Mémoire présenté devant le jury de l'EURIA en vue de l'obtention du Diplôme d'Actuaire EURIA et de l'admission à l'Institut des Actuaires

le 6 Septembre 2023

Par : Christian Borel WAFO KANKEU
Titre : Impact du montant de rente sur la longévité au sein d'un portefeuille de rentiers : Une approche par les modèles additifs généralisés

Confidentialité : Non

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

Membre présent du jury de l'Institut des Actuaires :
Marine HABART
Yassin NAJID
Signature :

Entreprise :
ACS LinkPact
Signature :

Membres présents du jury EURIA :
Franck VERMET
Signature :

Directeur de mémoire entreprise :
Guillaume BIESSY
Signature :

Invité :

Signature :

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

Signature du responsable entreprise :

Signature du candidat :

EURIA
EURo Institut
d'Actuariat

6, avenue le Gorgeu
CS 93837
29238 Brest Cedex 3

T +33 (0)2 98 01 66 55
euria@univ-brest.fr

Résumé

Le risque longévité pour un assureur disposant d'un portefeuille de rentiers consiste en une sous-estimation de l'espérance de vie de ses assurés. Un aspect sous-jacent de ce risque longévité est l'hétérogénéité des dynamiques de survie due aux différences de niveau de vie. Ce phénomène se traduit par le fait que les rentiers les plus riches aient une espérance de vie plus élevée que les moins riches. Ainsi dans ce mémoire, il est entrepris la construction d'une table de mortalité prospective pour un portefeuille de rentiers de l'assureur Swiss Life, en tenant compte du montant de rente perçu par les individus dans l'évaluation des probabilités de décès.

Une approche par les modèles additifs généralisés est utilisée pour la construction de la table. Il est capté à l'issue de la construction de cette dernière une décroissance du niveau de mortalité avec le montant de rente. Cet effet du montant de rente est confirmé par deux autres modèles : le modèle de Cox et les forêts aléatoires de survie.

La table construite a ensuite été utilisée pour le calcul de provisions. Ces provisions se sont révélées êtres plus prudentes que les provisions obtenues avec les tables réglementaires et que celles obtenues avec le modèle utilisé jusqu'ici par Swiss Life.

Mots clés : risque longévité, modèles additifs généralisés, montant de rente, table de mortalité, tables réglementaires, modèle de Cox, forêts aléatoires de survie.

Abstract

Longevity risk for an insurer with a portfolio of annuitants consists of underestimating the life expectancy of its insured individuals. An underlying aspect of this longevity risk is the heterogeneity of survival dynamics due to differences in the standard of living. This phenomenon results in the fact that the wealthiest annuitants have a higher life expectancy than the less affluent ones. Thus, in this paper, the construction of a prospective mortality table is undertaken for a portfolio of annuitants of the insurer Swiss Life, taking into account the amount of annuity received by individuals in the assessment of death probabilities.

An approach using generalized additive models is used for table construction. At the end of this construction, a decrease in mortality level is captured with the amount of annuity. This effect of the annuity amount is confirmed by two other models: the Cox model and random survival forests.

The constructed table was then used for reserve calculations. These reserves turned out to be more cautious than the reserves obtained with regulatory tables and than those obtained with the model previously used by Swiss Life.

Keywords: longevity risk, generalized additive models, annuity amount, mortality table, regulatory tables, Cox model, random survival forests.

Remerciements

Je tiens à remercier :

— Mon tuteur en entreprise Guillaume BIESSY, Manager et Responsable du département de Recherche et Développement au sein du cabinet ACS LinkPact. Pour son accompagnement et son implication active à la réalisation de ce mémoire, ses conseils avisés, son incroyable expertise technique ainsi que ses nombreuses relectures de ce document.

— Michel MORCOS EL DOUAIHY et Guillaume RAMOND les Fondateur et Associés du cabinet ACS LinkPact, pour m'avoir permis d'effectuer mon alternance au sein de leur cabinet.

— Nos interlocuteurs chez Swiss Life en les personnes d'Olivier REVERCHON, Anthony DEBERNARDI et Henintsoa RAJAONAH. Pour avoir fourni les données sur lesquels a porté l'étude menée dans ce mémoire, leur présence aux différentes réunions et leur réactivité lors de nos différentes sollicitations.

— L'ensemble du personnel de ACS LinkPact pour leur acceuil au sein du cabinet.

— Les corps professoral et administratif de l'EURIA pour les deux années de formation que j'y ai reçues.

Synthèse

Impact du montant de rente sur la longévité au sein d'un portefeuille de rentiers : Une approche par les modèles additifs généralisés

Christian Borel WAFO KANKEU

problématique et objectifs

Le risque longévité pour un assureur disposant d'un portefeuille de rentiers consiste en une sous-estimation de l'espérance de vie de ses assurés. Une espérance de vie plus élevée par rapport aux prévisions de l'assureur aboutit à des pertes financières pour ce dernier, vu qu'il doit payer des primes plus longtemps que ce qui était prévu, et donc en plus grande quantité que ce qui avait été provisionné au départ. La prise en compte de ce risque longévité est donc un enjeu important pour les assureurs couvrant ce type de garanties. Ceci d'autant plus que de nombreuses études tendent à montrer que l'espérance de vie humaine devrait continuer d'augmenter pendant plusieurs années encore, notamment dans les pays développés (Kontis et al. 2017). Pour tenir compte de cette espérance de vie qui évolue, des tables de mortalité prospectives sont utilisées par les assureurs pour le calcul des provisions. Leur aspect prospectif permet d'imprimer une dynamique de hausse quasi linéaire de l'espérance de vie chaque année (Debonneuil, Loisel, and Planchet 2018).

Ces améliorations de l'espérance de vie ne se font cependant pas au même rythme pour l'ensemble de la population. Il existe une certaine hétérogénéité notamment due au niveau de vie qui fait que les individus les plus aisés vivent en moyenne plus longtemps que les plus modestes (INSEE 2018). Dans le cas d'un portefeuille de rentiers, ce phénomène se traduit par le fait que les rentiers les plus riches aient une espérance de vie plus élevée que les moins riches.

L'objectif principal de l'étude menée dans ce mémoire sera ainsi de tenir compte du niveau de vie (au moyen du montant de la rente perçue) dans la construction d'une table de mortalité d'expérience prospective (calcul des probabilités de décès) pour des assurés d'un portefeuille de rentiers de l'assureur Swiss Life.

Deux principales approches sont généralement utilisées pour la construction des tables prospectives. Une approche intrinsèque basée uniquement sur l'utilisation des données d'expérience, et une approche par référence externe basée sur le positionnement de la mortalité du groupe étudié par rapport à une mortalité de référence. C'est cette dernière approche qui est la plus

couramment utilisée en actuariat, car elle est particulièrement adaptée lorsque l'on ne dispose pas de données en grandes quantités, ce qui est généralement le cas pour des portefeuilles individuels. De très nombreux mémoires d'actuariat ont déjà traité la question de la construction des tables de mortalité prospectives pour des portefeuilles de rentiers par l'approche référence externe. On a ainsi Fall (2019) qui a élaboré une table prospective pour un portefeuille d'épargne retraite du groupe Malakoff, Yikmis (2020) pour un portefeuille de rente viagère du groupe Generali, ou encore Guez (2018) sur des données de rentiers de Groupama Gan Vie. Il est à noter tout de même des approches plus originales comme Durieux and Samba (2013) qui ont exploré les apports de la théorie des copules dans la modélisation de la mortalité, ou encore Bastien (2020) qui a utilisé la théorie de la crédibilité pour construire une table de mortalité prospective.

Des études qui prennent en compte des caractéristiques autres que l'âge et le sexe dans les modèles de mortalité sont plus rares. Une problématique assez similaire à la notre a tout de même déjà été traitée dans Ziegelmeyer (2015). L'auteur avait exploré trois méthodes pour tenir compte du montant de rente dans l'évaluation des probabilités de décès, dont deux consistaient en une segmentation de la population assurée suivant le montant de rente de manière à obtenir des sous populations homogènes du point de vue de la mortalité. Il avait ensuite construit des tables différentes pour chacune de ces sous populations.

L'approche de modélisation qui sera utilisée dans ce mémoire sera celle des **modèles additifs généralisés**, elle ne nécessite pas une segmentation a priori de la population des assurés. Ce type de modèles est encore assez peu utilisé en Actuariat. Néanmoins, Côté (2016) présente une application intéressante de ces derniers à de la tarification en assurance automobile.

Description du portefeuille à l'étude

Les données utilisées dans ce mémoire proviennent d'un portefeuille de rentes viagères de l'assureur Swiss Life. Il a été arrêté pour l'étude une période d'observation de 6 ans, allant du 1^{er} Janvier 2017 au 31 Décembre 2022.

Au total, on dénombre 56 220 assurés qui ont été observés lors de cette période d'observation, avec 37 491 hommes (67%) pour 18 729 femmes (33%). Il a été enregistré un total de 6 716 décès dont 63% d'hommes (4 254) et 37% de femmes (2 462).

Âges en sortie d'observation

L'âge moyen de sortie d'observation est de 76 ans pour les femmes contre $73,5$ ans pour les hommes. Un pic de densité est cette fois ci observé à l'âge de 69 ans (Figure 1). Il est toujours observé une densité plus importante aux grands âges chez les femmes par rapport aux hommes. A noter qu'il y a de la censure dans les observations, ces âges ne reflètent donc

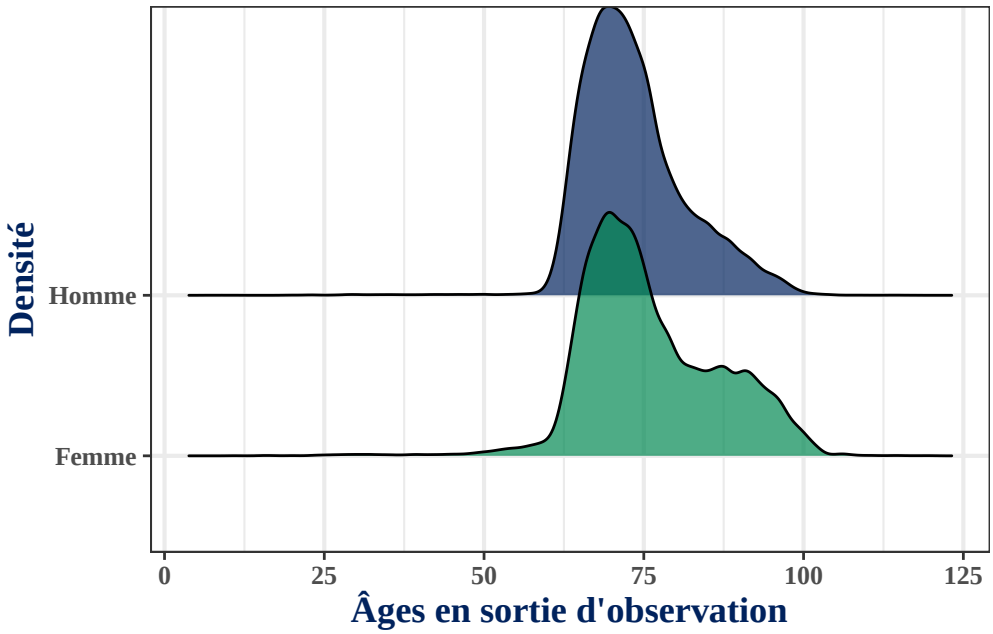

Figure 1: Distribution des ages en sortie d'observation

pas la distribution ultime des âges de sortie. **L'âge moyen au décès** est de 87 ans chez les femmes et de 81 ans chez les hommes.

Montant de rente

La distribution des montants de rente (Figure 2) aussi bien pour les hommes que pour les femmes est fortement asymétrique à droite. Au global, 97% des assurés perçoivent un montant de rente annuel inférieur à 10 000€. On dénombre une trentaine d'assurés qui touchent des rentes supérieures à 50 000€, avec un maximum à près de 650 000€.

Positionnement du portefeuille par rapport à la référence réglementaire

Il sera question ici de comparer le niveau de mortalité observé dans le portefeuille avec celui des tables réglementaires **TGH05** et **TGF05**. Nous partirons ici des données sous leur forme la plus agrégée. L'idée est de calculer les décès théoriques (d_{th}) qu'on aurait observé si la mortalité du portefeuille était telle que décrite par les tables réglementaires. On compare ensuite ces décès théoriques aux décès effectivement observés dans le portefeuille.

Les tables **TGH05** (pour les hommes) et **TGF05** (pour les femmes) sont des tables générationnelles qui donnent les niveaux de mortalité sur plusieurs générations, de 1900 à 2005 pour

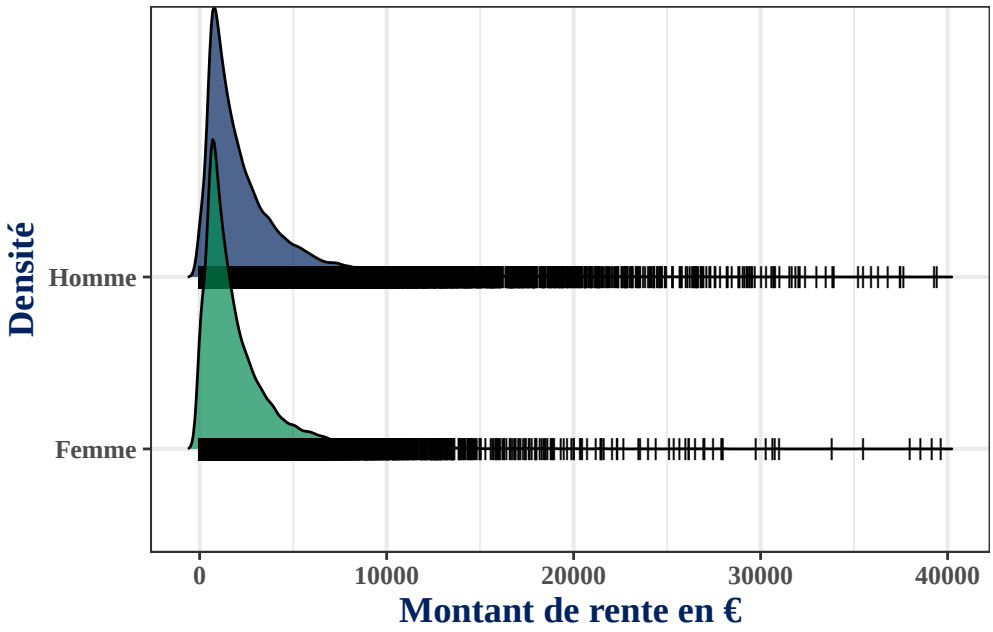

Figure 2: Distribution des montants de rente

des ages allant de 0 à 120 ans. Elles se présentent (pour chaque génération) comme un écoulement d'une population de taille initiale (fictive) 100 000, qui diminue au fil des années jusqu'à extinction complète (décès de l'ensemble des individus). Pour un **sexe donné**, le nombre d'individus survivants de la génération t, d'age x est alors donné par L_x^t. Le nombre de décès théorique à l'age x des individus de la génération t est donné par :

$$\begin{aligned} d_{th}^{t,x} &= ec_x^t \times h_x(t) \\ &= -ec_x^t \times \log(1 - L_{x+1}^t / L_x^t) \end{aligned}$$

Avec :

- ec_x^t l'exposition du portefeuille pour les individus de la génération t d'âge x
- $h_x(t)$ la force de mortalité de la table réglementaire pour les individus de la génération t d'âge x.

Une fois ces décès théoriques obtenus par âge, sexe et génération, il est possible de les sommer par âge, par sexe, voire même par année (génération).

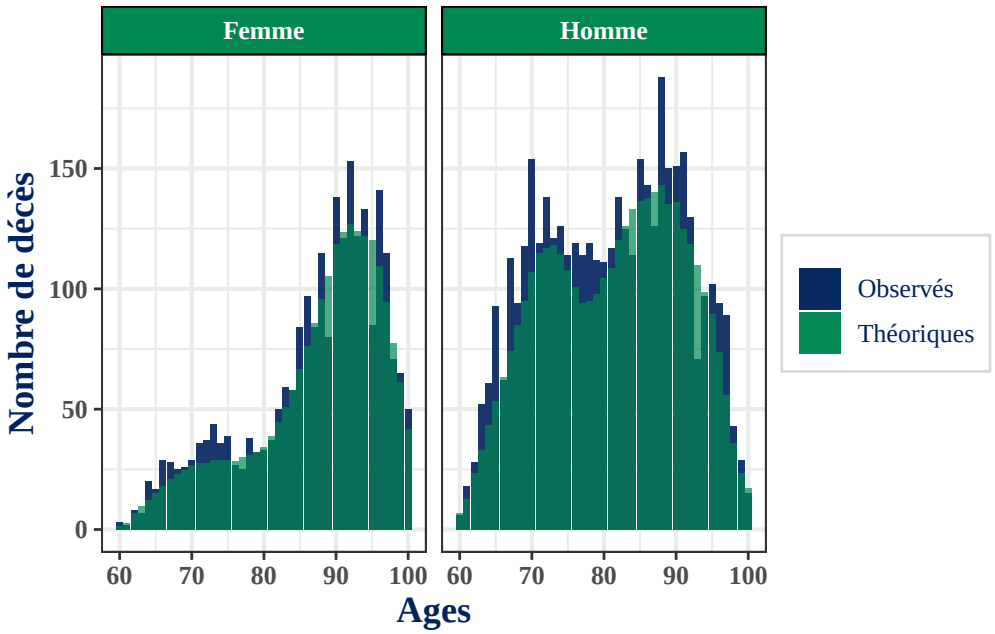

Figure 3: Décès observés et décès prédits par les tables règlementaires par âge et par sexe

Par âge et par sexe

Quasiment à tous les âges entre 60 et 100 ans, aussi bien pour les hommes que pour les femmes, les décès observés dans le portefeuille sont supérieurs à ceux prévus par les tables réglementaires (Figure 3). Au global on constate ainsi une surmortalité dans le portefeuille par rapport aux tables réglementaires de l'ordre de 10%. Guez (2018) dans une étude similaire aboutissait également à une surmortalité dans le portefeuille à l'étude par rapport aux tables réglementaires.

Par classe de montant de rente

Sur la Figure 4, en dehors des classes **"500-700"** et **"1,6K-2K"** qui présentent des pics assez singuliers, on peut discerner une tendance à la baisse de l'écart de mortalité par rapport aux tables réglementaires à mesure que la classe de montant de rente augmente.

Cette relation entre mortalité et montant de rente sera investiguée de manière plus rigoureuse par la suite avec les différents modèles qui seront calibrés.

