de prévoyance, où s'ajoutent d'autres garanties telles que l'arrêt de travail ou la dépendance. Il impacte de manière significative la tarification et la rentabilité, c'est pourquoi il convient de le mesurer avec précision.

I.2.1) Les tables de mortalité

Une table de mortalité est une table indiquant, pour chaque âge, la probabilité annuelle de décès d'un individu. Elles peuvent également être construites sur plusieurs axes (sexe, ancienneté, catégorie socioprofessionnelle, etc.)

Utilisation des tables de mortalité en entreprise

Dans le cadre de l'activité d'assurance, les tables de mortalité sont très importantes car ce sont elles qui permettent d'estimer précisément le risque de mortalité. Ces tables impactent la tarification des produits d'assurance, le provisionnement ainsi que les ratios de solvabilité.

Types de tables de mortalité

Il existe deux types de tables de mortalité :

1. Les tables du moment : elles adoptent une perspective transversale. La mortalité d'une population fictive est étudiée, sans tenir compte de l'année de naissance des individus qui la composent. Ainsi, pour tout âge x, les individus dont la mortalité est étudiée ont bien l'âge x, mais sont nés à des moments différents : en 1970, en 1984, etc.
2. Les tables générationnelles : elles adoptent une perspective longitudinale. Les mortalités de groupes d'individus **nés la même année** sont étudiées tout au long de leur existence. Il en résulte la création d'un tableau contenant pour chaque année de naissance une table de mortalité du moment. Ces tables sont souvent prospectives par génération, car par exemple il n'est pas encore possible de connaître de manière certaine la proportion de survivants à 70 ans de personnes nées en 1990. Il est alors nécessaire d'extrapoler les taux de mortalité.

Dans le cadre de ce mémoire, nous allons construire des tables du moment. Capturer l'amélioration de l'espérance de vie n'a que peu d'intérêt en prévoyance individuelle. Effectivement, les garanties de prévoyance individuelle concernent le risque de décès, contrairement à l'épargne par exemple où il y a un risque de survie avec l'utilisation de rentes viagères. En outre, construire des tables générationnelles demanderait une très grande quantité de données pour s'assurer d'avoir assez d'exposition pour chaque génération et chaque âge, ce que nous n'avons pas.

I.2.2) Provenance des tables de mortalité

Les tables de mortalité peuvent être réglementaires ou d'expérience.

Tables de mortalité réglementaires

Les tables de mortalité réglementaires ont été réalisées en étudiant la mortalité de la population française. Ce sont des tables du moment et des tables générationnelles.

Les tables TH00-02 et TF00-02 sont des **tables de mortalité du moment réglementaires**. Ces tables reposent sur les observations INSEE de la mortalité entre 2000 et 2002 (de la population masculine pour la TH et féminine pour la TF). Par l'arrêté de 20 décembre 2005[Jou05a], ces tables ont été homologuées pour la tarification et le provisionnement en cas de décès et en cas de vie. Elles sont applicables pour les contrats d'assurance vie (sauf rentes viagères) depuis le 1er juillet 2006.[Jou05b]

Les tables TGH05 et TGF05 sont des **tables de mortalité générationnelles réglementaires**. Ces tables fournissent les taux de mortalité pour toutes les générations entre 1900 et 2005 (de la population masculine pour la TGH et féminine pour la TGF). Par l'arrêté du 1er août 2006[Jou06], ces tables ont été homologuées pour la tarification et le provisionnement des contrats de rentes viagères. Ces tables sont utilisables depuis le 1er janvier 2007.[Jou06]

Tables de mortalité d'expérience

Les tables de mortalité d'expérience sont des tables construites sur une population spécifique. En effet, lorsque la population d'un portefeuille d'assurés adopte un comportement trop éloigné de celui prédit par les tables réglementaires, les compagnies d'assurance ont la possibilité de construire leurs propres tables, adaptées à leurs portefeuilles d'assurés. Il faut cependant pour cela disposer d'un volume suffisant de données.

Utiliser une table d'expérience à la place d'une table réglementaire confère un avantage compétitif à l'assureur. En effet, la table de mortalité ainsi construite reflétera le risque intrinsèque au portefeuille couvert et autorisera une tarification plus précise. Les tables de mortalité d'expérience permettent également d'optimiser le calcul des provisions *Best Estimate* dans le cadre des référentiels prudentiels Solvabilité II et IFRS17.

I.2.3) Cadre prudentiel

La réglementation précise dans quelles conditions l'assureur peut utiliser ses propres tables d'expérience construites à partir des données de la population concernée par son portefeuille.[Cod17] Ces tables doivent être certifiées par un actuaire indépendant agréé par l'Institut des Actuaires.

La table d'expérience est valide deux ans sans suivi annuel et jusqu'à cinq ans dans le cas d'un suivi annuel.[Insa] En effet, les articles A335-1 du Code des Assurances, A212-10 du Code de la Mutualité et A931-10-10 du Code de la Sécurité Sociale permettent aux assureurs, dans certaines conditions, de construire des tables d'expérience sur la base des données propres au portefeuille. Ceci en remplacement des tables officiellement en vigueur, pour autant qu'elles permettent de mieux estimer les engagements contractuels.

I.3) Construction de la base d'exposition

La construction de la base d'exposition s'appuie sur les recommandations de l'Institut des Actuaires pour la certification de tables, qui a publié une méthodologie générale[Insb] ainsi qu'une plus détaillée[Ins06].

I.3.1) Choix de la période d'observation

Afin d'estimer au mieux la mortalité, il faut choisir judicieusement la fin de la période d'observation. Pour ce faire, il convient de s'assurer que la quasi-totalité des décès survenus au cours de la période d'observation ont bien été déclarés. En choisissant une date de fin d'observation trop récente, le **phénomène de déclaration tardive des décès** peut en effet conduire à sous-estimer le taux de mortalité. Des décès seraient absents de notre base de données, car tous les sinistres ne sont pas déclarés immédiatement après leur survenue, cela peut prendre plusieurs mois.

La figure I.2 montre le taux de décès reportés selon le temps de déclaration. Pour réaliser ce graphique, les décès s'étant produits de 2016 à 2021 dans notre portefeuille ont été utilisés. L'année 2022 a été exclue pour s'assurer de la justesse des répartitions. En

effet, tous les décès de celle-ci peuvent ne pas encore avoir été déclarés à l'heure actuelle et cela viendrait fausser les résultats.

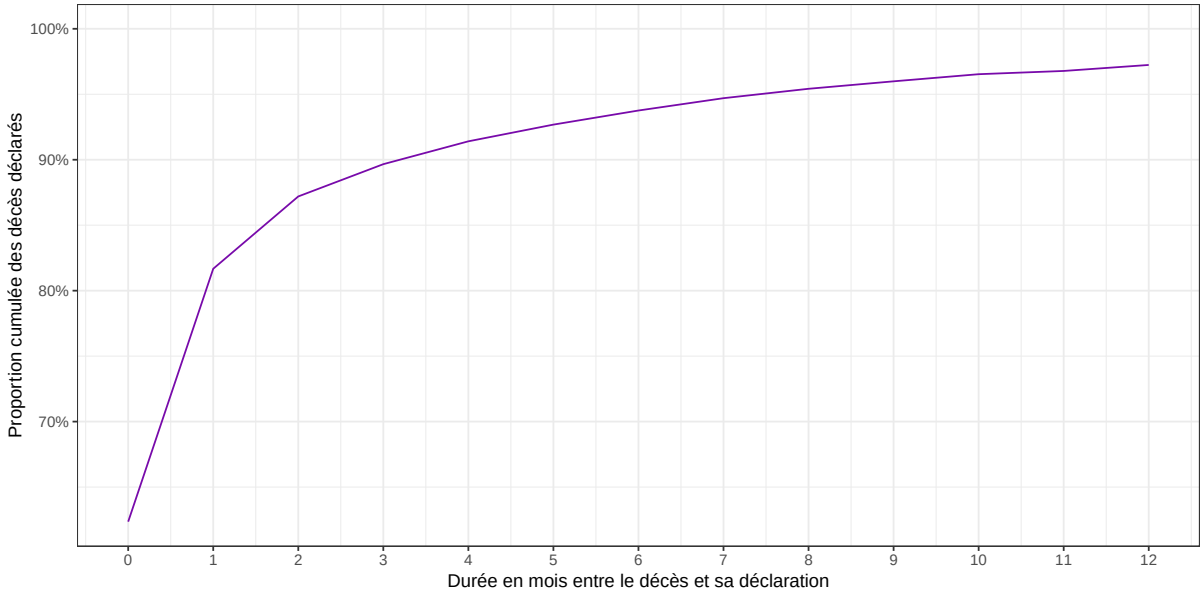

Figure I.2 — *Répartition des durées de déclaration*

La figure I.2 permet de constater que 95% des décès sont déclarés dans les 7 mois. Par conséquent, pour la construction de la base d'exposition, la date de fin d'observation sera mise au minimum à 6 mois avant la date d'extraction des données.

L'Institut des Actuaires recommande dans sa méthodologie détaillée de choisir une durée d'observation qui soit un multiple de douze mois. En effet, il existe un **phénomène de saisonnalité des décès** au cours d'une année. Ainsi, certaines périodes de l'année sont plus touchées par les décès que d'autres.

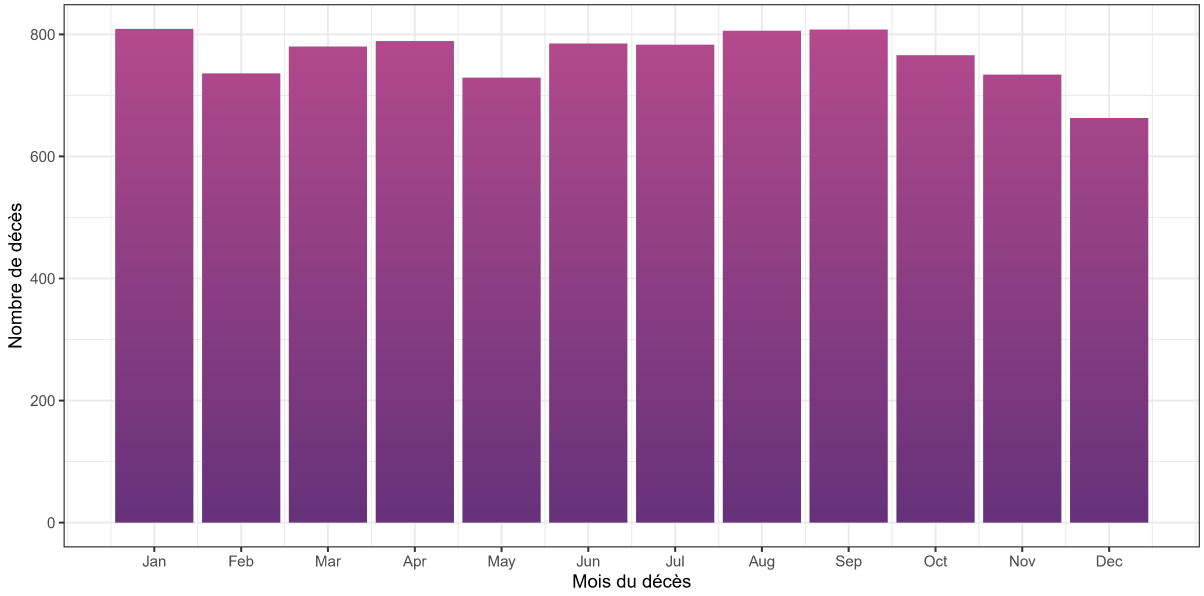

Figure I.3 — *Saisonnalité des décès*

Ce phénomène se retrouve effectivement dans nos données comme l'illustre la figure I.3. Utiliser un multiple de douze mois permet de se prémunir de ce phénomène et c'est ce que nous ferons par la suite.

De plus, une **période d'observation de trois à cinq ans** est recommandée par l'Institut des Actuaires. Celle-ci doit être suffisamment élevée afin d'avoir assez de données et que la loi des grands nombres s'applique. D'un autre côté, la mortalité évoluant selon les générations, il ne faut pas que la période d'observation soit trop étalée dans le temps. Dans notre cas, la période d'observation choisie est une période de 5 ans, de 2018 à 2022.

Enfin, il est légitime de se demander si la Covid-19 a eu un impact sur les taux de mortalité, et donc s'il faut conserver ou non les années touchées. Pour cela, les taux bruts des années avant la Covid-19 et des années Covid-19 ont été tracés sur la figure I.4. Celle-ci montre donc les taux bruts par âge et selon l'année d'observation. Dans le cas d'un **effet Covid-19**, des courbes différentes entre les années pré-Covid-19 et les années Covid-19 sont attendues.

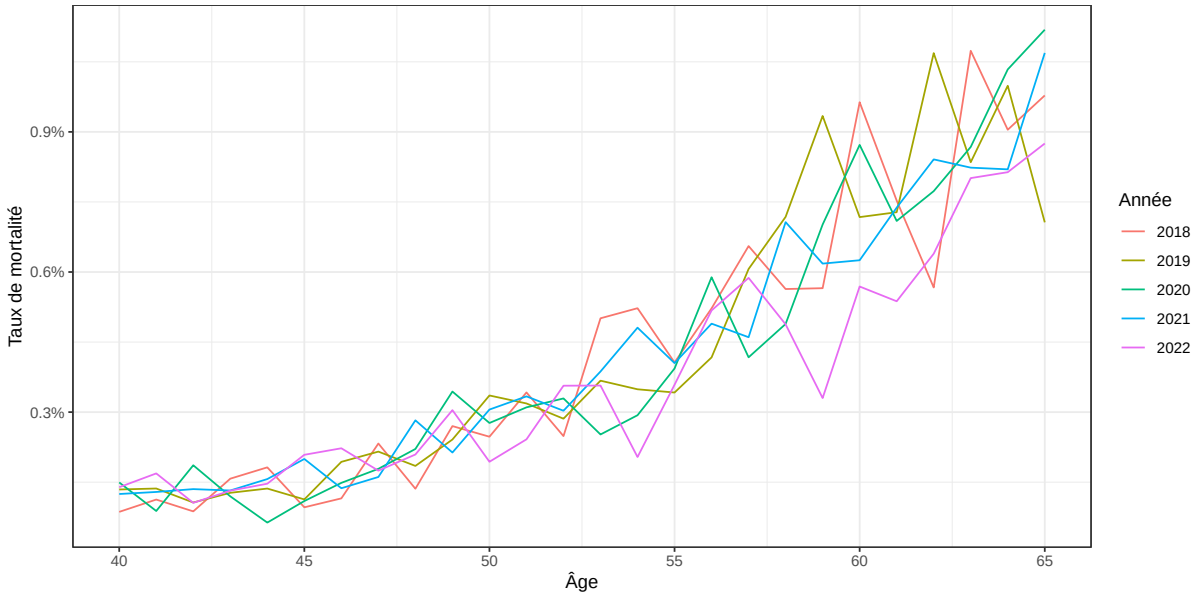

Figure I.4 — *Taux de mortalité bruts par année*

Il ne semble donc pas y avoir de changements particuliers en 2020 ou 2021. Ce résultat a aussi été observé par les autres assureurs en prévoyance pour le même type d'offre. Les années correspondant au Covid-19 peuvent donc être conservées.

I.3.2) Dédoublonnage

Par dédoublonnage, on entend le prétraitement de la base d'exposition permettant de ne pas compter plusieurs fois un assuré ayant plusieurs contrats.

La base d'exposition d'origine est à la maille client-contrat. Un client peut posséder plusieurs contrats et donc avoir plusieurs lignes. Pour le risque décès toutes causes, 90% des assurés de la base n'ont qu'un contrat, 8% en ont 2, et les 2 derniers pour cents en ont 3 ou plus. Par conséquent, la base d'exposition a été dédoublonnée afin de ne pas comptabiliser plusieurs fois un même assuré. Cela a diminué de 10% le nombre de lignes de la base d'exposition pour passer de la base à la maille client-contrat à la base à la maille client.

Pour les assurés possédant plusieurs contrats, la ligne représentant l'assuré a été définie selon les règles suivantes :

- le minimum des dates d'effet est retenu ;
- le minimum des dates de décès est retenu en cas de décès ;
- le maximum des dates de clôture est retenu.

I.3.3) Taux avant ou après refus

Faut-il construire une table de mortalité avant, ou après refus de certains dossiers de sinistres par l'assureur ? Tout d'abord, ce sont les taux de mortalité avant refus qui sont certifiés, les taux après refus ne le sont pas. En effet, les taux avant refus correspondent au niveau de risque du portefeuille, et c'est ce qui est certifié. Les taux après refus dépendent quant à eux des exclusions, franchises et autres règles de refus instaurées par l'assureur. Celles-ci peuvent varier au fil du temps et selon les produits. Ainsi, du fait de cette variabilité, les taux après refus ne sont pas certifiables.

Les taux après refus sont toutefois très utiles, notamment pour calculer le *Best Estimate* de la manière la moins biaisée possible. Ce sera utile dans le cadre de Solvabilité II ainsi que d'IFRS17.

Dans le cadre de ce mémoire, nous allons d'abord nous concentrer sur les taux avant refus. Cela permettra de comparer les taux construits par les méthodes classiques et les processus gaussiens avec la table de mortalité certifiée. Ensuite, dans le chapitre IV d'application pratique, une table de mortalité après refus sera construite avec les processus gaussiens. De cette manière, il sera possible d'analyser l'impact de la nouvelle table sur le *Best Estimate* et le choc de mortalité.

I.4) Statistiques descriptives sur le portefeuille

Dans cette section sont présentées quelques statistiques descriptives afin d'avoir un aperçu des principales caractéristiques du portefeuille. Ces résultats peuvent ensuite nous orienter dans les démarches de modélisation ou aider à expliquer des résultats.

I.4.1) Nombre d'assurés et de décès

Il y a environ 3 800 000 lignes dans la base d'observation (après dédoublonnage), dont **2 200 000 lignes pour le risque décès toutes causes**, le reste étant du décès accidentel. Les statistiques suivantes seront sur la base d'exposition filtrée sur le décès toutes causes.

I.4.2) Proportions par sexe et réseau

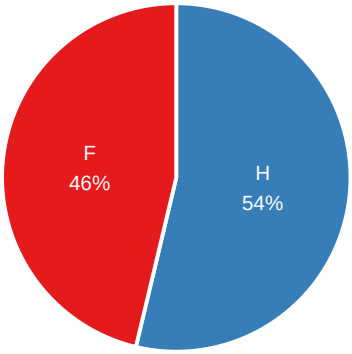

Figure I.5 — *Répartition par sexe*

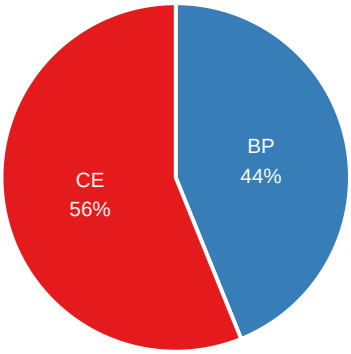

Figure I.6 — *Répartition par réseau*

Les figures I.5 et I.6 montrent une plus grande proportion d'hommes et de contrats CE. Une analyse plus fine en figure I.7 par année de souscription met en lumière la plus forte souscription des contrats dans le réseau CE que BP. Il est à noter que les contrats CE ont été rajoutés en 2016, ce qui explique pourquoi les années précédentes il n'y a pas de données pour ce réseau.

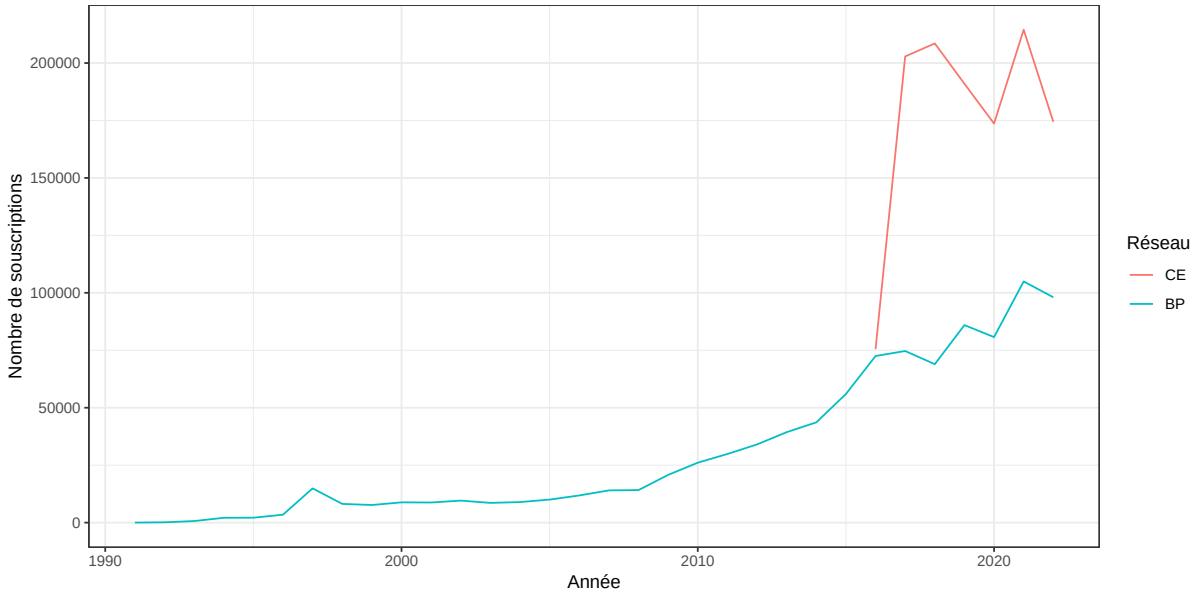

Figure I.7 — *Nombre de souscriptions annuelles selon le réseau*

I.4.3) Distribution des âges

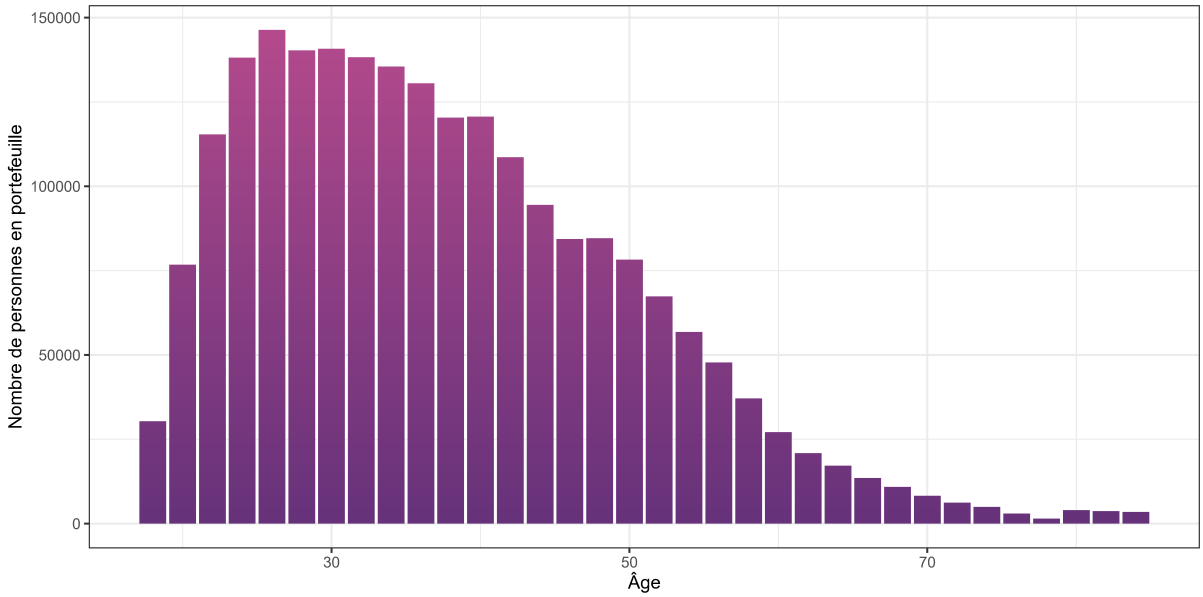

Figure I.8 — *Répartition du stock des assurés par âge*

Comme le montre la figure I.8, les personnes d'environ 30 ans sont les plus nombreuses. En outre, le tableau I.1 met en lumière que les garanties de nos contrats décès s'achèvent

souvent à 85 ou 89 ans. Cependant, les conditions générales des contrats précisent que la date limite d'adhésion est plus tôt, à 65 ans.

Âge de fin de couverture	Proportion de contrats
55 ans	9%
65 ans	1%
70 ans	7%
75 ans	2%
85 ans	45%
89 ans	36%

Tableau I.1 — *Âges de fin de couverture pour la garantie décès*

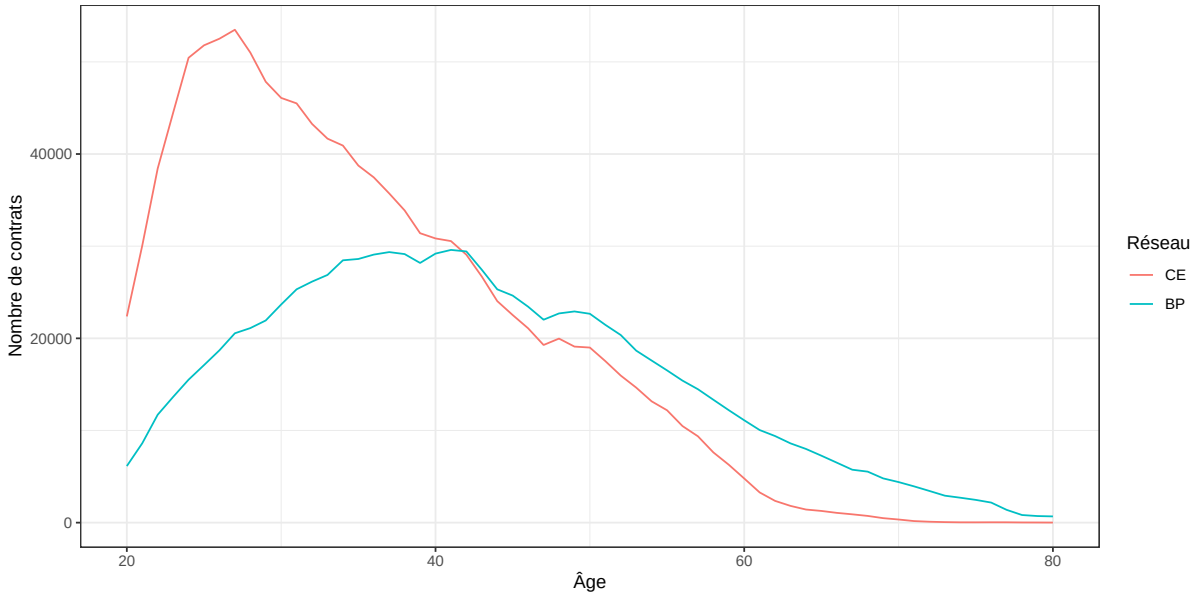

Figure I.9 — *Nombre de contrats du stock par âge selon le réseau*

La figure I.9 donne des informations plus précises sur la répartition des âges en portefeuille. Elle met en lumière la grande variabilité selon le réseau, avec un pic d'âges vers 25 ans pour les CE tandis que celui des BP est plus tard, vers 40 ans, et avec une décroissance moins rapide que pour les CE. C'était un résultat attendu, car le portefeuille CE est plus jeune en termes de lancements de produits.

I.4.4) Complétude et exactitude des variables

Dans la base d'exposition utilisée, la majorité des variables sont 100% exhaustives. Seule la variable CSP a 20% de valeurs manquantes, car cette information n'est pas toujours demandée selon les produits.

L'exactitude des variables a également été contrôlée afin de supprimer les lignes ayant des valeurs aberrantes. Afin de vérifier l'exactitude des variables, il a notamment été constaté que :

- Aucun client n'a plusieurs dates de naissance ;
- Aucun client n'a plusieurs dates de décès ;
- Aucune date de naissance n'est supérieure à la date d'effet ;
- L'âge de l'assuré à l'adhésion est supérieur à 18 ans, sauf pour 69 assurés ;
- La date d'effet est inférieure à la date de clôture ;
- La date d'effet est inférieure à la date de survenance, sauf pour 7 assurés ;
- Il n'y a pas de date de naissance ni de date d'effet vide parmi les données utilisées.

Il convient aussi de s'assurer que les sinistres comptés soient les **sinistres survenus avant résiliation**. Dans le cas où la date de décès est supérieure à la date de clôture du contrat, le décès n'est pas comptabilisé, il y a simplement une censure à droite. En effet, l'individu est censé quitter l'observation dès sa date de clôture.

Chapitre II

Approche classique

II.1) Calcul des taux bruts

II.1.1) Concepts fondamentaux de l'analyse de survie

Le terme d'analyse de survie est employé de manière générale pour désigner le temps X qui s'écoule jusqu'à la survenance d'un évènement particulier. Dans le cadre de ce mémoire, X représente une durée de vie humaine.

Deux phénomènes distincts expliquent l'incomplétude des données de survie : la censure et la troncature.[Hub94] Ceux-ci sont causés par diverses perturbations qui affectent la variable aléatoire X, perturbations qui peuvent être indépendantes ou non du phénomène étudié. Avec la troncature, l'information est complètement perdue car elle n'est pas détectée, tandis qu'avec la censure on possède au moins l'information que la variable observée X est supérieure (ou inférieure) à un seuil C.

Ignorer les censures et troncatures conduit à utiliser des données contenant des biais. Par conséquent, il est nécessaire de prendre en compte ces phénomènes pour s'assurer de la justesse de nos estimations.

Censures

Une durée de vie aléatoire X est dite censurée par une variable aléatoire de censure C si l'on observe parfois C au lieu de X. L'information donnée par C sur la valeur de X est :

- $X < C$ s'il y a censure à gauche ;
- $X > C$ s'il y a censure à droite.

La censure à gauche se produit notamment quand l'instrument de mesure détecte les observations où X est plus petite qu'une valeur minimale C, mais n'est pas capable de les mesurer plus précisément. La valeur minimale C est alors attribuée à ces observations, alors que la vraie valeur est $X < C$. Concernant le décès, il y a censure à gauche quand l'individu décède avant la date de début d'étude. Nous savons alors uniquement que le décès a eu lieu avant l'étude, sans précision supplémentaire. Ainsi, X est inconnue et nous savons seulement que $X < C$ avec C la date de début de l'étude.

X n'est pas observée, la seule information connue est C quand $X < C$.