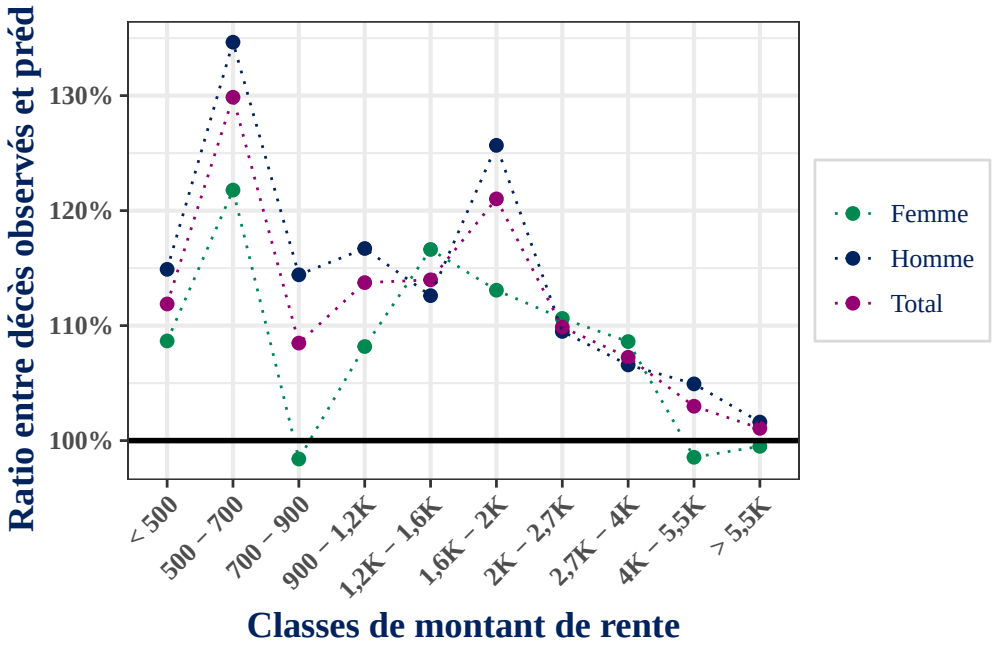

Figure 4: Décès observés et décès prédits par les tables règlementaires par classe de montant de rente

Les modèles additifs généralisés

Un modèle additif généralisé en abrégé **GAM** (Generalized Additive Model) est un modèle linéaire généralisé avec des prédicteurs linéaires qui sont des fonctions **lisses** des variables explicatives.

$$g(\mu_i) = A_i\theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}) + ... \tag{1}$$

Avec:

- $\mu_i \equiv \mathbb{E}(Y_i)$, Y_i étant la variable d'intérêt de loi appartenant à la famille exponentielle
- g une fonction de lien
- $A_i\theta$ correspondant à la partie paramétrique du modèle, où A_i est une ligne de la matrice de design, et θ le vecteur de paramètre associé
- f_j une fonction lisse associée à la covariable x_j

Ce type de modèle permet une certaine flexibilité (comparativement aux modèles totalement paramétriques) dans la matérialisation des dépendances entre les différentes covariables et la variable d'intérêt.

Construction de la table de mortalité prospective

Les modèles additifs généralisés ont été calibrés sur le nombre de décès, en supposant que ce dernier suit une loi de Poisson . Nous mettrons en **offset l'exposition** et la **force de mortalité** des tables réglementaires. Pour une loi de poisson, la fonction de lien g est la fonction **logarithme népérien**.

Le modèle calibré s'écrit comme suit :

$$\log(E(D) = d) = \alpha + S_1(x) + S_2(m_2) + S_3(x, m_2) + \beta_1 \times s + \log(ec) + \log(\mu) \qquad (2)$$

Avec:

- α, β_1 : des coefficients qui seront estimés par le modèle
- x, d, m_2, s, ec, μ : l'âge, le nombre de décès, le montant de rente, le sexe, l'exposition au risque et la force de mortalité des tables réglementaires
- S_1, S_2 des fonctions lisses à estimer pour l'âge et le montant de rente respectivement
- S_3 la fonction lisse à estimer pour l'interaction entre l'âge et le montant de rente
- m_2 la médiane de la classe de montant de rente.

Effet de l'âge

Sur la Figure 5, on observe un effet décroissant de l'âge. Ainsi l'écart entre les décès observés dans le portefeuille et ceux prédits par les tables réglementaires tend à décroître à mesure que l'âge augmente. Une interprétation de l'effet représenté sur ce graphique est la suivante : Toute chose égale par ailleurs, pour un individu du portefeuille âgé de 60 ans, il faudrait multiplier la force de mortalité des tables réglementaires par 132% pour obtenir la force de mortalité de cet individu dans le portefeuille.

Effet du montant de rente

On constate sur la Figure 6 une tendance à la baisse des écarts par rapport aux tables réglementaires à mesure que la médiane de la classe de montant de rente augmente. L'effet capturé ici est beaucoup plus net que ce que l'on observait avec le modèle précédent dans lequel la rente était traitée comme une variable catégorielle. De plus, le lissage élimine les fluctuations erratiques que l'on pouvait observer d'une classe de montant de rente à l'autre, ce qui semble plus proche de la réalité du phénomène étudié. L'effet capturé ici tend à se courber légèrement vers les montants de rente les plus élevés avec une pente de moins en moins raide. Ceci traduit

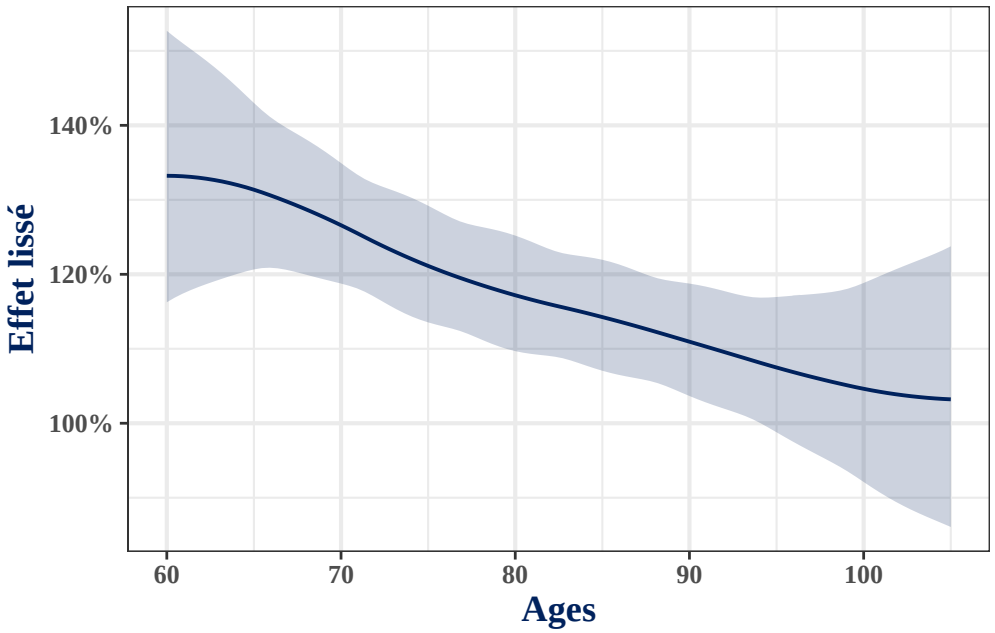

Figure 5: Effet lisse de l'âge

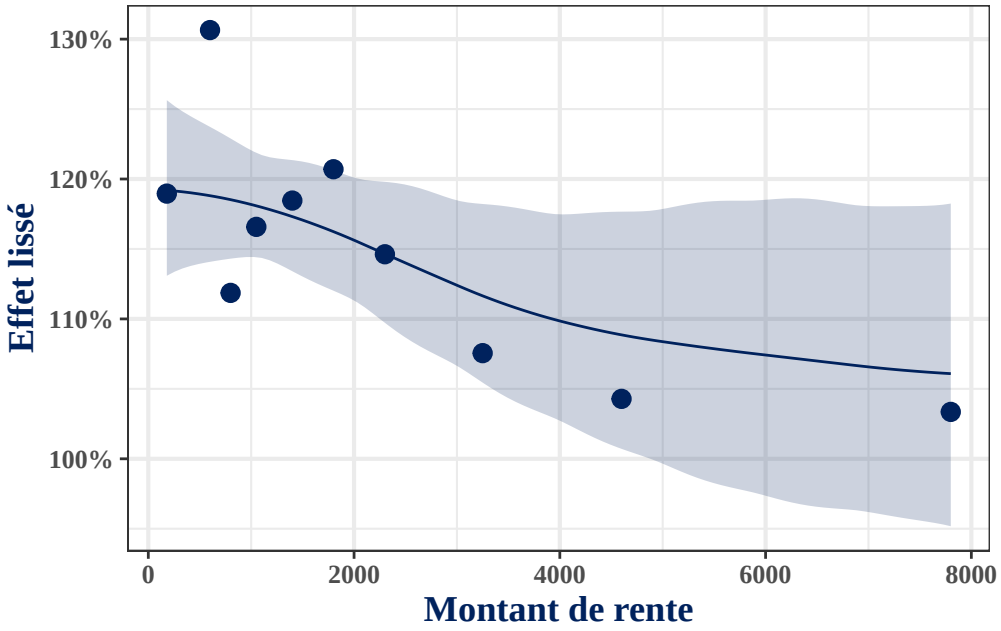

Figure 6: Effet lisse du montant de rente

une certaine atténuation de l'effet du montant de rente vers les montants élevés. Ceci est plutôt en accord avec la littérature sur le sujet qui stipule que les améliorations de mortalité liées au niveau de vie sont de plus en plus faibles à mesure que le niveau de vie augmente.

Effet d'interaction

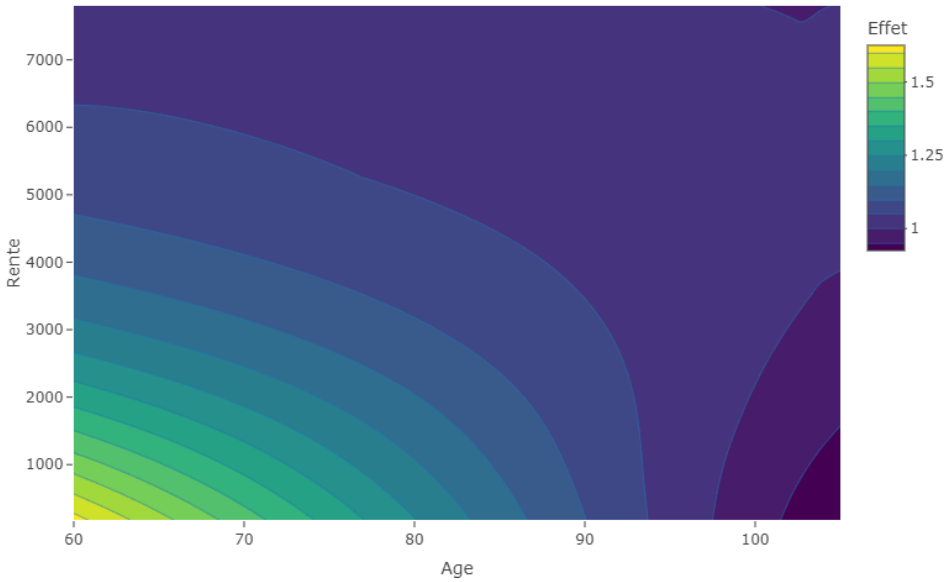

Figure 7: Effet d'interaction entre âge et montant de rente

La Figure 7 présente les effets combinés des variables âge et montant de rente prises individuellement ainsi que l'effet de leur interaction. Il apparaît ici que plus l'âge augmente et plus un effet éventuel du montant de rente tend à disparaître (de moins en moins d'effets verticaux vers les grands âges sur le graphique), ce qui semble logique car au-delà d'un certain âge la mortalité est déterminée exclusivement par des facteurs biologiques liés au vieillissement.

L'effet du montant de rente a été confirmé par un modèle de Cox et par des forêts aléatoires de survie.

Calcul des provisions

Pour évaluer l'impact de la table construite, les provisions ont été calculées avec quatre modèles différents :

- Le modèle GAM ayant servi à la construction de la table
- Un modèle GAM ne prenant pas en compte le montant de rente

- Le modèle utilisé par Swiss Life
- Les tables réglementaires.

Ecarts globaux

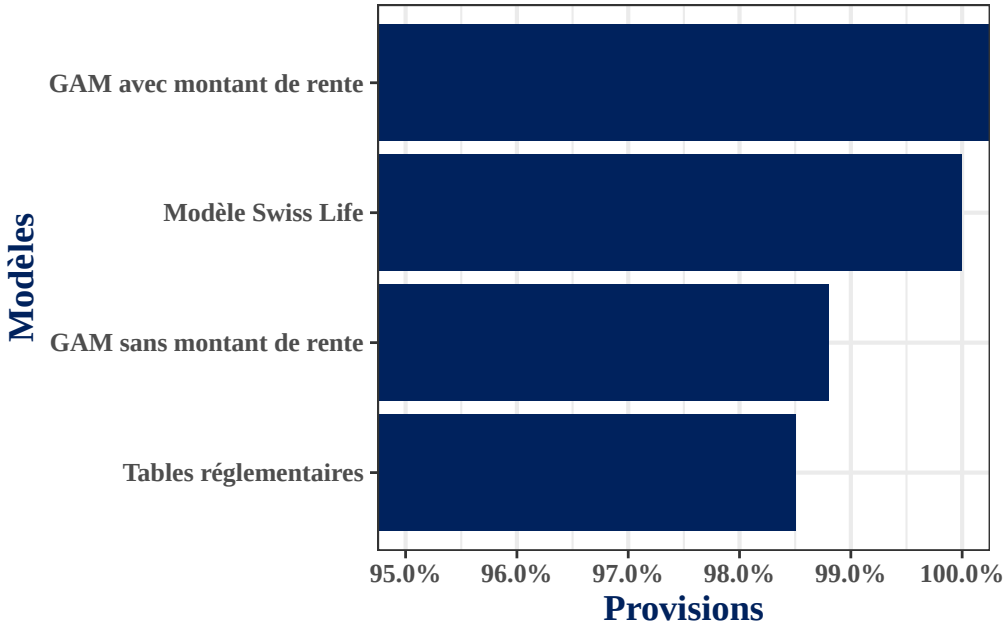

Figure 8: Provisions globales par modèle exprimées en terme de pourcentage des provisions du modèle Swiss Life

la Figure 8 présente les provisions totales évaluées par les différents modèles en pourcentage des provisions évaluées par le modèle interne. Le premier constat ici est que les provisions évaluées suivant les deux modèles GAM et le modèle Swiss Life sont plus prudentes que les tables réglementaires. Ensuite, le modèle Swiss Life se révèle moins prudent que le modèle GAM avec montant de rente , mais plus prudent que le modèle GAM sans montant de rente.

Impact du montant de rente

Pour isoler l'effet du montant de rente, nous allons comparer les provisions du modèle GAM incluant le montant de rente et celles du modèle GAM n'incluant pas le montant de rente. Ceci permet d'éliminer les interférences qui pourraient être liées aux différences intrinsèques entre les modèles GAM et le modèle Swiss Life. La Figure 9 présente les provisions par montant de rente (pour le modèle GAM incluant le montant de rente) en pourcentage des provisions

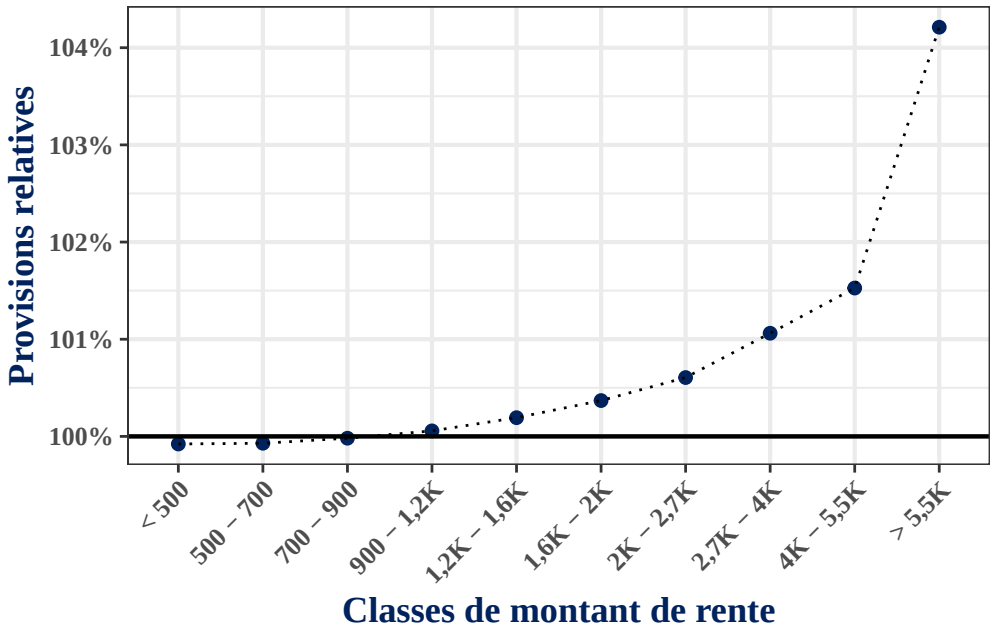

Figure 9: Provisions par montant de rente exprimées en terme de pourcentage des provisions GAM sans prise en compte de la rente

fournies par le modèle GAM n'incluant pas le montant de rente. On constate des écarts de provisions de plus en plus élevés à mesure que le montant de rente augmente. Ceci est assez logique vu que le modèle GAM avec montant de rente comme vu dans le chapitre précédent, décrit une diminution de la mortalité avec le montant de rente. De ce fait, à âge et sexe inchangés, il fournit des provisions d'autant plus élevées (relativement au modèle ne prenant pas en compte le montant de rente) que le montant de rente est élevé. On constate un écart maximale de plus de 4% pour la classe de montant de rente la plus élevée.

Conclusion

L'objectif principal de l'étude menée dans ce mémoire était la construction pour un portefeuille de rentiers d'une table de mortalité prospective qui tienne compte du montant de rente perçu par les individus dans l'évaluation des probabilités de décès. Pour ce faire, des modèles additifs généralisés ont été utilisés. Il a été observé pour le portefeuille à l'étude une décroissance du niveau de mortalité avec l'augmentation du montant de rente. C'est un résultat cohérent au vu de la littérature sur le sujet de la longévité et de l'espérance de vie qui stipule que les individus les plus aisés vivent en moyenne plus longtemps que les individus les plus modestes.

Des provisions mathématiques ont été calculées avec la table prospective ainsi construite. Au

global, les provisions obtenues via la table construite sont plus prudentes que celles obtenues avec le modèle Swiss Life ou encore les tables réglementaires. Cette étude présente un certains nombre de limites dont la principale est sans doute le fait que le montant de la rente perçue par les individus dans le portefeuille étudié ici n'est pas forcément représentatif de leur niveau de vie. en effet il est possible par exemple que des assurés ayant des montants de rente faibles dans le portefeuille étudié ici aient d'autres contrats de rente dans d'autres compagnies avec des rentes bien plus élevées. Il faudrait alors utiliser la somme des montants sur les différents contrats pour être parfaitement exhaustif, chose qui nous est impossible malheureusement vu que cette information est inaccessible.