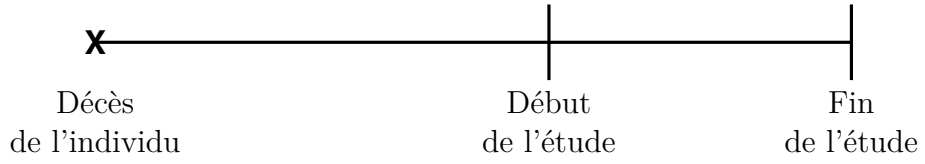

Figure II.1 — *Illustration d'une censure à gauche*

La censure à droite est quant à elle le cas où le décès ne se produit pas durant la période d'observation de l'individu. Ceci soit parce que l'individu sort du portefeuille avant la fin de la période d'étude, soit parce qu'il est toujours en vie à la fin de celle-ci. Dans le cas d'une censure à droite, l'individu a survécu au moins jusqu'à la date où il n'est plus observé mais aucune information sur ce qu'il lui arrive par la suite n'est disponible. Ainsi, X est inconnue et l'on sait seulement que $X > C$, C étant la date de sortie de l'individu (avec $C \in]$*début de l'étude*, *fin de l'étude*$]$).

X n'est pas observée, la seule information connue est C quand $X > C$.

Figure II.2 — *Illustration d'une censure à droite*

Les causes d'une censure à droite se produisant avant la fin de l'étude peuvent être la résiliation du contrat ou encore la perte de l'individu dans le système informatique. Il y a également censure à droite quand l'individu est encore vivant à la fin de la période d'étude : dans ce cas, la seule information disponible est que $X >$ *date de fin*. La plupart des bases de données en assurance vie sont largement censurées, car la majorité des assurés résilient ou restent en vie pendant la période d'étude.

Ne pas prendre en compte la censure à droite introduit un biais car les seules dates disponibles pour les individus censurés sont les dates de censure. Leur date de décès intervient nécessairement après leur date de sortie, mais est inconnue. Utiliser leur date de censure en tant que date de décès (qui n'est pas disponible) reviendrait à sous-estimer X.

Troncatures

La troncature se produit lorsque la variable d'intérêt n'est pas observable en-dessous d'un seuil T (cas des troncatures à gauche) ou au-dessus de ce seuil T (cas des troncatures à droite).[Sta16] Plus généralement, il y a troncature si l'observation de la variable d'intérêt X n'a lieu que conditionnellement à un événement B. Lorsqu'il y a troncature, les valeurs de X qui sont observées n'appartiennent donc qu'à un sous-ensemble des valeurs réellement prises par X.

La troncature à gauche est le cas où la variable X n'est observée que si elle est supérieure à T. Par exemple, si l'on étudie la taille de poissons, la troncature à gauche serait le fait de n'avoir que des poissons de taille supérieure à T la taille des trous dans le filet qui a été utilisé pour les attraper. De même, lorsqu'il y a une franchise en jours, les sinistres (comme l'arrêt de travail) ne sont reportés que s'ils sont supérieurs à la franchise, d'où une troncature à gauche. Enfin, dans le cas de la mortalité, la troncature à gauche intervient notamment avec l'âge minimum de souscription : la durée de vie des individus du portefeuille est nécessairement supérieure à cet âge minimum.

$$\boxed{X \text{ n'est observée que si } X > T.}$$

La troncature à droite est quant à elle le cas où la variable d'intérêt X n'est observable que si elle est inférieure à T. Par conséquent, il y a troncature à droite quand les seuls individus observables sont ceux ayant expérimenté l'événement d'intérêt avant la date T.

$$\boxed{X \text{ n'est observée que si } X < T.}$$

II.1.2) Estimateur binomial

Cet estimateur a l'avantage d'être facile à calculer. Toutefois, il n'est utilisable que lorsque tous les individus sont totalement observables sur la période $]x, x+1]$.[Pla11] En particulier, **le décès doit être l'unique cause possible de sortie** pour tout individu en vie à l'âge x. Ce modèle est donc difficilement utilisable en pratique, car dans la majorité des cas le décès n'est pas la seule cause de sortie (il y a notamment la résiliation).

Notations

Soient les notations suivantes :

- n_x : le nombre d'individus en vie de l'âge x à l'âge $x+1$;
- D_x : la variable aléatoire représentant le nombre de décès observés à l'âge x (i.e. en $]x,\ x+1]$) ;
- d_x : la réalisation de D_x ;
- q_x : la probabilité de décéder dans l'année pour un individu d'âge x.

Hypothèses

Deux hypothèses sont faites dans le cadre de cet estimateur :

1. Chaque décès est indépendant des autres ;
2. D_x, le nombre de décès à l'âge x, suit une loi $B(n_x,\ q_x)$.

Calcul de l'estimateur binomial

$\hat{q_x}$ est déterminé par la méthode du maximum de vraisemblance.

$$\mathbb{P}(D_x = d_x) = \binom{n_x}{d_x} \times q_x^{d_x} \times (1-q_x)^{n_x - d_x}.$$

Ainsi, la vraisemblance est :

$$L(q_x) = K \times q_x^{d_x} \times (1-q_x)^{n_x - d_x},$$

avec K une constante indépendante de q_x.

En notant l le logarithme de L :

$$l(q_x) = ln(K) + d_x \times ln(q_x) + (n_x - d_x) \times ln(1 - q_x).$$

Pour déterminer le $\hat{q_x}$ maximisant la vraisemblance, on dérive la log-vraisemblance puis égalise à 0 pour isoler $\hat{q_x}$:

$$\begin{aligned}
\frac{\partial l}{\partial q_x}(\hat{q_x}) &= \frac{d_x}{\hat{q_x}} + (n_x - d_x) \times \frac{-1}{1 - \hat{q_x}} \\
&= \frac{d_x \times (1 - \hat{q_x}) - (n_x - d_x) \times \hat{q_x}}{\hat{q_x} \times (1 - \hat{q_x})} \\
&= \frac{d_x - \hat{q_x} \times (d_x + n_x - d_x)}{\hat{q_x} \times (1 - \hat{q_x})} \\
&= \frac{d_x - \hat{q_x} \times n_x}{\hat{q_x} \times (1 - \hat{q_x})} \\
&= 0.
\end{aligned}$$

Par conséquent, l'estimateur binomial est :

$$\boxed{\hat{q}_x^{Bin} = \frac{d_x}{n_x}}.$$

Notons que cet estimateur est également celui obtenu par la méthode des moments. En effet, comme le nombre de décès suit une loi $B(n_x,\ q_x)$, alors $\mathbb{E}(D_x) = n_x \times q_x$, d'où $q_x = \frac{\mathbb{E}(D_x)}{n_x} \approx \frac{d_x}{n_x}$.

II.1.3) Estimateur de Hoem

L'estimateur de Hoem (1976)[Hoe76] généralise l'estimateur binomial en introduisant des censures et troncatures. Il prend en compte le fait qu'un assuré i n'est exposé dans $[x,\ x+1]$ qu'entre les dates α_i et β_i. Ainsi, le risque auquel est soumis l'assuré se trouve uniquement dans la période $[\alpha_i,\ \beta_i] \subset [x,\ x+1]$.

Pour pouvoir utiliser cet estimateur, il faudra disposer des dates d'entrée et de sortie des individus afin de calculer leur exposition au risque. Hoem repose donc sur la notion d'exposition au risque. C'est un estimateur paramétrique[Pla11], car il repose sur une **hypothèse de distribution infra-annuelle des décès** : l'hypothèse de Balducci.

Notations

Soient les notations suivantes :

- n_x : le nombre d'individus en vie à l'âge x ;
- $[\alpha_i,\ \beta_i]$: l'intervalle inclus dans $[x,\ x+1]$ dans lequel l'assuré i est observé ;
- $X_1,\ \ldots,\ X_{n_x}$: n_x variables de Bernoulli de paramètre ${}_{\beta_i-\alpha_i}q_{x+\alpha_i}$, indépendantes et identiquement distribuées, valant 1 si l'individu décède dans l'année et 0 sinon ;
- $D_x = \sum_{i=1}^{n_x} X_i$: la variable aléatoire représentant le nombre de décès observés en $]x,\ x+1]$;
- d_x : la réalisation de D_x.

Hypothèses

Deux hypothèses sont faites dans le cadre de cet estimateur :

1. Chaque décès est indépendant des autres, i.e. les X_i sont indépendants ;
2. Hypothèse de Balducci : ${}_{\beta_i-\alpha_i}q_{x+\alpha_i} = (\beta_i - \alpha_i)\, q_x$.

Calcul de l'estimateur de Hoem

Pour rappel, $D_x = \sum_{i=1}^{n_x} X_i$.

$$\begin{aligned}
\mathbb{E}[D_x] &= \sum_{i=1}^{n_x} \mathbb{E}[X_i] \\
&= \sum_{i=1}^{n_x} {}_{\beta_i-\alpha_i}q_{x+\alpha_i} \\
&= \sum_{i=1}^{n_x} (\beta_i - \alpha_i)\, q_x \text{ d'après l'hypothèse de Balducci} \\
&= q_x \times \sum_{i=1}^{n_x} (\beta_i - \alpha_i).
\end{aligned}$$

Donc $q_x = \frac{\mathbb{E}[D_x]}{\sum_{i=1}^{n_x}(\beta_i-\alpha_i)}$.
En remplaçant $\mathbb{E}[D_x]$ par son estimateur d_x[Lai18], l'estimateur de Hoem est :

$$\hat{q}_x = \frac{d_x}{\sum_{i=1}^{n_x}(\beta_i - \alpha_i)}.$$

En notant $E_x = \sum_{i=1}^{n_x} (\beta_i - \alpha_i)$ l'exposition au risque à l'âge x, l'estimateur peut se réécrire ainsi : [Ndi16]

$$\boxed{\hat{q}_x^{Hoem} = \frac{d_x}{E_x}}.$$

L'estimateur de Hoem correspond donc au ratio entre le nombre de décès à l'âge x et l'exposition au risque de mortalité pour ce même âge. L'exposition est la somme, sur les individus d'âge x, de la quantité en fraction d'années pendant laquelle ils ont été observés.

Propriétés de l'estimateur de Hoem

L'estimateur de Hoem présente plusieurs caractéristiques.[Bal13] Ainsi, cet estimateur est sans biais, c'est-à-dire que $\mathbb{E}\left[\hat{q}_x\right] = q_x$. Il est également convergent (parfois aussi appelé *consistant*). Un estimateur $\hat{\theta}_n$ de θ est convergent si $\forall \epsilon > 0,\ \lim_{n \to +\infty} \mathbb{P}\left(\left|\hat{\theta}_n - \theta\right| > \epsilon\right) = 0$.

L'espérance et la variance de cet estimateur sont les suivantes :

$$\mathbb{E}\left(\hat{q}_x^{Hoem}\right) = q_x \quad \text{et} \quad \text{Var}\left(\hat{q}_x^{Hoem}\right) = \frac{q_x\,(1-q_x)}{E_x}.$$

En outre, quand l'exposition E_x est suffisamment grande, et donc que $D_x = \sum_{i=1}^{n_x} X_i$ est la somme de suffisamment de termes, alors la variable aléatoire D_x peut être approximée par une loi normale. En effet, en utilisant le théorème central limite :

$$\frac{D_x - \mathbb{E}(D_x)}{Var(D_x)} \sim \mathcal{N}(0,1)$$

$$\implies \frac{\hat{q}_x^{Hoem} \times E_x - q_x \times E_x}{\sqrt{E_x \times q_x \times (1-q_x)}} \sim \mathcal{N}(0,1)$$

$$\implies \frac{E_x \times (\hat{q}_x^{Hoem} - q_x)}{\sqrt{E_x} \times \sqrt{q_x \times (1-q_x)}} \sim \mathcal{N}(0,1)$$

$$\implies \frac{\sqrt{E_x} \times (\hat{q}_x^{Hoem} - q_x)}{\sqrt{q_x \times (1-q_x)}} \sim \mathcal{N}(0,1)$$

$$\implies \hat{q}_x^{Hoem} - q_x \sim \frac{\sqrt{q_x \times (1-q_x)}}{\sqrt{E_x}} \times \mathcal{N}(0,1)$$

$$\implies q_x \sim \hat{q}_x^{Hoem} - \sqrt{\frac{q_x \times (1-q_x)}{E_x}} \times \mathcal{N}(0,1).$$

Comme q_x est inconnu, son estimateur $\hat{q}_x^{Hoem}$ peut être utilisé dans la formule précédente. Cela permet d'obtenir un intervalle de confiance asymptotique pour l'estimateur

de Hoem :

$$\boxed{IC_{1-\alpha}\left(q_x\right) = \left[\hat{q}_x^{Hoem} \pm \phi_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{q}_x^{Hoem} \times (1-\hat{q}_x^{Hoem})}{E_x}}\right]},$$

où $\phi_{1-\frac{\alpha}{2}}$ représente le quantile d'ordre $1-\frac{\alpha}{2}$ de la loi normale centrée réduite $\mathcal{N}(0,1)$. Ce quantile vaut environ 1,96 pour un intervalle de confiance à 95% ($\alpha = 0{,}05$).

Cette formule met en lumière que le corridor formé par les intervalles de confiance sera beaucoup plus large aux âges ayant peu d'exposition. En effet, la construction de ces intervalles dépend de la racine de la variance de l'estimateur $\hat{q}_x^{Hoem}$, qui a le terme E_x au dénominateur. Plus l'exposition est faible, plus la variance est forte et donc plus les intervalles de confiance seront larges.

II.1.4) Estimateur de Kaplan-Meier

L'estimateur de Kaplan-Meier (1958)[Kap58] ne fait **aucune hypothèse**[Hoe84] sur les décès infra-annuels (comme l'hypothèse de Balducci, de répartition uniforme ou encore de constance par morceaux des décès sur l'année).[Ins06] C'est donc un estimateur non paramétrique.

Celui-ci permet d'estimer la fonction de survie S. Il repose sur l'idée qu'être en vie à l'instant t, c'est être en vie juste avant t et de ne pas décéder en t. Cet estimateur fait donc intervenir les probabilités de survie en t, conditionnellement au fait d'être en vie juste avant. Il est également appelé estimateur produit-limite, car il est calculé par produit de ces probabilités de survie en chaque instant t. Ainsi, la probabilité de survivre au moins jusqu'en t_3 vaut $S(t_3) = \mathbb{P}(X > t_3) = (1 - \mathbb{P}(X = t_1 | X \geq t_1)) \times (1 - \mathbb{P}(X = t_2 | X \geq t_2)) \times (1 - \mathbb{P}(X = t_3 | X \geq t_3))$, les décès ne se produisant pas entre les dates t_1, t_2, t_3, etc.

La fonction de survie obtenue par Kaplan-Meier est constante par morceaux. C'est une courbe en escalier qui ne "saute" qu'en présence d'une observation de décès. Par conséquent, il y a autant de sauts de marche dans la fonction de survie que de dates de décès uniques. Comme toutes les fonctions de survie, elle démarre à 1 et décroît vers 0.

Un point important est que cet estimateur prend en compte les censures. Néanmoins, celles-ci doivent être non informationnelles, c'est-à-dire que la variable aléatoire de censure C ne doit pas dépendre de X ou inversement. Si aucune censure n'intervient, l'estimateur de Kaplan-Meier est équivalent à la fonction de survie empirique $\widehat{S}_n(t) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{t_k > t}$.

Notations

Soient les notations suivantes :

- n_i : le nombre d'individus en vie juste avant la date t_i ;
- d_i : le nombre d'individus décédés à la date t_i ;
- t_i : les dates, triées par ordre croissant, pour lesquelles il y a au moins un décès.

Formule de l'estimateur de Kaplan-Meier

L'estimateur de Kaplan-Meier estime la fonction de survie S :

$$\boxed{\hat{S}^{KM}(t) \;=\; \mathbb{P}(X > t) \quad = \prod_{i \;tq\; t_i \leq t} \Big(1 - \mathbb{P}(X = t_i | X \geq t_i)\Big) \;=\; \prod_{i \;tq\; t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)}.$$

Cela permet d'en déduire les q_x :

$$\boxed{\hat{q}_x^{KM} \;=\; 1 - \frac{\hat{S}^{KM}(x+1)}{\hat{S}^{KM}(x)}}.$$

Propriétés de l'estimateur de Kaplan-Meier

L'estimateur de Kaplan Meier possède les propriétés suivantes : [Pla22]

- il est convergent ;
- il est asymptotiquement gaussien ;
- il est un estimateur du maximum de vraisemblance.

Cependant, cet estimateur est biaisé positivement. Cela signifie que la mortalité peut être surestimée.

L'estimateur de Greenwood permet d'estimer la variance de l'estimateur :

$$\widehat{\mathrm{Var}}\Big(\hat{S}^{KM}(t)\Big) = \hat{S}^{KM}(t)^2 \times \sum_{t_i < t} \frac{d_i}{n_i \,(n_i - d_i)}.$$

Il est alors possible d'obtenir un intervalle de confiance asymptotique pour l'estimateur de Kaplan-Meier :

$$IC_{1-\alpha}\Big(S(t)\Big) = \Big[\hat{S}^{KM}(t) \;\pm\; \phi_{1-\frac{\alpha}{2}} \times \widehat{\mathrm{Var}}\Big(\hat{S}^{KM}(t)\Big)\Big],$$

où $\phi_{1-\frac{\alpha}{2}}$ représente le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale centrée réduite $\mathcal{N}(0, 1)$.

Mise en pratique

Pour appliquer Kaplan-Meier, la base d'exposition doit contenir les colonnes suivantes :

- date de naissance ;
- date d'effet du contrat ;
- date de clôture du contrat (= date de décès si décès) ;
- décès (booléen : True ou False).

Cela permet alors de calculer pour chaque ligne de la base les informations suivantes :

- date d'entrée en observation = max(date de début d'observation, date effet du contrat) ;
- date de sortie d'observation = min(date de fin d'observation, date de clôture du contrat) ;
- non censuré = True si décès et False sinon, car toutes les données sont censurées sauf en cas de décès.

Ensuite, la date de naissance permet de déterminer l'âge en entrée d'observation et à la sortie d'observation.

Finalement, il suffit d'appliquer Kaplan-Meier en fournissant les colonnes suivantes :

- âge d'entrée en observation ;
- âge de sortie d'observation ;
- non censuré.

Par exemple, avec le package *survival* sur *R*, l'appel à la fonction se fait ainsi :

```
wx_km <-
  survfit(
    Surv(
      age_entree_obs,
      age_sortie_obs,
      non_censure,
      type = "counting"
    ) ~ 1,
    type = "kaplan-meier",
    error = "greenwood",
    data = base_exposition,
    conf.int = .95,
    conf.type = "plain"
  )
```

La sortie de ce code fournit plusieurs listes de valeurs dont le nombre d'éléments est le même. Celui-ci est égal au nombre d'événements (clôture du contrat ou décès) dédoublonnés de la base d'exposition. Les *ex-aequo* ne sont donc pas pris en compte : s'il y a deux décès en âge 18.017796, ou trois clôtures de contrat, ou encore un décès et une clôture, alors la valeur 18.017796 n'apparaîtra qu'une fois.

Ainsi, *wx_km$time* contient les âges exacts où s'est produit un (ou plusieurs s'il y avait des *ex-aequo*) événement, comme 18.017796, 18.026010, 18.028747, etc. De plus, *wx_km$surv* représente les valeurs de la fonction de survie de Kaplan-Meier à ces âges d'événement. Pour résumer, nous ne possédons que les valeurs de $S(x)$ avec x un âge exact où s'est produit un décès ou une clôture dans la base d'exposition.

Pour obtenir les valeurs $S(x)$ avec x un âge spécifique (entier dans notre cas), voici la méthode utilisée en R :

$$S(x) = min(wx\_km\$surv[wx\_km\$time <= x]).$$

La valeur de S en un certain x est donc approximée par la dernière valeur de S connue juste avant l'âge x. Les $S(x)$ pouvant maintenant être calculés pour tous les âges voulus, il ne reste plus qu'à en déduire les q_x avec la formule définie précédemment. Notons qu'il est nécessaire d'utiliser $S(x) = \mathbf{max}(wx\_km\$surv[wx\_km\$time>=x])$ dans le cas particulier où aucun décès ne s'est encore produit, notamment pour calculer $S(18)$. Ne pas prendre en compte ce cas particulier donnerait un taux de mortalité nul à 18 ans.

II.2) Lissage des taux bruts

II.2.1) Métriques de comparaison des lissages

Il existe plusieurs métriques permettant de comparer des lissages, détaillées notamment par Koye[Koy19] et Guillon[Gui15]. Dans notre cas, les bornes sur lesquelles seront appliquées ces métriques sont $x_{min} = 18$ ans et $x_{max} = 65$ ans. Voici les métriques qui seront utilisées par la suite :

Ratio observés sur attendus

Aussi appelé *SMR* (Standardized Mortality Ratio), c'est une méthode simple et pragmatique qui permet de vérifier la fidélité de la table construite à la mortalité observée. Les décès attendus correspondent à l'exposition multipliée par le taux de mortalité. Un ratio $\frac{\text{Décès observés}}{\text{Décès attendus}}$ proche de 1 montre une bonne capacité de la table à prédire la mortalité d'expérience. S'il est supérieur à 1, alors les taux de mortalité sont sous-estimés, tandis qu'ils sont surestimés dans le cas où il est inférieur à 1.

$$\boxed{O/A(x) = \frac{D_x}{E_x \times q_x} \to 1}.$$

Fidélité des taux lissés aux taux bruts

Le critère de fidélité consiste à calculer la distance entre les taux lissés et les taux bruts. Il est également connu sous le nom de SCR (Somme des Carrés des Résidus, ou *Sum of Squared Residuals* en anglais). L'objectif est que cette distance soit petite pour que les taux lissés ne diffèrent pas trop des taux bruts. Plus la valeur $\sum_{(\text{âges x})} \left(q_x^{\text{Lissés}} - q_x^{\text{Bruts}}\right)^2$ est proche de 0, plus les taux lissés sont proches des taux bruts.

$$\boxed{F = \sum_{x=x_{min}}^{x_{max}} \left(q_x^{\text{Lissés}} - q_x^{Bruts}\right)^2 \to 0}.$$

Régularité des taux lissés

Le critère de régularité calcule la somme des accroissements de mortalité par âge x pour s'assurer que les taux de mortalité n'oscillent pas trop brusquement. C'est donc une mesure d'à quel point les taux sont « lisses ». Plus la valeur $\sum_{(\text{âges x})} \left(q_x^{\text{Lissés}} - q_{x+1}^{\text{Lissés}}\right)^2$ est proche de 0, plus les taux lissés sont réguliers.

$$\boxed{R = \sum_{x=x_{min}}^{x_{max}-1} \left(q_x^{\text{Lissés}} - q_{x+1}^{\text{Lissés}}\right)^2 \to 0}.$$

Statistique du R²

Le R carré, aussi connu sous le nom de *coefficient de détermination*, représente la proportion de la variance capturée par le modèle. Il est calculé comme la variation expliquée divisée par la variation totale. Ses valeurs sont entre 0 et 1 et plus le R carré est proche de 1, plus le modèle explique la variabilité des données brutes.

$$\boxed{\begin{aligned} R^2 &= \frac{\text{Variance expliquée}}{\text{Variance totale}} \\ &= \frac{\text{Variance totale - Variance résiduelle}}{\text{Variance totale}} \\ &= 1 - \frac{\sum_{x=x_{min}}^{x_{max}} \left(q_x^{Lissés} - q_x^{Bruts}\right)^2}{\sum_{x=x_{min}}^{x_{max}} \left(q_x^{Bruts} - \bar{q_x}^{Bruts}\right)^2} \\ &\to 1 \end{aligned}},$$

avec $\bar{q_x}^{Bruts} = \frac{1}{x_{max}-x_{min}+1} \times \sum_{x=x_{min}}^{x_{max}} q_x^{Bruts}$ la moyenne des taux bruts.

Statistique MAPE

La statistique $MAPE$ (*Mean Absolute Percentage Error*) mesure l'exactitude du lissage par rapport aux observations. Elle correspond à la moyenne des écarts relatifs entre les taux bruts et les taux lissés. Le modèle ayant la valeur $MAPE$ la plus faible sera privilégié.

$$\boxed{MAPE = \frac{1}{x_{max} - x_{min} + 1} \sum_{x=x_{min}}^{x_{max}} \left| \frac{q_x^{Bruts} - q_x^{Lissés}}{q_x^{Bruts}} \right|}.$$

Test du χ^2 d'adéquation aux taux bruts

La statistique ci-dessous doit suivre une loi du Khi-deux à (nombre d'âges) − 1 − (nombre de paramètres estimés) degrés de liberté. Dans notre cas, le nombre de degrés de liberté pour chacun des lissages que nous utiliserons est de $48 - 1 - 0 = 47$. En effet, il y a $65 - 18 + 1 = 48$ âges pour lesquels les taux bruts seront lissés, et aucun paramètre de modèle n'aura besoin d'être estimé car tous les lissages choisis par la suite sont des lissages non paramétriques. Le modèle ayant la valeur du χ^2 la plus faible sera privilégié.

$$\boxed{\chi^2_{Lissage} = \sum_{x=x_{min}}^{x_{max}} \frac{\left(n_x q_x^{\text{Bruts}} - n_x q_x^{\text{Lissés}}\right)^2}{n_x q_x^{\text{Lissés}}}},$$

avec n_x le nombre d'individus en vie à l'âge x.

II.2.2) Méthodes de lissage

Moyennes mobiles

Le lissage par moyennes mobiles est une méthode non paramétrique qui lisse localement chaque taux de mortalité. Ainsi, le taux en âge x se voit attribuer la valeur de la moyenne des taux sur la fenêtre $[x-a,\, x+b]$. La plupart du temps, $a = b$ pour prendre autant d'âges à gauche de x qu'à sa droite. Il existe plusieurs façons de calculer une moyenne : par moyenne arithmétique (somme des données, divisée par leur nombre n), moyenne géométrique (produit des données, pris à la racine n-ième), etc. La moyenne arithmétique étant celle usuellement utilisée, voici l'expression du taux en âge x lissé par une moyenne mobile arithmétique :

$$\boxed{q_x^{MM} = \frac{1}{1+a+b} \times \sum_{y=x-a}^{x+b} q_y}.$$

Notons n le nombre d'âges dans la fenêtre de la moyenne mobile : avec les notations précédentes, $n = 1+a+b$. La valeur de n sera souvent choisie impaire car cela permettra, en plus du taux en x, de prendre autant d'âges d'un côté que de l'autre de l'âge x : $\frac{n-1}{2}$. Plus n sera pris grand, plus les taux seront lissés, mais moins ils seront fidèles. À l'inverse, une faible valeur de n donnera une bonne fidélité mais pas nécessairement une bonne régularité. Par conséquent, le choix de n permet d'arbitrer entre la fidélité (pour un n faible) et la régularité (pour un n élevé) aux taux bruts.

Il convient aussi de remarquer que plus n est pris élevé, moins il y aura de taux pouvant être lissés par les moyennes mobiles. En effet, les âges extrêmes ne peuvent pas être lissés par moyenne avec les taux des âges avant et après eux, puisque par définition ils sont les âges extrêmes. Ainsi, un lissage par moyennes mobiles en prenant k âges de part et d'autre de chaque âge x ne pourra s'appliquer que pour $x \in [x_{min}+k,\, x_{max}-k]$. Dans le cas plus général d'une moyenne avec a âges avant et b âges après, le lissage ne s'appliquera que pour $x \in [x_{min}+a,\, x_{max}-b]$.

Whittaker-Henderson

Le lissage de Whittaker-Henderson a été proposé par Whittaker en 1922 pour construire des tables de mortalité, puis a été amélioré par les travaux de Henderson en 1924.[Bie23] Cette méthode étant non paramétrique, elle évite les problématiques du calibrage des paramètres, en particulier dans le cas de données peu volumineuses. Ce lissage consiste à déterminer les taux q_x^{WH} minimisant une fonction de coût M. Celle-ci est une combinaison linéaire $M = F + h \times R$ entre le critère de fidélité F et le critère de régularité R.

Le paramètre h permet de contrôler l'importance accordée à la fidélité et à la régularité dans la fonction de coût, et donc si l'une des deux fonctions doit être plus favorisée. En particulier, si $h = 0$, il n'y a pas de lissage car $q_x^{WH} = q_x^{Bruts}$. À partir de $h > 1$, le lissage de Whittaker-Henderson accorde un poids supérieur au lissage plutôt qu'à la fidélité. Concernant les autres nouveaux termes introduits dans la formule de coût ci-dessous, ils seront explicités dans les paragraphes suivants.

$$\boxed{M = F + h \times R = \sum_{x=x_{min}}^{x_{max}} w_x \left(q_x^{WH} - q_x^{Bruts}\right)^2 + h \times \sum_{x=x_{min}}^{x_{max}-z} \delta^z (q_x^{WH})^2}.$$

Dans le critère de fidélité, des poids w_x sont ajoutés. Cela permet de ne pas donner la même importance à la fidélité selon l'âge. Dans notre cas, ces poids dépendent de l'exposition. Ainsi, plus l'exposition est faible, plus le poids est faible. Le poids w_x en l'âge x est alors défini tel que :

$$w_x = \frac{E_x}{\max\limits_{x_{min} \leq x \leq x_{max}} (E_x) - \min\limits_{x_{min} \leq x \leq x_{max}} (E_x)}.$$

Un paramètre z est introduit dans le critère le régularité. Il représente le nombre de fois où est appliqué l'opérateur δ des *différences avant*.[Les11] Cet opérateur est la généralisation de la mesure de régularité définie précédemment $\delta(q_x^{WH}) = q_x^{WH} - q_{x+1}^{WH}$. Ainsi, δ^z est une fonction calculant les *différences avant d'ordre z*. Lorsque $z = 2$, $\delta^2(q_x^{WH}) = \delta \circ \delta(q_x^{WH}) = \delta(q_x^{WH} - q_{x+1}^{WH}) = (q_x^{WH} - q_{x+1}^{WH}) - (q_{x+1}^{WH} - q_{x+2}^{WH}) = q_x^{WH} - 2 \times q_{x+1}^{WH} + q_{x+2}^{WH}$. Plus généralement, $\delta^n(q_x^{WH}) = \delta \circ \delta^{n-1}(q_x^{WH}) = \ldots = \sum_{j=0}^{n} \binom{n}{j} (-1)^{n-j} q_{x+n-j}^{WH}$.

Passons maintenant sous forme matricielle pour plus de simplicité. Posons n le nombre d'âges à lisser, avec $n = x_{max} - x_{min} + 1$. Soient Q^{Bruts} et Q^{WH} les vecteurs de taille $(n \times 1)$ respectivement des q_x^{Bruts} et des q_x^{WH}. Notons également W la matrice diagonale de taille $(n \times n)$ contenant les poids w_x. Le lissage de Whittaker-Henderson consiste à déterminer la valeur du vecteur Q^{WH}.

$$Q^{Bruts} = \begin{pmatrix} q_{x_{min}}^{Bruts} \\ \ldots \\ q_{x_{max}}^{Bruts} \end{pmatrix}, \; Q^{WH} = \begin{pmatrix} q_{x_{min}}^{WH} \\ \ldots \\ q_{x_{max}}^{WH} \end{pmatrix}$$

$$\text{et } W = \begin{pmatrix} w_{x_{min}} & 0 & 0 & 0 & 0 \\ 0 & w_{x_{min}+1} & 0 & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & w_{x_{max}-1} & 0 \\ 0 & 0 & 0 & 0 & w_{x_{max}} \end{pmatrix}.$$

Le critère de fidélité se réécrit $F = \left(Q^{WH} - Q^{Bruts}\right)^T \times W \times \left(Q^{WH} - Q^{Bruts}\right)$. En outre, posons Δ^z la matrice telle que le critère de régularité $R = \sum_{x=x_{min}}^{x_{max}-z} \delta^z(q_x^{WH})^2$ soit égal à ${}^T\left(\Delta^z \times Q^{WH}\right) \times \left(\Delta^z \times Q^{WH}\right)$. Lorsque $z = 2$, comme calculé précédemment, $\delta^2(q_x^{WH}) = q_x^{WH} - 2 \times q_{x+1}^{WH} + q_{x+2}^{WH}$, ce qui donne la matrice Δ^2 de taille $(n-2, n)$ suivante :

$$\Delta^2 = \begin{pmatrix} 1 & -2 & 1 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \dots & 0 & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & \dots & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 1 & -2 & 1 \end{pmatrix}.$$

Plus généralement, Δ^n est de taille $(n - z, n)$. Ses termes sont les coefficients binomiaux d'ordre z avec alternance des signes et commençant positivement pour z pair.[Les11]

Finalement, en notations matricielles, la fonction de coût M se réécrit :

$$\boxed{\begin{aligned} M = {}&{}^T\left(Q^{WH} - Q^{Bruts}\right) \times W \times \left(Q^{WH} - Q^{Bruts}\right) \\ &+ h \times {}^T\left(\Delta^z \times Q^{WH}\right) \times \left(\Delta^z \times Q^{WH}\right) \end{aligned}}.$$

En développant :

$$\begin{aligned} M = {}&{}^TQ^{WH} \times W \times \left(Q^{WH} - Q^{Bruts}\right) - {}^TQ^{Bruts} \times W \times \left(Q^{WH} - Q^{Bruts}\right) \\ &+ h \times {}^TQ^{WH} \times {}^T\Delta^z \times \Delta^z \times Q^{WH} \\ = {}&{}^TQ^{WH} \times W \times Q^{WH} - {}^TQ^{WH} \times W \times Q^{Bruts} \\ &- {}^TQ^{Bruts} \times W \times Q^{WH} - {}^TQ^{Bruts} \times W \times Q^{Bruts} \\ &+ h \times {}^TQ^{WH} \times {}^T\Delta^z \times \Delta^z \times Q^{WH} \\ = {}&{}^TQ^{WH} \times W \times Q^{WH} - 2 \times {}^TQ^{WH} \times W \times Q^{Bruts} - {}^TQ^{Bruts} \times W \times Q^{Bruts} \\ &+ h \times {}^TQ^{WH} \times {}^T\Delta^z \times \Delta^z \times Q^{WH}. \end{aligned}$$

Pour la minimiser, M est dérivée par rapport au vecteur des taux lissés Q^{WH} : [Orf18]

$$\begin{aligned} \frac{\partial M}{\partial Q^{WH}} &= 2 \times W \times Q^{WH} - 2 \times W \times Q^{Bruts} - 0 + 2h \times {}^T\Delta^z \times \Delta^z \times Q^{WH} \\ &= 2 \times W \times \left(Q^{WH} - Q^{Bruts}\right) + 2h \times {}^T\Delta^z \times \Delta^z \times Q^{WH}. \end{aligned}$$

Le minimum se trouve en la valeur des taux lissés Q^{WH} annulant cette dérivée :

$$2 \times \left(W \times \left(Q^{WH} - Q^{Bruts}\right) + h \times {}^{T}\Delta^{z} \times \Delta^{z} \times Q^{WH}\right) = 0$$

$$\implies W \times \left(Q^{Bruts} - Q^{WH}\right) = h \times {}^{T}\Delta^{z} \times \Delta^{z} \times Q^{WH}$$

$$\implies W \times Q^{Bruts} = \left(W + h \times {}^{T}\Delta^{z} \times \Delta^{z}\right) \times Q^{WH}$$

$$\implies \boxed{Q^{WH} = \left(W + h \times {}^{T}\Delta^{z} \times \Delta^{z}\right)^{-1} \times W \times Q^{Bruts}}.$$

Lissage par noyaux discrets

Ce lissage utilise des techniques de noyaux discrets beta et a été implémenté en un package *R* sous le nom *DBKGrad*.[Maz14] Les noyaux discrets ont été choisis par les auteurs, car dans le contexte des tables de mortalité les variables telles que l'âge, l'année calendaire et la durée sont « pragmatiquement considérées comme discrets ». De plus, l'utilisation de noyaux bêta est motivée par « la réduction des biais aux frontières ».

Notons $\overset{\circ}{q}_y$ le taux de mortalité à l'âge y à lisser. Dans ce cas, les taux lissés par les noyaux discrets se calculent ainsi :

$$\boxed{\widehat{q}_x = \sum_{y \in \mathcal{X}} \frac{k_h(y; m = x)}{\sum_{j \in \mathcal{X}} k_h(j; m = x)} \overset{\circ}{q}_y = \sum_{y \in \mathcal{X}} K_h(y; m = x) \overset{\circ}{q}_y, \quad x \in \mathcal{X}},$$

où $k_h(\cdot; m)$ est la fonction de noyaux discrets, $m \in \mathcal{X}$ est le mode unique du noyau, $h > 0$ est l'intervalle (fixe) gouvernant le compromis biais-variance, et $K_h(\cdot; m)$ est le noyau normalisé. Comme l'âge est considéré discret avec des valeurs uniformément espacées, le lissage par noyaux défini précédemment est équivalent à un lissage par moyennes mobiles pondérées (Gavin et al. 1995).

$$k_h(x; m) = \left(x + \frac{1}{2}\right)^{\frac{m+\frac{1}{2}}{h(\omega+1)}} \left(\omega + \frac{1}{2} - x\right)^{\frac{\omega+\frac{1}{2}-m}{h(\omega+1)}}, \quad x \in \mathcal{X}.$$

Version normalisée :

$$K_h(x; m) = \frac{k_h(x; m)}{\sum_{y \in \mathcal{X}} k_h(y; m)}, \quad x \in \mathcal{X}.$$

II.3) Positionnement par modèle de Cox

Le modèle de Cox (1972)[Cox72], aussi appelé *modèle à risques proportionnels*, permet de déterminer la relation entre la survie et des variables explicatives. Dans sa forme la plus simple avec une seule variable explicative, ce modèle fait l'hypothèse que le taux instantané de mortalité $\mu(x)$ s'écrit sous la forme suivante :

$$\boxed{\mu(x|y) = \mu_0(x) \times exp(\beta_y)}.$$

Ici, y est la modalité d'une variable explicative Y et μ_0 est une fonction indépendante de Y. Par exemple, dans le cas d'une seule variable explicative Y étant le sexe, le modèle de Cox donne $\mu(x|H) = \mu_0(x) \times exp(\beta_H)$ et $\mu(x|F) = \mu_0(x) \times exp(\beta_F)$. Ainsi, avec Cox, le taux instantané de mortalité est décomposé comme une fonction de hasard de référence ne dépendant que de l'âge x, multipliée par l'exponentielle d'un coefficient ne dépendant que de la valeur de la variable Y.

Pour les exemples suivants avec plusieurs variables explicatives $X_1, \ldots, X_n$, supposons pour simplifier que celles-ci soient binaires : elles n'ont que deux modalités 0 et 1. Dans ce cas, le modèle de Cox donne $\mu(x|X_1, \ldots, X_n) = \mu_0(x) \times exp(\sum_{i=1}^{n} \beta_i X_i)$. Le taux instantané de mortalité est donc ici encore le produit d'un taux de mortalité de base $\mu_0(x)$ avec des exponentielles de coefficients ne dépendant que de la valeur des variables explicatives.

Enfin, si l'on prend deux individus avec les mêmes valeurs de $X_1, \ldots, X_n$ sauf en X_j, où $X_j = 1$ pour le premier individu et $X_j = 0$ pour le second, alors le ratio de leurs taux instantanés de mortalité sera constant en tout âge x. En effet, il vaudra :

$$\frac{\mu_0(x) \times exp(\sum_{i=1, i\neq j}^{n} \beta_i x_i) \times exp(\beta_j \times 1)}{\mu_0(x) \times exp(\sum_{i=1, i\neq j}^{n} \beta_i x_i) \times exp(\beta_j \times 0)} = exp(\beta_j).$$

Un effet multiplicatif des covariables sur une fonction de hasard de base est donc introduit. Le modèle est semi-paramétrique, car la fonction de hasard de base μ_0 n'a pas de forme prédéterminée, les seuls paramètres à déterminer sont les coefficients β_i des covariables.

Pour déduire les q_x des μ_x de Cox, il faut d'abord déterminer la fonction de risque cumulée $H(x) = \int_0^x \mu_t dt$. Ensuite, la valeur de la fonction de survie S est calculée avec la formule $S(x) = exp(-H(x))$. Enfin, il suffit d'appliquer comme avec Kaplan-Meier la formule $q_x = 1 - \frac{S(x+1)}{S(x)}$ pour trouver les q_x.

Le modèle de Cox est particulièrement utile dans le cas où l'on ne dispose pas de beaucoup de données pour certaines modalités de la variable explicative. Il permet alors de réaliser des abattements sur des populations sur certains produits, catégorie socioprofessionnelle, etc. En particulier, c'est le modèle de Cox qui avait été utilisé pour calculer les taux du réseau CE lors de la dernière certification de table. En effet, il n'y

avait pas assez de données CE pour pouvoir construire de manière fiable une table de mortalité pour ce réseau.

II.3.1) Hypothèses du modèle de Cox

Ce modèle fait deux hypothèses sur les données :

- Hypothèse des risques proportionnels : le rapport des risques instantanés (*hazard rate* en anglais) pour deux modalités de la variable explicative est indépendant du temps.
- Hypothèse de log-linéarité : $\log\left(h\left(t \mid Z_{i1}, \ldots, Z_{ip}\right)\right) = \log\left(h_0(t)\right) + \theta_0 Z_i$. Le logarithme du risque instantané est donc une fonction linéaire des Z_{ij}.

En particulier, l'hypothèse du risque proportionnel au cours du temps entre les différentes modalités de la variable explicative Y est forte et n'est pas toujours vérifiée. Elle doit donc être validée.

Pour s'assurer de la validité d'un positionnement utilisant le modèle de Cox, il est possible d'utiliser différents outils :

- les résidus de Schoenfeld : pour vérifier l'hypothèse de risque proportionnel au cours du temps ;
- les résidus de martingales : pour vérifier l'hypothèse de log-linéarité ;
- le test du log rank : pour vérifier que les fonctions de survie positionnées sont différentes et donc que les modalités de la variable explicative jouent un rôle sur la mortalité.

II.3.2) Test du log-rank

Le test du log-rank, parfois aussi appelé *test de Mantel–Cox*, est un test statistique permettant de comparer des fonctions de survie. Il a été proposé pour la première fois par Nathan Mantel en 1966[Man66] et fut ensuite nommé *test du log-rank* par les deux frères Richard et Julian Peto.

Ce test détermine si deux (ou plus) fonctions de survie diffèrent significativement ou non, leurs différences pouvant en effet n'être dues qu'à des fluctuations d'échantillonnage et donc au hasard. C'est un test non-paramétrique, car aucune hypothèse sur la distribution

des temps de survie n'est nécessaire. Il est largement utilisé dans les essais cliniques afin de mesurer l'efficacité de nouveaux traitements lorsque la mesure est un temps de survenance avant un événement.

L'hypothèse nulle de ce test est $\mathbf{H_0}$: *Les fonctions de survie sont les mêmes.* Ainsi, sous H_0, $\forall t > 0, S_A(t) = S_B(t)$. Cela signifie qu'en toute date t, le nombre de décès observés dans les groupes A et B doit être proportionnel au nombre de personnes à risque dans le groupe.

Formule générale de la statistique

Posons O_{A_i} et O_{B_i} le nombre de décès observés respectivement dans les groupes A et B à la date t_i. Le nombre de décès observés dans le groupe A est alors $O_A = \sum_i O_{A_i}$, et de même $O_B = \sum_i O_{B_i}$ pour le groupe B. De plus, soit N_i le nombre de personnes exposées au risque à la date t_i, qui se décompose ainsi : $N_i = N_{A_i} + N_{B_i}$. Enfin, notons E_{A_i} et E_{B_i} le nombre de décès estimés dans les deux groupes respectifs à la date t_i. Le nombre de décès estimés dans le groupe A est alors $E_A = \sum_i E_{A_i}$, et de même $E_B = \sum_i E_{B_i}$ pour le groupe B.

Pour toute date t_i où se produit un décès dans au moins l'un des groupes, O_{A_i} et O_{B_i} sont connus car c'est le nombre de décès dans chacune des deux sous-populations. De même, N_{A_i} et N_{B_i} sont connus : c'est le nombre de personnes à risque dans les sous-populations à la date t_i où s'est produit O_{A_i} et O_{B_i} décès dans les deux groupes.

Concernant E_{A_i} et E_{B_i}, ceux-ci **sont calculés sous l'hypothèse que H_0 soit vraie**. Ainsi, si l'hypothèse nulle est vérifiée, alors le nombre de décès pour un groupe est simplement le ratio du nombre de personnes décédées sur le nombre de personnes exposées au risque, multiplié par le nombre d'exposés au risque dans le groupe. En d'autres termes, sous H_0, le nombre de décès estimés pour le groupe A vaudra $E_{A_i} = N_{A_i} \times \frac{O_i}{N_i} = N_{A_i} \times \frac{O_{A_i}+O_{B_i}}{N_{A_i}+N_{B_i}}$.

Enfin, O_{A_i} (et O_{B_i} aussi par un raisonnement analogue) suit une loi hypergéométrique de paramètres $(N, K, n) = (N_i, N_{A_i}, O_i)$. Grâce aux formules de la moyenne et de la variance d'une loi hypergéométrique, il en découle que $E_{A_i} = \mathbb{E}(O_{A_i}) = O_i \times \frac{N_{A_i}}{N_i}$ et que $Var(O_{A_i}) = E_{A_i} \times \frac{N_i - O_i}{N_i} \times \frac{N_i - N_{A_i}}{N_i - 1}$.

$$\boxed{\text{Statistique du log-rank } Z = \frac{(E_A - O_A)^2}{Var(E_A - O_A)} = \frac{(E_B - O_B)^2}{Var(E_B - O_B)}},$$

$$
\begin{aligned}
\text{où } Var(E_A - O_A) &= Var\left(\sum_i (E_{A_i} - O_{A_i})\right) \\
&= \sum_i \left(Var\left(E_{A_i} - O_{A_i}\right)\right) \textit{ par indép. des v.a. } X_i = E_{A_i} - O_{A_i} \\
&= \sum_i \left((-1)^2 \times Var\left(O_{A_i}\right)\right) \textit{ car } Var(b + aX) = 0 + a^2 \times Var(X) \\
&= \sum_i \left(E_{A_i} \times \frac{N_i - O_i}{N_i} \times \frac{N_i - N_{A_i}}{N_i - 1}\right) \textit{ Var d'une loi hypergéo.} \\
&= \sum_i \left(N_{A_i} \times \frac{O_i}{N_i} \times \frac{N_i - O_i}{N_i} \times \frac{N_i - N_{A_i}}{N_i - 1}\right) \mathbb{E} \textit{ d'une loi hypergéo.} \\
&= \sum_i \left(\frac{N_{A_i} \times O_i \times (N_i - O_i) \times (N_i - N_{A_i})}{N_i^2 \times (N_i - 1)}\right) \\
&= \sum_i \left(\frac{N_{A_i} \times O_i \times (N_i - O_i) \times N_{B_i}}{N_i^2 \times (N_i - 1)}\right).
\end{aligned}
$$

La statistique du log-rank peut se calculer de deux manières : avec $Var(E_A - O_A)$ ou avec $Var(E_B - O_B)$. Comme le montre le développement précédent, $Var(E_A - O_A)$ fait aussi intervenir le groupe B, car il dépend à la fois du nombre de personnes à risque dans le groupe A et de ceux à risque dans le groupe B.

Formule simplifiée de la statistique

La formule exacte de la statistique du log-rank nécessite le calcul de la variance de la différence $(E_A - O_A)$. Cette opération peut être chronophage si l'on ne dispose pas d'un logiciel statistique. Elle peut aussi être compliquée à réaliser lorsqu'on compare plus de deux courbes de survie.

Il est alors possible d'utiliser une formule simplifiée qui donne un résultat avec une précision acceptable. Dans ce cas, la valeur obtenue est généralement un peu sous-estimée.[Agr11]

$$\boxed{\text{Statistique du log-rank simplifiée } \tilde{Z} = \frac{(E_A - O_A)^2}{E_A} + \frac{(E_B - O_B)^2}{E_B}}.$$

Interprétation de la statistique

Sous H_0, il est possible d'approximer la loi de Z ainsi : $\sqrt{Z} \sim \mathcal{N}(0,1)$ et $Z \sim \chi^2_1$.[Tia14] Afin de déterminer si l'hypothèse H_0 est vérifiée, le test du log-rank utilise donc la statistique du Khi-deux à 1 degré de liberté [1]. En effet, sous H_0, la statistique du log-rank suit approximativement une loi du Khi-deux à 1 degré de liberté.

Comme dans tous les tests statistiques, le test du log-rank retourne une p-valeur p qui est la probabilité que la loi du Khi-deux à 1 degré de liberté ait une valeur plus extrême que la valeur calculée par la statistique du log-rank. Si $p < 0{,}05$, alors l'hypothèse nulle est rejetée et on en conclut que les fonctions de survie diffèrent significativement.

Validité du test du log-rank

Plusieurs points sont à noter à propos du test du log-rank.[Alb05] D'abord, comme pour l'estimation des taux bruts avec Kaplan-Meier, le test du log-rank n'est valide que sous l'hypothèse de censure non informative. Ainsi, les dates de censure ne doivent pas dépendre de l'événement observé. Par exemple, un assuré pourrait être censuré parce qu'on n'a plus de nouvelles de lui, avec cependant une censure qui provient en réalité du fait qu'il soit décédé sans qu'on ne le sache.

De plus, dans le cas d'un croisement des fonctions de survie à partir d'un certain point, le test du log-rank peut tout de même conclure à l'égalité de ces fonctions. C'est pour cela qu'il convient d'effectuer également une analyse graphique en plus du test du log-rank.

Enfin, le test du log-rank se fonde sur une statistique attribuant des poids égaux à toutes les observations. Il existe d'autres tests qui donnent plus de poids aux décès précoces qu'aux décès tardifs dans la comparaison. Ces tests sont ceux de Wilcoxon (souvent rebaptisé test de Gehan) et de Peto-Prentice. Le test de Gehan dépend plus de la distribution des censures que le test de Peto-Prentice, son emploi n'est donc pas recommandé.[Alb05]

1) Pus généralement, s'il y a n fonctions de survie à comparer, la statistique utilisée est une statistique du Khi-deux à $n-1$ degrés de liberté.

II.4) Prolongement de table

Le prolongement de table permet d'extrapoler les taux de mortalité sur des périodes où il y a peu d'observation. Dans le cas où l'on veut prolonger la table de mortalité jusqu'à l'âge ultime ω où la survie est nulle, on parle de *fermeture de table*.

Dans notre portefeuille, il y a de moins en moins d'exposition dès l'âge de 65 ans, et une exposition quasi-nulle avant la date de fin maximale des garanties décès des contrats, fixée à 89 ans. Par conséquent, réaliser à un prolongement de table est nécessaire. Pour cela, la méthode de Coale et Kisker a été choisie, mais il en existe beaucoup d'autres.[Qua05] Les individus du portefeuille qui sont d'un âge élevé étant peu nombreux, l'importance du prolongement de table est de toute manière moindre. En outre, l'écart entre deux méthodes de fermetures en termes de provisionnement ne devient véritablement significatif qu'à des âges très élevés.[Pla11]

La méthode de Coale et Kisker (1990) est une méthode de fermeture de table de mortalité. Les taux de mortalité instantanés sont extrapolés à l'aide de la formule suivante :

$$\hat{\mu}_x = \mu_{65} \times e^{(x-65)\times g_x} \quad \forall x \geq 65.$$

Dans cette formule, g_x désigne le taux moyen de croissance exponentiel des μ_x entre 65 ans et x ans. Ainsi, $g_x = \frac{1}{x-65+1} \times \sum_{i=65}^{x} \frac{log\left(\frac{\mu_i}{\mu_{65}}\right)}{65-i}$ pour $x \geq 65$.

Enfin, comme l'on ne possède pas toujours les q_x après 80 ans, ceux extrapolés par la méthode de Coale et Kisker ne les nécessitent pas car il est supposé que :

$$\hat{\mu}_x = \hat{\mu}_{x-1} \times e^{(x-80)\times s + g_{80}} \quad \forall x \geq 80,$$

avec :

$$s = -\frac{ln(\hat{\mu}_{79}) + 31 \times g_{80}}{465} \quad \text{et} \quad g_{80} = \frac{1}{15} \times ln\left(\frac{\hat{\mu}_{80}}{\hat{\mu}_{65}}\right).$$

Lorsque les μ_x sont déterminés, il est facile d'en déduire les q_x. En effet, $\mu_x = -ln(1-q_x)$. Par conséquent,

$$q_x = 1 - e^{-\mu}.$$

II.5) Application des méthodes classiques

Les données extraites sont celles du deuxième trimestre 2023. Cependant, la période d'observation choisie s'arrête fin 2022. Cela laisse 6 mois de délai pour les déclarations tardives et devrait garantir d'avoir presque tous les décès s'étant produits pour ne pas sous-estimer la mortalité. Une période d'observation de cinq ans est alors choisie : du 01/01/2018 au 31/12/2022.

II.5.1) Construction des taux bruts

Pour construire les taux bruts, il est nécessaire de disposer des données suivantes pour chaque individu du portefeuille :

- date de naissance (pour calculer l'âge) ;
- date d'effet du contrat (date d'entrée de l'individu dans l'effectif exposé au risque) ;
- date de clôture du contrat (date de sortie de l'individu dans l'effectif exposé au risque) ;
- date de décès si un décès s'est produit.

À l'aide du package *survival* de *R*, les méthodes de Kaplan-Meier et de Hoem ont été appliquées à notre portefeuille. Le choix de faire deux méthodes et non de se limiter à une seule permet de comparer leurs résultats. Les taux obtenus par ces deux moyens sont censés être très proches, ce qui permettra de contrôler si les résultats sont cohérents.

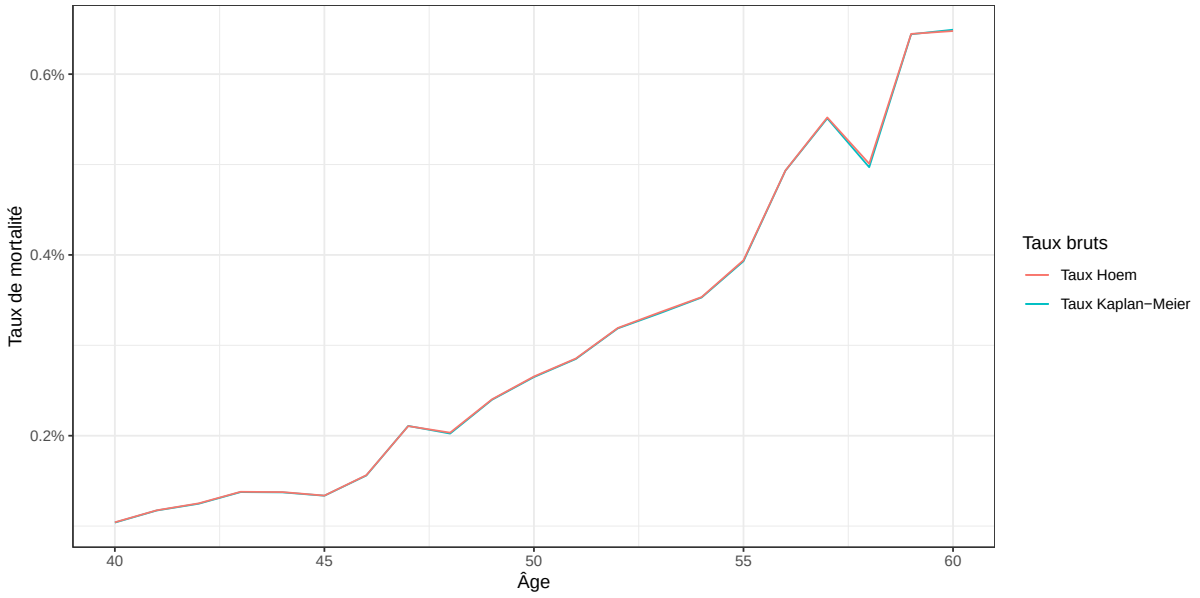

Figure II.3 — *Comparaison des taux bruts de Hoem et de Kaplan-Meier*

La figure II.3 montre des taux quasiment identiques, les courbes étant presque confondues. Les taux obtenus avec l'estimateur de Kaplan-Meier seront choisis pour la suite car celui-ci possède de meilleures propriétés.

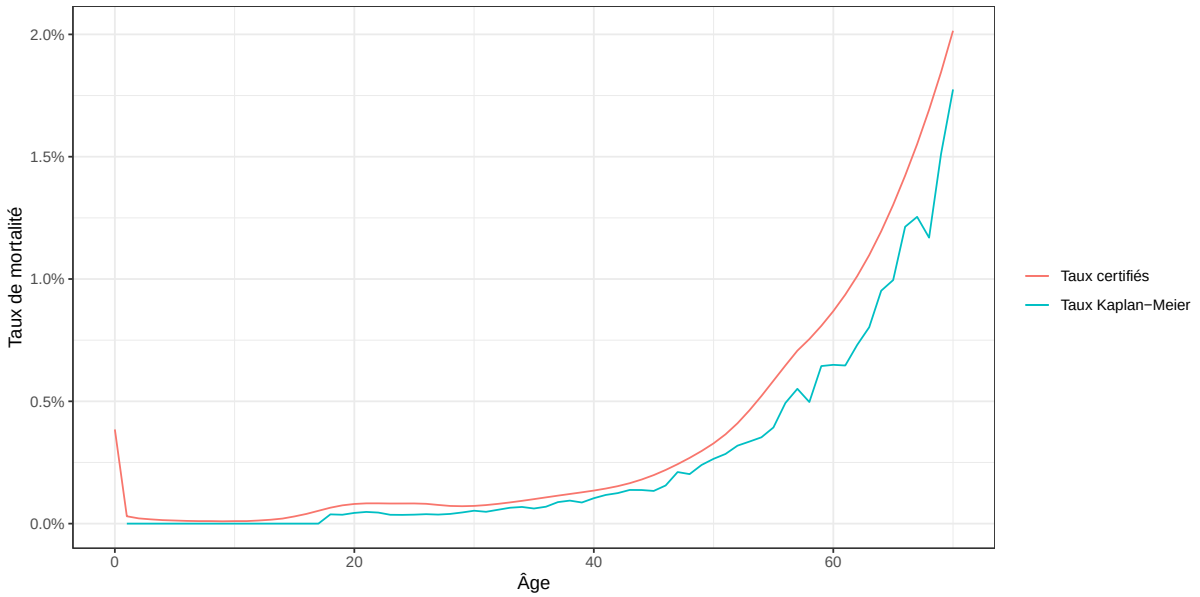

Figure II.4 — *Comparaison des taux Kaplan-Meier avec ceux certifiés*

Enfin, la figure II.4 montre que les taux de Kaplan-Meier sont relativement proches de ceux qui ont été certifiés. La certification précédente a eu lieu en 2021 : les années d'exposition utilisées pour calculer les taux de mortalité ne sont pas les mêmes, ce qui peut expliquer ces différences. D'autant que la table certifiée en 2021 a uniquement été réalisée sur les Banques Populaires, contrairement à nos taux qui incorporent les données

des deux réseaux (BP et CE). Enfin, les taux certifiés sont déjà lissés, contrairement à ceux Kaplan-Meier que nous venons de calculer et sur lesquels nous n'avons pas encore appliqué de lissage. Le lissage utilisé par les taux certifiés était celui des noyaux discrets, qui a tendance à surestimer la mortalité en suivant les pics supérieurs de la courbe.

II.5.2) Lissage

Les différents lissages ont été appliqués : les moyennes mobiles avec une fenêtre de 3 ans en figure II.5, Whittaker-Henderson en figure II.6 ainsi que les noyaux discrets en figure II.7.