Summary

Impact of the annuity amount on longevity within a portfolio of annuitants : An approach using Generalized Additive Models

Christian Borel WAFO KANKEU

Problem and objectives

The longevity risk for an insurer with a portfolio of annuitants consists of underestimating the life expectancy of its insured individuals. A higher life expectancy compared to the insurer's forecasts results in financial losses for the insurer, as it must pay premiums for a longer period than anticipated, and therefore in larger amounts than initially provisioned. Taking this longevity risk into account is therefore an important challenge for insurers covering this type of guarantee. This is especially true given that many studies tend to show that human life expectancy is expected to continue increasing for several more years, particularly in developed countries (Kontis et al. 2017). To account for this evolving life expectancy, prospective mortality tables are used by insurers for reserve calculations. Their prospective aspect allows for a nearly linear increase in life expectancy each year (Debonneuil, Loisel, and Planchet 2018).

However, these improvements in life expectancy do not occur at the same rate for the entire population. There is a certain heterogeneity, particularly due to the standard of living, which results in wealthier individuals living longer on average than those with more modest means (INSEE 2018). In the case of a portfolio of annuitants, this phenomenon results in wealthier annuitants having a higher life expectancy than less affluent ones.

The main objective of the study conducted in this paper is to take into account the standard of living (through the amount of received annuity) in the construction of a prospective experience mortality table (calculating death probabilities) for insured individuals in a portfolio of annuitants of the insurer Swiss Life.

Two main approaches are generally used for constructing prospective tables. An intrinsic approach based solely on the use of experience data, and an external reference approach based on the positioning of the studied group's mortality compared to a reference mortality. The latter approach is the most commonly used in actuarial science, as it is particularly suitable when there is not a large amount of data available, which is usually the case for individual portfolios.

Many actuarial theses have already addressed the question of constructing prospective mortality tables for annuitant portfolios using the external reference approach. For example, Fall (2019) developed a prospective table for a retirement savings portfolio of the Malakoff group, Yikmis (2020) for an annuity portfolio of the Generali group, or Guez (2018) on annuitant data from Groupama Gan Vie. There are also more original approaches, such as Durieux and Samba (2013), which explored the contributions of copula theory in mortality modeling, or Bastien (2020), who used credibility theory to construct a prospective mortality table.

Studies that take into account characteristics other than age and sex in mortality models are rarer. A somewhat similar issue to ours has already been addressed in Ziegelmeyer (2015). The author had explored three methods to account for the amount of annuity in assessing death probabilities, two of which involved segmenting the insured population according to the amount of annuity to obtain homogeneous subpopulations in terms of mortality. Different tables were then constructed for each of these subpopulations.

The modeling approach used in this paper will be that of **generalized additive models**. This approach does not require a priori segmentation of the insured population. This type of model is still relatively underused in actuarial science. However, Côté (2016) presents an interesting application of these models to automobile insurance pricing.

Description of the study portfolio

The data used in this paper come from a portfolio of life annuities from the insurer Swiss Life. An observation period of 6 years was chosen for the study, ranging from January 1, 2017, to December 31, 2022.

In total, there are $56,220$ insured individuals who were observed during this observation period, with $37,491$ males (67%) and $18,729$ females (33%). A total of $6,716$ deaths were recorded, with 63% being males $(4,254)$ and 37% being females $(2,462)$.

Ages at eExit from observation

The average age at exit from observation is 76 years for females compared to $73,5$ years for males. A density peak is observed at the age of 69 years this time (Figure 1). A higher density at older ages in females compared to males is still observed. Note that there is censoring in the observations, so these ages do not reflect the ultimate distribution of exit ages. The **average age at death** is 87 years for females and 81 years for males.

Figure 1: Distribution of ages at exit from observation

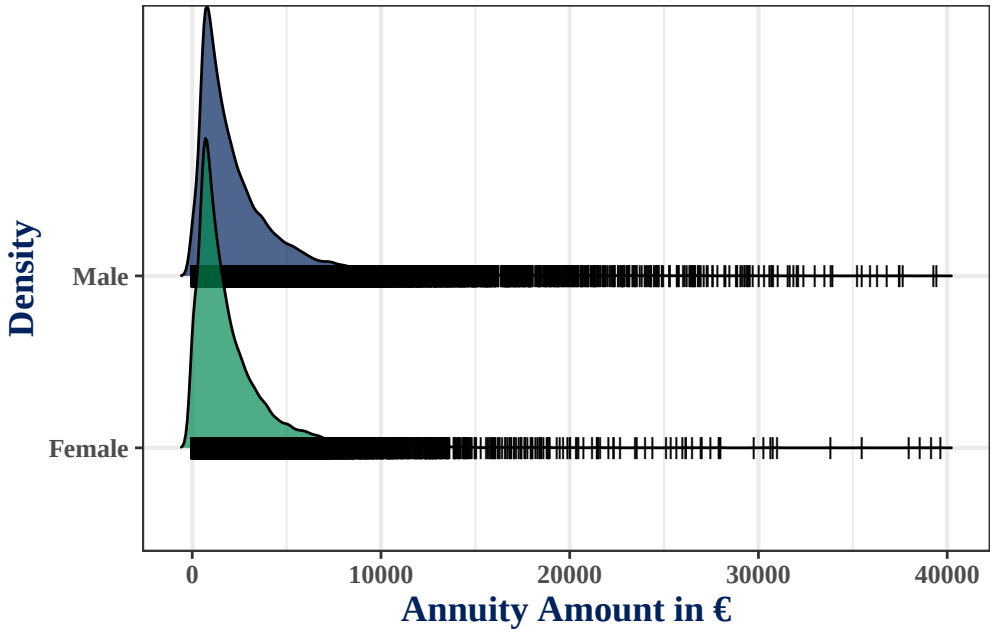

Figure 2: Distribution of annuity amount

Annuity amount

The distribution of annuity amounts (Figure 2) for both males and females is strongly right-skewed. Overall, 97% of insured individuals receive an annual annuity amount of less than €10 000. There are around thirty insured individuals receiving annuities higher than €50 000, with a maximum near €650 000.

Portfolio Positioning Relative to Regulatory Reference

Here, we will compare the observed mortality level in the portfolio with that of the regulatory tables **TGH05** and **TGF05**. We will start with the data in their most aggregated form. The idea is to calculate the theoretical deaths (d_{th}) that would have been observed if the portfolio's mortality were as described by the regulatory tables. We then compare these theoretical deaths to the deaths actually observed in the portfolio.

The **TGH05** (for males) and **TGF05** (for females) tables are generational tables that provide mortality levels over multiple generations, from 1900 to 2005, for ages ranging from 0 to 120 years. They are presented (for each generation) as the flow of an initial (fictional) population size of 100 000, which decreases over the years until complete extinction (death of all individuals). For a given **sex**, the number of surviving individuals from generation t, aged x, is then given by L_x^t. The number of theoretical deaths at age x for individuals of generation t is given by:

$$\begin{aligned} d_{th}^{t,x} &= ec_x^t \times h_x(t) \\ &= -ec_x^t \times \log(1 - L_{x+1}^t / L_x^t) \end{aligned}$$

Where:

- ec_x^t is the portfolio exposure for individuals of generation t aged x
- $h_x(t)$ is the force of mortality from the regulatory table for individuals of generation t aged x.

Once these theoretical deaths are obtained by age, sex, and generation, it is possible to sum them by age, by sex, or even by year (generation).

By Age and Sex

At nearly all ages between 60 and 100 years, for both males and females, the deaths observed in the portfolio are higher than those expected by the regulatory tables (Figure 3). Globally, a mortality excess of approximately 10% is observed in the portfolio compared to the regulatory tables. Guez (2018) in a similar study also found higher mortality in the studied portfolio compared to the regulatory tables.

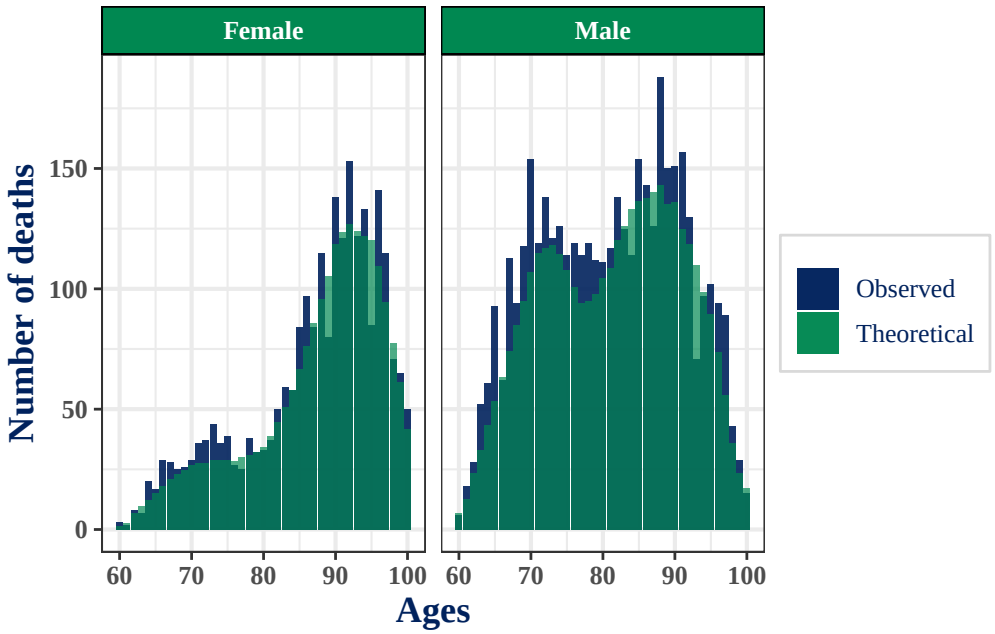

Figure 3: Deaths observed and deaths predicted by the regulatory tables by age and sex

By Annuity Amount Category

On Figure 4, aside from the **"500-700"** and **"1,6K-2K"** categories, which show quite unique peaks, there is a tendency for the mortality gap compared to the regulatory tables to decrease as the median annuity amount increases.

This relationship between mortality and annuity amount will be further rigorously investigated with the different models that will be calibrated.

Generalized Additive Models (GAMs)

A Generalized Additive Model, abbreviated as **GAM**, is a generalized linear model with smooth predictor functions of explanatory variables.

$$g(\mu_i) = A_i\theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}) + ... \tag{1}$$

Where:

- $\mu_i \equiv \mathbb{E}(Y_i)$, Y_i being the variable of interest following an exponential family distribution

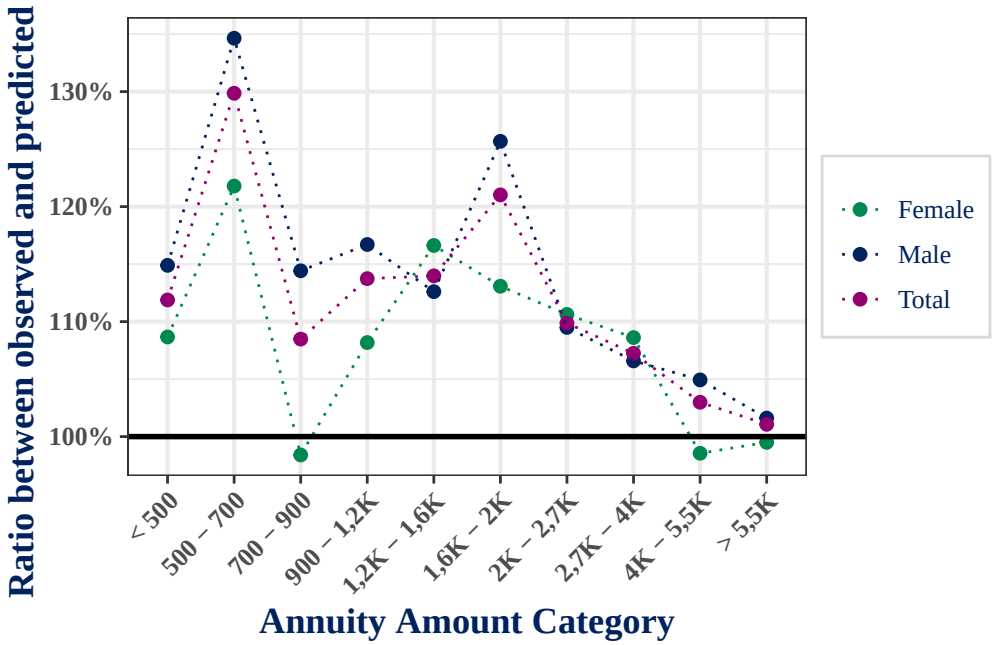

Figure 4: Deaths observed and deaths predicted by the regulatory tables by annuity amount category

- g a link function
- $A_i\theta$ corresponding to the parametric part of the model, where A_i is a row of the design matrix, and θ is the parameter vector associated
- f_j a smooth function associated with covariate x_j

This type of model offers some flexibility (compared to fully parametric models) in capturing dependencies between different covariates and the variable of interest.

Construction of the Prospective Mortality Table

The generalized additive models were calibrated on the number of deaths, assuming that it follows a Poisson distribution. We will put the exposure and the force of mortality from the regulatory tables in the **offset**. For a Poisson distribution, the link function g is the **natural logarithm** function.

The calibrated model is written as follows:

$$\log(E(D) = d) = \alpha + S_1(x) + S_2(m_2) + S_3(x, m_2) + \beta_1 \times s + \log(ec) + \log(\mu) \qquad (2)$$

Where:

- α, β_1 are coefficients to be estimated by the model
- x, d, m_2, s, ec, μ represent age, number of deaths, annuity amount, sex, exposure to risk, and force of mortality from the regulatory tables
- S_1, S_2 are smooth functions to be estimated for age and annuity amount respectively
- S_3 is the smooth function to be estimated for the interaction between age and annuity amount
- m_2 is the median of the annuity amount category.

Age effect

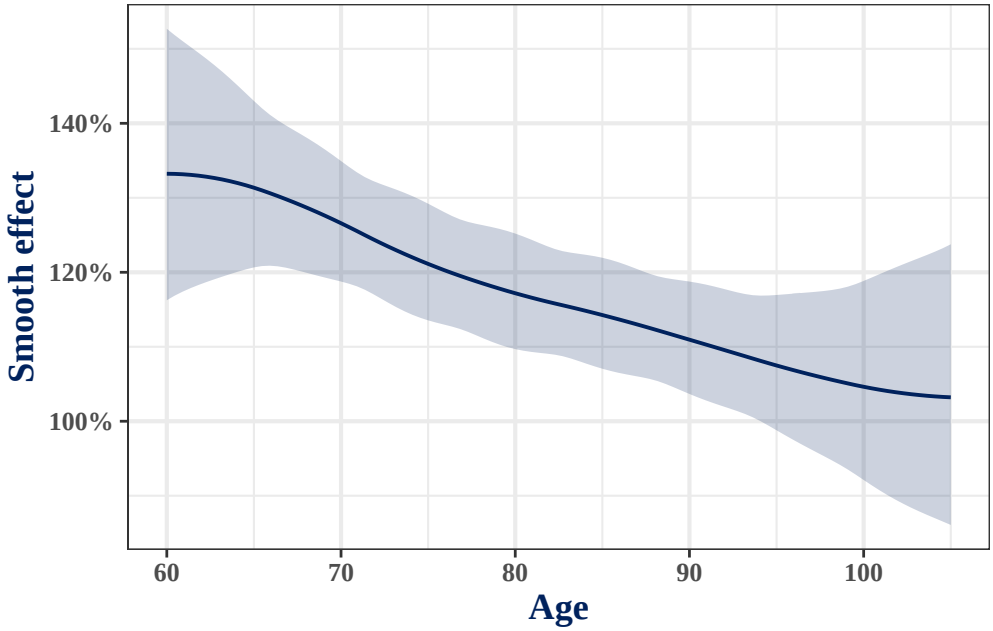

Figure 5: Smooth effect of age

On Figure 5, a decreasing effect of age is observed. Thus, the gap between deaths observed in the portfolio and those predicted by the regulatory tables tends to decrease as age increases. An interpretation of the effect shown in this graph is as follows: All else being equal, for a 60-year-old individual in the portfolio, the force of mortality from the regulatory tables would need to be multiplied by 132% to obtain the force of mortality for that individual in the portfolio.

Annuity amount effect

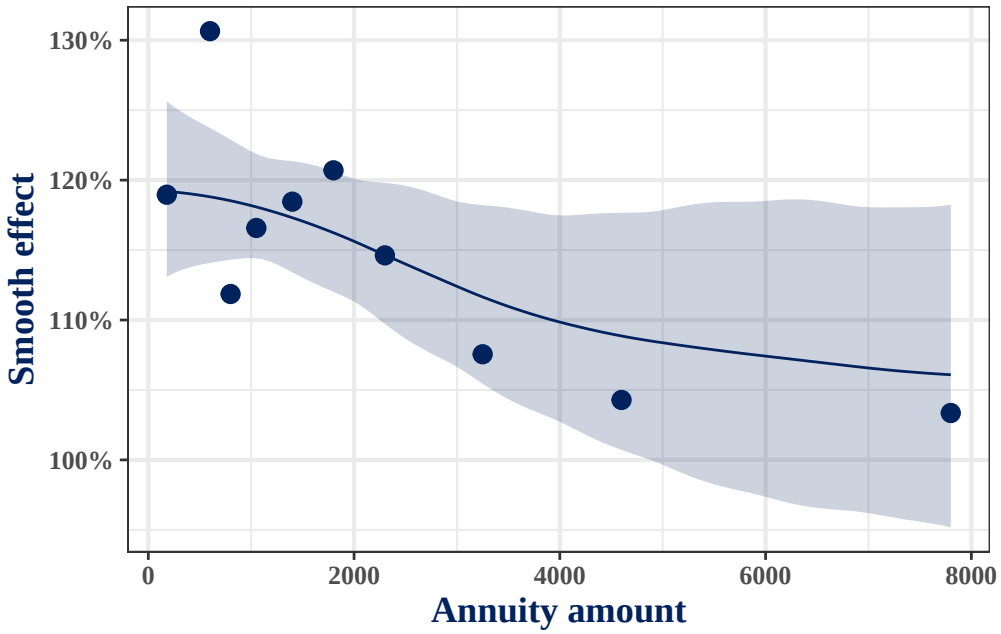

Figure 6: Smooth effect of annuity amount

On Figure 6, a trend of decreasing gaps compared to the regulatory tables is observed as the median annuity amount increases. The effect captured here is much clearer than what was observed with the previous model in which annuity amount was treated as a categorical variable. Furthermore, smoothing eliminates the erratic fluctuations that could be observed from one annuity amount category to another, which seems closer to the reality of the studied phenomenon. The effect captured here tends to curve slightly toward higher annuity amounts with a less steep slope. This reflects that improvements in mortality due to the standard of living are increasingly weaker as the standard of living increases.

Interaction effect

Figure 7 presents the combined effects of the age and annuity amount variables taken individually, as well as the effect of their interaction. It appears here that as age increases, any potential annuity amount effect tends to disappear (fewer vertical effects towards older ages on the graph), which is logical since beyond a certain age, mortality is exclusively determined by biological factors related to aging.

The annuity amount effect was confirmed by a Cox model and random survival forests.

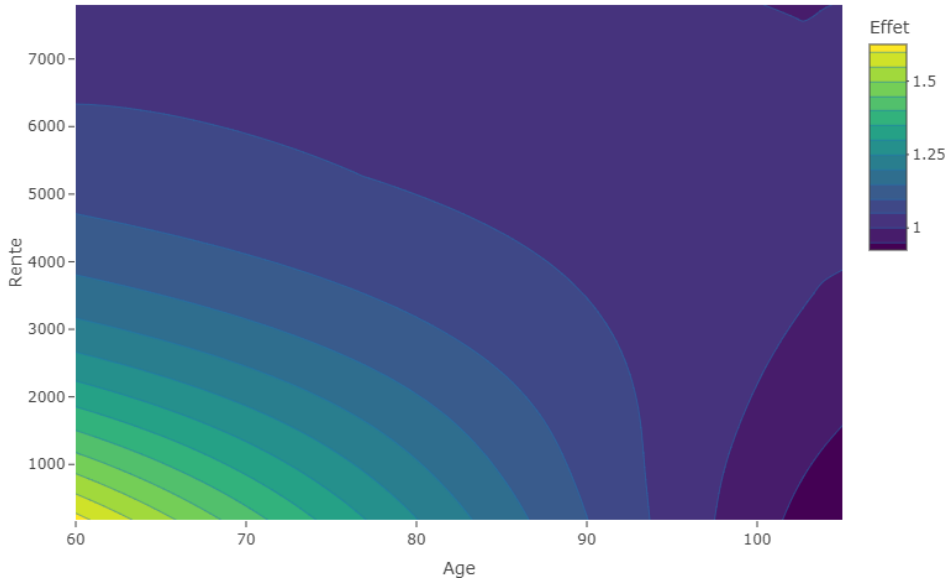

Figure 7: Interaction effect between age and annuity amount

Reserve Calculations

To assess the impact of the constructed table, reserves were calculated using four different models:

- The GAM model used for table construction
- A GAM model not taking annuity amount into account
- The model used by Swiss Life
- Regulatory tables.

Overall Differences

Figure 8 presents the total reserves evaluated by the different models as a percentage of the reserves evaluated by the internal model. The first observation here is that the reserves evaluated using both GAM models and the Swiss Life model are more cautious than the regulatory tables. Additionally, the Swiss Life model is less cautious than the GAM model with annuity amount, but more cautious than the GAM model without annuity amount.

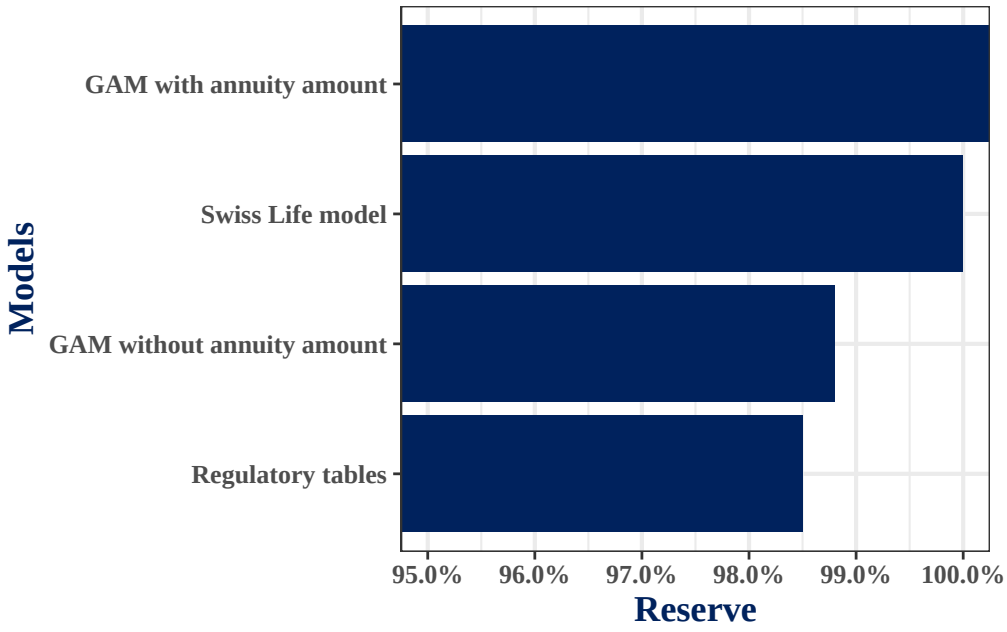

Figure 8: Global reserve by model

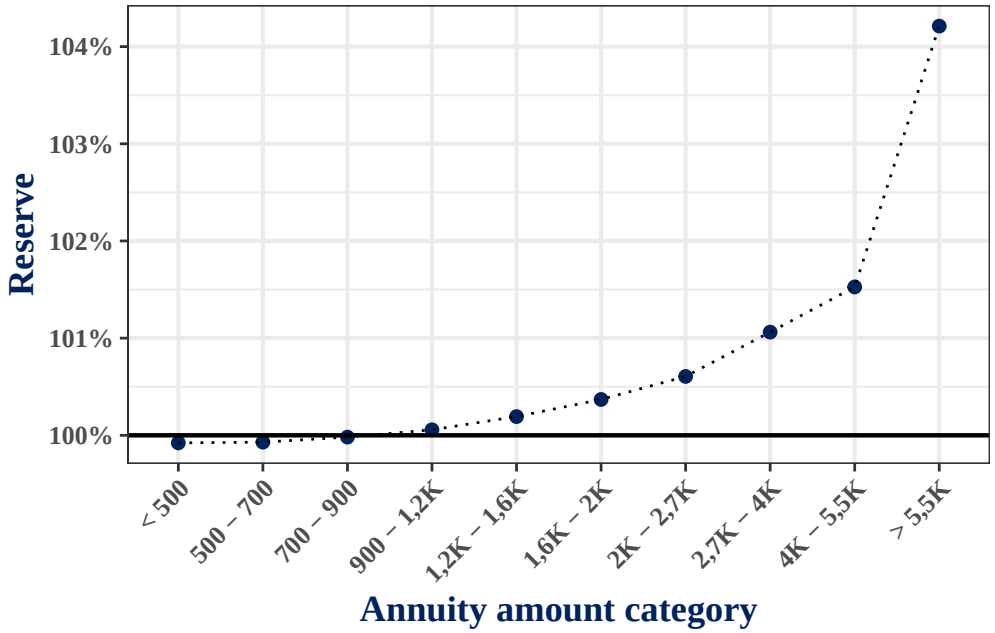

Figure 9: Reserve by annuity amount

Annuity amount impact

To isolate the annuity amount effect, we will compare the reserves from the GAM model including annuity amount with those from the GAM model not including annuity amount. This eliminates interferences that could be due to intrinsic differences between the GAM models and the Swiss Life model. Figure 9 presents the reserves per annuity amount (for the GAM model including annuity amount) as a percentage of the reserves provided by the GAM model not including annuity amount. Increasing gaps in reserves are observed as the median annuity amount increases. This is quite logical, as the GAM model with annuity amount, as seen in the previous chapter, describes a decrease in mortality with annuity amount. Therefore, at unchanged age and sex, it provides relatively higher reserves (compared to the model not including annuity amount) as the annuity amount increases. A maximum difference of over 4% is observed for the highest annuity amount category.

Conclusion

The main objective of the study in this paper was to construct a prospective mortality table for a portfolio of annuitants, taking into account the annuity amount received by individuals when evaluating death probabilities. To achieve this, generalized additive models were used. A decrease in mortality with increasing annuity amount was observed in the studied portfolio. This result is consistent with the literature on longevity and life expectancy, which states that more affluent individuals tend to live longer than more modest individuals.

Mathematical reserves were calculated using the constructed prospective table. Overall, the reserves obtained from the constructed table are more cautious than those obtained from the Swiss Life model or the regulatory tables. This study has some limitations, with the main one probably being that the annuity amount received by the individuals in the portfolio studied here is not necessarily representative of their standard of living. indeed it is possible for example that policyholders with low annuity amounts in the portfolio studied here have other annuity contracts in other companies with much higher annuities. It would then be necessary to use the sum of the amounts on the different contracts to be perfectly exhaustive, something which is unfortunately impossible for us since this information is inaccessible.

Table des matières

Liste des Figures

Liste des Tables

Introduction

Le risque longévité pour un assureur disposant d'un portefeuille de rentiers consiste en une sous-estimation de l'espérance de vie de ses assurés. Une espérance de vie plus élevée par rapport aux prévisions de l'assureur aboutit à des pertes financières pour ce dernier, vu qu'il doit payer des primes plus longtemps que ce qui était prévu, et donc en plus grande quantité que ce qui avait été provisionné au départ. La prise en compte de ce risque longévité est donc un enjeu important pour les assureurs couvrant ce type de garanties. Ceci d'autant plus que de nombreuses études tendent à montrer que l'espérance de vie humaine devrait continuer d'augmenter pendant plusieurs années encore, notamment dans les pays développés (Kontis et al. 2017). Pour tenir compte de cette espérance de vie qui évolue, des tables de mortalité prospectives sont utilisées par les assureurs pour le calcul des provisions. Leur aspect prospectif permet d'imprimer une dynamique de hausse quasi linéaire de l'espérance de vie chaque année (Debonneuil, Loisel, et Planchet 2018).

Ces améliorations de l'espérance de vie ne se font cependant pas au même rythme pour l'ensemble de la population. Il existe une certaine hétérogénéité notamment due au niveau de vie qui fait que les individus les plus aisés vivent en moyenne plus longtemps que les plus modestes (INSEE 2018). Dans le cas d'un portefeuille de rentiers, ce phénomène se traduit par le fait que les rentiers les plus riches aient une espérance de vie plus élevée que les moins riches.

L'objectif principal de ce mémoire sera ainsi pour un portefeuille de rentiers de l'assureur Swiss Life, de construire une table de mortalité prospective (évaluer des probabilités de décès) qui tienne compte (en plus de l'âge et du sexe) du niveau de vie des assurés. Le niveau de vie étant approché ici par le montant de la rente perçue. Le cadre de modélisation retenu pour mener à bien cet objectif est celui les modèles additifs généralisés, il sera challengé par deux approches alternatives.

Le mémoire va s'organiser en six chapitres. Le premier sera dédié à la mise en contexte de l'étude et à la présentation des concepts de tables de mortalité et de rentes. Le deuxième chapitre présentera quelques concepts probabilistes de base indispensables à la compréhension des principes de fonctionnement des modèles qui seront calibrés dans ce mémoire. Dans le troisième chapitre, seront présentées les données utilisées, quelques statistiques descriptives ainsi que le positionnement de la mortalité observée dans le portefeuille par rapport à la référence réglementaire. Le chapitre quatre permettra de présenter le cadre théorique des modèles additifs généralisés qui seront utilisés pour la construction de la table prospective. Le chapitre cinq sera dédié à la construction de

la table proprement dite, à travers la modélisation de la mortalité du portefeuille par les modèles additifs généralisés. Cette modélisation sera challengée par deux approches alternatives par les modèles de Cox et les forêts aléatoires de survie. Enfin, dans le chapitre six, des provisions seront calculées en utilisant les probabilités de décès fournies par la table construite. L'impact de la prise en compte du montant de rente sur le calcul de ces provisions sera isolé.

Chapitre 1

Mise en contexte

Il sera question dans ce chapitre de préciser le contexte de l'étude qui sera menée tout au long de ce mémoire. Ceci se fera par la présentation et la définition de quelques concepts clés utiles pour la compréhension du sujet.

1.1 Présentation générale

Les assureurs Vie, particulièrement ceux qui proposent des produits avec prestations garanties tant qu'un assuré ou un bénéficiaire tierce est en vie (typiquement des **rentes viagères**) sont exposés au **risque longévité**. Le risque longévité est un risque de long terme qui correspond au risque financier associé au fait que les individus vivent en moyenne significativement plus longtemps que prévu. Plus concrètement, le risque longévité c'est le risque pour un assureur sur un portefeuille donné, d'observer une augmentation non anticipée de l'espérance de vie aux grands âges.

La durée de vie augmente de génération en génération (INSEE 2022b) : les individus nés en 1900 n'ont vécu en moyenne que 56 ans pour les femmes et 48 ans pour les hommes, quand ceux âgés de 20 ans en 2022 vivraient en moyenne 91 ans pour les femmes et 88 ans pour les hommes. Mieux encore, l'espérance de vie à la naissance des individus nés en 2022 serait de 93 ans pour les femmes et de 90 ans pour les hommes.

Un aspect sous-jacent du risque de longévité est l'hétérogénéité au sein d'une population en terme de dynamique de longévité ou de survie. En effet les améliorations en terme d'espérance de vie peuvent ne pas être de la même ampleur dans toutes les couches de la population, par exemple il est constamment observé une longévité plus élevée chez les femmes que chez les hommes. Cette différence dans les améliorations de mortalité au sein d'une population est désignée sous le nom de **risque de base**, et au-delà du genre elle peut être due à des réalités économiques, sociales ou encore géographiques différentes.

Le risque de base peut être décomposé en deux composantes :

- Le **niveau** : il s'agit de la différence actuelle en terme d'espérance de vie au sein d'une population hétérogène ;

- La **tendance** : c'est l'évolution de ces différences dans le temps, une approche de modélisation de cette dernière est présentée dans Kabore (2017).

L'étude menée ici aura pour but principal de prendre en compte l'effet du montant de rente sur la mortalité globale au sein d'un portefeuille de rentiers. On s'intéressera donc au risque de base lié à des considérations économiques, et plus au niveau (différences de survie actuelles) plutôt qu'à une éventuelle évolution dans le temps.

Les conditions économiques peuvent être une source importante de risque de base au sein d'une population. En effet, l'espérance de vie varie en fonction du revenu, plus on est aisé et plus on vit longtemps. En France, parmi les 5% d'individus les plus aisés, l'espérance de vie à la naissance des hommes est de $84,4$ ans contre $71,7$ ans parmi les 5% les plus pauvres, soit presque 13 ans d'écart (INSEE 2018). Chez les femmes cet écart est plus faible mais toujours très élevé, 8 ans d'espérance de vie séparent les plus aisées des plus pauvres.

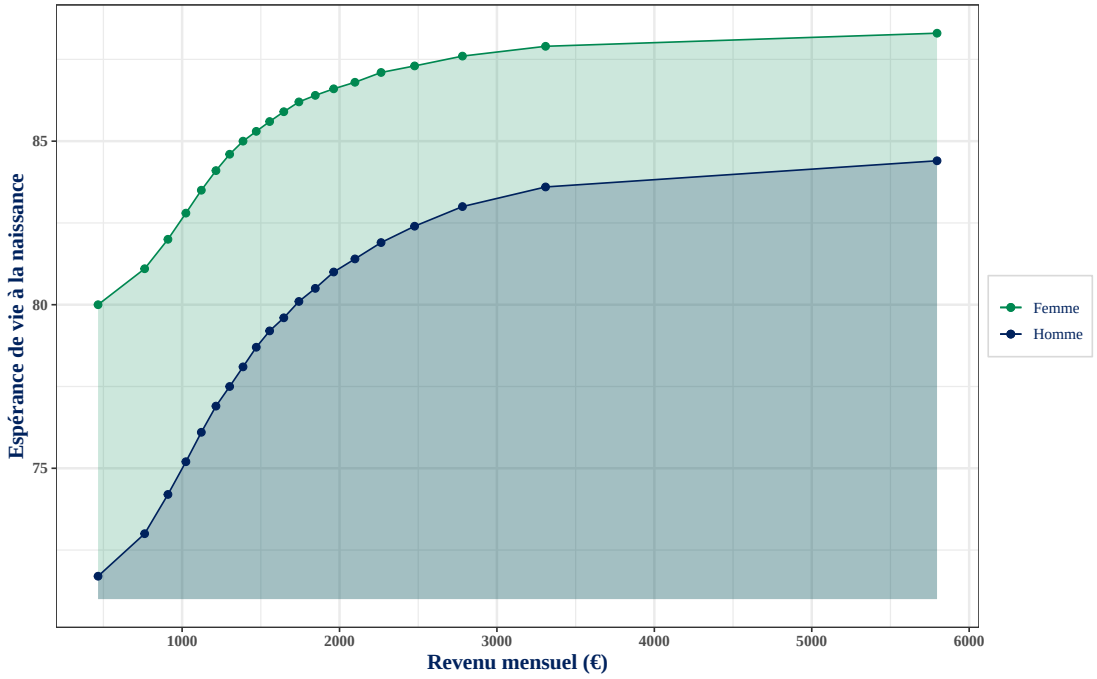

Figure 1.1 : Espérance de vie à la naissance selon le niveau de vie (Source: INSEE)

Sur la Figure 1.1, pour un revenu aux alentours de 1000€ par mois, 100€ supplémentaires sont associés à $0,9$ an d'espérance de vie en plus chez les hommes et $0,7$ an chez les femmes. Aux alentours de 2000€ par mois, cet écart n'est plus que de $0,3$ an chez les hommes et $0,2$ an chez les femmes (INSEE 2018).

L'espérance de vie augmente donc de moins en moins rapidement avec le niveau de vie. Le gain d'espérance de vie associé à une légère augmentation du niveau de revenu

s'atténue à mesure que le niveau de vie de base est élevé.

Cette différence de longévité entre les individus les plus aisés et ceux qui le sont le moins peut s'expliquer par divers éléments :

Tout d'abord des inégalités au niveau de la santé. Des difficultés financières peuvent limiter l'accès aux soins pour les individus les plus démunis. D'après l'enquête santé et protection sociale 2014 (ESPS), 11% des adultes parmi les 20% les plus modestes disent avoir renoncé pour des raisons financières à consulter un médecin au cours des 12 derniers mois, contre 1% des adultes parmi les 20% les plus aisés. D'autres éléments liés au niveau de revenu comme la catégorie socioprofessionnelle ou encore la région de résidence ont aussi une influence sur la santé des individus. Les cadres par exemple sont moins soumis aux risques professionnels (accidents du travail, exposition à des produits toxiques) que des ouvriers.