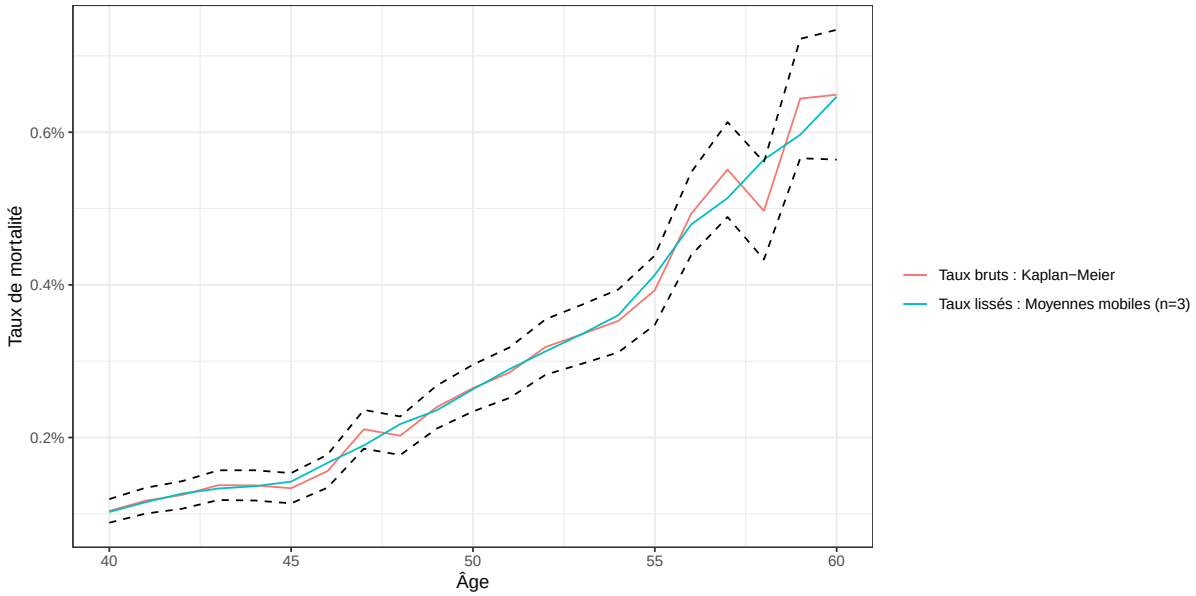

Figure II.5 — *Lissage des taux de Kaplan-Meier par les moyennes mobiles (n=3)*

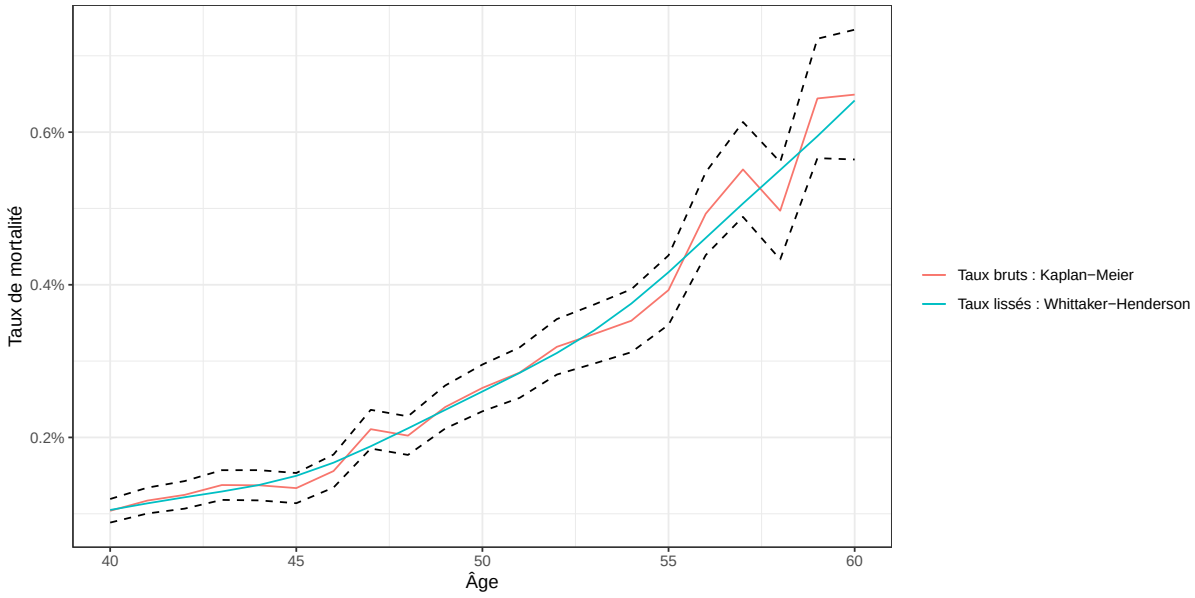

Figure II.6 — *Lissage des taux de Kaplan-Meier par Whittaker-Henderson*

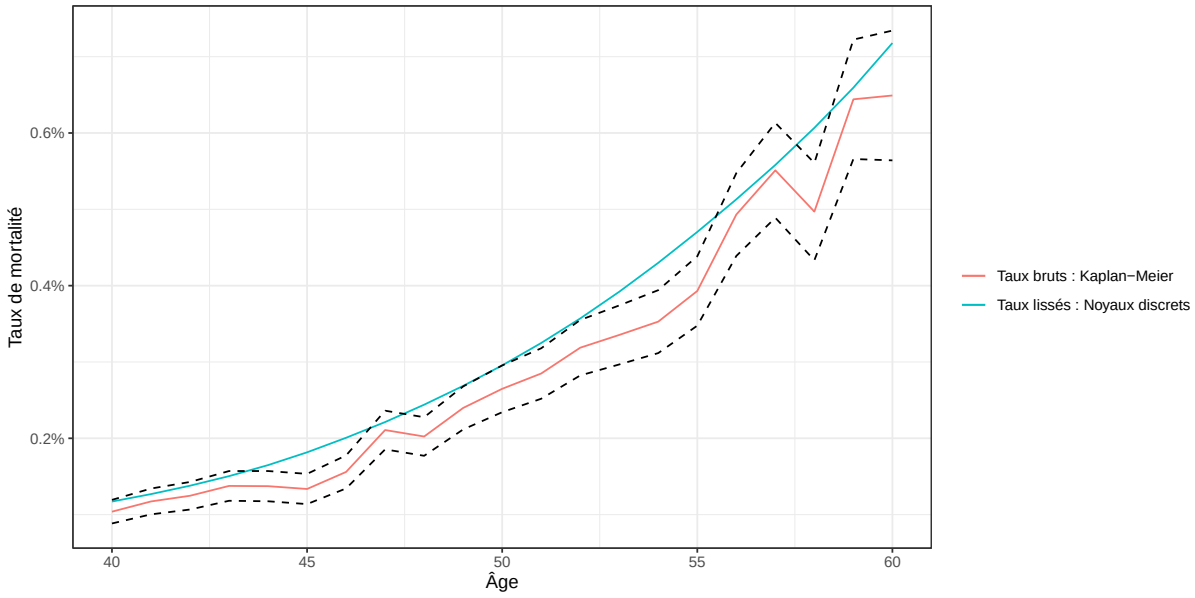

Figure II.7 — *Lissage des taux de Kaplan-Meier par les noyaux discrets*

Le graphique comparant ces différents lissages est en figure II.8. Dans celui-ci, les taux des moyennes mobiles n'ont pas été tracés, car ils sont très proches de ceux de Whittaker-Henderson mais en moins lisses, en témoigne la figure II.9.

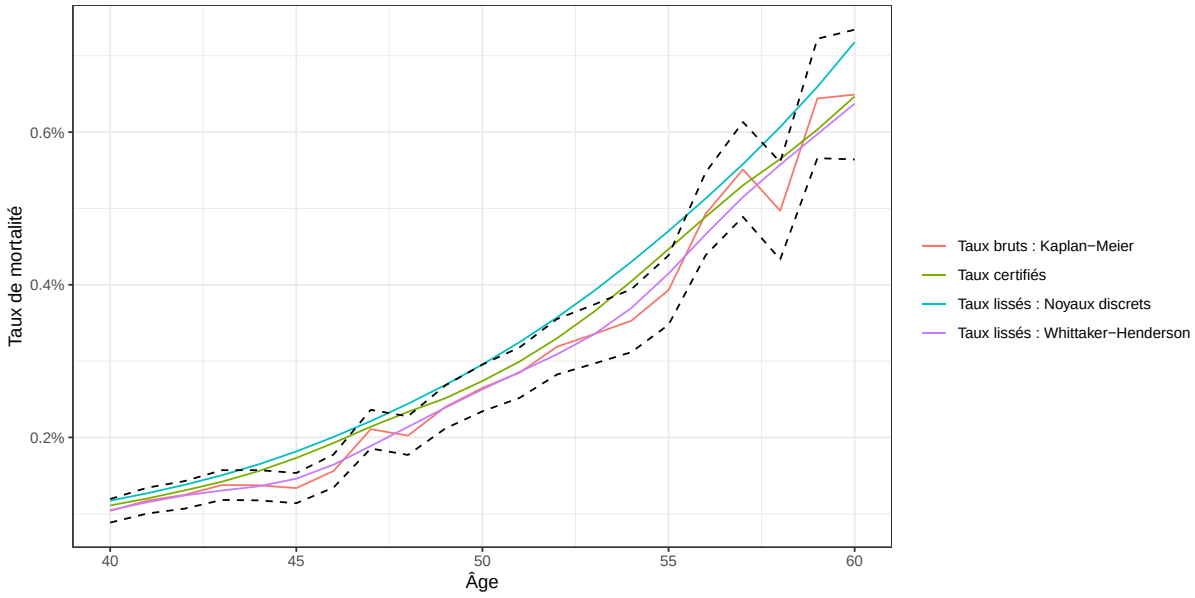

Figure II.8 — *Comparaison des différents lissages des taux de Kaplan-Meier*

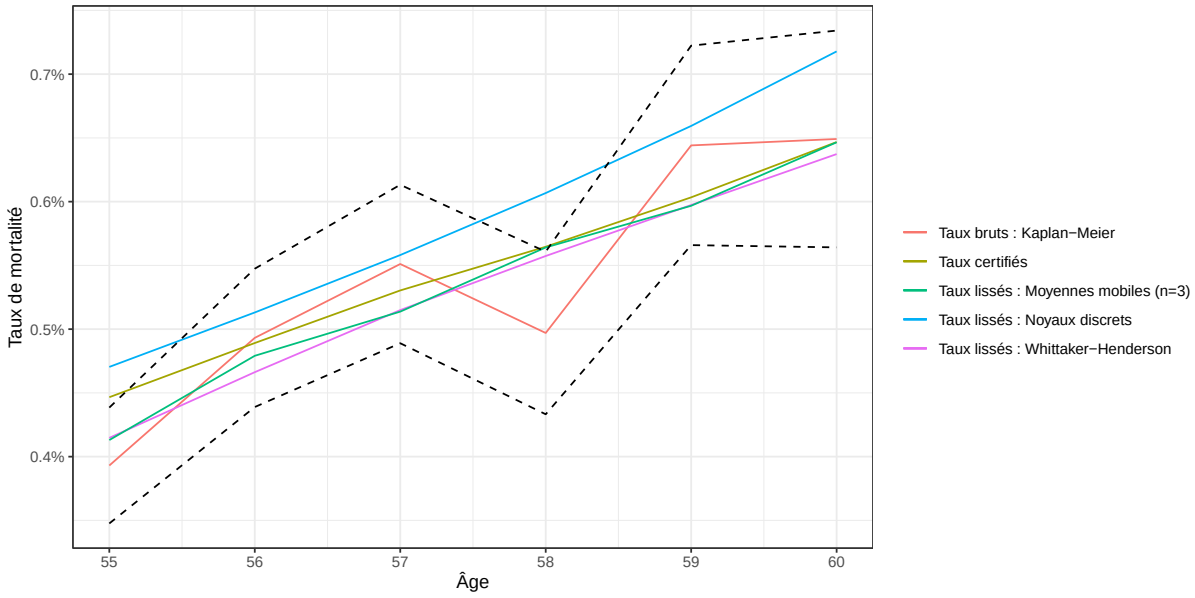

Figure II.9 — *Zoom sur la comparaison des différents lissages des taux de Kaplan-Meier*

Vérifions maintenant les valeurs des différents indicateurs de lissage (régularité, fidélité, etc.) Ceux-ci sont présentés dans le tableau II.1. Pour rappel, les critères de fidélité et de régularité doivent être proches de 0, tandis que le R^2 doit être proche de 1. De plus, le ratio $\frac{\text{Observés}}{\text{Attendus}}$ doit également être proche de 1, voire un peu en-dessous à titre de prudence. La statistique $MAPE$ doit quant à elle être la plus faible possible. Enfin, la valeur de la statistique du χ^2 doit être la plus faible possible, ce qui garantira que sa p-valeur sera $> 0{,}05$.

Lissage	$\frac{\text{Obs.}}{\text{Att.}}$	Fidélité	Régularité	R^2	MAPE	χ^2	p-val. du χ^2
Hoem	0,999	0	$8{,}3 \times 10^{-6}$	1	0	0	1
Kaplan Meier	1,004	0	$8{,}6 \times 10^{-6}$	1	0	0,4	1
Moy. mobiles (n=3)	1,012	$1{,}6 \times 10^{-6}$	$6{,}1 \times 10^{-6}$	0,995	5,5%	33	0,939
Moy. mobiles (n=5)	1,014	$1{,}9 \times 10^{-6}$	$5{,}5 \times 10^{-6}$	0,994	7,6%	46	0,514
Moy. mobiles (n=7)	1,013	$2{,}0 \times 10^{-6}$	$4{,}7 \times 10^{-6}$	0,994	8,8%	57	0,151
Whittaker Henderson	1,002	$1{,}2 \times 10^{-5}$	$4{,}7 \times 10^{-6}$	0.995	5,5%	35	0,902
Noyaux discrets	0,901	$1{,}6 \times 10^{-6}$	$5{,}7 \times 10^{-6}$	0,962	11,7%	129	10^{-9}

Tableau II.1 — *Comparaison des lissages*

Le tableau II.1 montre que les noyaux discrets est le seul lissage à avoir un ratio $\frac{\text{Observés}}{\text{Attendus}} < 1$. Cependant, les taux lissés de cette manière ne sont alors plus suffisamment proches des taux bruts, en témoigne la p-valeur du test du χ^2 qui est la seule à être inférieure au seuil de 0,05.

Le lissage de Whittaker-Henderson semble quant à lui donner d'excellents résultats sur toutes les métriques. C'est donc celui-ci qui sera choisi pour la suite.

II.5.3) Prolongement de table

Un prolongement de table a donc été réalisé avec Coale et Kisker. Il se trouve en figure II.10. Les taux de mortalité prolongés suivent bien la condition d'égalité à 1 à l'âge limite $\omega = 120$ ans.

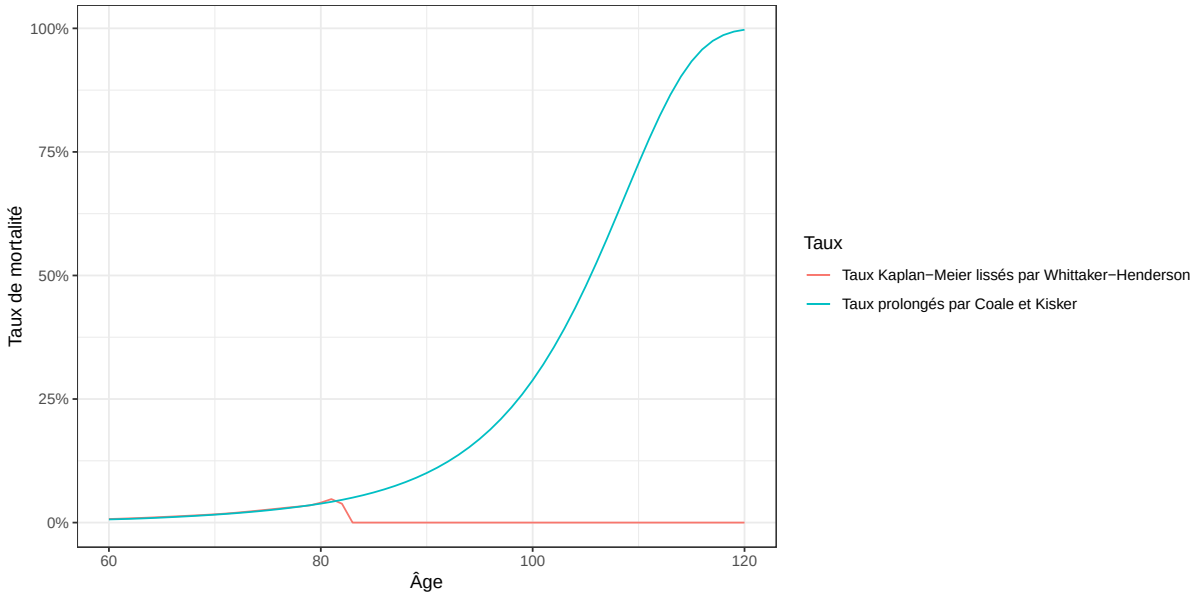

Figure II.10 — *Prolongement de table par Coale et Kisker*

II.5.4) Positionnements

Enfin, le modèle de Cox a été appliqué pour positionner la mortalité selon le sexe et selon le réseau. Nous allons d'abord vérifier que les hypothèses du modèle sont valides, puis calculer les taux de mortalité positionnés.

Vérification des hypothèses

La figure II.11 montre que les résidus de Schoenfeld pour le positionnement par sexe sont assez bien alignés, validant l'hypothèse des hasards proportionnels pour le sexe. En outre, la statistique du log-rank admet une valeur de 835,9 équivalente à une p-valeur $< 2 \times 10^{-16}$ et donc inférieure au seuil de 0,05. L'hypothèse H_0 d'égalité des courbes est rejetée et on en déduit que les courbes de survie par sexe sont significativement distinctes.

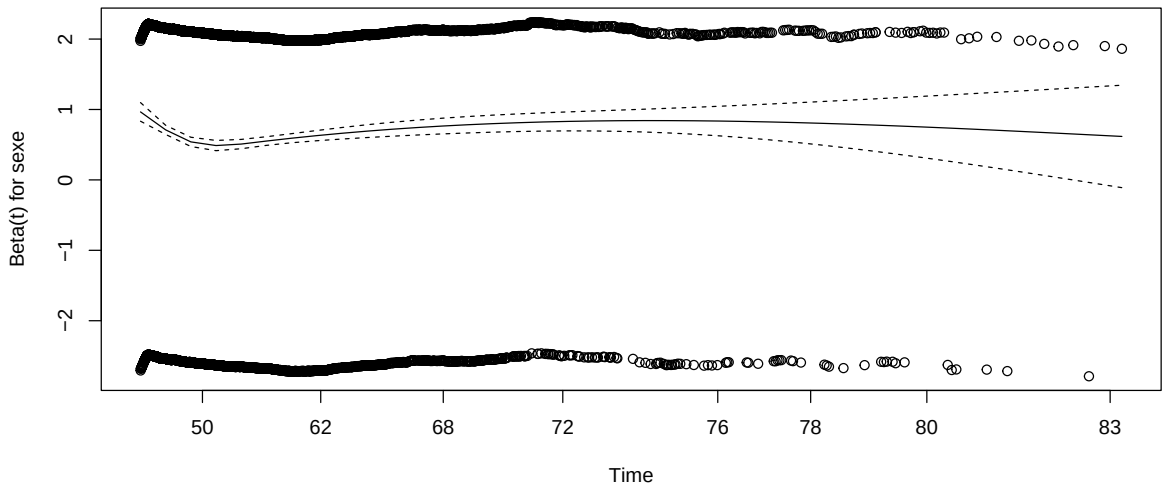

Figure II.11 — *Résidus de Schoenfeld du positionnement de Cox par sexe*

Le positionnement par sexe détermine un coefficient pour les hommes de 0,647 et l'exponentielle de celui-ci vaut 1,910. Cela signifie que, selon le modèle de Cox, **la mortalité instantanée des hommes est 1,9 fois celle des femmes**, qui est ici la mortalité instantanée de base. C'était un résultat attendu, les hommes ayant effectivement à âge égal un taux de mortalité supérieur aux femmes.

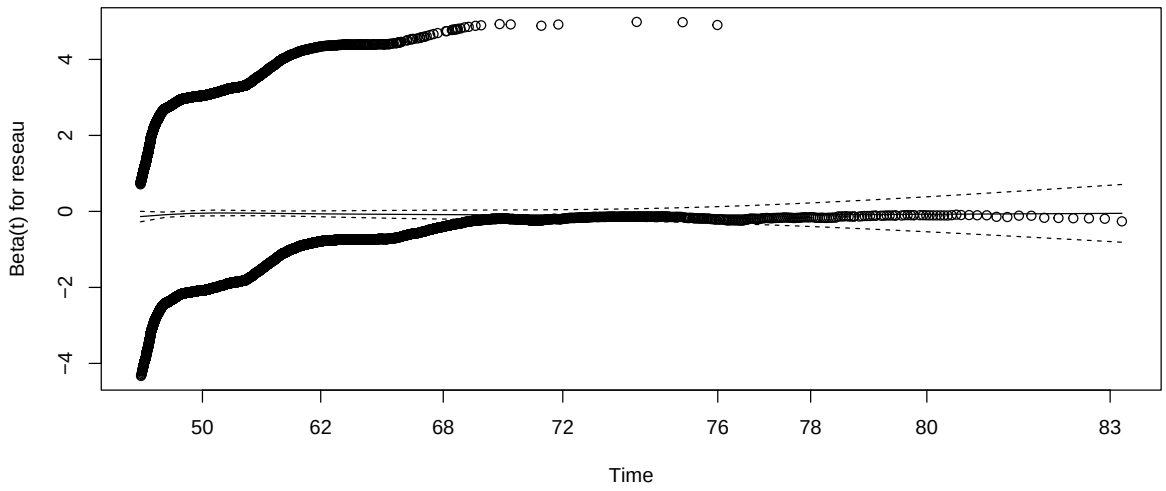

Figure II.12 — *Résidus de Schoenfeld du positionnement de Cox par réseau*

Concernant le réseau, la figure II.12 montre que les résidus de Schoenfeld sont là aussi alignés, validant l'hypothèse des hasards proportionnels. Il est toutefois à noter qu'aucun décès n'est observé pour le réseau Caisse d'Epargne après environ 70 ans, ce qui explique

pourquoi il n'y a plus de points sur la courbe à partir de cet âge pour le réseau CE contrairement au réseau BP.

En outre, la statistique du log-rank admet une valeur de 8,71 équivalente à une p-valeur de 0,003 et donc inférieure au seuil de 0,05. L'hypothèse H_0 d'égalité des courbes est rejetée, permettant d'en déduire que les courbes de survie par sexe sont significativement distinctes. Comme l'on aurait pu s'en douter, les courbes de survie par sexe diffèrent statistiquement plus entre elles que celles par réseau, la p-valeur pour le sexe étant extrêmement petite, largement plus petite que celle pour le réseau.

Le coefficient trouvé pour le positionnement des CE est de $-0{,}070$ et l'exponentielle de celui-ci vaut 0,932. Cela signifie que le modèle de Cox prédit que **la mortalité instantanée des Caisses d'Epargne est 0,9 fois celle des Banques Populaires**, qui correspond ici à la mortalité instantanée de base. La mortalité des Caisses d'Epargne est donc plus faible que celle des Banques Populaires.

Tracé des taux de Cox

Le tracé des taux positionnés par sexe se trouve en figure II.13. De même, la figure II.14 représente les taux positionnés par réseau. Ces positionnements seront comparés aux taux bruts ainsi qu'aux taux des processus gaussiens dans la dernière section du chapitre III.

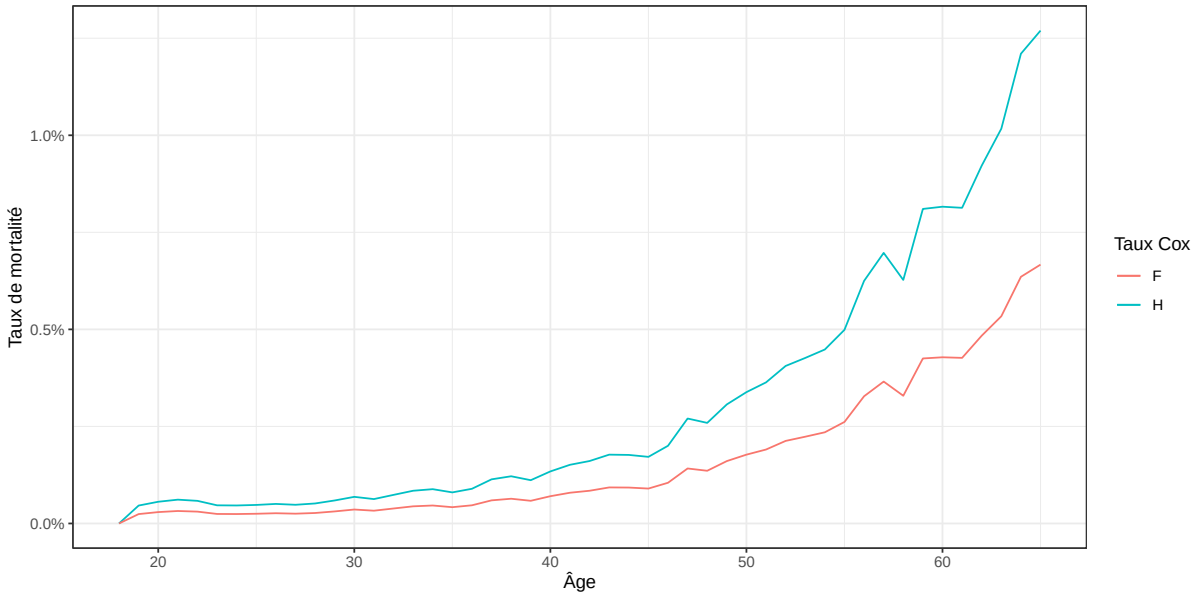

Figure II.13 — *Positionnement de Cox par sexe*

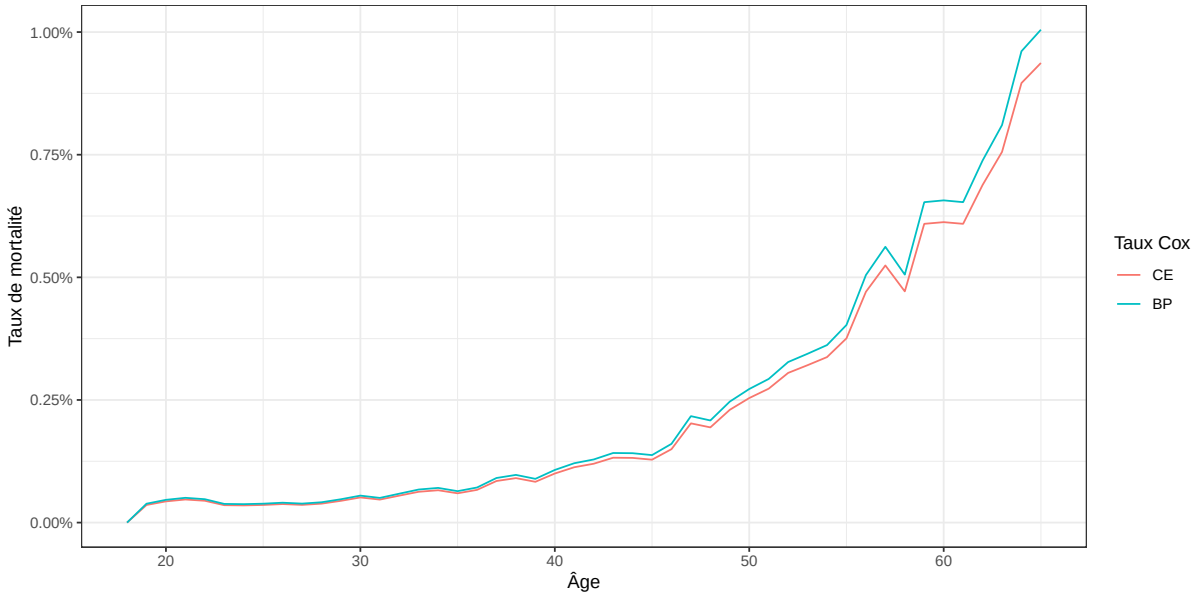

Figure II.14 — *Positionnement de Cox par réseau*

CHAPITRE III

PROCESSUS GAUSSIENS

III.1) Présentation générale

III.1.1) Généralités

Loi normale

Définition 1 — Loi normale

La variable aléatoire X suit une loi normale (aussi appelée *gaussienne*) $\mathcal{N}(\mu, \sigma^2)$ si sa densité de probabilité vaut :

$$f(x) = \frac{1}{\sqrt{2\pi} \times \sigma} \times exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

où μ représente la moyenne et σ^2 la variance.

Loi normale multivariée

Définition 2 — Loi normale multivariée

La loi normale multivariée est une distribution de probabilité généralisant la distribution normale à plusieurs dimensions. Une variable aléatoire Y suivant une loi normale multivariée est caractérisée par un vecteur de moyennes et une matrice de covariance. La densité de probabilité de la loi normale multivariée est définie comme suit :

$$f(Y) = \frac{1}{(2\pi)^{d/2} \times |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(Y-\mu)^T \Sigma^{-1} (Y-\mu)\right),$$

où :

- Y est le vecteur aléatoire multivarié ;
- d est la dimension du vecteur aléatoire ;
- $\mu = \mathbb{E}[Y] \in \mathbb{R}^d$ est le vecteur des moyennes ;
- Σ est la matrice (symétrique) de covariance.

Par analogie avec la loi normale, la loi normale multivariée est notée $\mathcal{N}(\mu, \Sigma)$.

Processus stochastiques

Définition 3 — Processus stochastique[Bre06]

Un processus stochastique $X = (X_t)_{t \in T}$ est une collection de variables aléatoires X_t indexée par un ensemble T.

En général, $T = \mathbb{R}$ ou $\mathbb{R}^+$ et l'on considère que le processus est indexé par le temps t. Si T est un ensemble fini, alors le processus est un vecteur aléatoire. Si $T = \mathbb{N}$, alors le processus est une suite de variables aléatoires. Plus généralement, quand $T \subset \mathbb{Z}$, le processus est dit discret. Enfin, quand $T \subset \mathbb{R}^d$, on parle de champ aléatoire (et de *drap* quand $d = 2$).

Il existe de nombreuses classes de processus particuliers : les processus de Markov (et notamment les chaînes de Markov quand T est discret), les martingales, les processus gaussiens, les processus de Poisson, les processus stables ou encore les processus de Lévy. Les lois de ces processus vérifient parfois des propriétés qui peuvent être utiles pour modéliser des phénomènes réels.

Processus gaussiens

Définition 4 — Processus gaussien[Sel06]

Un processus $(Y_x)_{x \in \chi}$ est dit gaussien si : $\forall n, \forall x_1, \ldots, x_n \in \chi$, le vecteur $Y = (Y_{x_1}, \ldots, Y_{x_n})$ est gaussien, c'est-à-dire qu'il suit une loi normale multivariée.

Parmi les cas particuliers de processus gaussiens, il y a notamment le mouvement brownien, le pont brownien, le processus d'Ornstein-Uhlenbeck, le brownien géométrique, le bruit blanc gaussien, etc.

III.1.2) Régression par processus gaussiens

La régression par processus gaussiens, en anglais *Gaussian Process Regression* ou *GPR*, est une technique de régression spatiale, aussi appelée *kriging*. Elle est utilisée pour effectuer des prédictions sur de nouvelles données à partir d'un ensemble de données d'entraînement, qui peuvent être des observations bruitées. C'est un modèle non paramétrique, il va donc s'adapter aux données sans avoir besoin d'imposer une certaine forme au préalable. Cette méthode est particulièrement utile lorsque les relations entre les données ne sont pas linéaires et pour que l'incertitude des prédictions soit prise en compte.

L'idée fondamentale derrière les processus gaussiens est de modéliser les relations entre les données en supposant que la valeur en chaque point suit une distribution gaussienne (normale). Ainsi, plutôt que de faire des prédictions précises à partir de fonctions mathématiques spécifiques, les processus gaussiens permettent de modéliser l'ensemble des valeurs possibles en un point qui sont cohérentes avec les données observées. En d'autres termes, au lieu de déterminer une seule fonction f qui ajuste les données en donnant la valeur de $f(x)$ en tout point x, on modélise plutôt la distribution de toutes les fonctions f potentielles qui pourraient les générer. Les processus gaussiens fournissent donc non seulement des prédictions, mais aussi une estimation de l'incertitude associée à chaque prédiction.

Un processus gaussien est entièrement défini par une fonction de moyenne et une fonction de covariance. La moyenne représente la tendance centrale des données, tandis que la covariance indique comment les différentes valeurs de données sont liées les unes aux autres. En ajustant les fonctions de moyenne et de covariance du processus gaussien aux données d'entraînement, il est ensuite possible d'effectuer des prédictions sur de nouvelles données non observées en générant une distribution de probabilité de leurs valeurs. Cela sera réalisé en utilisant la moyenne et la matrice de covariance du processus gaussien.

III.1.3) Exemple simple de régression par processus gaussiens

Supposons que nous ayons quelques points de données dans un espace à 3 dimensions, avec leurs valeurs cibles :

Point	Dimension 1	Dimension 2	Dimension 3	Valeur cible
A	1.0	2.0	3.0	5.0
B	2.5	1.5	2.0	7.2
C	3.0	3.0	1.0	4.8

Tableau III.1 — *Exemple simple de données d'apprentissage d'un processus gaussien*

Prédisons la valeur cible pour un nouveau point $T = [2.0, 2.0, 2.0]$.