1.2 Problématique et objectifs

Considérons un assureur Vie de la place, disposant d'un portefeuille d'assurés pour lesquels il est engagé pour le versement de rentes viagères. L'assureur doit donc évaluer ses engagements (provisions), pour ce faire il lui faut estimer la durée de vie résiduelle des individus qui constituent son portefeuille. A cette fin, les assureurs ont généralement recours aux tables de mortalité réglementaires $TGF/TGH05$ ou encore à des tables de mortalité d'expériences. Mais dans la plupart des cas, ces tables ne sont déclinées que par âge et par sexe, et ne tiennent donc pas compte de la différence de longévité entre individus due au niveau de vie. Dans ce cas, si on considère le montant de la rente perçue par un individu comme indicateur de son niveau de vie, les rentiers les plus riches sont ceux qui vivraient le plus longtemps. La non prise en compte par l'assureur du niveau de vie dans l'estimation de la durée de vie résiduelle des individus peut avoir des conséquences financières non négligeables. En effet, le pire des scénarios pour l'assureur est de sous estimer la longévité des individus les plus riches, et donc d'être amené à payer plus longtemps que prévu les montants de rente les plus élevés.

L'objectif principal de l'étude menée dans ce mémoire sera ainsi de tenir compte du niveau de vie (au moyen du montant de la rente perçue) dans la construction d'une table de mortalité d'expérience prospective (calcul des probabilités de décès) pour des assurés d'un portefeuille de rentiers.

Deux principales approches sont généralement utilisées pour la construction des tables prospectives. Une approche intrinsèque basée uniquement sur l'utilisation des données d'expérience, et une approche par référence externe basée sur le positionnement de la mortalité du groupe étudié par rapport à une mortalité de référence. C'est cette dernière approche qui est la plus couramment utilisée en actuariat, car elle est particulièrement adaptée lorsque l'on ne dispose pas de données en grandes quantités, ce qui est généralement le cas pour des portefeuilles individuels. De très nombreux mémoires

d'actuariat ont déjà traité la question de la construction des tables de mortalité prospectives pour des portefeuilles de rentiers par l'approche référence externe. On a ainsi Fall (2019) qui a élaboré une table prospective pour un portefeuille d'épargne retraite du groupe Malakoff, Yikmis (2020) pour un portefeuille de rente viagère du groupe Generali, ou encore Guez (2018) sur des données de rentiers de Groupama Gan Vie. Il est à noter tout de même des approches plus originales comme Durieux et Samba (2013) qui ont exploré les apports de la théorie des copules dans la modélisation de la mortalité, ou encore Bastien (2020) qui a utilisé la théorie de la crédibilité pour construire une table de mortalité prospective.

Des études qui prennent en compte des caractéristiques autres que l'âge et le sexe dans les modèles de mortalité sont plus rares. Une problématique assez similaire à la notre a tout de même déjà été traitée dans Ziegelmeyer (2015). L'auteur avait exploré trois méthodes pour tenir compte du montant de rente dans l'évaluation des probabilités de décès, dont deux consistaient en une segmentation de la population assurée suivant le montant de rente de manière à obtenir des sous populations homogènes du point de vue de la mortalité. Il avait ensuite construit des tables différentes pour chacune de ces sous populations.

L'approche de modélisation qui sera utilisée dans ce mémoire sera celle des **modèles additifs généralisés**, elle ne nécessite pas une segmentation a priori de la population des assurés. Ce type de modèles est encore assez peu utilisé en Actuariat. Néanmoins, Côté (2016) présente une application intéressante de ces derniers à de la tarification en assurance automobile.

1.3 Les tables de mortalité

1.3.1 Définition et historique

Une **table de mortalité** est une construction qui détermine les futures évolutions démographiques d'une population, elle permet de prévoir les probabilités de décès des individus selon l'âge et le sexe généralement.

L'invention de la table de mortalité a joué un rôle majeur dans l'histoire de la démographie et dans les progrès du calcul des probabilités. Jusqu'au 17^{ime} siècle, la mort était considérée soit comme une punition du ciel pour des péchés individuels ou collectifs, soit comme la manifestation d'un hasard pur (Dupâquier et Dupâquier 1985). L'idée que la mortalité obéisse à des lois impliquait une grande évolution conceptuelle. Jusqu'à la fin du Moyen Âge, la conception chrétienne traditionnelle de la mort interdit la spéculation à ce sujet, et donc l'idée qu'il puisse y avoir des lois (autres que la loi divine) qui puissent l'expliquer.

L'origine politique des tables de mortalité est anglaise, mais la première apparition à des fins économiques est observée aux Pays-Bas avec Johan de Witt en 1671. Les premiers

résultats statistiques sérieux furent les tables de Northampton conçues par Price (1780) à partir de registres paroissiaux. Mais la première table qui est devenue la norme des compagnies d'assurance britanniques et américaines pendant près d'un siècle est la table de Carlisle, construite par Milne (1815).

1.3.2 Types de table de mortalité

Comme indiqué dans Planchet (2022), les tables de mortalité peuvent être classées suivant des considérations **réglementaires** et des considérations **techniques**.

Les considérations techniques renvoient à la nature même de la table et à la manière dont elle est construite. Suivant ces considérations donc, on distingue les tables **périodiques** ou tables du **moment** et les tables **prospectives**.

Les considérations réglementaires quant à elles font référence au cadre d'utilisation de la table. On distingue alors suivant ces considérations les tables de mortalité **réglementaires** et les tables de mortalité d'**expérience**. Tous ces types de tables seront présentés par la suite.

Les tables périodiques ou tables du moment

Les tables de mortalité périodiques permettent de caractériser la mortalité d'une population à un instant précis, toutes générations confondues. Elles supposent une certaine stabilité des décès dans le futur, elles sont généralement construites sur la base de 3 années afin de couvrir les aléas qui pourraient subsister sur une seule année. Leur validité est limitée car la mortalité évolue dans le temps.

Les tables prospectives ou tables générationnelles

Les tables de mortalité prospectives intègrent l'aspect dynamique de la mortalité dans l'évaluation des probabilités de décès. Elles sont établies à partir d'une génération réelle par observation du niveau de mortalité réel en fonction de l'année de naissance. Ces tables prospectives sont a priori plus représentatives du niveau réel de la mortalité que les tables du moment, elles sont aussi plus contraignantes au niveau de leur construction vu qu'elles impliquent en théorie le suivi d'une génération jusqu'à son extinction complète. En pratique, des modèles de durée prospectifs permettant d'extrapoler la mortalité future sont utilisés pour leur construction.

Les tables de mortalité réglementaires

Les tables de mortalité réglementaires sont des tables construites sur la base des données de la population française et/ou des données de plusieurs portefeuilles d'assurés regroupés auprès des acteurs majeurs du secteur de l'assurance en France.

On distingue deux grands groupes de tables réglementaires : les tables réglementaires du moment et les tables réglementaires prospectives.

- **Les tables réglementaires du moment**

Il s'agit des tables **TH** et **TF** $00-02$ pour les assurances en cas de décès. Homologuées par l'arrêté du 20 Décembre 2005, ces tables ont été établies à partir des données de l'INSEE issues d'observations réalisées entre 2000 et 2002 et sont applicables aux contrats d'assurance vie souscrits depuis le 1^{er} Juillet 1993. La table **TF** décrit la mortalité féminine et la table **TH** la mortalité masculine. Ces tables sont utilisées pour des assurances en cas de décès. Ayant été établies à partir des données de la population générale qui est moins riche et moins bien couverte que la population assurée, elles prédisent plus de décès et sont donc prudentes. Ces tables peuvent aussi être utilisées avec des décalages d'âge (pour les rendre prudentes en terme de longévité) pour les assurances en cas de vie, à l'exclusion des rentes.

- **Les tables réglementaires prospectives**

Il s'agit des tables générationnelles **TGH** et **TGF** 05. Elles viennent répondre à la nécessité de prendre en compte les évolutions futures de la mortalité dans le calcul des provisions des rentes viagères. Ces tables ont été obtenues sur la base de la mortalité de la population des bénéficiaires de contrats de rente entre 1993 et 2005, et l'aspect prospectif a été introduit grâce aux données sur la population générale française de 1962 à 2000. Ces tables servent depuis le 1^{er} Janvier 2007 à la tarification et au provisionnement des contrats de rentes viagères immédiates ou différées.

Les tables d'expérience

Dans certains cas, lorsqu'un assureur dispose d'un portefeuille d'une taille assez conséquente, il peut souhaiter utiliser l'expérience observée sur ce portefeuille pour construire une table de mortalité qui lui est propre. Cette table peut ensuite être utilisée en lieu et place des tables réglementaires : Il s'agit d'une table de mortalité d'**expérience**. Cette démarche est rendue possible par l'article $A.132-18$ du code des assurances, et permettrait de mieux cerner les comportements de la population assurée qui seraient significativement différents de ceux de la population des tables réglementaires.

En pratique, la mise en place et l'autorisation d'utilisation d'une table d'expérience comporte 3 étapes :

- La construction de la table
- La certification initiale
- Le suivi destiné à assurer la pérennité du droit d'utilisation de la table.

Le suivi des tables d'expérience doit être annuel, à défaut les tables cessent d'être valides 2 années après leur certification initiale. Au global la validité de ces tables de mortalité est limitée à 5 ans. La certification ne concerne pas une table dans l'absolue, mais une table utilisée pour un contrat ou un groupe de contrats particuliers.

1.4 Les rentes viagères en assurance

1.4.1 Généralités

La **rente viagère** est un revenu à vie, issu de la transformation d'un capital en pension par un assureur. Elle offre à l'assuré ou à tout autre bénéficiaire la certitude de bénéficier d'une série de versements périodiques jusqu'à sa mort. La rente est revalorisée chaque année par l'assureur en fonction des gains techniques et financiers éventuellement réalisés (Vial 2012). Du point de vue de l'assuré, la rente viagère est un excellent moyen de se couvrir contre le risque de longévité, avec en contrepartie une immobilisation et donc une certaine renonciation à un capital.

Il existe plusieurs types de rentes viagères :

- Les rentes viagères simples obligatoires versées dans le cadre des produits de retraite pure
- Les rentes viagères immédiates : Elles donnent droit au versement d'une rente garantie à vie dès le dépôt du capital constitutif de la rente
- Les rentes viagères différées : Pour ce type de rente, chaque cotisation versée donne droit à une fraction de rente viagère dont le montant est connu immédiatement. Au fur et à mesure des versements, l'assuré cumule des droits estimés en euros
- Les rentes viagères issues de la conversion à l'échéance du capital placé sur un contrat d'assurance vie. Tout contrat peut ainsi prévoir que le bénéficiaire pourra opter au terme du contrat à une sortie en rente.

1.4.2 Transformation d'un capital en rente viagère

Pour transformer un capital en rente viagère, les assureurs appliquent un taux de conversion au capital à transformer, ce taux est fonction de l'espérance de vie. Le montant de la rente viagère à percevoir dépend en réalité de :

- Le montant de capital immobilisé
- La date de naissance et le sexe de l'assuré : Ce sont là deux éléments déterminants, le taux de conversion du capital en rente est d'autant plus faible que le rentier est jeune. Les hommes ayant une espérance de vie plus courte, ils devraient percevoir en théorie des montants de rente plus élevés que les femmes, mais la directive européenne unisexe de 2012 a interdit l'usage du sexe comme variable discriminante pour ce type de produit.
- La table de mortalité en vigueur au moment de la transformation
- Le taux technique retenu : Il s'agit d'une anticipation sur les produits financiers futurs qui sont susceptibles de venir revaloriser la rente. Un taux technique à zéro revient à minimiser les arrérages de départ en négligeant tout gain futur potentiel, mais avec une progression future rapide de ces arrérages .
- Les frais de gestion appliqués sur les arrérages de la rente
- Des éventuelles options de rente et garanties choisies par l'assuré pour limiter l'aliénation de son capital : Le niveau de la rente dépend des options et garanties que l'assuré est susceptible de choisir pour trouver des parades à l'immobilisation de son capital.
  - La réversion : C'est l'option la plus courante, ici la rente est versée à l'assuré principal jusqu'à son décès . Ensuite, un pourcentage (taux de réversion) de la rente de départ continue à être versée à un assuré secondaire (généralement le conjoint) tant que celui-ci est en vie.
  - Les annuités garanties : Elles permettent à l'assuré de déterminer un nombre d'années pendant laquelle la rente est acquise quoiqu'il arrive . En cas de décès de l'assuré pendant la période garantie , les annuités restantes sont attribuées au bénéficiaire désigné dans le contrat. Si l'assuré est toujours vivant à l'issue de la période garantie, la rente viagère continue à lui être versée jusqu'à son décès.
  - Les annuités certaines : cette option vient casser le caractère viager du versement de la rente. Ici, une date limite de versement des rentes est fixée. La rente n'est plus versée à vie, mais prévue pour une durée limitée. Contrairement aux annuités garanties, au-delà de la date limite fixée, plus aucune rente n'est versée même si l'assuré est encore en vie. En cas de décès

de l'assuré avant la fin de la période d'annuité certaine, la rente restante est versée à un bénéficiaire mentionné dans le contrat.

1.5 Rentes éducation

La rente éducation est un type de contrat d'assurance prévoyance qui permet en cas de décès ou d'invalidité totale du parent ayant souscrit l'assurance, de verser une rente aux enfants désignés comme bénéficiaires. Elle a pour but principal de financer les études des enfants et peut être versée jusqu'à 18 voire 28 ans selon le contrat.

En général, il existe deux principales conditions au versement d'une rente éducation :

- La rente est versée jusqu'à un certain age, quelle que soit la situation de l'enfant (18 ans par exemple)
- Le versement de la rente est conditionné par la poursuite des études. A titre d'exemple, la rente peut être versée aux enfants jusqu'à leurs 26 ans, tant qu'ils poursuivent leurs études.

Ces deux conditions sont généralement jumelées, c'est-à-dire que la rente est versée sans condition jusqu'à 18 ans, et puis jusqu'à 26 ans sous condition de la poursuite des études. Dans certains cas, la rente peut être versée à vie pour les enfants handicapés.

Chapitre 2

Rappels actuariels

Nous présenterons ici des concepts probabilistes de base nécessaires à la compréhension du principe de fonctionnement des différents modèles qui seront calibrés plus tard dans ce mémoire.

2.1 Notations des modèles de durées

Il est de bon usage de commencer par préciser ce que l'on entend ici par **durée**.

Une **durée** par définition est une valeur positive ou nulle modélisée par une variable aléatoire (qu'on notera T ici) à valeurs dans $\mathbb{R}^+$. On peut ainsi parler de durée avant une ruine, durée d'un arrêt de travail, durée entre deux sinistres et bien d'autres. Mais ce qui nous intéressera dans ce mémoire c'est bien évidement la **durée de vie humaine.**

Les observations de durées sont souvent données en semaines, mois ou années et peuvent ainsi être considérées comme des données discrètes, mais il ne s'agit là que d'approximations. En pratique, la variable aléatoire T est considérée comme absolument continue, c'est-à-dire comme admettant une densité de probabilité.

2.1.1 La fonction de survie

La durée de vie T en tant que variable aléatoire est caractérisée par sa fonction de répartition $F(t) = \mathbb{P}[T \leq t]$ pour $t \geq 0$. La **fonction de survie** se définit comme le complémentaire de cette quantité : $S(t) = \mathbb{P}[T > t] = 1 - F(t)$, avec $S(0) = 1$, $\lim_{t\to+\infty} S(t) = 0$. $S(t)$ représente ainsi la probabilité de survivre jusqu'à un instant t .

2.1.2 La densité de probabilité

La densité de probabilité associée à T est définie comme suit :

$$f(t) = \frac{1}{h} \lim_{h \to 0} \mathbb{P}[t \leq T < t + h] = \frac{d}{dt} F(t) = -\frac{d}{dt} S(t) \tag{2.1}$$

2.1.3 La force de mortalité

La force de mortalité encore appelée **fonction de hasard** est définie par :

$$h(t) = \frac{f(t)}{S(t)} = -\frac{S^{'}(t)}{S(t)} = -\frac{d}{dt} \log(S(t)) \tag{2.2}$$

Cette fonction de hasard détermine entièrement la loi de T, en effet :

$$S(t) = \exp(-\int_0^s h(s)ds) \tag{2.3}$$

On a également :

$$h(t) = \frac{F^{'}(t)}{S(t)} = \lim_{u \to 0} \frac{\mathbb{P}[t < T \leq t + u]}{uS(t)} = \lim_{u \to 0} \frac{\mathbb{P}[T \leq t + u | T > t]}{u}$$

Il vient alors que :

$$\mathbb{P}[T \leq t + u | T > t] = h(t)u + o(u). \tag{2.4}$$

$h(t)$ peut alors être interprété dans un contexte de mortalité comme un taux instantané de décès .

Il s'agit d'un indicateur de choix pour étudier l'impact de variables explicatives comme l'âge ou le sexe sur la mortalité.

On définit la fonction de hasard cumulé par :

$$H(t) = \int_0^t h(s)ds \tag{2.5}$$

L'objet des modèles de durées est ainsi de caractériser la variable aléatoire de durée T, ceci par la modélisation de sa fonction de survie, de sa densité ou encore de sa force de mortalité.

Il est courant cependant de considérer des durées conditionnellement au fait qu'une durée se soit déjà écoulée. Ainsi pour une durée de vie, on notera T_x la durée de vie d'un individu conditionnellement au fait qu'il soit vivant à l'age x . Il est alors possible de définir la probabilité de survie conditionnelle entre x et $x+t$ par :

$$ {}_tp_x = \mathbb{P}[T_x > t] = \mathbb{P}[T > x+t | T > x] = \frac{S(x+t)}{S(x)} \tag{2.6} $$

On définit par la même occasion le taux de mortalité entre x et $x+t$:

$$ {}_tq_x = 1 - {}_t\,p_x = \frac{S(x) - S(x+t)}{S(x)} \tag{2.7} $$

2.2 Le maximum de vraisemblance

2.2.1 Principe

On se place dans le cadre d'observation d'un échantillon $(T_1, ..., T_n)$ de n variables aléatoires indépendantes et identiquement distribuées de même loi que notre variable de durée T . On supposera ici que la famille de loi de T est connue, mais qu'un paramètre (ou vecteur de paramètres) θ est inconnu et doit être estimé. De nombreuses méthodes existent pour l'estimation de ce paramètre, la plus répandue est sans doute la méthode du **maximum de vraisemblance**.

La fonction de vraisemblance est donnée par :

$$ L(t_1, ..., t_n; \theta) = \prod_{i=1}^{n} f_\theta(t_i) \tag{2.8} $$

Avec : $(t_1, ..., t_n)$ une réalisation de $(T_1, ..., T_n)$.

Le principe de la méthode du maximum de vraisemblance est de trouver le paramètre θ qui maximise cette expression de la vraisemblance et par conséquent rend les observations $(t_1, ..., t_n)$ les plus vraisemblables, les plus probables au regard de la distribution considérée.

Dans la pratique il est souvent plus commode de maximiser la **log-vraisemblance**, le logarithme de la vraisemblance :

$$ l(t_1, ..., t_n; \theta) = \log(L(t_1, ..., t_n; \theta)) = \sum_{i=1}^{n} \log(f_\theta(t_i)) \tag{2.9} $$

La maximisation de cette logvraisemblance permet dans les cas les plus simples d'obtenir une expression explicite du paramètre θ. La plupart du temps par contre il faut utiliser des méthodes numériques d'optimisation comme la méthode de Newton-Raphson.

2.2.2 Prise en compte de la censure et de la troncature

Dans l'idéal pour des études de mortalité, il devrait être possible de s'appuyer sur l'observation complète des durées de vie des individus depuis leur naissance jusqu'à leur décès. Mais dans la pratique, l'observation des durées de vie est souvent partielle et sélective du fait des phénomènes de **censure** et de **troncature (Biessy 2022)**.

La censure

Soient U et C deux durées de vie (ou variables de durées plus généralement). U est dite **censurée** par C lorsque l'on n'observe pas directement U mais plutôt la variable aléatoire $T = \max(U, C)$ dans le cas de la censure à gauche et $T = \min(U, C)$ pour la censure à droite. L 'observation de la variable T est de plus complétée par celle d'une indicatrice $D = 1_{(U \geq C)}$ dans le cas de la censure à gauche et $D = 1_{(U < C)}$ dans le cas de la censure à droite.

La censure à gauche suppose la présence d'individus pour lesquels le décès est survenu mais dont l'age de décès est inconnu. Ce cas de figure est assez rare néanmoins dans les études de mortalité.

La censure à droite quant à elle est bien plus présente dans la réalité, on en distingue deux types :

- **La censure fixe** : C'est lorsque la variable de censure est déterministe, la censure est la même et connue a priori pour tous les individus. La date de fin d'observation des données pour une étude de mortalité constitue une censure fixe.
- **La censure aléatoire** : C'est lorsque la variable de censure est de nature aléatoire, et peut donc ainsi être différente pour tous les individus observés. Pour chaque individu la valeur de la censure ne sera connue que si il est effectivement censuré. En assurance la résiliation d'un contrat constitue une censure aléatoire. La date de résiliation n'est pas connue a priori et ne sera connue que si la résiliation est survenue avant le décès ou l'arrêt des données.

Dans ce cas de censure aléatoire , il est nécessaire de faire l'hypothèse de **censure non informative**. Elle postule l'indépendance conditionnellement aux caractéristiques des individus entre la survenance de l'évènement d'intérêt et la censure.

La troncature

Soient U et Q deux durées de vie. U est dite **tronquée** par Q lorsque la variable U n'est pas observée si $U < Q$ dans le cas de la troncature à gauche ou si $U \geq Q$ dans le cas de la troncature à droite. On observe une troncature à gauche dès lors que l'observation des individus ne démarre pas systématiquement à leur naissance, et une troncature à droite lorsque par exemple on ne dispose pour un individu donné des informations le concernant que lorsqu'il est décédé.

Classiquement dans les études de mortalité, et de façon plus spécifique dans le cas de l'étude menée dans le cadre de ce mémoire, l'observation est confrontée à :

- **La troncature à gauche** : Elle est liée au fait que les individus ne commencent à être observés dans le meilleur des cas qu'à leur age d'entrée (souscription) dans le portefeuille.
- **La censure à droite** : Elle est liée au caractère limité de l'observation dans le temps. Pour les individus encore en vie à la date d'extraction des données par exemple, le décès n'est pas observé.

Pour tenir compte ces deux phénomènes, la fonction de vraisemblance de l'Équation 2.8 se réécrit de la façon suivante :

$$L(t_1, ..., t_n; \theta) = \prod_{i=1}^{n} (f_\theta(t_i) S_C(t_i))^{\delta_i} (f_C(t_i) S_\theta(t_i))^{1-\delta_i} \tag{2.10}$$

Avec :

- S_C et f_C les fonctions de survie et de densité de la variable de censure C .
- δ_i l'indicatrice de censure pour l'individu i

En considérant l'hypothèse de **censure non informative** définie plus haut, il est possible de simplifier cette expression en éliminant la densité et la fonction de survie de C , vu que cette dernière n'apporte aucune information sur le paramètre θ.

$$L(t_1, ..., t_n; \theta) = c \prod_{i=1}^{n} (f_\theta(t_i))^{\delta_i} (S_\theta(t_i))^{1-\delta_i} \tag{2.11}$$

La constante c contient tout ce qui ne dépend pas de θ et ne joue donc aucun rôle dans la maximisation de la vraisemblance ou de la logvraisemblance.

2.3 Comparaisons de modèles

En analyse statistique il existe deux grandes approches pour comparer des modèles (Biessy 2022) :

- L'approche par les **tests d'hypothèses** qui permet de justifier l'utilisation d'un modèle plus complexe. Cette approche n'est valable cependant que pour des modèles imbriqués.
- L'approche par les **critères statistiques**, les plus utilisées sont le AIC et le BIC . Cette approche peut être utilisée pour comparer des modèles quelconques non nécessairement imbriqués, à condition qu'ils soient appliqués aux mêmes données.

$$AIC = dev + 2 \times edf;\ BIC = dev + \log(n) \times edf$$

Avec :

- dev : la déviance du modèle, $dev = 2[l_{max} - l]$. l représente la logvraisemblance du modèle, et l_{max} la logvraisemblance du modèle saturé (modèle avec autant de paramètres que d'observations);
- n : le nombre d'observations;
- edf : le nombre de degrés de liberté du modèle, il représente la dimension effective de l'espace des prédicteurs linéaires. Il s'agit du nombre de paramètres strictement indépendants dont dépend le modèle. En présence de contraintes absolues (relation linéaire entre paramètres) ou relatives (une pénalisation des écarts entre paramètres), ce nombre sera plus faible que le nombre total de paramètres.

Plus les valeurs de ces critères sont faibles et meilleurs sont les modèles correspondants. Le critère BIC en plus d'être défini dans un cadre bayésien, est plus stricte que le AIC en ce sens qu'il pénalise d'autant plus un modèle que les données sur lesquelles il a été calibré sont nombreuses. Pour ces raisons le critère AIC sera privilégié ici.

Un peu plus loin dans ce mémoire, il sera présenté une approche ad hoc permettant de comparer des modèles calibrés sur des données de formats différents.

Chapitre 3

Présentations des données et analyses descriptives

Dans ce chapitre, il sera présenté le jeu de données qui a servi de base à toutes les analyses effectuées dans ce mémoire. La population étudiée sera ensuite clairement définie, et des statistiques descriptives la concernant seront présentées. Enfin on terminera par la comparaison de la mortalité observée dans notre portefeuille avec celle prévue par les tables réglementaires $TGF/TGH05$.

3.1 Présentation des bases de données

Les données utilisées dans ce mémoire proviennent d'un portefeuille de rentes viagères de l'assureur **Swiss Life**. Il s'agit d'un groupe fondé en Suisse en 1857, il est en Europe l'un des principaux fournisseurs de produits d'assurance Vie, prévoyance et retraite. Swiss Life intervient principalement sur les marchés suisse, français et allemand. En France, Swiss Life est un acteur majeur sur le marché de l'assurance patrimoniale et est depuis plusieurs années le leader en prévoyance et retraite pour les collaborateurs des grandes firmes multinationales. Le groupe affichait au 1^{er} Janvier 2023 un chiffre d'affaires de $19,6$ milliards de francs suisse.

Les données fournies se présentent sous la forme de deux bases de données distinctes, une base de **stocks** et une base de **flux**.

- **La base de stock** : Elle présente les différentes évaluations faites sur les contrats entre Janvier 2014 et Janvier 2023. Elle contient 12 variables pour plus de 410 000 lignes d'observations. Chaque ligne ici représentant une année d'observation pour un contrat donné, chaque contrat est lié à une **tête principale** et éventuellement une **tête secondaire** (voire plusieurs) à laquelle une proportion (taux de réversion) de la rente est versée si elle survit plus longtemps que la tête principale. Chaque contrat apparaît ainsi dans cette base autant de fois que d'années durant lesquelles il a été observé. Pour un contrat donné, les informations principales suivantes sont renseignées :

  - La date de souscription
  - La date de naissance (tête principale et tête secondaire éventuellement)
  - La date d'évaluation
  - Le sexe (tête principale et tête secondaire éventuellement)
  - Le montant de la rente annuelle
  - Le code postal (tête principale et tête secondaire éventuellement)
  - Le taux de réversion
  - Le type de rente
  - La pays de résidence
  - Le montant des provisions mathématiques associées au contrat l'année en cours.

- **La base de flux** : Elle renseigne les différents décès survenus au sein de la population de rentiers au cours de la période d'évaluation (Janvier 2014 à Avril 2023), ces décès sont dénombrés à plus de 8 700.

Il est à noter que ces bases dans leurs formes initiales présentaient quelques anomalies mineures, des dates de naissances différentes attribuées à un même assuré, des décès différents associés à un même individu, et enfin des décès (70) recensés dans la base de flux qui ne correspondaient à aucun assuré dans la base de stock. A l'exception de cette dernière anomalie mentionnée, les autres ont été résolues par une seconde extraction de données par l'assureur propriétaire du portefeuille. Les 70 décès non identifiés ont tout simplement été exclus du champ des analyses. Certaines variables présentaient des valeurs manquantes (code postal, provisions mathématiques notamment), mais les variables clés qui allaient servir aux analyses (date de souscription, date de naissance, sexe, montant de rente, date de décès) étaient parfaitement renseignées.

3.2 Population étudiée

Les analyses qui seront effectuées par la suite ne seront menées que sur une partie des individus présents dans nos deux bases de données.

3.2.1 Têtes secondaires et têtes principales

Déjà, un premier questionnement est suscité au niveau de la prise en compte ou non des têtes secondaires. Dans l'idéal, il faudrait procéder à des analyses séparées pour les têtes secondaires et pour les têtes principales. Mais ici, ce n'est qu'une proportion marginale des individus observés ($2,8\%$) qui sont des têtes secondaires, pour seulement 4% des décès observés. Il n'est donc pas envisageable de mener une étude uniquement sur ces têtes secondaires.

Dans le même temps, il n'est non plus envisageable de mener une analyse incluant à la fois les têtes secondaires et les têtes principales, car il y a un risque de travailler sur deux sous-populations hétérogènes. La seconde tête qui est généralement le/la conjoint(e) de la tête principale peut être victime du *syndrome du cœur brisé*, qui stipule que, après le décès du premier conjoint, le second est exposé à un risque de décès plus élevé (Gourieroux et Lu 2015).

Une dernière difficulté avec les têtes secondaires est que les informations relatives à leurs décès ne sont pas enregistrées par l'assureur si la tête principale est encore en vie. Il y a donc un risque de sous estimation de la mortalité du fait de l'incomplétude des informations les concernant.

Nous partons donc sur une exclusion des têtes secondaires du champ des analyses à venir.

3.2.2 Résidents français

Parmi les assurés présents dans les bases de données, il y a une faible proportion (2%) de résidents de pays étrangers. Ces individus seront exclus du champ des analyses par la suite pour ne conserver que les résidents français.

3.2.3 Type de rente

Dans le portefeuille à disposition, il est recensé une immense majorité de rentes viagères (97%), soit simple, soit avec annuités garanties, et une minorité de rentes à annuités certaines et de rentes éducations (3%). Dans le cadre de cette étude, seules les rentes viagères seront retenues dans le champ d'analyse. A noter en ce qui concerne la nature des produits, que le portefeuille est constitué aux deux tiers de contrats collectifs (principalement des articles 83[1]) et d'un tiers de contrats individuels (principalement des Madelin[2]). Cette information est fournie par l'assureur détenteur du portefeuille, mais elle n'est pas disponible ligne à ligne dans les bases de données à disposition.

[1]Contrat de retraite à cotisations définies réservé aux salariés, il est financé conjointement par l'entreprise et le salarié

[2]Contrat réservé aux professionnels indépendants afin de leur permettre de constituer un revenu viager à la retraite

3.2.4 Période d'observation

L'observation des individus se fait a priori sur une période allant de la première date de souscription au produit d'assurance jusqu'à la date d'extraction des données. Mais dans les faits, il sera retenue une période d'observation plus restreinte, ceci pour les raisons suivantes :

- Le potentiel caractère non exhaustif des données les mois précédant la date d'extraction. Il se peut en effet que des décès survenus à proximité de la date d'extraction des données n'aient pas encore été remontés à l'assureur. Pour éviter une sous estimation de la mortalité , la date de fin des observations doit être antérieure à la date d'extraction des données.
- Les données recueillies peu après le lancement du produit peuvent également ne pas être exhaustives. De plus des évolutions de la politique de souscription peuvent conduire à avoir une population observée lors des premières années du produit assez différente de la population observée ultérieurement. Par conséquent, la date de début des observations sera bien ultérieure à la date de lancement du produit. A noter également dans notre cas que les décès n'ont commencé à faire l'objet d'un reporting régulier de la part de l'assureur que à partir de l'année 2014.

Concernant le choix de la période d'observation en elle même, elle devra idéalement s'étaler sur un nombre entier d'années pour limiter l'impact du caractère saisonnier de la mortalité qui présente une surmortalité en hiver et en été par rapport au reste de l'année (Biessy 2022). L'objectif étant la construction d'une table de mortalité **prospective**, cette période d'observation devra être la plus étendue possible . Pour cette étude, le choix est porté sur une période d'observation de 6 ans, allant du 1^{er} Janvier 2017 au 31 Décembre 2022.

3.2.5 Quid des années Covid

La France comme l'ensemble du monde d'ailleurs a été sévèrement touchée en 2020 et 2021 par la pandémie du Covid 19. De Mars 2020 à Décembre 2021, entre 130 000 et 146 000 français sont décédés du Covid, il est ainsi estimé pour ces deux années une surmortalité de 95 000 décès due à cette pandémie (INSEE 2022a). Une étude de mortalité ne saurait donc se faire sans quelques précautions par rapport à cette période trouble pour s'assurer qu'elle n'induit pas un biais trop grand dans les analyses.

Sur la Figure 3.1, il est observé des différences marquantes pour les décès entre 70 et 80 ans pour les années $2020 - 2021$ par rapport aux autres années d'observation. Globalement il est constaté une surmortalité de l'ordre de 9% pour les années Covid par rapport aux années dites normales. Dans la suite de cette étude, cet écart de mortalité sera assumé, et aucun traitement spécifique ne sera apporté aux données des années Covid. Ceci du fait que cet écart est d'une ampleur relativement modeste et qu'il n'est

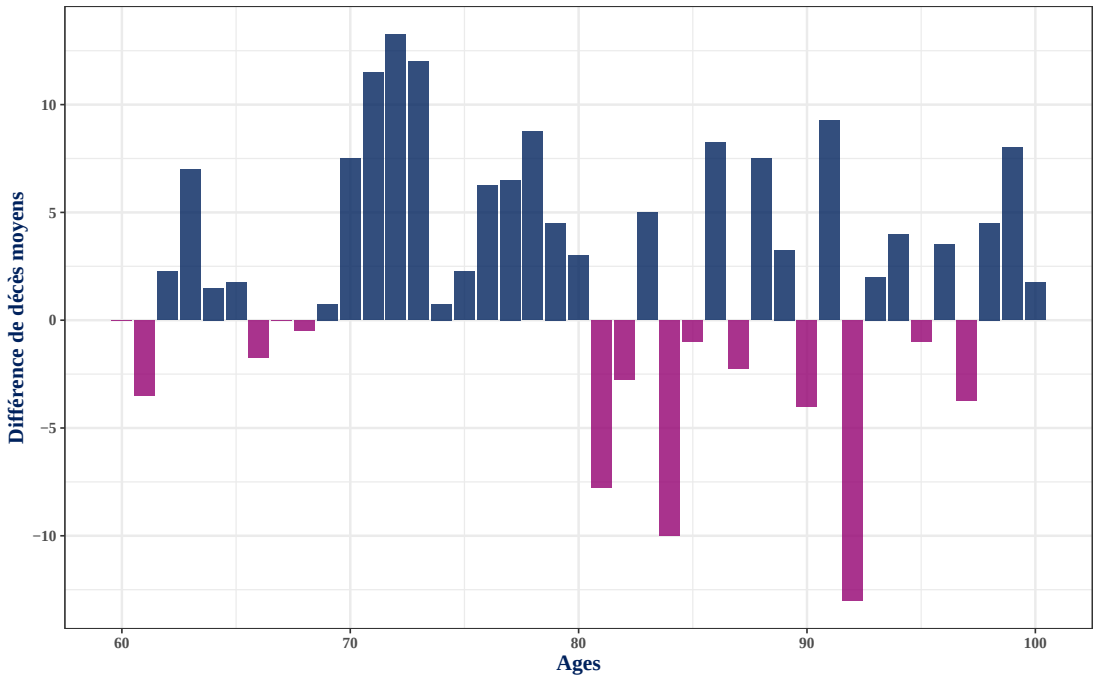

Figure 3.1 : Différences de décès moyens par âge entre les années Covid(2020-2021) et les années normales

pas envisageable de supprimer ces années Covid de l'étude, car cela entraînerait une perte d'information trop importante.

3.3 Mise en forme des différentes bases de données

Après avoir clairement défini le périmètre de l'étude, les données doivent être mises dans un format adéquat pour être analysées convenablement.

Pour évaluer l'influence du montant de rente sur la mortalité des individus dans le portefeuille, trois modèles seront calibrés sur les données à disposition : des modèles additifs généralisés (GAM), des forêts aléatoires de survie et des modèles de Cox. Chacun de ces modèles nécessite une mise en forme particulière des données en entrée.

Avant toute chose, on procède à une jointure entre les bases de stock et les bases de flux. Ceci est fait de façon à ce que pour chaque individu présent dans la base de stock, lui soit associé sa date de décès éventuelle.

3.3.1 Base agrégée

Les caractéristiques des individus qui seront prises en compte dans la modélisation de la mortalité sont le sexe, l'âge et le montant de la rente perçue. Dans l'optique d'une mise en relation du niveau de mortalité observé dans le portefeuille avec celle décrite par les tables réglementaires, il sera également intégré une dimension supplémentaire pour l'année calendaire. Nous avons donc procédé à une agrégation des données à disposition suivant ces différentes caractéristiques. Le montant de rente a été scindé en 10 classes (proches des déciles) pour faciliter les calculs. Pour conserver la possibilité de traiter le montant de rente comme une variable numérique, il a été créé une variable correspondant à la médiane des classes de montant de rente (m_2). Ainsi pour chacune des combinaisons des caractéristiques sexe (s), âge (x), classe de montant de rente (m) et année calendaire (y), il est calculé **l'exposition centrale au risque** (ec) et le **nombre de décès** (d) correspondants. Ce format de données est adapté pour l'application des modèles additifs généralisés, et offre également la possibilité de calculer des taux bruts de décès assez facilement en rapportant le nombre de décès à l'exposition au risque. Cette base est présentée dans la Table 3.1.