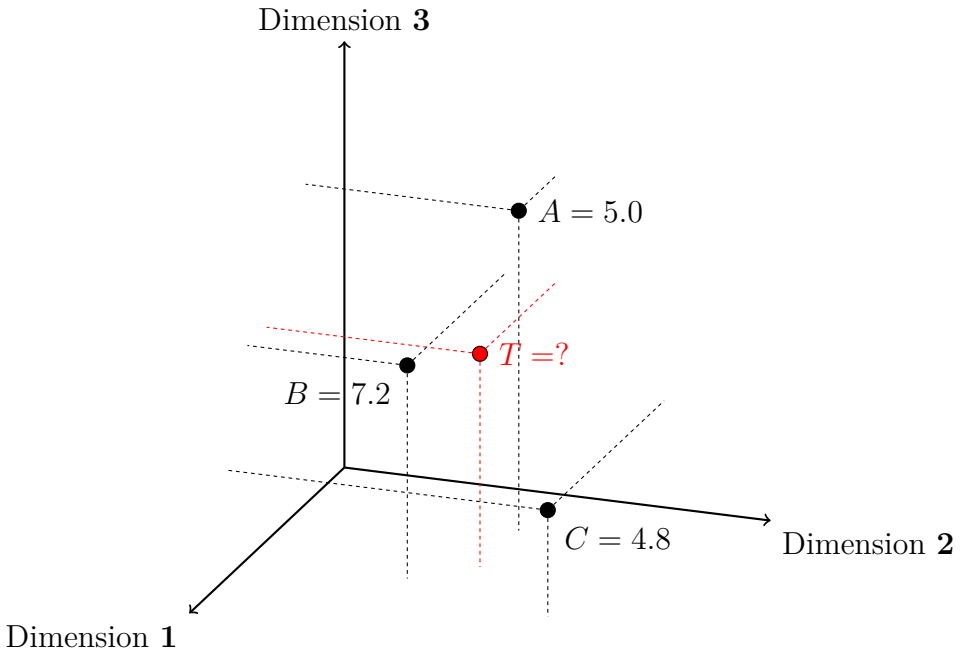

Figure III.1 — *Représentation des points dans un espace 3D avec T le point inconnu*

Données d'entraînement bruitées

Cet exemple est inspiré de l'explication réalisée par Ebden qui incorpore le terme de bruit σ_n^2 directement dans la fonction de covariance pour plus de simplicité. [Ebd08] Les données d'entraînement sont effectivement supposées bruitées par un bruit gaussien de variance σ_n^2. Ainsi, en tout point x où l'on mesure la valeur de la variable d'intérêt, la mesure réalisée n'est pas exactement la vraie valeur, mais plutôt la somme de la vraie valeur avec un terme aléatoire correspondant au bruit :

$$y = f(x) + \mathcal{N}\left(0, \sigma_n^2\right).$$

En tout point inconnu x_*, nous allons prédire la valeur y_* et non $f(x_*)$. En effet, les espérances de ces deux termes sont les mêmes, l'espérance du bruit gaussien étant nulle.

Cette espérance sera la valeur prédite et la variance permettra quant à elle de donner des intervalles de confiance.

Calcul de la similarité

Pour appliquer les processus gaussiens, il convient de choisir une fonction de covariance qui va permettre de mesurer la similarité entre deux points x_i et x_j de l'espace. Dans cet exemple, prenons la fonction RBF, à laquelle nous incorporons un terme contenant le bruit des observations σ_n^2 :

$$k\left(x, x'\right) = \sigma_f^2 \exp\left[\frac{-\left\|x - x'\right\|^2}{2l^2}\right] + \sigma_n^2 \delta\left(x, x'\right).$$

Dans cette formule, σ_f^2 et l sont des constantes à fixer, dont la valeur optimale peut être déterminée par une optimisation des hyperparamètres. De plus, δ est la fonction delta de Kronecker, valant 1 si $x = x'$ et 0 sinon. Ainsi, on a $\forall x, k(x, x) = \sigma_f^2 + \sigma_n^2$ qui est le maximum que peut prendre la fonction k.

Pour chaque point d'entraînement A, B et C, sa similarité avec le point inconnu T est calculée. La similarité entre chaque paire de points d'entraînement est également déterminée. Pour cela, la fonction k définie précédemment qui permet de donner une mesure de la similarité entre deux points est utilisée. Cela donne le vecteur K_* et la matrice K :

$$K_* = \left[k(T, A),\, k(T, B),\, k(T, C)\right],$$

$$K = \begin{bmatrix} k(A, A) & k(A, B) & k(A, C) \\ k(B, A) & k(B, B) & k(B, C) \\ k(C, A) & k(C, B) & k(C, C) \end{bmatrix}.$$

Loi suivie par la valeur y_* en T

L'hypothèse clé des processus gaussiens est que les valeurs cibles peuvent être représentées comme un échantillon d'une distribution gaussienne multivariée. Par conséquent :

$$\begin{pmatrix} y \\ y_* \end{pmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K & K_*^{\mathrm{T}} \\ K_* & k(T,T) \end{bmatrix}\right),$$

avec $y = [y_A, y_B, y_C]$ dans cet exemple. Il correspond au vecteur des sorties pour les points d'entraînement. La valeur en T que l'on veut déterminer est notée y_*. La moyenne μ de cette loi normale est 0, d'où le 0 dans le $\mathcal{N}$. Il est possible de choisir une fonction de moyenne non nulle, mais cela n'est pas nécessaire dans cet exemple simplifié.

L'utilisation de la formule de la loi conditionnelle d'une loi normale multivariée permet d'en déduire [1] la loi suivie par y_* sachant y :

$$y_* \mid y \sim \mathcal{N}\left(K_* K^{-1} y, \quad k(T,T) - K_* K^{-1} K_*^{\mathrm{T}}\right).$$

Ainsi, le meilleur estimateur de y_*, qui est la valeur cible au point T, est l'espérance :

$$\boxed{y_* = K_* K^{-1} y}.$$

De plus, l'incertitude de l'estimation est contenue dans sa variance :

$$\boxed{Var(y_*) = k(T,T) - K_* K^{-1} K_*^{\mathrm{T}}}.$$

Conclusion

Estimer la valeur y_* au point T revient donc à un simple produit matriciel de 3 termes. Celui-ci nécessite de calculer le vecteur de similarité des données d'entraînement avec le point T ainsi que d'inverser la matrice K de similarité entre les données d'entraînement.

C'est le calcul de cette inversion de matrice qui est le plus coûteux en temps. [Del03] Ce calcul est effectivement en $O(N^3)$, avec N le nombre de données d'entraînement. Heureusement, ce calcul n'a besoin d'être réalisé qu'une fois, lors de l'entraînement du modèle. Ensuite, la prédiction est en $O(N)$ et le calcul des intervalles de confiance est en $O(N^2)$.

Il est à noter qu'il n'est pas nécessaire de faire ces calculs point par point. Effectivement, il est possible de remplacer y_* par un vecteur pour prédire plusieurs points d'un coup. Les formules restent les mêmes, sauf pour $k(T,T)$ qui devient un vecteur K_{**} :

$$K_{**} = \left[k(x_*^1, x_*^1),\ k(x_*^2, x_*^2),\ \ldots\ ,\ k(x_*^n, x_*^n)\right].$$

Dans ce cas, la formule se réécrit donc :

$$y_* \mid y \sim \mathcal{N}\left(K_* K^{-1} y, \quad K_{**} - K_* K^{-1} K_*^{\mathrm{T}}\right).$$

1) La démonstration de cette formule est assez longue. Pour une démonstration étape par étape, se référer au lien https://statproofbook.github.io/P/mvn-cond. L'idée est d'utiliser la définition des probabilités conditionnelles $\mathbb{P}(y_* \mid y) = \frac{\mathbb{P}(y_*,\, y)}{\mathbb{P}(y)}$. Ainsi, $\mathcal{L}(y_* \mid y) = \frac{\mathcal{L}(y_*, y)}{\mathcal{L}(y)} = \frac{\mathcal{N}(\mu(y, y_*),\, \Sigma)}{\mathcal{N}(\mu(y),\, K)}$ où $\Sigma = \begin{bmatrix} K & K_*^{\mathrm{T}} \\ K_* & k(T,T) \end{bmatrix}$.
On passe alors en fonction de densité $f_{y_*|y}(x) = \frac{f_{y_*, y}(x)}{f_y(x)} = \ldots$ où la formule de la densité de probabilité de la loi normale multivariée (cf. la définition 2) est utilisée au numérateur et au dénominateur. Après plusieurs développements, la formule de la densité de probabilité d'une autre loi normale multivariée est obtenue : $\mathcal{N}\left(K_* K^{-1} y, \quad k(T,T) - K_* K^{-1} K_*^{\mathrm{T}}\right)$.

III.2) Processus gaussiens pour la mortalité

Dans cette section, la régression par processus gaussiens dans le cas particulier de la mortalité est présentée en détail. Cette présentation est inspirée de celle rédigée par Ludkovski et al.[Lud18]

Ces chercheurs ont utilisé les processus gaussiens pour modéliser la mortalité en fonction de l'âge et de l'année calendaire. Dans leur article, les processus gaussiens sont appliqués avec succès sur les données CDC (Centers for Disease Control), qui sont des données d'excellente qualité avec beaucoup d'exposition par âge.

Tranche d'âge	Exposition CDC	Exposition	Coefficient
20 – 30 ans	220 000 000	1 000 000	220
30 – 40 ans	220 000 000	1 200 000	175
40 – 50 ans	200 000 000	1 000 000	200
50 – 60 ans	210 000 000	500 000	400
60 – 70 ans	120 000 000	120 000	1 000

Tableau III.2 — *Différences d'exposition (arrondies) avec la CDC*

Comme le montre le tableau III.2, les données utilisées dans le cadre ce mémoire sont très différentes des données CDC, avec beaucoup moins d'exposition par âge. La question est donc de savoir si les processus gaussiens demeurent pertinents dans un cas plus réaliste avec les données d'un assureur, où les expositions sont plus faibles.

III.2.1) Données nécessaires

Les variables d'entrée choisies sont l'âge et l'année calendaire, donnant de ce fait un espace à 2 dimensions. Il est cependant tout à fait possible d'ajouter des variables pour avoir un espace d'entrée à 3 dimensions ou plus. Un espace unidimensionnel est aussi possible pour obtenir une table de mortalité uniquement selon l'âge par exemple.

La première étape pour faire des estimations par processus gaussiens est de fournir deux matrices. À noter que ce sont des matrices, car l'espace d'entrée est à 2 dimensions, sinon ce seront des vecteurs de dimensions supérieures.

- **D** (Deaths) contenant le nombre de décès en fonction des variables d'entrée : l'âge i et l'année calendaire j ;

- **E** (Exposure) contenant l'exposition en fonction des variables d'entrée : l'âge i et l'année calendaire j. Il est aussi possible d'utiliser une matrice **L** à la place, qui compte la population de milieu d'année.

Notons $N = N_{ages} \times N_{annees}$ le nombre de cellules dans D et E. Tout d'abord, posons $X = (x^n)_{n\in[\![1,N]\!]}$ un vecteur de taille N contenant les variables d'entrée. Dans notre cas, ce sont des couples (âge, année) :

$$\begin{aligned}\forall n \in [\![1, N]\!],\ x^n &= (x^n_{ag},\ x^n_{yr}) \\ &= (\text{âge de la } n^e\text{cellule, année de la } n^e \text{ cellule}) \\ &= \text{par exemple } (78,\ 2016) \\ &\quad \text{pour les personnes ayant 78 ans (âge) en 2016 (année).}\end{aligned}$$

Ensuite, le modèle gaussien postule que $\frac{D_{i,j}}{E_{i,j}} \sim \mathcal{N}\left(e^{\mu_{i,j}},\ \sigma^2 E_{i,j}\right)$. Il est alors possible d'en déduire la valeur du taux de mortalité grâce à l'espérance $e^{\mu_{i,j}}$ et d'obtenir des intervalles de confiance grâce à la variance $\sigma^2 E_{i,j}$. Cependant, dans notre cas nous allons travailler sur la log-mortalité $\mu_{i,j} = log(\frac{D_{i,j}}{E_{i,j}})$ et passer en vecteur de taille N. Ainsi, posons $\boxed{y^n = log\left(\frac{D^n}{E^n}\right)}$ la log-mortalité de la n^e cellule, c'est-à-dire à l'âge x^n_{ag} et l'année x^n_{yr}. En découle la création d'un vecteur $Y = (y^n)_{n\in[\![1,N]\!]}$.

En conclusion, à tout couple $x^n = (x^n_{ag},\ x^n_{yr})$ correspond un y^n qui est la valeur de la log-mortalité observée sur cet âge et année. X correspond aux entrées et Y aux sorties (la log-mortalité). Ce sont ces deux vecteurs de dimension 1 et de taille N qui seront utilisés dans la suite.

Pour généraliser à un espace d'entrée de dimension $p \geq 3$, D et E seraient alors ici de dimension p. De plus, $N = N_{var_1} \times N_{var_2} \times ... \times N_{var_p}$. Enfin, pour $n \in [\![1,N]\!]$, $x^n = (x^n_{var_1},\ x^n_{var_2},\ ...\ , x^n_{var_p})$ et $y^n = log\left(\frac{D^n}{E^n}\right)$.

Finalement, la forme de la base à fournir en entrée des processus gaussiens est donnée dans le tableau III.3. Les colonnes *Année* et *Âge* seront utilisées en entrée X et la colonne *Log(Taux)* sera utilisée en sortie Y. Les autres colonnes sont les données des étapes intermédiaires ayant permis de calculer le log-taux.

Année	**Âge**	**Exposition**	**Décès**	**Taux** $= \frac{Décès}{Exposition}$	**Log(Taux)**
2016	45	43 182,21	34	$7{,}9 \times 10^{-4}$	$-3{,}1$
2016	62	29 586,41	46	$1{,}6 \times 10^{-3}$	$-2{,}8$
2017	27	62 097,11	18	2.9×10^{-4}	$-3{,}54$

Tableau III.3 — *Extrait des données en entrée des processus gaussiens*

III.2.2) Fonction f suivant une loi normale multivariée

En deuxième étape, et comme dans les approches traditionnelles de régression de la mortalité, on suppose qu'il existe une fonction paramétrique f telle que $Y = f(X) + \epsilon$. La variable ϵ correspond au terme d'erreur, c'est un bruit gaussien de moyenne 0 et de variance $\sigma^2(X)$. Ainsi, pour chaque cellule y^n de Y, on a $y^n = f(x^n) + \epsilon^n$. La fonction f fait donc le lien entre les entrées $(x^n_{ag},\ x^n_{yr})$ et la log-mortalité bruitée $y^n - \epsilon^n$.

Avec les processus gaussiens, la fonction f est de plus modélisée comme une variable aléatoire vérifiant la propriété suivante : $\forall X = (x^{i_1},\ \dots\ , x^{i_n})$ un ensemble de couples d'âges et d'années,

$$\Big(f(x^{i_1}),\ \dots\ , f(x^{i_n})\Big) \sim \mathcal{N}\left(moyenne = \mu(X),\ covariance = k(X, X)\right).$$

Ainsi, avec les processus gaussiens, il existe des fonctions μ et k telles que tout ensemble fini de $f(x^i)$ suive une loi gaussienne multivariée de fonction de moyenne $\mu(\cdot)$ et covariance $k(\cdot, \cdot)$. Cette propriété est abrégée sous la notation $f(X) \sim \mathcal{GP}\left(\mu(X),\ k(X, X)\right)$.

Comme expliqué dans l'exemple simplifié de la section précédente, la covariance va mesurer la similitude entre deux points. Avec les processus gaussiens, les taux de mortalité sont tous corrélés entre eux. $k((i_1, j_1),\ (i_2, j_2))$ est la covariance entre l'âge i_1 à l'année j_1 et l'âge i_2 à l'année j_2. Plus (i_1, j_1) et (i_2, j_2) sont proches, plus leur covariance $k((i_1, j_1),\ (i_2, j_2))$ sera élevée car la connaissance de y_{i_1,j_1} va grandement influencer l'estimation de y_{i_2,j_2}.

III.2.3) Choix de la fonction de covariance

En troisième étape, il convient de choisir une fonction de covariance, aussi appelée *noyau* ou *kernel* en anglais. Cette fonction va exprimer la relation entre les différentes valeurs d'input.

La fonction de covariance doit vérifier certaines propriétés.[Bas19] Ainsi, elle doit être symétrique : le lien entre x_1 et x_2 doit être le même que celui entre x_2 et x_1. En outre, si ces deux points sont similaires, alors ils doivent avoir une corrélation forte, donc $k(x_1, x_2)$ grand. Cette propriété implique que les taux de mortalités seront lissés, car deux points voisins auront toujours une valeur proche,. Au contraire, si ces deux points sont éloignés, alors ils doivent avoir une corrélation faible, donc $k(x_1, x_2)$ petit. Cette dernière propriété implique que l'impact d'un point x_0 éloigné d'un point à prédire x_* doit être négligeable sur la prédiction y_*. Enfin, la matrice de la fonction de covariance doit être définie positive, ce qui garantit son inversibilité pour calculer K^{-1}.

Les noyaux les plus couramment utilisés pour la modélisation par processus gaussien sont les suivants : [ENS20]

- Noyau gaussien, aussi appelé RBF (Radial Basis Function) ou Exponentielle quadratique, de grandeur $l > 0$:

$$k\left(x_1, x_2\right) = \exp\left(-\frac{\left\|x_1, x_2\right\|^2}{2l^2}\right).$$

  La formule ci-dessus est une forme simplifiée qui peut être complexifiée en ajoutant des hyperparamètres.

- Noyau Matérn, de grandeur $l > 0$ et de paramètre ν :

$$k\left(x_1, x_2\right) = \frac{2^{1-\nu}}{\Gamma(\nu)} * \left(\frac{\sqrt{2\nu} * \left\|x_1, x_2\right\|}{l}\right)^{\nu} * K\left(\frac{\sqrt{2\nu} * \left\|x_1, x_2\right\|}{l}\right),$$

  avec K la fonction Bessel modifiée de second genre et d'ordre ν. Il y a 3 valeurs de ν principalement utilisées : $\nu = 1/2$, $\nu = 3/2$ et $\nu = 5/2$.

- Noyau polynomial, de degré alpha entier naturel :

$$k\left(x_1, x_2\right) = \left(x_1^T x_2 + c\right)^{\alpha},$$

  avec $c \geq 0$ un paramètre libre (quand c = 0, on parle de noyau polynomial homogène).

Rasmussen cite quant à lui d'autres fonctions de covariance : [Ras06]

- Constante : $k_{\mathrm{C}}\left(x, x'\right) = C$.
- Linéaire : $k_{\mathrm{L}}\left(x, x'\right) = x^{\top} x'$.
- Bruit blanc gaussien : $k_{\mathrm{GN}}\left(x, x'\right) = \sigma^2 \delta_{x,x'}$.
- Ornstein-Uhlenbeck : $k_{\mathrm{OU}}\left(x, x'\right) = \exp\left(-\frac{|d|}{\ell}\right)$.
- Périodique : $k_{\mathrm{P}}\left(x, x'\right) = \exp\left(-\frac{2\sin^2\left(\frac{d}{2}\right)}{\ell^2}\right)$.
- Quadratique rationnelle : $k_{\mathrm{RQ}}\left(x, x'\right) = \left(1 + |d|^2\right)^{-\alpha}, \quad \alpha \geq 0.$

La fonction de covariance la plus adaptée dépend du phénomène physique à modéliser, et donc des connaissances *a priori* à son sujet. Dans le cas de la mortalité, nous utiliserons la covariance exponentielle quadratique tout comme Ludkovski et al.[Lud18] La fonction exponentielle quadratique s'écrit de cette manière dans le cas de données d'entrée en deux dimensions (âge et année) :

$$\boxed{k\left(x^i, x^j\right) = \eta^2 \exp\left(-\frac{\left(x_{ag}^i - x_{ag}^j\right)^2}{2\theta_{ag}^2} - \frac{\left(x_{yr}^i - x_{yr}^j\right)^2}{2\theta_{yr}^2}\right).}$$

III.2.4) Détermination des hyperparamètres

Enfin, en utilisant les données d'entraînement et la fonction de covariance choisie, il est possible de déterminer les hyperparamètres optimaux. Les hyperparamètres sont au nombre de quatre dans cette section : $\theta_{ag}, \theta_{yr}, \eta^2$ et la variance de bruit σ^2.

Pour les déterminer, il est possible d'utiliser le maximum de vraisemblance en maximisant $\mathbb{P}(\theta|x,y)$. Avec la formule de Bayes, cela revient à maximiser $\mathbb{P}(y|x,\theta)$, ou encore $log\left(\mathbb{P}(y|x,\theta)\right)$. Ici, θ représente l'ensemble des hyperparamètres : $\theta = \{\theta_{ag}, \theta_{yr}, \eta^2, \sigma^2\}$.

III.2.5) Prédiction de la mortalité

En conclusion, une fois les hyperparamètres optimaux déterminés, il ne reste plus qu'à entraîner le processus gaussien. Cet entraînement va notamment calculer l'inverse de la matrice $K = k(x,x)$ définie dans l'exemple introductif. Une fois les calculs effectués, les fonctions de moyenne et de covariance seront totalement déterminées. Il sera alors possible de calculer la moyenne et la matrice de covariance du processus gaussien en tout vecteur de points x_* grâce à la formule ci-dessous : [Mig21]

$$\boxed{f_*\left(x_* \mid x, y\right) \sim \mathcal{N}\Big(\begin{array}{r}\text{moyenne} = \mu(x_*) + k\left(x, x_*\right)^T \left(k(x,x) + \Sigma\right)^{-1}(y - \mu(x)), \\ \text{covariance} = k\left(x_*, x_*\right) - k\left(x, x_*\right)\left(k(x,x) + \Sigma\right)^{-1} k(x_*, x)^T\Big)\end{array}},$$

avec $\Sigma = diag(\sigma(x_i)^2)$ une matrice $N \times N$ contenant le bruit des observations aux points d'entraînement, μ la moyenne de la distribution a priori et k la fonction de covariance choisie précédemment.

La moyenne est souvent prise à 0, ce qui simplifie l'expression. En effet, Ludkovski et al. ont montré que le choix de la moyenne a un impact négligeable sur les prédictions des processus gaussiens en des points connus (i.e. sans faire d'extrapolation de taux).[Lud18] Des expressions différentes de μ seront essayées par la suite en application pratique dans le cas du positionnement par cohortes.

Grâce à l'hypothèse de normalité, déterminer la distribution a posteriori $\mathbb{P}(f|y)$ en un point revient à calculer la moyenne et la covariance en ce point. Par conséquent, il deviendra aussi possible de prédire en des points inconnus de l'espace des états.

Enfin, il est à noter qu'essayer de prédire les mêmes points que ceux utilisés pour l'entraînement du processus gaussien ne donnera pas les valeurs exactes d'entraînement, mais des valeurs lissées. Le seul cas particulier où ce ne sera pas le cas est si l'on suppose que les données d'entraînement ne sont pas bruitées : si $\sigma = 0$, alors $f_*(x_i) = y_i$. Ainsi, et sauf dans cette situation spécifique, les processus gaussiens vont lisser les taux de mortalité à travers les dimensions si l'on prédit sur les points d'entraînement.

III.3) Application pratique

Les processus gaussiens ont été appliqués en langage R grâce au package *DiceKriging*. Les hyperparamètres sont estimés grâce à la méthode du maximum de vraisemblance.

III.3.1) Mise au format de la base d'exposition

Afin de faire tourner le modèle de processus gaussiens, il est nécessaire de retraiter la base d'exposition fournie en entrée. D'abord, il faut évidemment comme expliqué dans la partie théorique précédente qu'elle soit sous la forme d'un tableau avec une colonne par variable d'entrée et une colonne pour le log-taux de mortalité. Cependant, il faut également procéder à d'autres retraitements, sans quoi il y aura des erreurs lors de l'entraînement du modèle qui ne pourra pas se réaliser.

Ainsi, il faut que les données fournies en entrée ne contiennent **pas de valeurs manquantes**. Heureusement, dans le cadre de ce mémoire, les valeurs manquantes ne concernaient que la CSP, qui était parfois vide pour certains clients. La solution a été de créer une classe « Inconnu » dans le cas de valeurs manquantes.

En outre, **le nombre de décès doit être non nul**. Effectivement, s'il y a 0 décès pour une certaine combinaison des variables d'entrée, le taux de mortalité vaudra 0, ce qui entraînera que le log-taux sera $-\infty$. La solution à ce problème a été de remplacer les valeurs nulles du nombre de décès par 10^{-3}. Des valeurs plus basses ont été testées telles que 10^{-6} et 10^{-9}, mais les résultats étaient quasiment identiques pour ces autres valeurs. Le choix de ce paramètre de remplacement des nombres de décès nuls semble donc secondaire.

Enfin, les variables en format texte telles que les variables catégorielles H/F ou le réseau BP/CE doivent être converties en nombres. Effectivement, tout comme en machine learning, il est nécesssaire de recoder ces variables pour fournir en entrée uniquement des nombres et pas du texte. Du **label encoding** a ainsi été appliqué pour transformer BP en 1 et CE en 2 par exemple.

III.3.2) Remarques quant aux calculs

Contrairement à ce qu'on pourrait penser, ce qui prend le plus de temps de calcul n'est pas d'ajuster le processus gaussien en calculant les hyperparamètres optimaux et l'inverse de la matrice de covariance. Dans notre cas, c'est plutôt l'étape de préparation des données d'entraînement pour les mettre au bon format qui est chronophage. Il faut en effet filtrer la base d'exposition, calculer l'exposition des personnes, regrouper des lignes entre elles, recoder les variables texte, faire la somme des expositions par groupe, etc.

Les données d'entrée des processus gaussiens dépendent des variables désirées. Par exemple, pour positionner le décès selon le sexe, les lignes d'exposition par âge et sexe doivent être groupées pour y effectuer tous les retraitements expliqués précédemment. Cependant, si l'on veut ensuite positionner par rapport au réseau, il faut tout recommencer et grouper par âge et réseau. En a découlé la création de plusieurs jeux de données pour l'entraînement des processus gaussiens.

```
config_pg <- list(
    date_solvency = "230603",
    nom_base = "base_exposition",
    variables_entree = c("age"),
    variable_sortie = "log_taux_brut",
    formule = ~ 1,
    covtype = "gauss",
    filtre_risque = c("DCTC"),
    filtre_reseau = c(),
    filtre_sexe = c(),
    filtre_annees = c(2018:2022),
    filtre_ages = c(18:65),
    groupement_ages = 1,
    apres_refus = FALSE,
    remplacement_deces_0 = 1 / (10 ^ 3),
    exposition_approchee = TRUE
)
```

Figure III.2 — *Exemple de configuration pour les processus gaussiens*

Des fonctions ont été développées pour automatiser la création des datasets ainsi que l'application des processus gaussiens. Elles permettent de spécifier les variables d'entrée voulues, les filtres opérés sur les données (par exemple n'apprendre que sur les BP, seulement sur les hommes, changer la période d'observation, exclure les âges avec peu d'exposition, etc.), un éventuel pré-lissage (utile s'il y a peu d'exposition), ainsi que les paramètres des processus gaussiens eux-mêmes. Tous ces paramètres étaient renseignés dans une liste de configuration comme dans la figure III.2.

Ces opérations prenaient à chaque fois du temps à être exécutées et le nombre de datasets créés était toujours plus large, car de nombreuses combinaisons de paramètres

ont été essayées. En a alors également découlé la création d'un mécanisme de cache des résultats déjà calculés pour n'avoir à appliquer ces fonctions que quand leur résultat n'était pas encore dans le cache. Chaque élément dudit cache est donc identifié de façon unique par sa liste de paramètres. Par conséquent, avant de lancer tous les retraitements de la base d'exposition et les calculs du processus gaussien, le cache est d'abord parcouru pour vérifier que la liste de paramètres en entrée n'est pas déjà calculée.

À noter qu'il n'aurait pas été judicieux de grouper les lignes sur trop de dimensions à la fois (par exemple âge, sexe et réseau). Effectivement, plus il y a de variables d'entrée, plus l'exposition par groupe sera faible. Un modèle à une dimension (par âge) a été réalisé ainsi que des modèles à 2 dimensions (par âge et une autre variable). Ces derniers ont permis le positionnement du décès par rapport à différentes variables dont le sexe, le réseau ou encore la catégorie socioprofessionnelle. L'avantage de ces positionnements est qu'aucune hypothèse de hasard proportionnel n'est nécessaire sur les données, contrairement au modèle de Cox. Cependant, même en deux dimensions seulement, l'exposition n'est parfois pas suffisante pour garantir des intervalles de confiance petits.

III.3.3) Filtres sur les données

Lors des premiers ajustements de processus gaussiens, les intervalles de confiance étaient extrêmement élevés, cf. la figure III.3. La réalisation d'un filtre sur l'âge entre 18 et 65 ans a permis de réduire drastiquement ceux-ci.

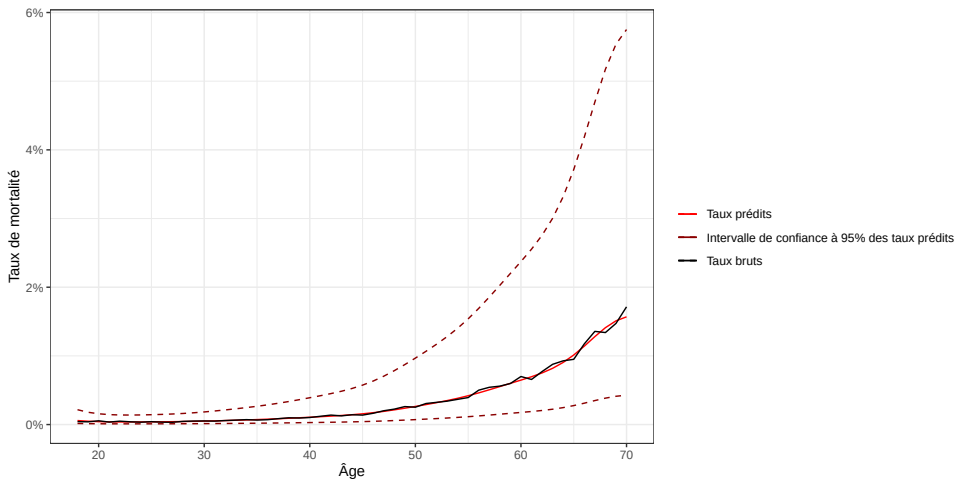

Figure III.3 — *Taux par âge prédits par les processus gaussiens sans filtre sur les âges des données d'entraînement*

En effet, les quelques données aux âges extrêmes (dans les 70 et 80 ans) impactaient sensiblement les résultats. Ainsi, l'exposition à partir de 65 ans est de moins de 1 000

personnes, sans même vouloir segmenter selon une autre variable pour faire un positionnement par rapport à celle-ci. À partir de 70 ans, l'exposition n'est plus que de quelques dizaines de personnes par année et le dernier âge disponible est de 86 ans. De 70 à 86 ans, aucun décès n'est constaté, générant des taux de mortalités nuls à ces grands âges, ce qui ne devrait pas être le cas.