Table 3.1 : Base agrégée

x	y	s	m	m2	ec	d
60	2017	Femme	2K - 2,7K	2300	19.48	0
60	2017	Femme	< 500	180	5.05	0
60	2017	Femme	500 - 700	600	13.49	0
60	2017	Femme	1,2K - 1,6K	1400	24.08	0
60	2017	Femme	900 - 1,2K	1050	21.27	0
60	2017	Femme	700 - 900	800	10.85	0

3.3.2 Base individuelle fractionnée

Un autre format de données qui sera utilisé dans le processus de modélisation, est un format individuel fractionné, où chaque ligne représentera un individu pour une année d'observation donnée (année calendaire). Il sera observé les mêmes variables que dans le cas des données agrégées à savoir le sexe (s), l'âge (x), l'année calendaire (y), la classe de montant de rente (m), l'exposition au risque (ec), le nombre de décès (d), et les valeurs originales des montants de rente pour chaque individu (m_2). Ce format de données sera utilisé pour calibrer les modèles de Cox. L'avantage de ce format par rapport au précédent est qu'il permet d' observer l'ensemble de l'étendue de la variable montant de rente, et permettrait donc a priori d'évaluer de façon plus fine son effet sur la mortalité. Cette base est présentée dans la Table 3.2.

Table 3.2 : Base individuelle fractionnée

x	y	s	m	m2	ec	d
90.24778	2017	Homme	2K - 2,7K	2098.11	1	0
91.24709	2018	Homme	2K - 2,7K	2098.11	1	0
92.24641	2019	Homme	2K - 2,7K	2098.11	1	0
93.24572	2020	Homme	2K - 2,7K	2098.11	1	0
94.24778	2021	Homme	2K - 2,7K	2098.11	1	0
95.24709	2022	Homme	2K - 2,7K	2098.11	1	0

3.3.3 Base individuelle

Les forêts aléatoires de survie nécessitent un format de données encore différent de ce qui a été présenté jusqu'ici. La disposition des données est telle que chaque individu est représenté par une et une seule ligne dans la base de données. En plus des caractéristiques sexe (s), âge (x), classe de montant de rente (m), montant de rente (m_2), une variable de durée (ec) correspondant au nombre d'années durant lesquelles chaque individu est observé est calculée, de même qu'une variable de censure (d) indiquant si le décès a été effectivement observé ou pas pour un individu au cours de la période d'observation (Table 3.3).

Table 3.3 : Base individuelle

x	y	s	m	m2	ec	d
96	2022	Homme	2K - 2,7K	2098.11	6.00	0
87	2019	Homme	$<$ 500	467.23	2.93	1
97	2021	Femme	1,6K - 2K	1673.01	4.81	1
100	2022	Femme	$>$ 5,5K	6525.13	6.00	0
100	2022	Femme	1,2K - 1,6K	1229.49	5.34	1
96	2022	Femme	$>$ 5,5K	12083.71	6.00	0

3.4 Statistiques descriptives

Le périmètre de l'étude ainsi que les différents formats de données qui seront utilisés dans les analyses ayant été explicités, il est à présent question de présenter des statistiques descriptives qui seront utiles pour la phase de modélisation. A noter que ces statistiques sont relatives à la population assurée du périmètre défini plus haut.

3.4.1 Répartition par sexe

Au total, on dénombre 56 220 assurés qui ont été observés lors de la période d'observation, avec 37 491 hommes (67%) pour 18 729 femmes (33%). Il a été enregistré un total de 6 716 décès dont 63% d'hommes (4 254) et 37% de femmes (2 462).

3.4.2 Répartition par âge

3.4.2.1 Âges à l'entrée sous observation

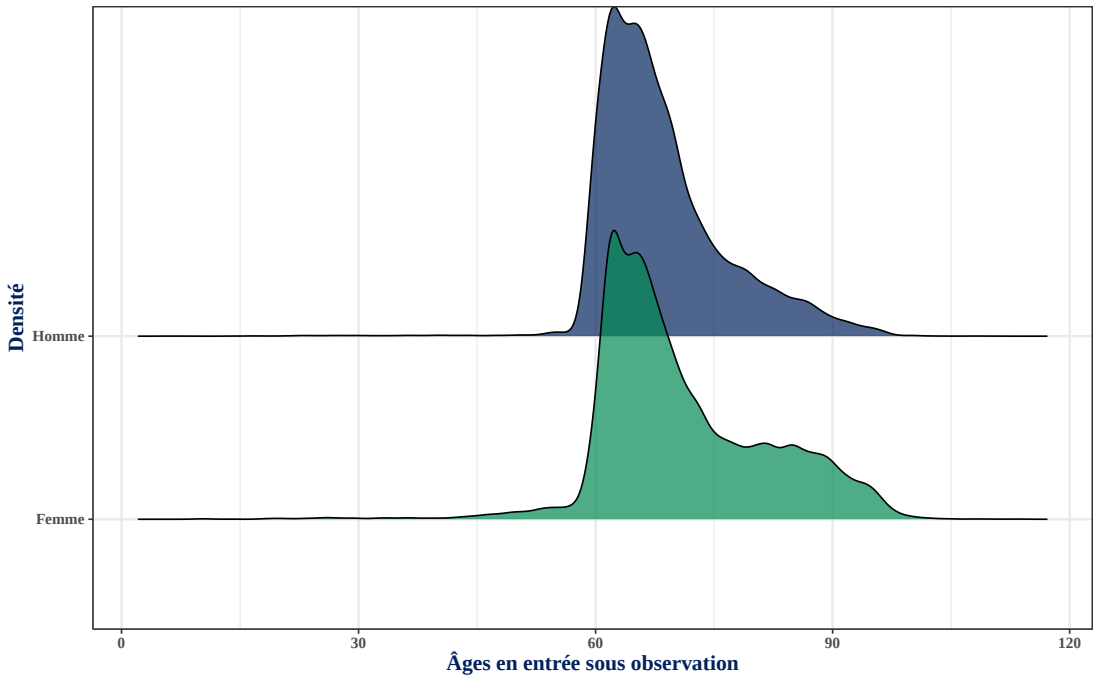

Figure 3.2 : Distribution des âges en entrée sous observation

L'âge moyen d'entrée sous observation est de 71 ans pour les femmes et de 68, 5 ans pour les hommes. Sur la Figure 3.2 on observe un pic de densité entre 62 et 63 ans aussi bien pour les hommes que pour les femmes. Aux grands âges, on observe une densité plus importante chez les femmes que chez les hommes. Globalement à l'entrée sous observation, les femmes sont donc plus âgées que les hommes.

3.4.2.2 Âges en sortie d'observation

L'âge moyen de sortie d'observation est de 76 ans pour les femmes contre 73, 5 ans pour les hommes. Un pic de densité est cette fois ci observé à l'âge de 69 ans (Figure 3.3). Il

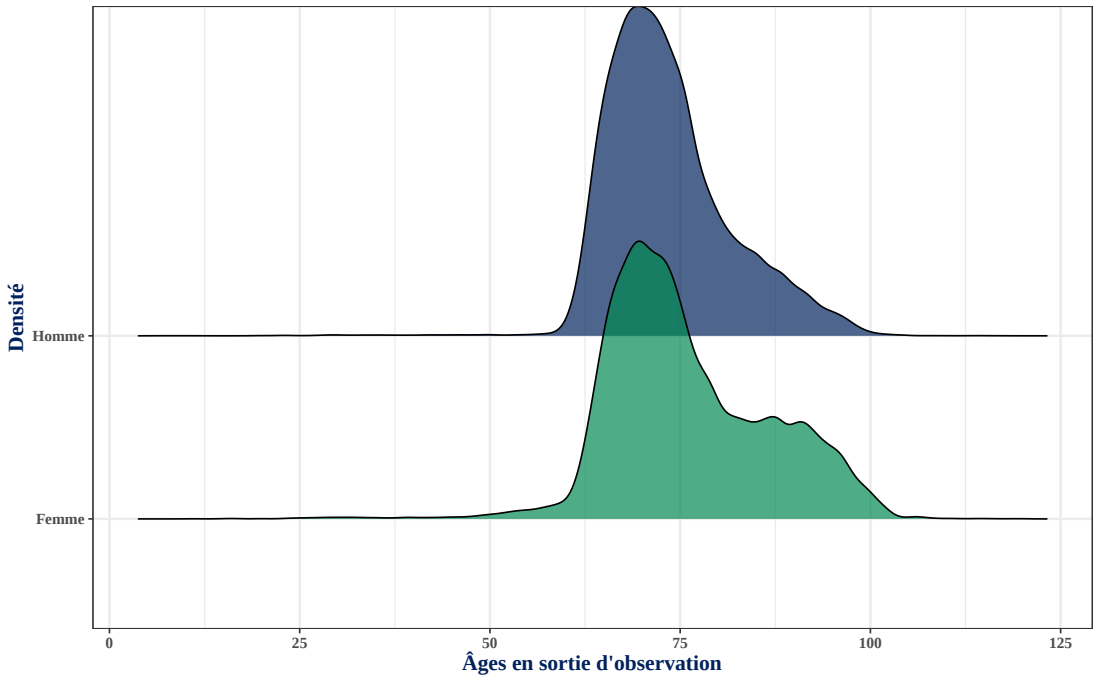

Figure 3.3 : Distribution des âges en sortie d'observation

est toujours observé une densité plus importante aux grands âges chez les femmes par rapport aux hommes. A noter qu'il y a de la censure dans les observations, ces âges ne reflètent donc pas la distribution ultime des âges de sortie. **L'âge moyen au décès** est de 87 ans chez les femmes et de 81 ans chez les hommes.

3.4.3 Montant de rente

La distribution des montants de rente (Figure 3.4) aussi bien pour les hommes que pour les femmes est fortement asymétrique à droite. Au global, 97% des assurés perçoivent un montant de rente annuel inférieur à 10 000€. On dénombre une trentaine d'assurés qui touchent des rentes supérieures à 50 000€, avec un maximum à près de 650 000€.

Le constat fait sur la Figure 3.5 est que dans les classes de montant de rente les plus basses (inférieurs à 1 600€) les proportions de femmes sont supérieures à celles des hommes. Cette tendance s'inverse pour les classes de montant de rente supérieures à ce seuil de 1 600€, les proportions d'hommes deviennent alors supérieures à celles des femmes. Les rentiers de sexe masculin du portefeuille perçoivent donc des montants de rente plus élevés que les femmes.

Sur la Figure 3.6 , on constate que la plus basse des classes de montant de rente (**"<500"**) semble être constituée d'individus particulièrement âgés (un pic de densité autour de 89

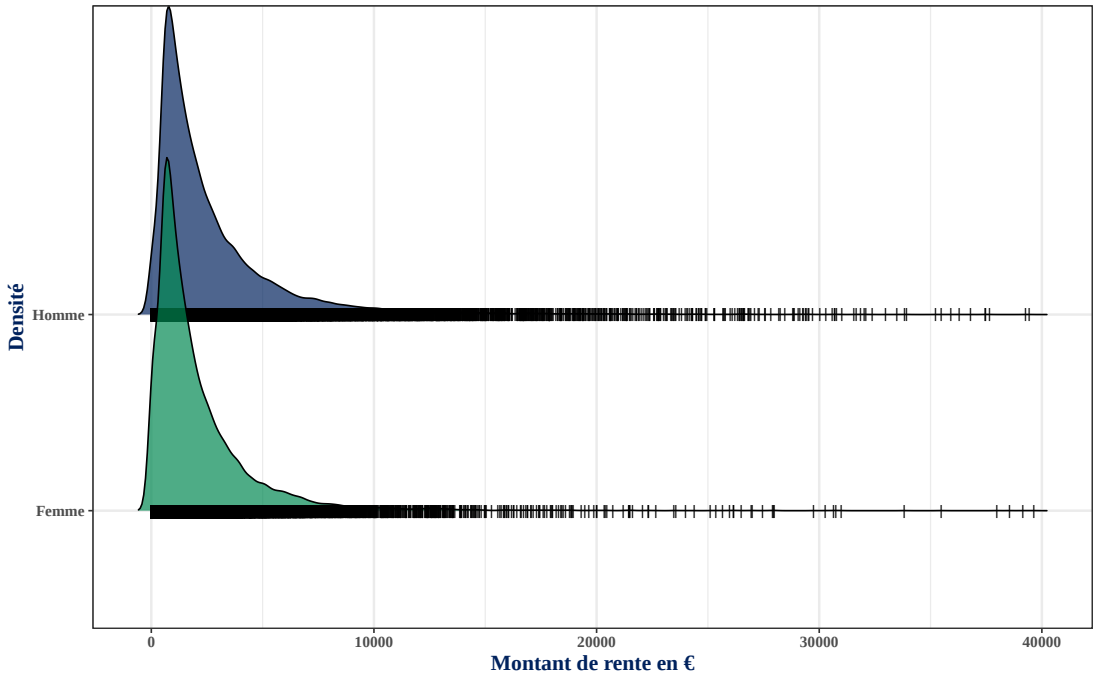

Figure 3.4 : Distribution des montants de rente

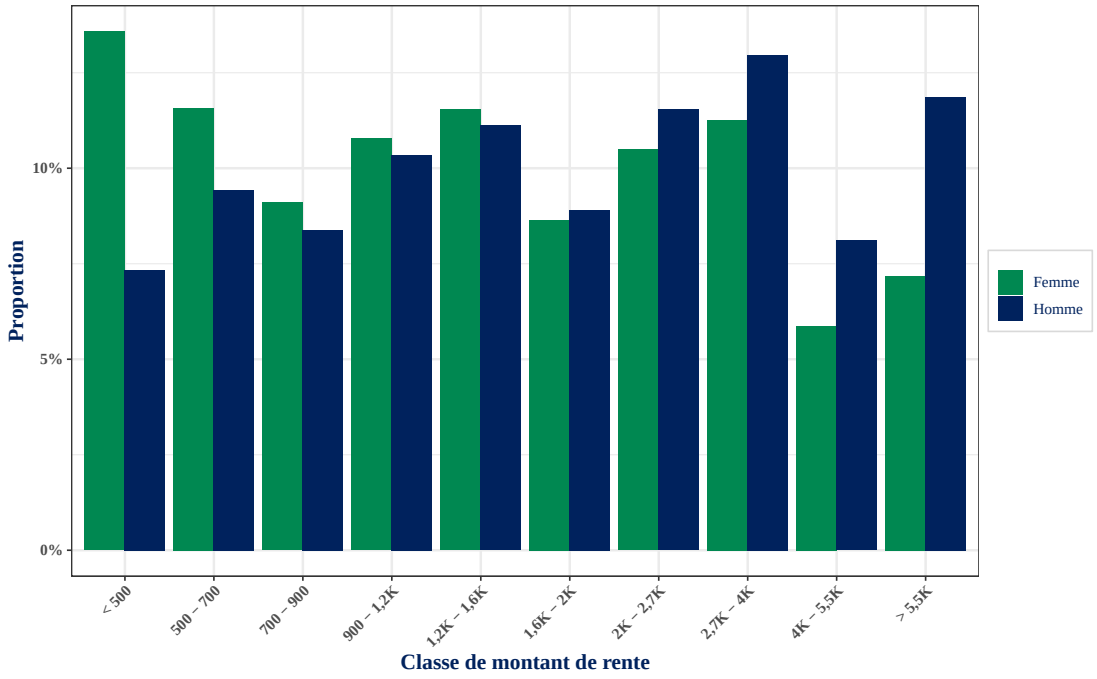

Figure 3.5 : Proportions d'individus de chaque sexe par classes de montant de rente

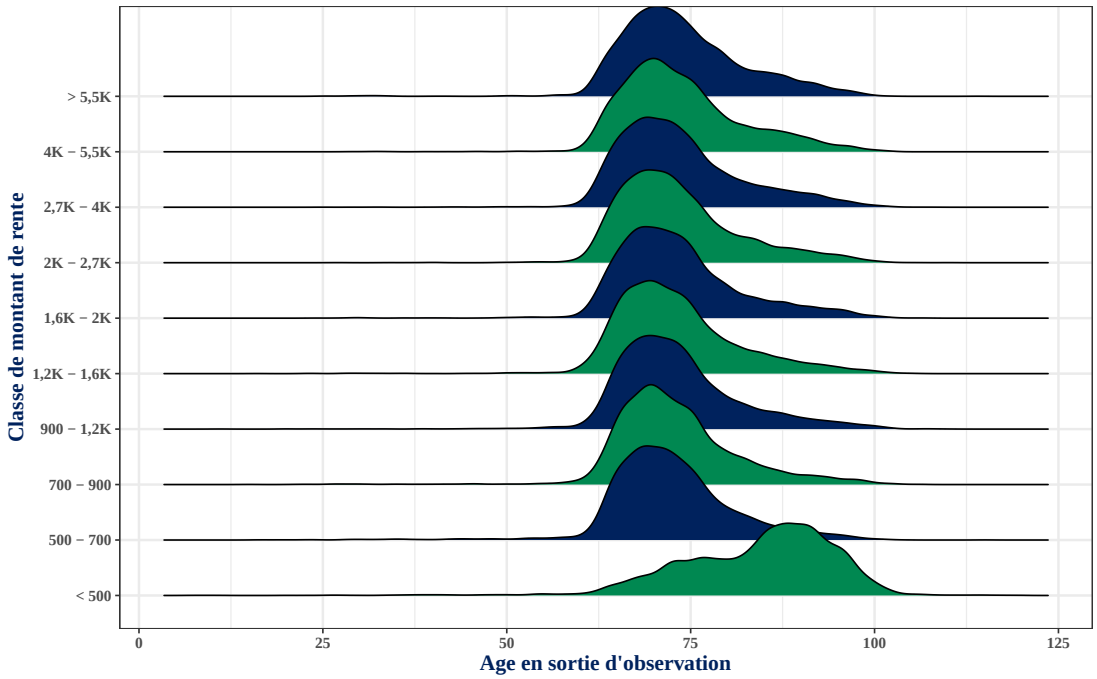

Figure 3.6 : Distribution des âges par classe de montant de rente

ans) comparativement aux autres classes. Les autres classes sont assez similaires en terme d'âges des individus qui les constituent , on observe un pic au niveau des densités de distribution des âges autour de 70 ans.

26% des décès sont enregistrés dans la classe de montant de rente la plus basse (Figure 3.7), c'est deux fois plus que pour toute autre classe. On observe cependant une exposition au risque assez équivalente entre les différentes classes. La surmortalité observée dans la classe **"< 500"** n'est manifestement pas due à une surexposition de cette dernière par rapport aux autres classes.