III.4) Résultats obtenus

III.4.1) Taux de mortalité par âge

Les processus gaussiens prédisent la log-mortalité et en fournissent des intervalles de confiance. Ceux-ci sont donnés en figure III.4.

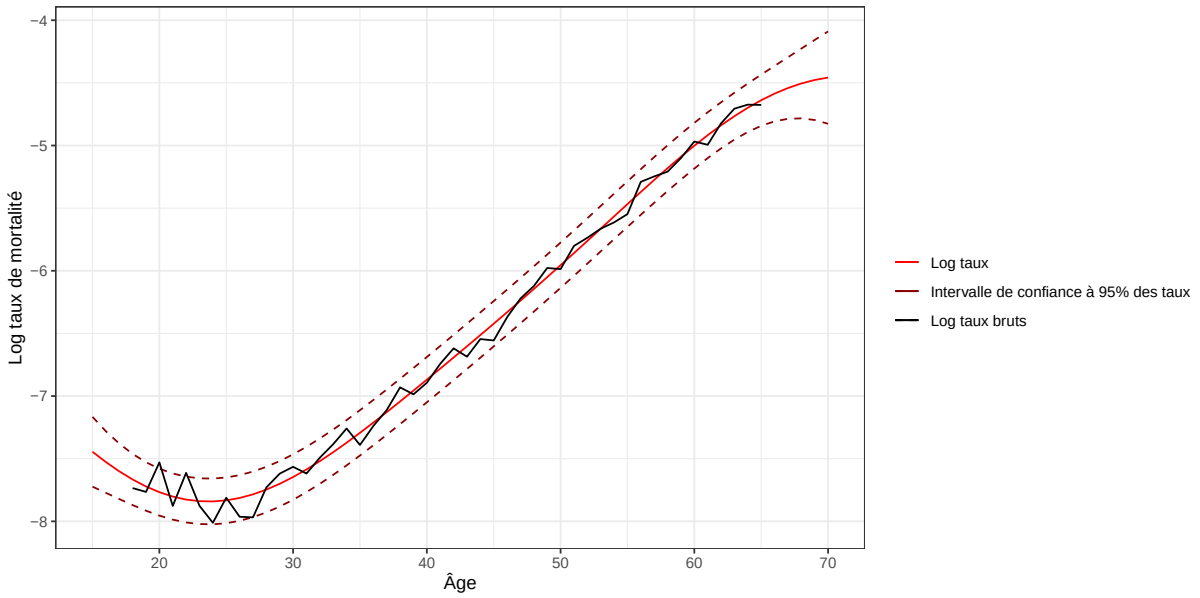

Figure III.4 — *Log-taux par âge prédits par les processus gaussiens*

Pour récupérer les taux de mortalité et les intervalles de confiance de ceux-ci, il faut ensuite passer à l'exponentielle :

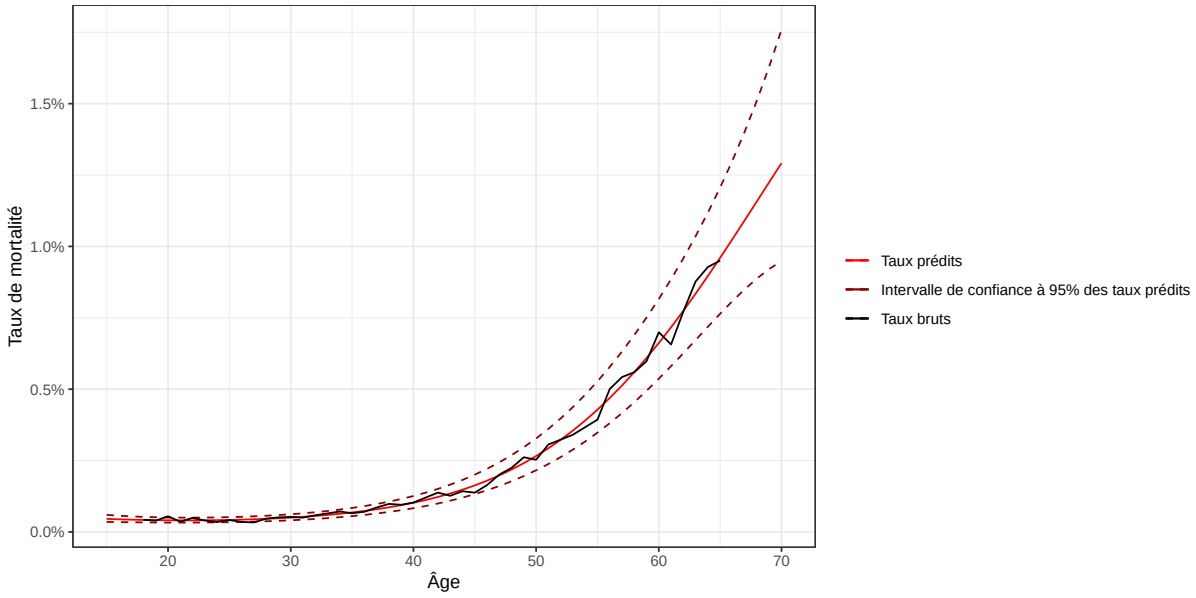

Figure III.5 — *Taux par âge prédits par les processus gaussiens*

Comme le montre la figure III.5, les intervalles de confiance sont beaucoup plus petits qu'en figure III.3 une fois le filtre par âge réalisé.

Utilisation d'autres fonctions de covariance

Plusieurs fonctions de covariance ont été essayées. Celle préconisée par Ludkovski et al. dans leur article est celle nommée *Gauss*. La fonction de covariance *Matern* est quant à elle celle utilisée par défaut par le package *DiceKriging*. Enfin, la fonction de covariance exponentielle a également été testée.

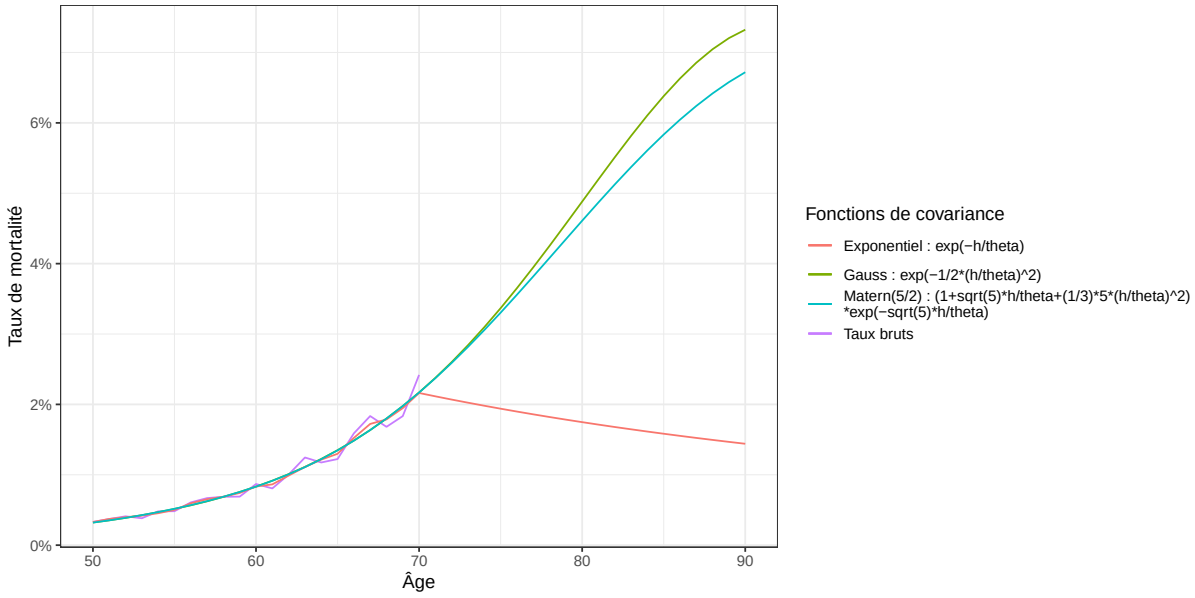

Figure III.6 — *Processus gaussiens avec d'autres fonctions de covariance*

Les résultats de cette comparaison sont dans la figure III.6. Le choix de la fonction de covariance ne semble donc pas avoir d'impact si l'on ne prédit pas en-dehors des données connues. En revanche, dès qu'il n'y a plus de taux bruts (à 70 ans ici), le processus gaussien utilisant la fonction de covariance exponentielle produit des prédictions qui décroissent avec le temps. Les deux autres fonctions de covariance comparées sont quant à elles bien croissantes avec le temps et sont proches l'une de l'autre. En conclusion, la fonction de covariance gaussienne sera celle qui sera utilisée.

Comparaison avec la méthode classique

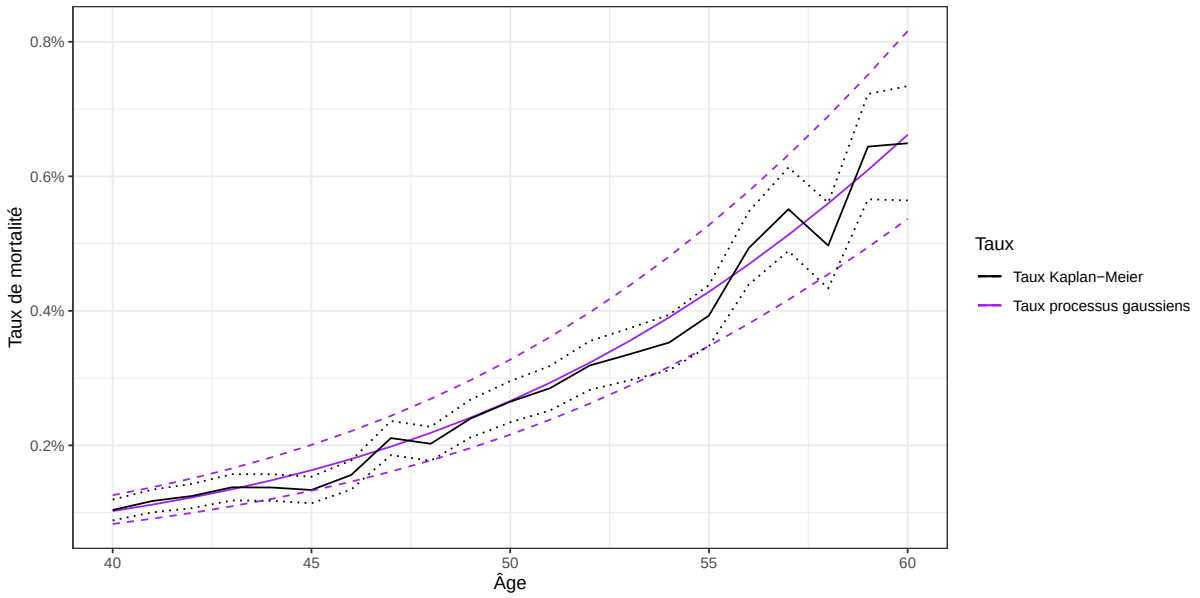

Figure III.7 — *Comparaison des processus gaussiens avec Kaplan-Meier*

Comme le montre la figure III.7, les taux des processus gaussiens et ceux de Kaplan-Meier sont proches, à la différence que les taux des processus gaussiens sont déjà lissés contrairement à ceux de Kaplan-Meier.

Les intervalles de confiance des taux de Kaplan-Meier sont cependant plus petits que ceux des processus gaussiens. Cela vient du fait que les processus gaussiens utilisent les valeurs des âges situés autour du point à prédire, en incorporant leur variance. Les points après 60 ans ayant de moins en moins d'exposition, la variance augmente pour chacun de ces points et une partie est incorporée aux prédictions des âges avant 60 ans.

III.4.2) Taux de mortalité par âge et année

Dans cette sous-section, nous allons positionner les taux de mortalité par année pour tenter de prédire les années suivantes. Les processus gaussiens vont donc lisser les taux de mortalité bruts jusqu'à la dernière année connue, ici 2022, et prédire les taux pour les années 2023 et 2024. Pour cela, seules les lignes Banque Populaire ont été conservées dans la base d'exposition. Cela permet de faire apprendre les processus gaussiens sur plus d'années, de 2010 à 2022, car les données CE ne remontent pas avant 2016.

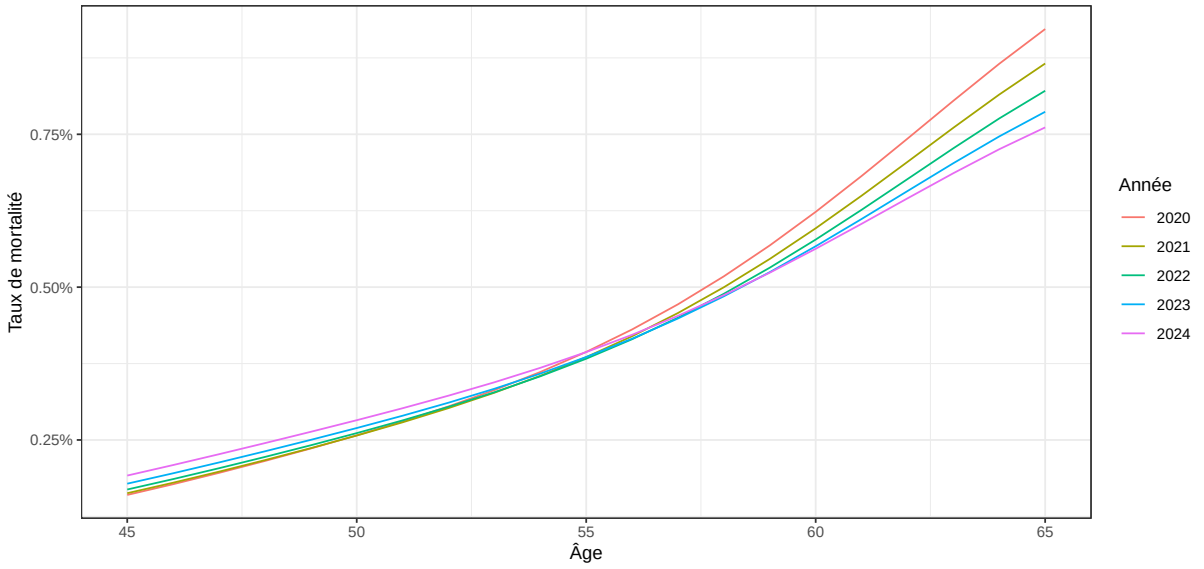

Figure III.8 — *Positionnement par année avec les processus gaussiens*

La figure III.8 prédit un abaissement de la mortalité dans le temps à âge constant. Cependant, une augmentation de la mortalité au fil du temps pour les âges vers 40 ans est également prédite. C'est un résultat qui ne semble pas conforme à l'évolution de la mortalité.

Utilisation d'autres formules de tendance

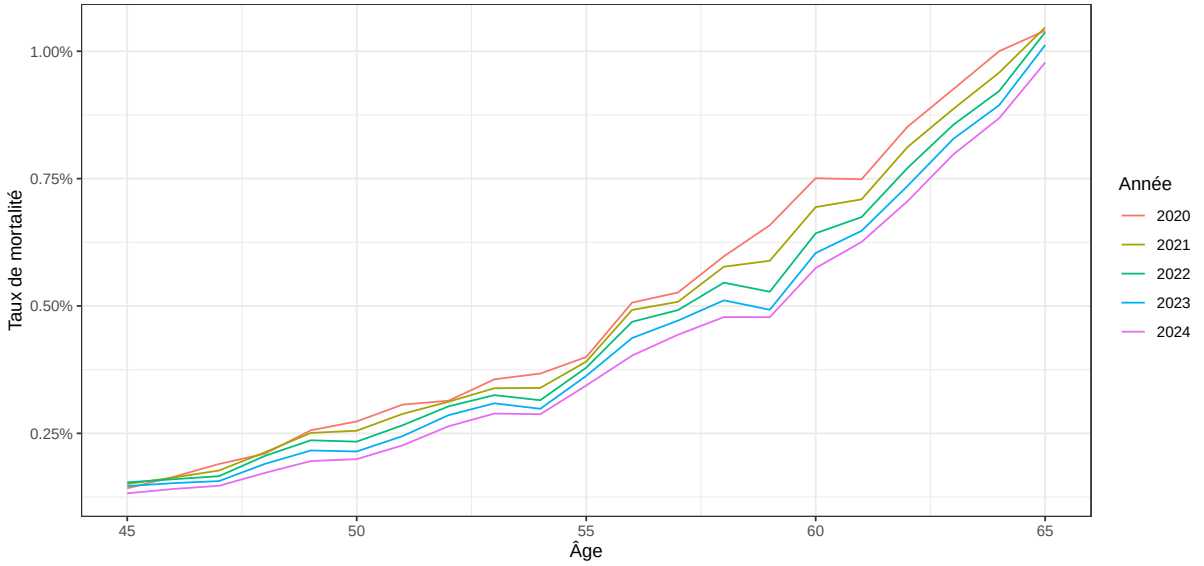

Figure III.9 — *Positionnement par année avec les processus gaussiens avec une tendance linéaire*

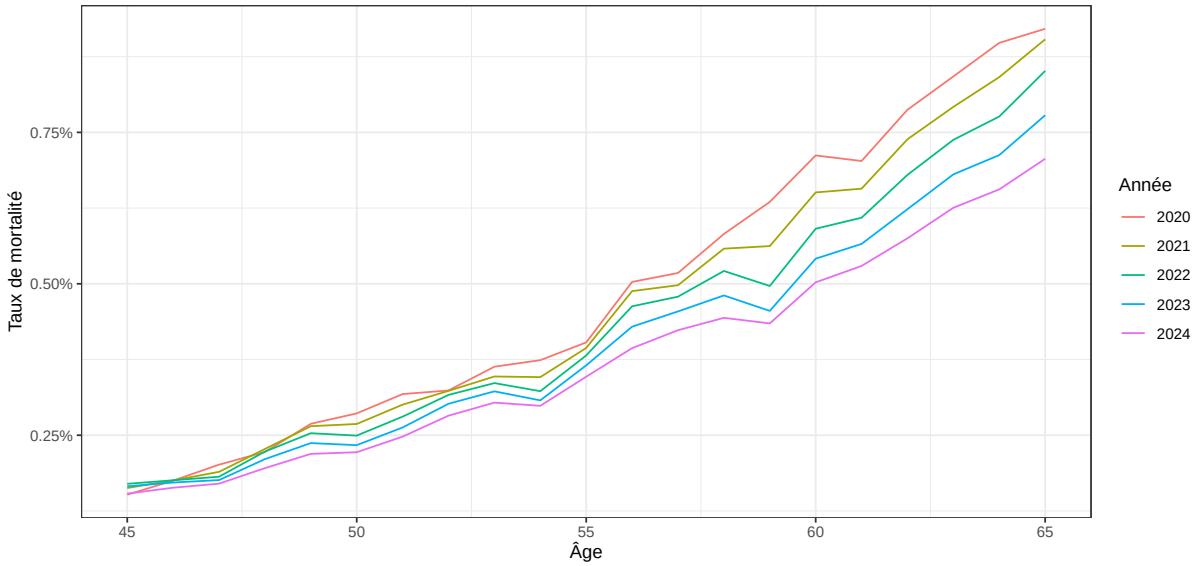

Figure III.10 — *Positionnement par année avec les processus gaussiens avec une tendance quadratique par âge*

La figure III.9 et la figure III.10 donnent quant à elles des résultats plus crédibles. Elles ont été obtenues en modifiant le paramètre *formula* dans la fonction permettant d'ajuster les processus gaussiens.

Le paramètre utilisé auparavant était simplement $formule = \sim 1$, ce qui signifie que la tendance du modèle est constante : $\mu(x^n) = \beta_0$. Dans la figure III.9, $formule = \sim x.age + x.annee$, c'est-à-dire que la tendance du modèle est linéaire : $\mu(x^n) = \beta_0 + \beta_1^{ag} x_{ag}^n + \beta_1^{yr} x_{yr}^n$. Enfin, la figure III.10 utilise $formule = \sim x.age + I(x.age \wedge 2) + x.annee$, signifiant que la tendance du modèle est quadratique par âge : $\mu(x^n) = \beta_0 + \beta_1^{ag} x_{ag}^n + \beta_1^{yr} x_{yr}^n + \beta_2^{ag} (x_{ag}^n)^2$.

Avec les nouvelles formules de tendance, les taux de mortalité sont beaucoup moins lisses qu'auparavant. La figure III.9 semble préférable à la figure III.10, car la mortalité baisse beaucoup trop rapidement vers 65 ans en fonction des années.

Utilisation d'autres fonctions de covariance

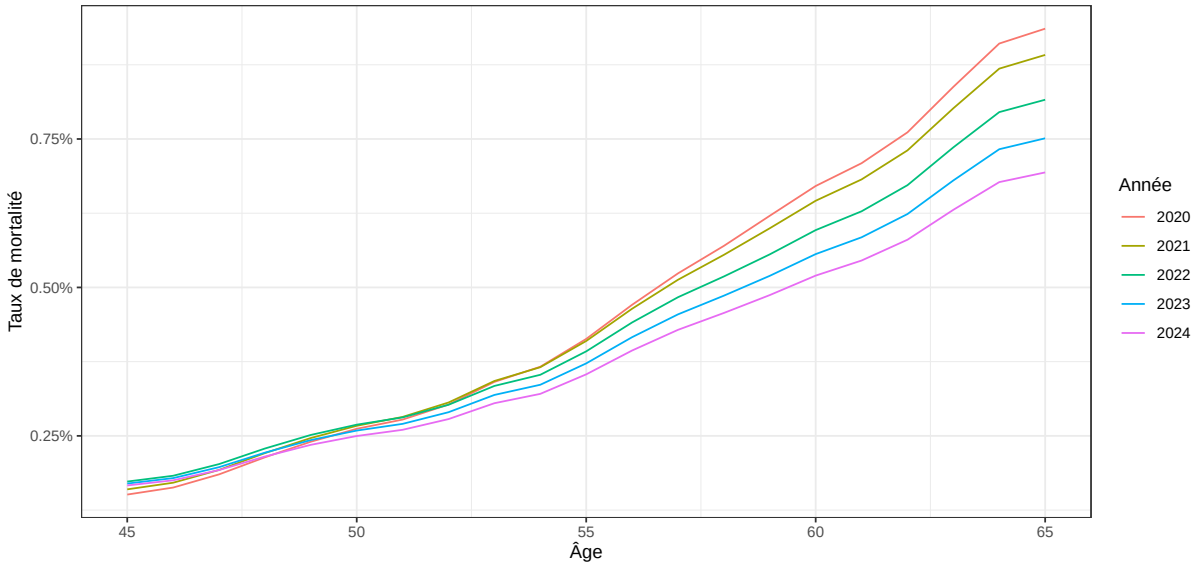

Figure III.11 — *Positionnement par année avec les processus gaussiens en utilisant la fonction de covariance Exponentielle*

En essayant une autre fonction de covariance et en reprenant la formule initiale $formule = \sim 1$, les résultats sont encore différents comme l'indique la figure III.11. Ce graphique est obtenu par la fonction de covariance Exponentielle et est déjà meilleur que les précédents. Les taux de mortalité sont bien décroissants avec le temps à âge constant, et il n'y a plus de problème de mortalité croissante vers les âges plus jeunes. Cependant, la mortalité semble décroître beaucoup trop rapidement selon les années.

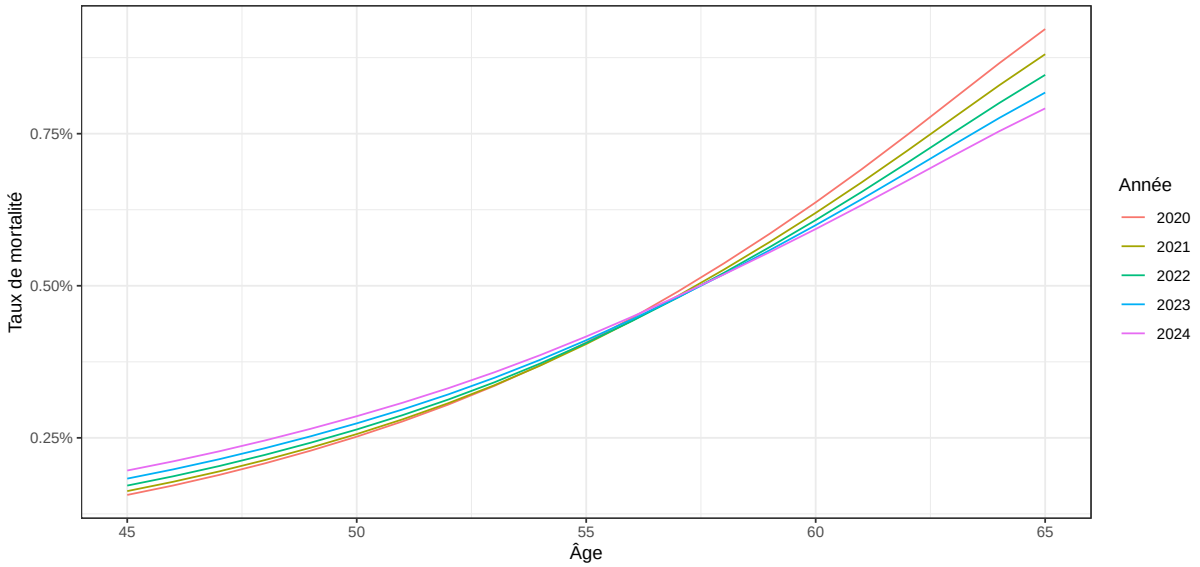

Figure III.12 — *Positionnement par année avec les processus gaussiens en utilisant la fonction de covariance Matern(5/2)*

La fonction de covariance Matern(5/2) a également été essayée en figure III.12. Néanmoins, les résultats sont quasiment les mêmes que pour ceux de la fonction de covariance utilisée en figure III.8.

Séparation en deux processus gaussiens

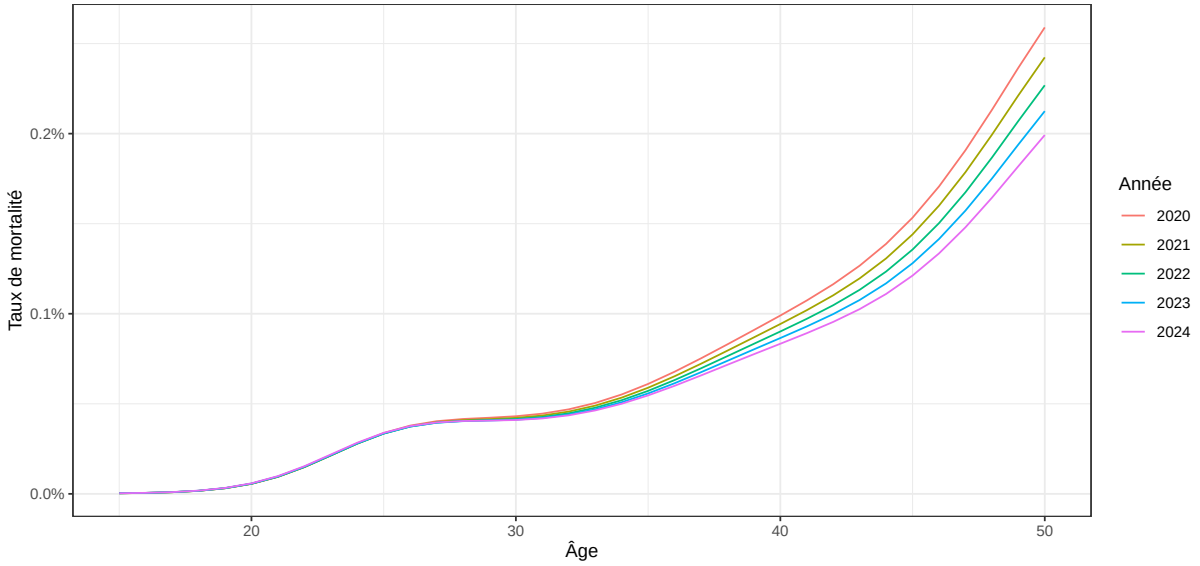

Figure III.13 — *Positionnement par année avec les processus gaussiens en utilisant les âges inférieurs à 50 ans*

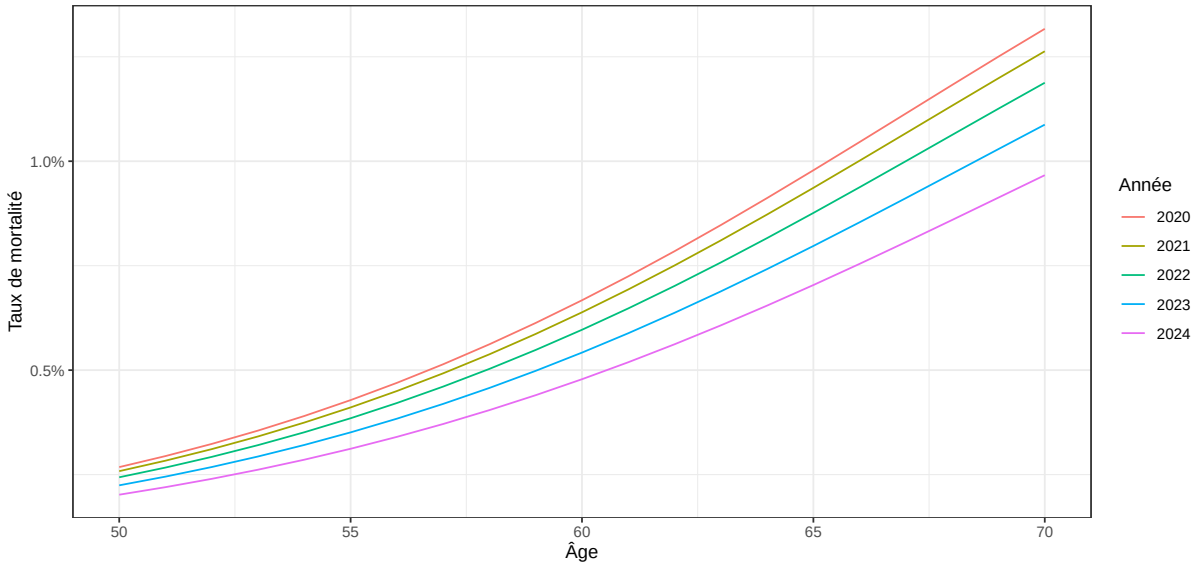

Figure III.14 — *Positionnement par année avec les processus gaussiens en utilisant les âges de 50 à 65 ans*

Enfin, Ludkovski et al.[Lud18] suggèrent d'ajuster les processus gaussiens en deux fois : une fois sur les âges jeunes, et une fois sur les âges plus vieux. Cette méthode est nécessaire quand la structure de la covariance n'est pas stationnaire : $C(x^i, x^j)$ ne dépend

pas de seulement $|x^i - x^j|$, mais aussi des valeurs des points x^i et x^j. La figure III.13 et la figure III.14 en sont le résultat.

Les résultats sont très satisfaisants, même si là encore la décroissance de mortalité en fonction des années semble trop rapide. Utiliser cette méthode introduit aussi la question de savoir comment joindre les deux tables de mortalité ainsi créées. En effet, il y a une discontinuité à l'âge de jointure comme l'atteste la figure III.15.

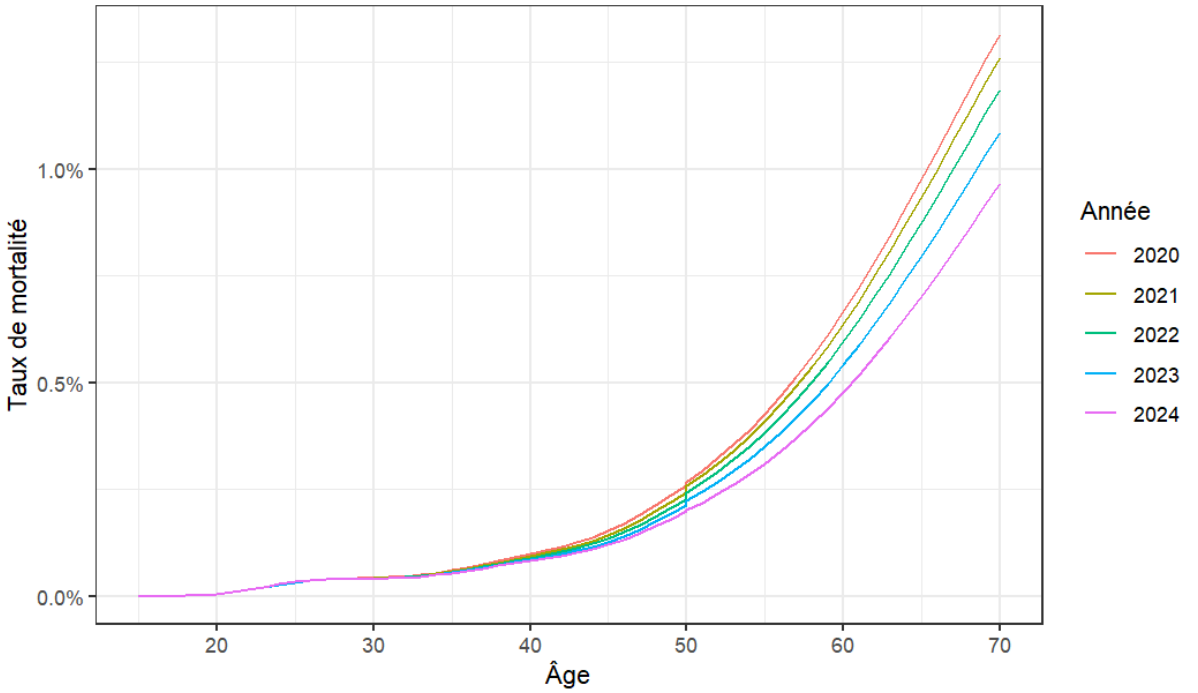

Figure III.15 — *Positionnement par année avec les processus gaussiens en fusionnant la figure III.13 et la figure III.14*

III.4.3) Taux de mortalité par âge et sexe

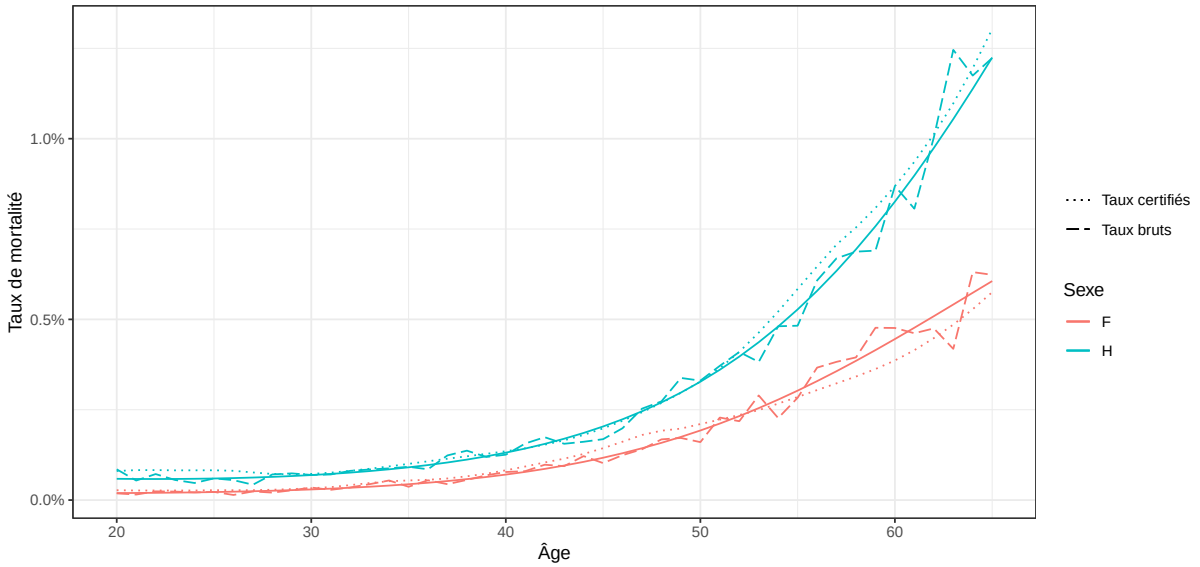

Figure III.16 — *Positionnement par sexe avec les processus gaussiens*

Un positionnement par sexe avec les processus gaussiens a également été réalisé. Comme le montre la figure III.16, les taux hommes et les taux femmes sont clairement distincts.

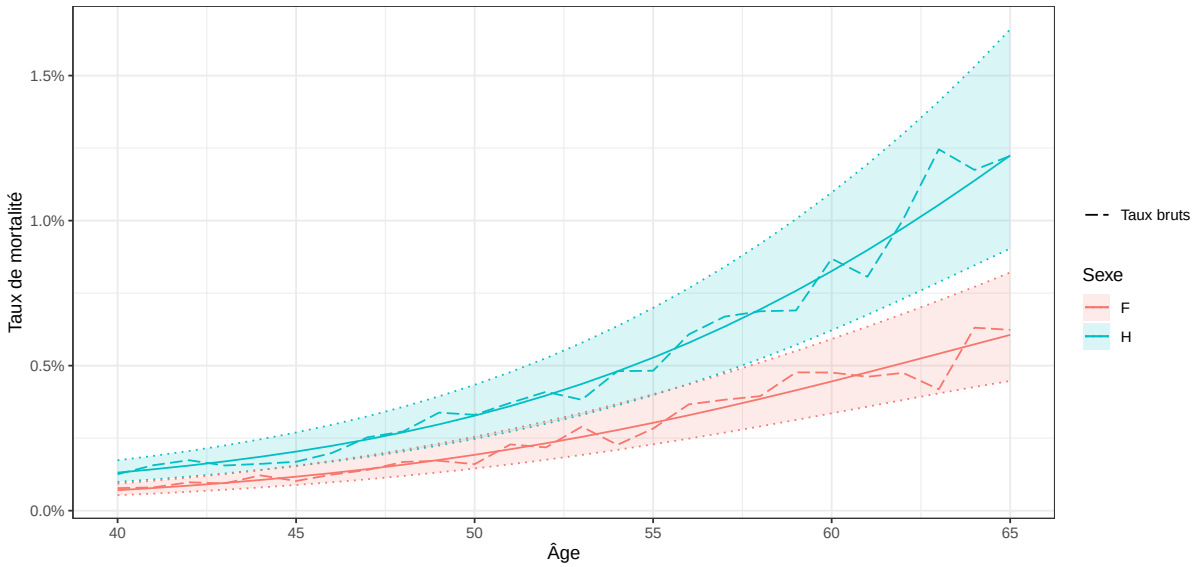

Figure III.17 — *Intervalles de confiance à 95% du positionnement par sexe avec les processus gaussiens*

La différence est encore plus visible sur la figure III.17 avec les intervalles de confiance à 95% des taux par sexe qui ne se chevauchent pas. Les taux bruts sont d'ailleurs toujours

contenus dans ces intervalles de confiance. De plus, ces derniers croissent avec l'âge, ce qui était attendu car l'exposition diminue sensiblement au fil des âges.

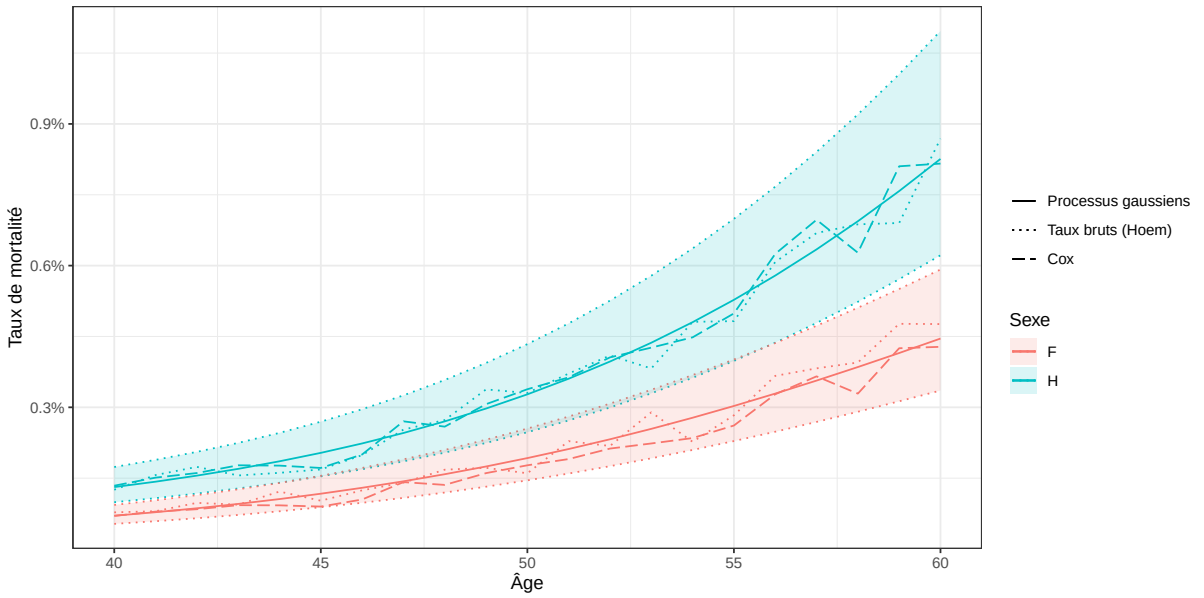

Figure III.18 — *Comparaison du positionnement par sexe entre les processus gaussiens et le modèle de Cox*

Enfin, les taux positionnés par sexe avec le modèle de Cox ont été comparés avec ceux des processus gaussiens en figure III.18. Une comparaison entre les taux bruts et ceux de Cox permet déjà de déceler des différences entre les deux courbes, qui proviennent de l'hypothèse forte des hasards proportionnels pour le positionnement de Cox.

Les processus gaussiens semblent être un positionnement intéressant. En effet, aucune hypothèse sur les taux de mortalité n'a besoin d'être faite, contrairement au modèle de Cox. Les positionnements des processus gaussiens ont aussi l'avantage d'être déjà lissés, même si ce lissage n'en est pas vraiment un : c'est simplement la suppression par les processus gaussiens du bruit dans les données d'entraînement.

III.4.4) Taux de mortalité par âge et réseau

Les taux de mortalité positionnés par réseau se trouvent en figure III.19. Dans le cas de la mortalité par réseau, les taux des processus gaussiens sont assez proches des taux bruts, malgré des fluctuations.

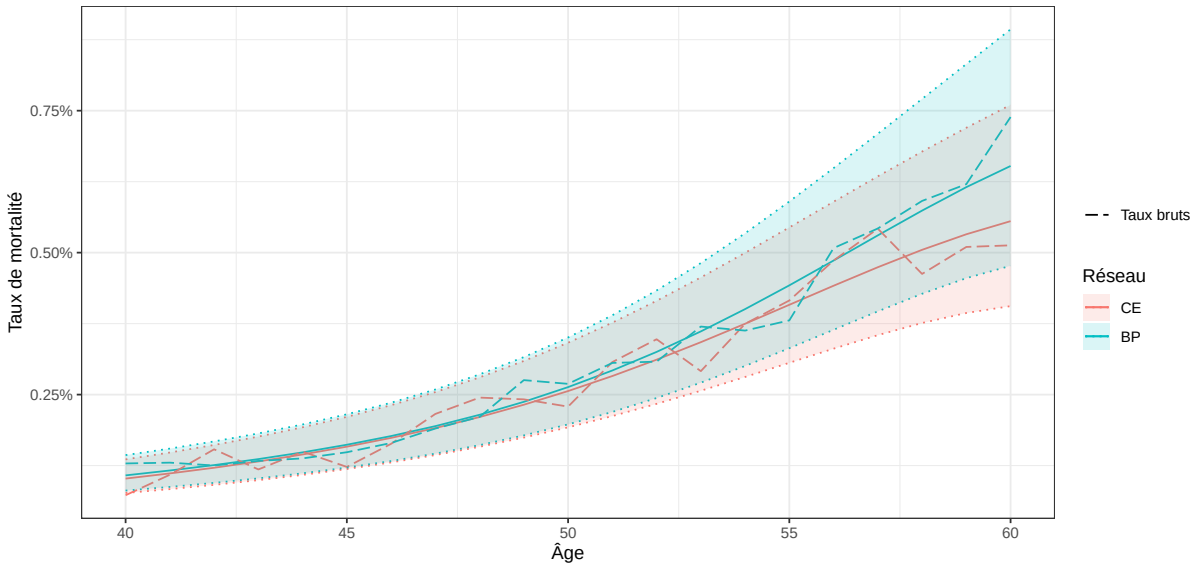

Figure III.19 — *Positionnement par réseau avec les processus gaussiens*

Les intervalles de confiance à 95% sont cependant beaucoup moins distincts que pour le positionnement par sexe de la figure III.17. Cela demeure un résultat cohérent, car le sexe est beaucoup plus discriminant en terme de mortalité que le fait de souscrire son contrat d'assurance dans un réseau BP ou CE.

Le portefeuille CE est plus récent que celui des BP, car il n'existe que depuis 2016 contrairement à celui des BP. La sélection médicale peut donc encore jouer un rôle, ce qui expliquerait pourquoi les taux CE sont plus faibles que ceux BP.

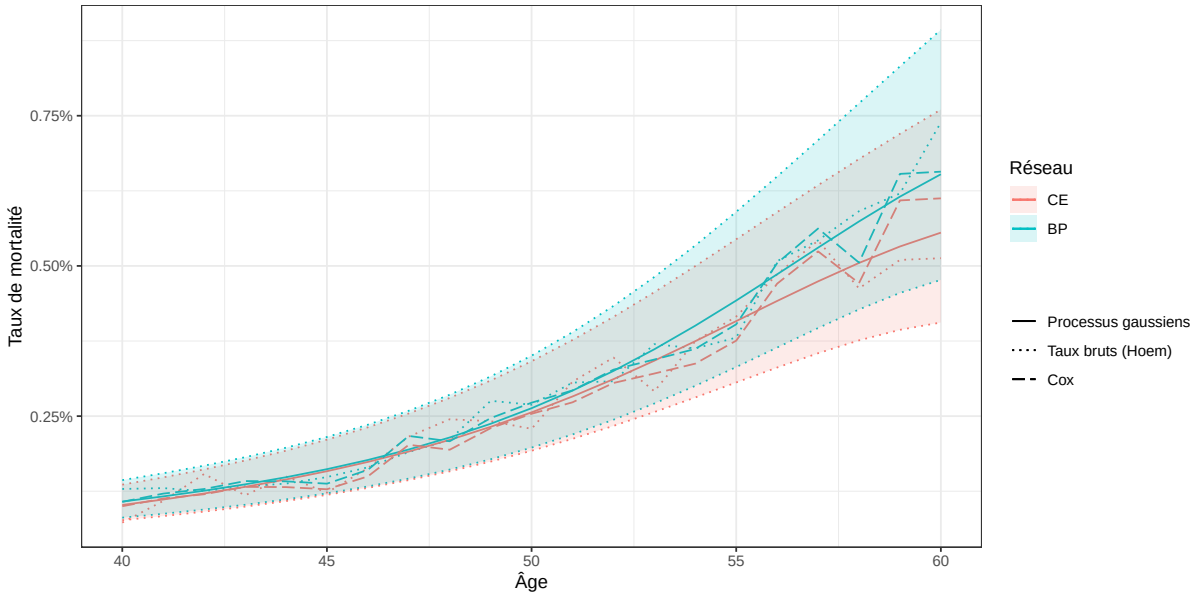

Figure III.20 — *Comparaison du positionnement par réseau entre les processus gaussiens et le modèle de Cox*

Comme le montre la figure III.20, le positionnement par processus gaussiens semble là encore très pertinent. Les taux bruts sont mieux suivis, car l'hypothèse de hasards proportionnels implique ici que les taux ne peuvent pas s'écarter au fil du temps, ce qui est pourtant le cas sur ce graphique vers 60 ans. Il en découle que les taux positionnés par Cox vers cet âge différent sensiblement des taux observés.

III.4.5) Taux de mortalité par âge et catégorie socioprofessionnelle

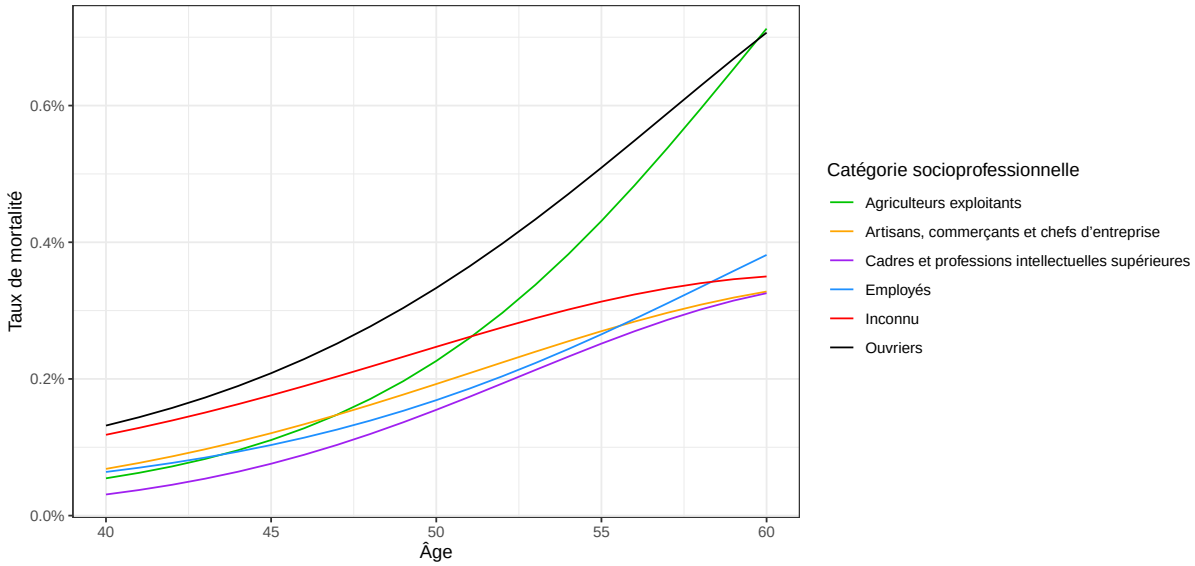

Figure III.21 — *Positionnement par catégorie socioprofessionnelle avec les processus gaussiens*

Enfin, les taux de mortalité ont été positionnés par rapport à la catégorie socioprofessionnelle dans la figure III.21. Les résultats obtenus sont cohérents. En effet, la mortalité des ouvriers est supérieure à celle des autres groupes, et de même la mortalité des cadres est plus faible.

Pour rappel, la variable catégorie socioprofessionnelle avait des valeurs manquantes (environ 20%). Cela vient du fait que cette information n'est pas toujours demandée selon les produits de prévoyance. Par conséquent, la classe *Inconnu* contient les personnes dont on ne connaît pas cette information et se révèle être positionnée entre les extrêmes des ouvriers et des cadres.

Chapitre IV

Application pratique

Dans ce court chapitre, nous allons construire une table de mortalité **après refus** avec les méthodes classiques ainsi qu'avec les processus gaussiens. Ensuite, une analyse de sensibilité sur le *Best Estimate* sera menée avec l'utilisation de ces lois afin de déterminer leur impact.

IV.1) Calcul des taux après refus

IV.1.1) Nécessité d'un lissage préalable

Les données après refus sont plus volatiles que celles avant refus. Effectivement, les sinistres acceptés par l'assureur ne sont qu'une sous-partie des sinistres qui ont été reportés ($Acceptés \subset Reportés$).

La plus grande volatilité des données a alors exigé de réaliser un lissage des données brutes avant l'entraînement des processus gaussiens. Le lissage choisi est un lissage par moyennes mobiles ($n = 7$). Si un lissage n'avait pas été effectué, les prédictions par processus gaussiens pour les âges avec peu d'exposition (après 65 ans) se révélaient décroissantes à cause des taux des dernières années connues. La forte variabilité de ceux-ci laissaient à penser que les taux décroissaient à ces âges là. C'est pourquoi un lissage a été mené, pour s'assurer de la croissance des taux de mortalité au fil du temps et pouvoir prédire les années ultérieures.

Cette nécessité d'avoir des données suffisamment lisses en entrée des processus gaussiens est un de ses désavantages. Les intervalles de confiance fournis par la suite par les

processus gaussiens ne seront plus corrects, car la régularité des taux aura été artificiellement améliorée. Les processus gaussiens ne faisant aucune hypothèse de croissance des taux au fil du temps, il est donc nécessaire de s'assurer que les données d'entraînement reflètent cette propriété afin qu'elle soit apprise par le modèle.

IV.1.2) Comparaison des taux des deux méthodes

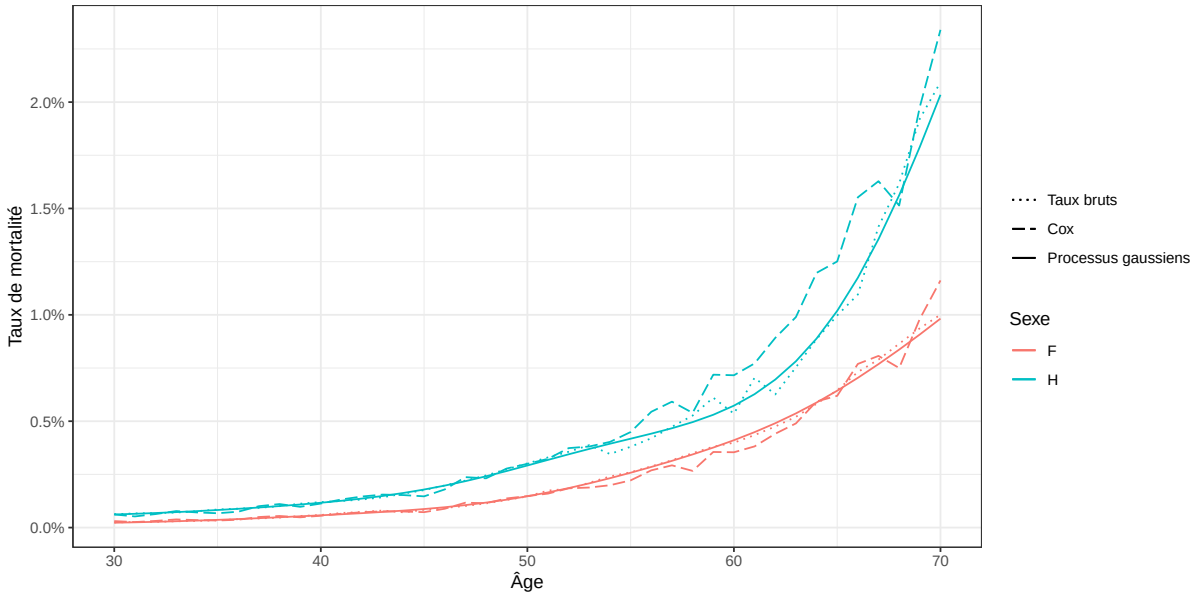

Figure IV.1 — *Comparaison des taux après refus des processus gaussiens et de Cox*

L'outil de calcul du *Best Estimate* de BPCE Assurances utilise, en tant qu'une de ses nombreuses entrées, une table de mortalité segmentée par âge, sexe et risque (décès toutes causes ou décès accidentel). Par conséquent, les taux de mortalité après refus ont été calculés par les processus gaussiens en dimension 2, ce qui a permis un positionnement par sexe. Ils ont aussi été déterminés par les méthodes classiques en utilisant le modèle de Cox. La comparaison des deux tables de mortalité ainsi obtenues se trouve en figure IV.1.

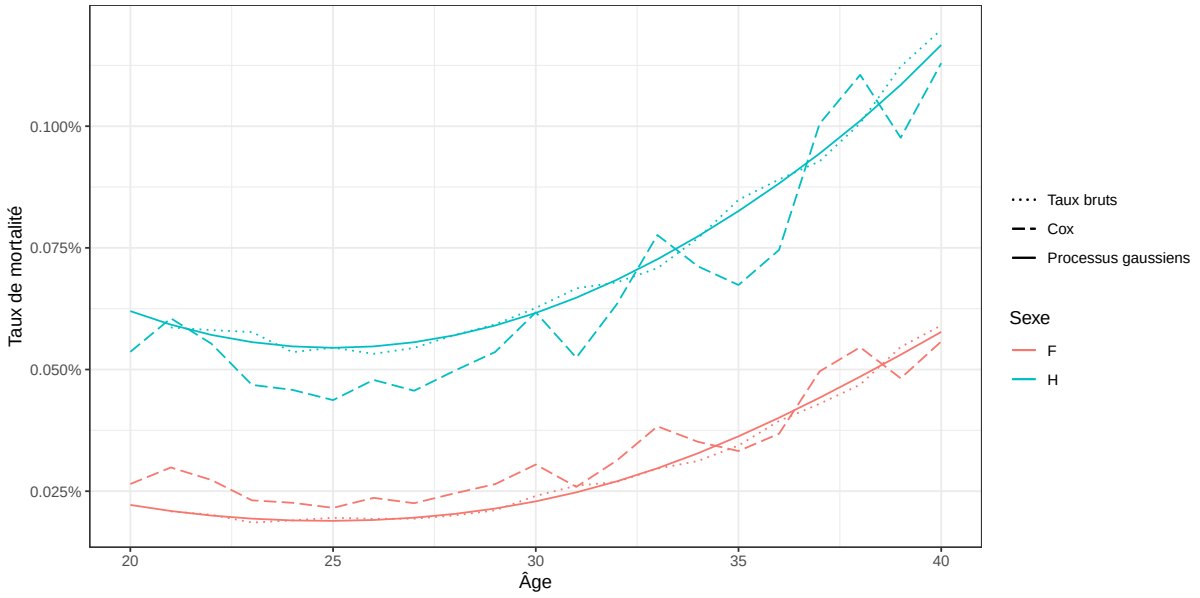

Figure IV.2 — *Zoom sur les taux après refus de 20 à 40 ans*

Zoomons sur ce graphique pour regarder plus précisément les différences entre les deux courbes. Tout d'abord, la figure IV.2 compare les taux des deux méthodes pour les âges entre 20 et 40 ans. Celle-ci met en lumière que les taux de mortalité des hommes jeunes sont sous-estimés avec Cox, alors que les processus gaussiens sont beaucoup plus proches des taux bruts. Concernant les femmes, c'est l'inverse pour Cox : les taux sont surestimés, mais ici aussi les processus gaussiens sont bien plus fidèles aux taux bruts.

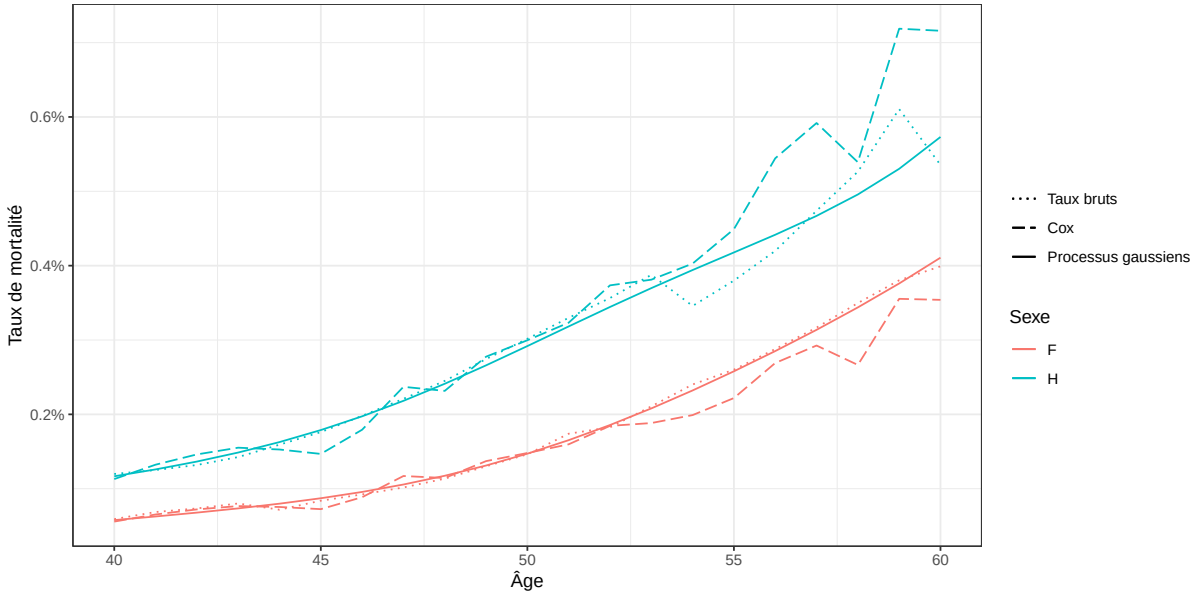

Figure IV.3 — *Zoom sur les taux après refus de 40 à 60 ans*

Ensuite, un zoom sur les âges de 40 à 60 ans est réalisé en figure IV.3. Entre 40 et 50 ans, les taux des processus gaussiens semblent relativement équivalents à ceux de

Cox. Cependant, à partir de 50 ans, la meilleure adéquation aux taux bruts des processus gaussiens par rapport à Cox devient clairement visible. En particulier, les taux des hommes sont surestimés avec Cox à partir de 50 ans, alors que les processus gaussiens sont plus fidèles. Concernant les taux des femmes, c'est l'inverse à partir de 50 ans : ils sont sous-estimés.

En conclusion, suite à ces analyses graphiques, l'écart de *Best Estimate* entre les deux méthodes devrait être modéré. En effet, aucune n'a des taux plus élevés que l'autre en tout âge, leurs différences pourraient donc se compenser. Les taux des processus gaussiens seront cependant plus adaptés dans le cas d'un calcul de *Best Estimate*, collant beaucoup plus aux taux bruts que ceux de Cox.

IV.1.3) Fermeture de table

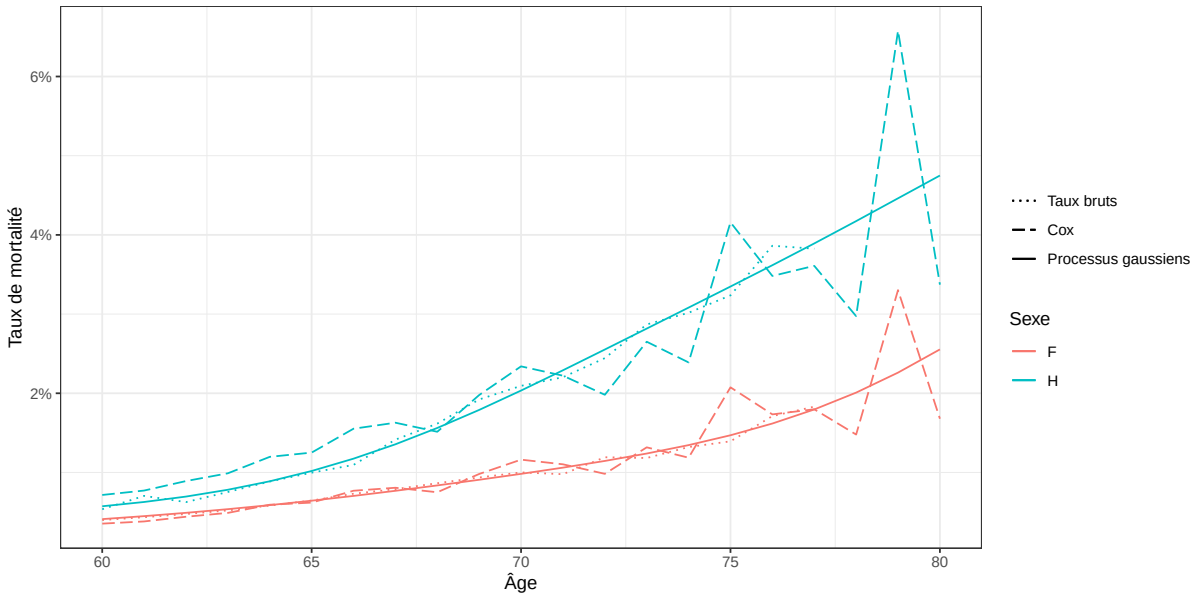

Figure IV.4 — *Zoom sur les taux après refus de 60 à 80 ans*

L'exposition du portefeuille se raréfie dès 65 ans, cependant l'outil de calcul du *Best Estimate* de BPCE Assurances nécessite de fournir tous les taux de mortalité, de 0 à 120 ans. Par conséquent, une fermeture de table par Coale et Kisker a été appliquée sur les taux de Cox, mais aussi sur ceux des processus gaussiens. Cela a permis de prolonger ces deux tables de 65 à 120 ans. Le résultat de cette fermeture de table se trouve en figure IV.5.

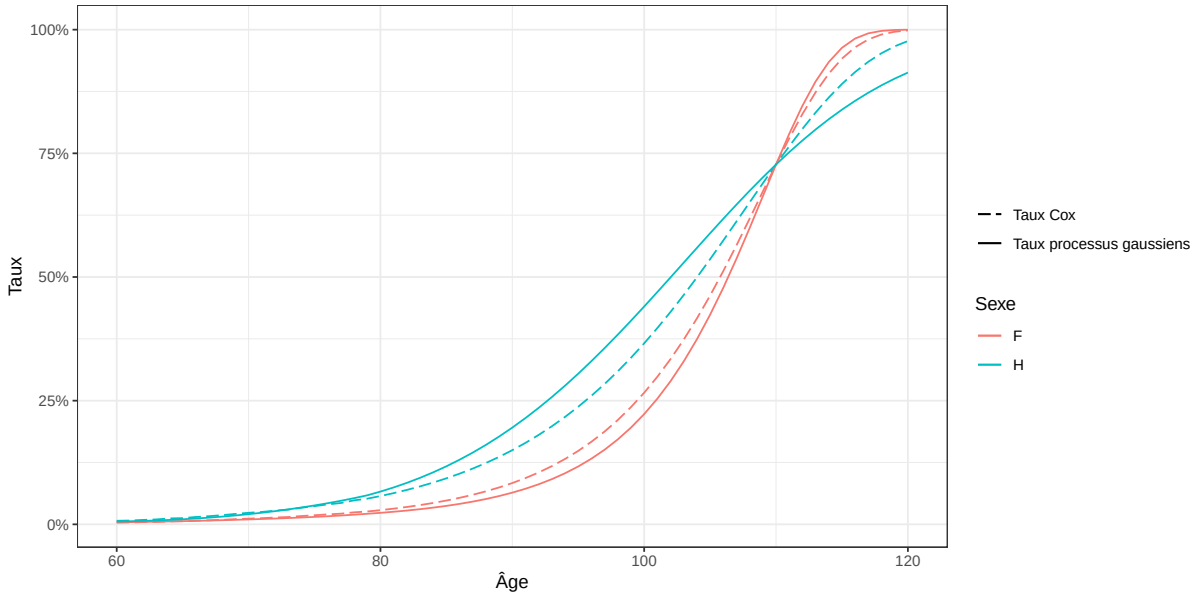

Figure IV.5 — *Fermeture de table des taux après refus*

IV.2) Sensibilité du Best Estimate

IV.2.1) Le Best Estimate

Le *Best Estimate*, ou « meilleure estimation du passif », constitue un élément essentiel dans le domaine de l'assurance. Il s'agit de l'évaluation la plus juste possible, se distinguant par une absence totale de marge pour le risque, celle-ci étant prise en compte dans un autre élément : la « Risk Margin » ou marge pour risque.

Le Comité européen des superviseurs d'assurance et de pensions professionnelles (CEIOPS) joue un rôle central dans la définition du *Best Estimate* dans le contexte européen. Dans sa communication CP numéro 26, le CEIOPS adopte la définition suivante, tirée des spécifications techniques du QIS4 (Quantitative Impact Study 4) : le *Best Estimate* repose sur la moyenne pondérée des futurs flux de trésorerie, en prenant en compte leur probabilité d'occurrence, tout en considérant la valeur temporelle de l'argent. Cette dernière est estimée en utilisant une courbe des taux sans risque pertinente.

Conformément à la directive européenne, le *Best Estimate* doit être calculé **brut de réassurance**. Cela signifie qu'il ne doit pas prendre en compte les effets de la réassurance lors de son calcul initial. Toutefois, il est important de noter qu'un actif de réassurance

est reconnu à l'actif de l'entreprise. Cette reconnaissance tient compte des probabilités de défaut du réassureur, contribuant ainsi à la gestion des risques.

Les hypothèses sous-jacentes à la détermination du *Best Estimate* reposent sur des données actuelles et crédibles. Une caractéristique fondamentale de ces hypothèses est leur réalisme. En d'autres termes, elles doivent **refléter fidèlement** les conditions économiques et les probabilités de manière pragmatique, garantissant ainsi une estimation solide et pertinente.

Le *Best Estimate* est donc un élément essentiel en assurance. Il fournit une base solide pour la prévision des flux de trésorerie futurs, tout en servant de fondement à la gestion des risques. Son caractère réaliste et sa neutralité en font un outil indispensable pour les professionnels de l'assurance et les régulateurs européens, garantissant ainsi la stabilité et la fiabilité du secteur.

IV.2.2) Calcul du Best Estimate

Le *Best Estimate* correspond à la moyenne, pondérée par leur probabilité, des flux de trésorerie futurs et actualisés. Une formule générale du BE est la suivante :

$$\begin{aligned} BE &= VAP(\text{Flux entrants} - \text{Flux sortants}) \\ &= VAP(\text{Primes TTC} - \text{Prestations} - \text{Commissions} - \text{Taxes} \\ &\qquad - \text{Frais généraux sur les primes} - \text{Frais généraux sur les prestations futures}) \end{aligned}$$

avec VAP la valeur actuelle probable, c'est-à-dire la somme des flux actualisés pondérés par leur probabilité d'occurrence.

IV.2.3) Résultats de la sensibilité

Réseau	BE modèle de Cox	BE processus gaussiens	Écart en %
BP	50 027 878	48 897 254	−2,26%
CE	33 629 413	33 548 977	−0,24%

Tableau IV.1 — *Sensibilité Best Estimate*

Le tableau IV.1 présente les résultats de la sensibilité du *Best Estimate* pour les deux réseaux. Les résultats sont proches, avec un écart de 2,26% pour le réseau BP et

de seulement 0,24% pour le réseau CE. Ces écarts sont faibles, bien que non négligeables dans le cas du réseau BP.

En tant que meilleure estimation, les processus gaussiens semblent plus adaptés pour le calcul du *Best Estimate*, étant plus proches des taux bruts. L'écart de BE avec les méthodes classiques reste toutefois faible. Par conséquent, les deux méthodes peuvent être utilisées pour le calcul du *Best Estimate* et être comparées l'une à l'autre afin de valider les résultats. Les processus gaussiens pourraient être utiles afin de confirmer les résultats obtenus par les méthodes classiques, en tant qu'outil supplémentaire de validation.

CHAPITRE V

CONCLUSION

L'objectif de ce mémoire était d'implémenter une méthode récente d'ajustement de loi de mortalité par processus gaussiens. En s'appuyant sur un papier théorique, le défi était de mettre en pratique cet algorithme et de le comparer aux méthodes actuellement utilisées par les actuaires du département.

Le passage de la théorie à la pratique a été source de satisfactions, bien que certaines limites aient été rencontrées. C'est ce que nous allons résumer dans cette conclusion.

V.1) Nombreux avantages des processus gaussiens

Nous avons pu constater dans ce mémoire les nombreux avantages des processus gaussiens. Tout d'abord, c'est une **méthode tout en un** : elle calcule les taux bruts, les lisse, donne des intervalles de confiance, permet du positionnement selon plusieurs variables explicatives, et enfin permet même de prédire en des points non connus.

Les intervalles de confiance fournis concernent certes les points d'entraînement, mais aussi les points inconnus pour lesquels l'exposition est nulle, ce qui est particulièrement utile. Les taux calculés par processus gaussiens sur des points connus sont proches de ceux qui seraient calculés par une méthode classique puis lissés.

Les intervalles de confiance sont cependant un peu plus grands que les méthodes classiques dans les âges élevés. En effet, les intervalles de confiance des processus gaussiens n'utilisent pas que l'exposition au point x, mais aussi les expositions des points aux alentours, qui sont encore plus faibles pour les âges élevés.

Les processus gaussiens permettent d'**extrapoler les taux aux âges où l'exposition est nulle**, comme en figure III.5 où les âges de 65 à 70 ans sont prédits. De façon logique, l'incertitude va en augmentant, les processus gaussiens utilisant les valeurs des points les plus proches connus.

Cette méthode offre un **réel atout quant au positionnement de la mortalité selon plusieurs variables**. En effet, pour réaliser une table de mortalité selon plusieurs variables, plusieurs solutions existent. L'une d'elles est de diviser la base d'exposition selon les modalités des variables, par exemple une base homme et une base femme. Cependant, cela diminue l'exposition dans chacune des deux bases, et donc augmente les intervalles de confiance. Une autre solution est le modèle de Cox, celui-ci utilisant toute la base d'exposition pour déterminer une fonction de hasard instantanée de base et des coefficients à lui appliquer selon les modalités des variables explicatives. Toutefois, procéder ainsi revient à faire une hypothèse forte sur les données : celle des risques proportionnels.

Les processus gaussiens ne font quant à eux aucune hypothèse de la sorte, tout en utilisant l'intégralité des données à disposition. Ils peuvent donc être très adaptés pour réaliser du positionnement. La tentative de positionnement par année a néanmoins donné des résultats mitigés, nous faisant essayer diverses fonctions de covariance, de formules de tendance, etc. Ces problèmes ne sont pas apparus pour les autres positionnements (par sexe, réseau ou catégorie socioprofessionnelle). Les résultats obtenus restent donc très encourageants vis-à-vis du positionnement.

Les résultats obtenus sont également **très facilement interprétables**, car la valeur prédite en chaque point suit une loi normale. Il devient alors facile d'obtenir des quantiles spécifiques, ce qui peut être utile pour les chocs de la directive Solvabilité II par exemple.

Enfin, les processus gaussiens sont **aisément implémentables**. En effet, des packages existent sur *Python* et sur *R* pour les utiliser. Une fois la base d'exposition au bon format, une simple fonction de l'un de ces packages permet l'entraînement du modèle, souvent en moins d'une minute. Les prédictions s'effectuent ensuite en fournissant au modèle les coordonnées des points sur lesquels on veut prédire la valeur moyenne et obtenir les intervalles de confiance. Ici aussi, le temps de calcul nécessaire est extrêmement faible.

L'implémentation des processus gaussiens dans ce mémoire a cependant été plus complexe que cela, car de nombreux essais ont été réalisés. Cela a nécessité d'automatiser toutes les étapes pour ne plus avoir qu'à fournir une liste de paramètres à la fonction de création des bases d'exposition et d'entraînement des processus gaussiens.

En conclusion, les processus gaussiens se révèlent être une méthode très intéressante pour la construction de tables de mortalité. Ils peuvent également être utiles dans une

optique confirmatoire des résultats des méthodes classiques. En ce sens, les processus gaussiens permettent de vérifier que les résultats obtenus par les méthodes classiques sont cohérents avec les données, et ce de manière très rapide. Cela peut notamment être utile pour confirmer que les hypothèses faites par les méthodes classiques sont bien vérifiées par les données.

V.2) Limites et pistes d'approfondissement

L'une des principales limites ayant été constatées avec les processus gaussiens appliqués à la modélisation de la mortalité réside dans leur capacité parfois limitée à projeter de manière fiable la mortalité à des âges inconnus. Selon les données les plus récentes disponibles dans l'ensemble des données d'entrée, les processus gaussiens peuvent dans certains cas générer une courbe de mortalité en forme de cloche. Cette représentation ne reflète évidemment pas la réalité, où la mortalité ne suit pas cette tendance. Une telle projection erronée se produit quand les données d'entraînement contiennent des taux de mortalité aux âges les plus avancés qui semblent décroître en raison de leur volatilité.

Ainsi, la qualité des données d'entrée fournies aux processus gaussiens revêt une importance cruciale lorsqu'il s'agit de faire des prédictions. Si ces données présentent une volatilité excessive, le modèle ne sera pas en mesure de fournir des prédictions fiables. C'est pourquoi, dans le chapitre IV d'application pratique, un prétraitement des données d'entraînement par lissage a été effectué pour garantir que les processus gaussiens puissent fournir des prédictions correctes. De même, il peut être plus intéressant de ne pas fournir les observations à partir d'un certain âge, même si elles sont disponibles, afin d'éviter que les processus gaussiens ne soient entraînés sur des données peu fiables.

Les processus gaussiens ne semblent donc pas permettre d'éviter la nécessité de procéder à des extrapolations par des modèles de fermeture de table en l'absence de points d'entraînement fiables pour les âges avancés. Une solution potentielle consisterait à ajouter manuellement des points d'entraînement à des valeurs spécifiques pour garantir que les prédictions générées par les processus gaussiens passent par ces points. En particulier, il serait envisageable de fixer le taux de mortalité à l'âge ω à 1 grâce à cette approche.

Dans la continuité du point précédent, les processus gaussiens n'intègrent aucune contrainte sur les taux. En particulier, leur évolution en fonction de l'âge ne suit aucune contrainte de monotonie. Pourtant, du point de vue actuariel, cette contrainte est par-

ticulièrement pertinente, par exemple dans le cadre de la tarification où l'on s'attend à observer une augmentation des taux en fonction de l'âge. Si cette contrainte était présente, les problèmes de prédiction pourraient se résoudre.

Riihimäki et Vehtari ont proposé une méthode intéressante pour intégrer cette contrainte dans les processus gaussiens, qu'il serait intéressant d'essayer de mettre en œuvre.[Rii10] Leur idée consiste à appliquer les processus gaussiens non pas à une fonction f telle que $f(x) = y$, mais plutôt au couple (f, f'). La contrainte de monotonie peut alors être exprimée comme une contrainte sur la fonction dérivée f', afin de garantir que toutes ses valeurs soient positives. Cette condition peut être introduite en ajoutant des points fictifs x' correspondant à la fonction dérivée dans l'ensemble de données d'entraînement, de manière à assurer la croissance de f aux endroits où elle ne l'est pas.

Une autre approche possible reviendrait à utiliser les processus gaussiens de la même manière qu'un modèle relationnel, comme le modèle de Brass dans les méthodes classiques. Les modèles relationnels sont particulièrement utiles lorsque les données disponibles sont limitées, en utilisant une table de mortalité de référence à laquelle des facteurs de correction sont appliqués. L'idée serait donc de fournir en entrée aux processus gaussiens à la fois les taux de mortalité bruts d'un portefeuille, mais aussi une table de mortalité de référence. Cela permettrait aux processus gaussiens d'apprendre les contraintes de monotonie sur les taux bruts à partir de la table de référence. Cette approche peut s'avérer très prometteuse pour l'extrapolation des taux de mortalité et la création de tables de mortalité par cohorte.

Enfin, en mai 2023, Mike Ludkovski et Jimmy Risk ont publié un article qui étend leur travail de 2018 sur les processus gaussiens en se penchant davantage sur le positionnement par cohorte.[Lud23] L'objectif de la méthode décrite dans leur article est d'utiliser de l'apprentissage automatique à base d'algorithme génétique pour déterminer la structure de la matrice de covariance la plus adaptée aux données d'entraînement.

En définitive, le domaine d'application des processus gaussiens pour la construction de tables de mortalité est en plein essor. De nouveaux articles sont régulièrement publiés à ce sujet, laissant présager des avancées dans les années à venir.

Liste des figures

Liste des tableaux

Bibliographie

Bibliographie du chapitre 1

[Lud18] Ludkovski M., Risk J., Zail H. *Gaussian process models for mortality rates and improvement factors*. Août 2018. URL : https://www.cambridge.org/core/journals/astin-bulletin-journal-of-the-iaa/article/gaussian-process-models-for-mortality-rates-and-improvement-factors/A2D48AFF8E32CEABF9B9DB899194D9C2 (cf. p. 10, 64, 67, 68, 80).

[Cor22] Corric J. *Quelle est la plus grande banque française ?* Juill. 2022. URL : https://www.agefi.fr/banque-assurance/actualites/quotidien/20220720/quelle-est-plus-grande-banque-francaise-347280 (cf. p. 11).

[Ass21] Assureur Pro. *Classement des compagnies d'assurance en France 2022*. Mai 2021. URL : https://assureurpro.com/classement-des-compagnies-dassurance-en-france/ (cf. p. 11).

[Dan22] Danton I. *Classement des bancassureurs 2022*. Avr. 2022. URL : https://www.argusdelassurance.com/classements/classements-marches/classement-des-bancassureurs-2022.199152 (cf. p. 11).

[Loi89] Loi Evin. *Loi n° 89-1009 du 31 décembre 1989*. Déc. 1989. URL : https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000709057 (cf. p. 12).

[Jou05a] Journal officiel. *Arrêté du 20 décembre 2005 relatif aux tables de mortalité, Article 5*. Déc. 2005. URL : https://www.legifrance.gouv.fr/jorf/article_jo/JORFARTI000002277632 (cf. p. 15).

[Jou05b] Journal officiel. *Arrêté du 20 décembre 2005 relatif aux tables de mortalité, Article 7*. Déc. 2005. URL : https://www.legifrance.gouv.fr/jorf/article_jo/JORFARTI000001106167 (cf. p. 15).

[Jou06] Journal officiel. *Arrêté du 1er août 2006 portant homologation des tables de mortalité pour les rentes viagères et modifiant certaines dispositions du code des assurances en matière d'assurance sur la vie et de capitalisation, Article 2*. Août 2006. URL : https://www.legifrance.gouv.fr/jorf/article_jo/JORFARTI000002073211 (cf. p. 15).

[Cod17] Code des Assurances. *Article A132-18*. Sept. 2017. URL : https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000035514715 (cf. p. 16).

[Insa] Institut des Actuaires. *Charte de la certification et de suivi des tables de mortalité et des lois de maintien en incapacité de travail et en invalidité*, p. 4. URL : http://www.ressources-actuarielles.net/EXT/ISFA/fp-isfa.nsf/0/1430ad6748ce3affc1256f130067b88e/%5C$FILE/ChartePublique.pdf (cf. p. 16).

[Insb] Institut des Actuaires. *Charte de la certification et de suivi des tables de mortalité et des lois de maintien en incapacité de travail et en invalidité*, p. 13. URL : http://www.ressources-actuarielles.net/EXT/ISFA/fp-isfa.nsf/0/1430ad6748ce3affc1256f130067b88e/%5C$FILE/ChartePublique.pdf (cf. p. 16).

[Ins06] Institut des Actuaires. *Lignes directrices mortalité de la Commission d'Agrément*. Juin 2006. URL : https://www.institutdesactuaires.com/global/gene/link.php?doc_id=173 (cf. p. 16, 32).

Bibliographie du chapitre 2

[Ins06] Institut des Actuaires. *Lignes directrices mortalité de la Commission d'Agrément*. Juin 2006. URL : https://www.institutdesactuaires.com/global/gene/link.php?doc_id=173 (cf. p. 16, 32).

[Hub94] Huber-Carol C. « Durées de survie tronquées et censurées ». In : Journal de la Société de statistique de Paris 135.4 (1994), p. 3-23. URL : http://www.numdam.org/item/JSFS_1994___135_4_3_0/ (cf. p. 25).

[Sta16] Stata Blog. *Understanding truncation and censoring*. 2016. URL : https://blog.stata.com/2016/12/13/understanding-truncation-and-censoring/ (cf. p. 27).

[Pla11] Planchet F., Therond P. *Modélisation statistique des phénomènes de durée*. Economica, 2011 (cf. p. 28, 29, 47).

[Hoe76] Hoem J. « The Statistical Theory of Demographic Rates : A Review of Current Developments [with Discussion and Reply] ». In : Scandinavian Journal of Statistics 3.4 (1976), p. 169-185. ISSN : 0303-6898. URL : https://www.jstor.org/stable/4615634 (cf. p. 29).

[Lai18] Laizet B. « Apports de la théorie de la crédibilité à l'estimation de la mortalité ». Mémoire d'actuariat, EURIA, sept. 2018. URL : https://www.

institutdesactuaires.com/docs/mem/124644cded736aba417bd405db6b8fbe.pdf (cf. p. 30).

[Ndi16] Ndiaye E. *Construction d'une table de mortalité et lissage par positionnement*. Oct. 2016, p. 35. URL : https://www.institutdesactuaires.com/docs/mem/6fa8f8f984a07c3fb391c97fc62608b6.pdf (cf. p. 31).

[Bal13] Baltesar B. « Construction d'une table de mortalité sur un portefeuille de temporaires décès ». Mémoire d'actuariat, ISFA, déc. 2013. URL : http://www.ressources-actuarielles.net/EXT/ISFA/1226-02.nsf/0/a843aa8832c0d8b8c1257c1c0069ed38/%5C$FILE/M%C3%A9moire%20Final%20Baltesar%20Benjamin%2002-01-14.pdf (cf. p. 31).

[Kap58] Kaplan E. L., Meier P. « Nonparametric Estimation from Incomplete Observations ». In : Journal of the American Statistical Association 53.282 (1er juin 1958), p. 457-481. ISSN : 0162-1459. DOI : 10.1080/01621459.1958.10501452. URL : https://www.tandfonline.com/doi/abs/10.1080/01621459.1958.10501452 (visité le 23/02/2023) (cf. p. 32).

[Hoe84] Hoem J. « A flaw in actuarial exposed-to-risk theory ». In : Scandinavian Actuarial Journal (1984). URL : https://www.tandfonline.com/doi/abs/10.1080/03461238.1984.10413766 (cf. p. 32).

[Pla22] Planchet F. *Modèles de durées, Statistique des modèles non paramétriques*. 2022. URL : http://www.ressources-actuarielles.net/C1256F13006585B2/0/1430AD6748CE3AFFC1256F130067B88E/%5C$FILE/Seance5.pdf (cf. p. 33).

[Koy19] Koye G. K. « Comparaison des méthodes classiques et alternatives avec le machine learning pour la construction d'une table de mortalité d'experience Best Estimate ». Mémoire d'actuariat, ENSAE, nov. 2019. URL : https://www.institutdesactuaires.com/docs/mem/8dd841db383e6891f9a410a1a1e5669a.pdf (cf. p. 35).

[Gui15] Guillon Verne T. « Construction de tables de mortalité d'expérience sur de petits échantillons pour l'estimation de la sinistralité décès ». Mémoire d'actuariat, ISFA, juill. 2015. URL : http://www.ressources-actuarielles.net/EXT/ISFA/1226-02.nsf/0/12d7471392b25cebc1257e1d001f0fa5/$FILE/M%C3%A9moire%20d%27actuaire%20Thomas%20Guillon%20Verne.002.pdf/M%C3%A9moire%20d%27actuaire%20Thomas%20Guillon%20Verne.pdf (cf. p. 35).

[Bie23] Biessy G. *Revisiting Whittaker-Henderson Smoothing*. Juin 2023. URL : https://arxiv.org/pdf/2306.06932.pdf (cf. p. 39).

[Les11] Leste-Lasserre C. « Construction de tables de mortalité d'expérience ». Mémoire d'actuariat, ISUP, 2011. URL : https://www.institutdesactuaires.com/docs/mem/3a16b895bd1415704a961b11b1456566.pdf (cf. p. 39, 40).

[Orf18] Orfanidis S. J. *Applied Optimum Signal Processing, Chap. 8 Whittaker-Henderson Smoothing*. 2018. URL : http://eceweb1.rutgers.edu/~orfanidi/aosp/aosp-ch08.pdf (cf. p. 40).

[Maz14] Mazza A., Punzo A. « DBKGrad : An R Package for Mortality Rates Graduation by Discrete Beta Kernel Techniques ». In : Journal of Statistical Software, Code Snippets 57.2 (2014), p. 1-18. DOI : 10.18637/jss.v057.c02. URL : https://www.jstatsoft.org/index.php/jss/article/view/v057c02 (cf. p. 41).

[Cox72] Cox D. R. « Regression Models and Life-Tables ». In : Journal of the Royal Statistical Society : Series B (Methodological) 34.2 (1972), p. 187-202. DOI : https://doi.org/10.1111/j.2517-6161.1972.tb00899.x (cf. p. 42).

[Man66] Mantel N. « Evaluation of survival data and two new rank order statistics arising in its consideration ». In : Cancer Chemotherapy Reports (mars 1966), p. 163-170. URL : https://pubmed.ncbi.nlm.nih.gov/5910392/ (cf. p. 43).

[Agr11] Agrinier N., Baumann C. *Recherche clinique et épidémiologique*. Fév. 2011. URL : https://fad.univ-lorraine.fr/pluginfile.php/23863/mod_resource/content/2/co/Realisation_Pratique_Test_Log-Rank.html (cf. p. 45).

[Tia14] Tian L. *Logrank Test*. Jan. 2014. URL : https://web.stanford.edu/~lutian/coursepdf/unitweek3.pdf (cf. p. 46).

[Alb05] Alberti C., Timsit J.-F., Chevret S. « Analyse de survie : le test du logrank ». In : Revue des Maladies Respiratoires 22.5, Part 1 (2005), p. 829-832. ISSN : 0761-8425. URL : https://www.sciencedirect.com/science/article/pii/S076184250585644X (cf. p. 46).

[Qua05] Quashie A., Denuit M. « Modèles d'extrapolation de la mortalité aux grands âges ». In : Scandinavian Actuarial Journal (fév. 2005). URL : http://www.planchet.net/EXT/ISFA/1226.nsf/769998e0a65ea348c1257052003eb94f/e2c194a584426385c125704200417aac/$FILE/Mod%C3%A8les%20d'extrapolation%20de%20la%20mortalit%C3%A9%20aux%20grands%20%C3%A2ges.pdf (cf. p. 47).

Bibliographie du chapitre 3

[Lud18] Ludkovski M., Risk J., Zail H. *Gaussian process models for mortality rates and improvement factors*. Août 2018. URL : https://www.cambridge.org/core/journals/astin-bulletin-journal-of-the-iaa/article/gaussian-process-models-for-mortality-rates-and-improvement-factors/A2D48AFF8E32CEABF9B9DB899194D9C2 (cf. p. 10, 64, 67, 68, 80).

[Bre06] Breton J. *Processus gaussiens*. 2006. URL : https://perso.univ-rennes1.fr/jean-christophe.breton/Fichiers/gauss_M2.pdf (cf. p. 59).

[Sel06] Selma B. *Introduction aux processus gaussiens*. 2006. URL : http://archives.univ-biskra.dz/bitstream/123456789/13656/1/boudib_salma.pdf (cf. p. 59).

[Ebd08] Ebden M. *Gaussian Processes for Regression*. Août 2008. URL : https://www.apps.stat.vt.edu/leman/VTCourses/GPtutorial.pdf (cf. p. 61).

[Del03] Delalleau O. *Introduction aux Processus Gaussiens*. 2003. URL : http://www.iro.umontreal.ca/~pift6266/A06/cours/030819_talk_lisa_gaussian-process.pdf (cf. p. 63).

[Bas19] Baskiotis N. *Processus Gaussien*. 2019. URL : https://dac.lip6.fr/wp-content/uploads/2021/04/ML-cours8-2020.pdf (cf. p. 66).

[ENS20] ENSIIE. *Algorithme Gaussian Process Upper Confidence Bound*. 2020. URL : http://skutnik.iiens.net/cours/3A/SGI/191018115914_PG.pdf (cf. p. 67).

[Ras06] Rasmussen C.E, Williams C.K.I. *Gaussian Processes for Machine Learning*. 2006. URL : http://gaussianprocess.org/gpml/ (cf. p. 67).

[Mig21] Miglietti R. *Modélisation de la mortalité à l'aide de modèles de régression gaussiens et de l'analyse en composantes principales fonctionnelle*. 2021. URL : https://dial.uclouvain.be/downloader/downloader.php?pid=thesis%3A32071&datastream=PDF_01&cover=cover-mem (cf. p. 68).

Bibliographie de la conclusion

[Rii10] Riihimäki J., Vehtari A. *Gaussian processes with monotonicity information.* 2010. URL : https://proceedings.mlr.press/v9/riihimaki10a.html (cf. p. 97).

[Lud23] Ludkovski M., Risk J. *Expressive Mortality Models through Gaussian Process Kernels.* Mai 2023. URL : https://arxiv.org/abs/2305.01728 (cf. p. 97).