Il est possible de calculer des taux bruts de décès par classe de montant de rente en rapportant les décès à l'exposition au risque. Sans surprise on observe un taux de décès assez élevé pour la première classe de montant de rente, les taux dans les autres classes étant nettement plus faibles (Figure 3.8). Il est assez difficile à ce stade de dégager une tendance (hormis éventuellement pour la première classe de montant de rente) d'évolution de la mortalité dans le portefeuille en fonction du montant de rente. Ceci est d'autant plus vrai qu'ici il y a certainement une interférence de la variable age dans ce qui est observé avec ce graphique, le pic de mortalité pour la première classe de rente s'expliquant en partie par le fait que les individus y sont plus âgés que dans les autres classes (Figure 3.6). Par la suite, pour avoir un aperçu plus juste (du moins sur le plan descriptif) du niveau mortalité selon le montant de rente, nous allons utiliser un indicateur qui neutralise l'effet de l'age. Cela se fera après avoir comparé la mortalité

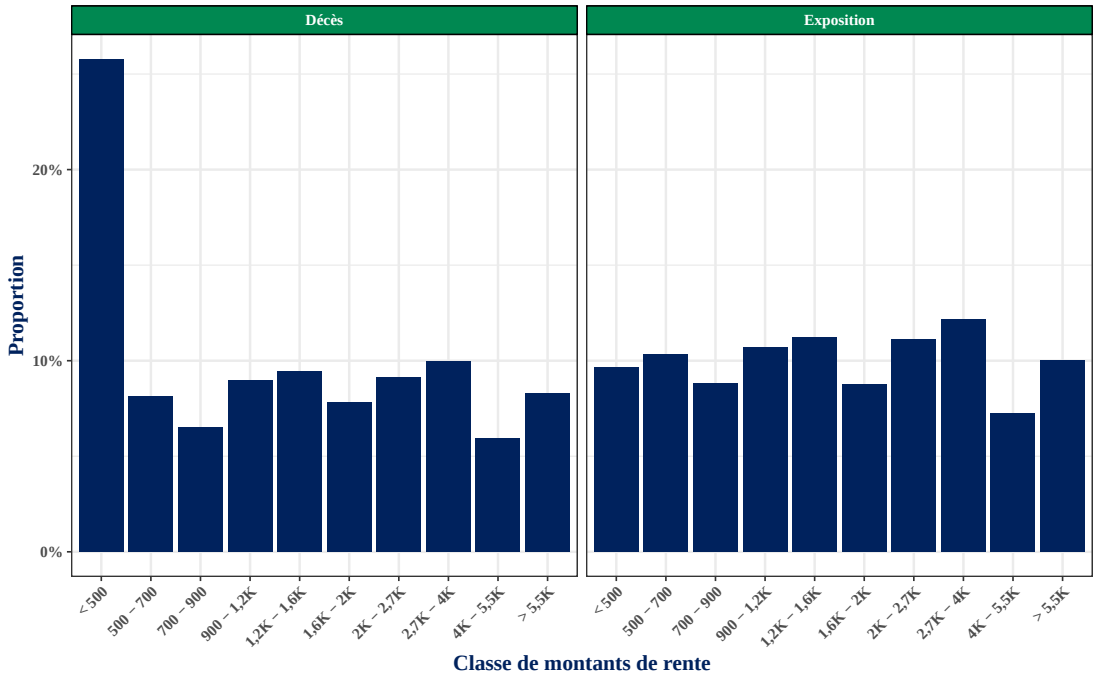

Figure 3.7 : Exposition et décès par classe de montant de rente

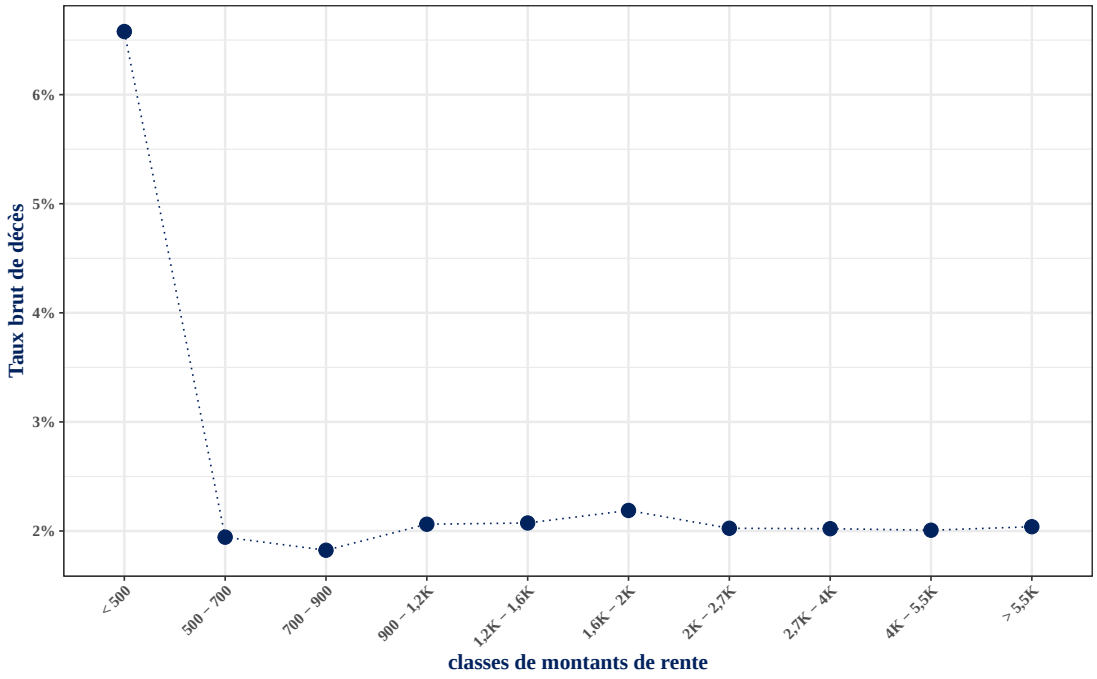

Figure 3.8 : Taux bruts de décès par classe de montant de rente

globale du portefeuille à celle des tables réglementaires $TGH/TGF05$.

3.5 Positionnement du portefeuille par rapport à la référence réglementaire

Il sera question ici de comparer le niveau de mortalité observé dans le portefeuille avec celui des tables réglementaires **TGH05** et **TGF05**. Nous partirons ici des données sous leur forme la plus agrégée. L'idée est de calculer les décès théoriques (d_{th}) qu'on aurait observé si la mortalité du portefeuille était telle que décrite par les tables réglementaires. On compare ensuite ces décès théoriques aux décès effectivement observés dans le portefeuille.

Les tables **TGH05** (pour les hommes) et **TGF05** (pour les femmes) sont des tables générationnelles qui donnent les niveaux de mortalité sur plusieurs générations, de 1900 à 2005 pour des ages allant de 0 à 120 ans. Elles se présentent (pour chaque génération) comme un écoulement d'une population de taille initiale (fictive) 100 000, qui diminue au fil des années jusqu'à extinction complète (décès de l'ensemble des individus). Pour un **sexe donné**, le nombre d'individus survivants de la génération t, d'age x est alors donné par L_x^t . Le nombre de décès théorique à l'age x des individus de la génération t est donné par :

$$\begin{aligned} d_{th}^{t,x} &= ec_x^t \times h_x(t) \\ &= -ec_x^t \times \log(1 - L_{x+1}^t / L_x^t) \end{aligned}$$

Avec :

- ec_x^t l'exposition du portefeuille pour les individus de la génération t d'âge x
- $h_x(t)$ la force de mortalité de la table réglementaire pour les individus de la génération t d'âge x.

Une fois ces décès théoriques obtenus par âge, sexe et génération, il est possible de les sommer par âge, par sexe, voire même par année (génération).

3.5.1 Par âge et par sexe

Quasiment à tous les âges entre 60 et 100 ans, aussi bien pour les hommes que pour les femmes, les décès observés dans le portefeuille sont supérieurs à ceux prévus par les tables réglementaires (Figure 3.9). Au global on constate ainsi une surmortalité dans le portefeuille par rapport aux tables réglementaires de l'ordre de 10%. Guez (2018) dans une étude similaire aboutissait également à une surmortalité pour un portefeuille de rentiers de Groupama Gan Vie par rapport aux tables réglementaires.

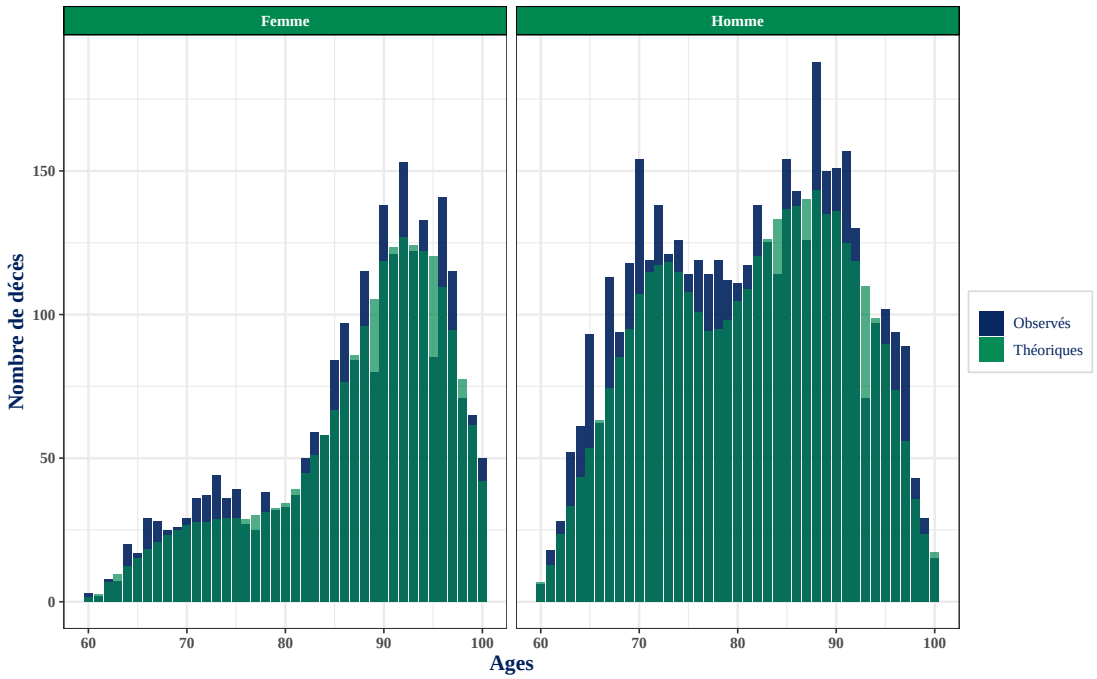

Figure 3.9 : Décès observés et décès prédits par les tables règlementaires par âge et par sexe

3.5.2 Par année et par sexe

La Figure 3.10 présente le ratio entre le nombre de décès observés dans le portefeuille et le nombre de décès théoriques prédit par les tables réglementaires pour les 6 années d'observation. Ce ratio est toujours supérieur à 100%, avec un pic en 2020 à 124%, celui-ci s'expliquant probablement par la pandémie de Covid 19. A noter que la chute observée en 2022 est probablement due au fait qu'elle soit relativement proche de la date d'extraction des données, et que potentiellement l'ensemble des décès correspondants n'aient pas encore été reportés.

3.5.3 Par classe de montant de rente

La Figure 3.11 dans un premier temps confirme le constat global qui avait été fait précédemment, à savoir des décès observés plus importants que les décès prédits par les tables réglementaires, et ce pour quasiment toutes les classes de montant de rentes. La classe **"< 500"** qui semblait nettement se démarquer des autres en terme de nombre de décès observés est un peu moins remarquable ici. Ceci confirme le fait que la surmortalité qui y est observée par rapport aux autres classes est en grande partie un effet de l'âge. En dehors des classes **"500-700"** et **"1,6K-2K"** qui présentent des pics assez singuliers

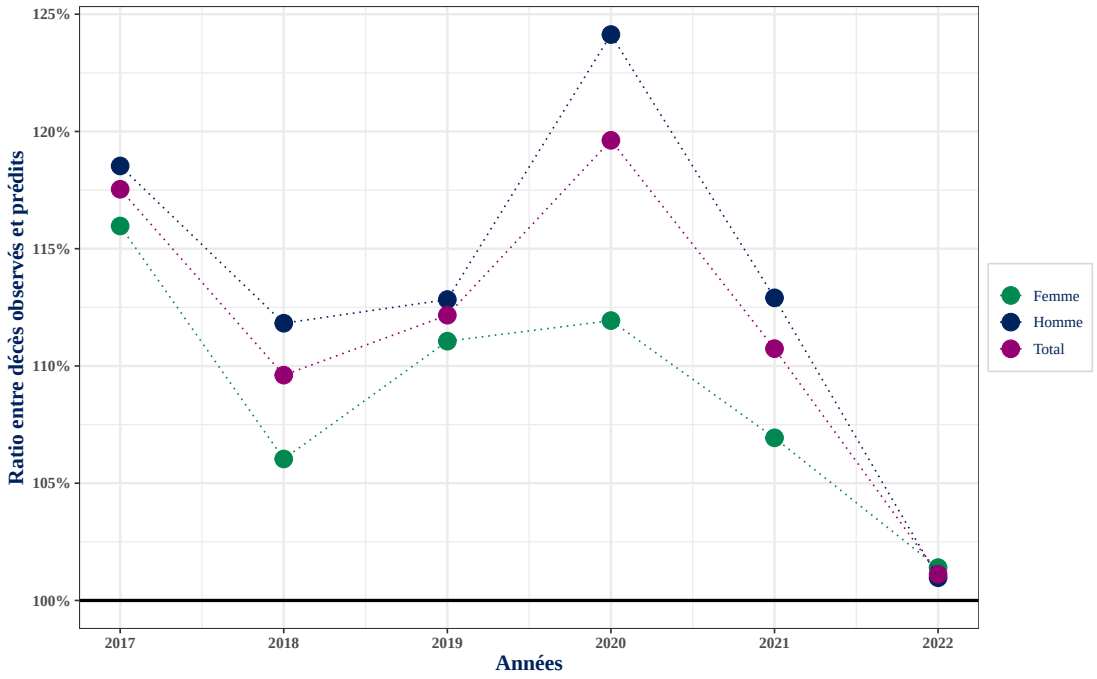

Figure 3.10 : Comparaison entre décès observés et décès prédits par les tables règlementaires par année et par sexe

, on peut discerner une tendance à la baisse de l'écart de mortalité par rapport aux tables réglementaires à mesure que la classe de montant de rente augmente.

Cette relation entre mortalité et montant de rente sera investiguée de manière plus rigoureuse par la suite avec les différents modèles qui seront calibrés.

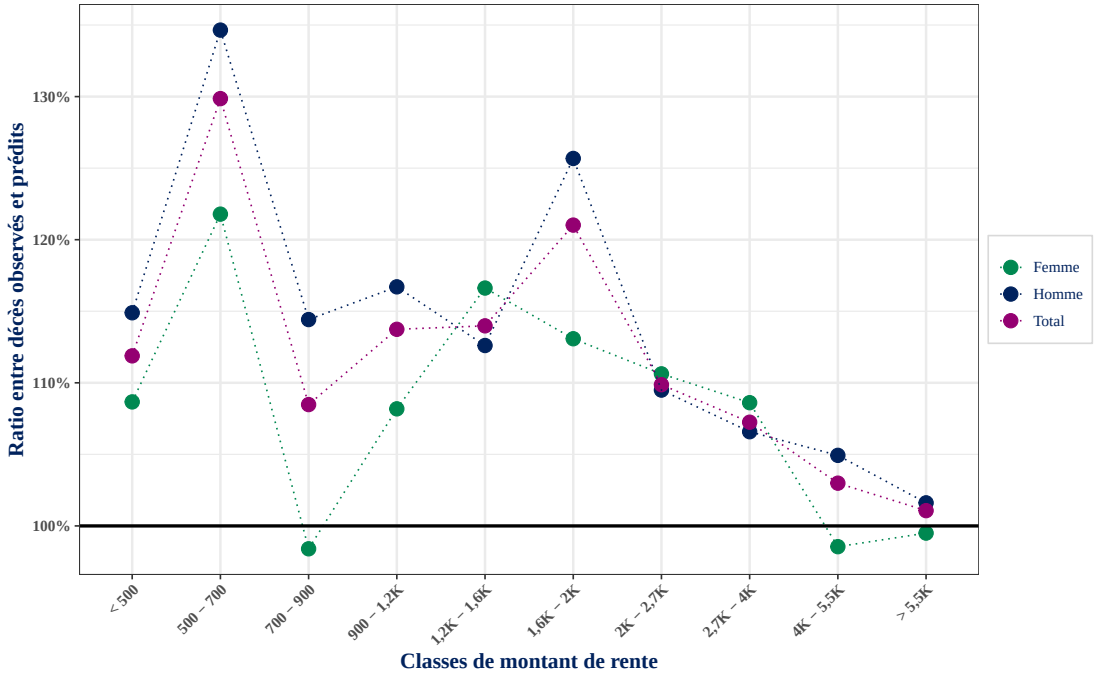

Figure 3.11 : Comparaison des décès observés et décès prédits par les tables règlementaires par classe de montant de rente

Chapitre 4

Les modèles additifs généralisés

Dans ce chapitre, il sera présenté les fondements théoriques des modèles additifs généralisés. Ce type de modèle étant encore très peu utilisé en Actuariat, nous avons jugé utile de consacrer tout un chapitre à son introduction.

4.1 Introduction

Sauf mention explicite du contraire, l'ensemble des éléments qui seront présentés ici sont issus de l'ouvrage Wood (2017) .

Un modèle additif généralisé en abrégé **GAM** (Generalized Additive Model) est un modèle linéaire généralisé avec des prédicteurs linéaires qui sont des fonctions **lisses** des variables explicatives.

$$g(\mu_i) = A_i\theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}) + ... \tag{4.1}$$

Avec:

- $\mu_i \equiv \mathbb{E}(Y_i)$, Y_i étant la variable d'intérêt de loi appartenant à la famille exponentielle
- g une fonction de lien
- $A_i\theta$ correspondant à la partie paramétrique du modèle, où A_i est une ligne de la matrice de design, et θ le vecteur de paramètre associé
- f_j une fonction lisse associée à la covariable x_j

Ce type de modèle permet une certaine flexibilité (comparativement aux modèles totalement paramétriques) dans la matérialisation des dépendances entre les différentes covariables et la variable d'intérêt. Cette flexibilité nouvelle s'accompagne de contraintes supplémentaires:

- La définition explicite de ce que sont les fonctions lisses f_j ;

- La définition du niveau de lissage optimal pour ces fonctions f_j, autrement dit, à quel point est ce qu'elles doivent être lisses?

Pour mieux cerner les contours de la notion de **fonctions lisses**, nous allons d'abord nous intéresser aux **modèles additifs** de structure plus simple avant de nous intéresser à leur généralisation aux **modèles additifs généralisés**.

4.2 Les modèles additifs

4.2.1 Modèle additif univarié

Le modèle additif le plus simple est le modèle **additif univarié**, il est constitué d'une seule variable explicative, et donc d'une seule fonction lisse associée à cette variable.

$$y_i = f(x_i) + \epsilon_i \tag{4.2}$$

Avec:

- y_i la variable réponse
- x_i la variable explicative
- f la fonction lisse de la covariable x_i
- ϵ_i un terme de nuisance aléatoire qui suit une loi normale centrée réduite de variance σ^2.

La question qui se pose légitimement ici est celle de savoir comment peut être représentée, comment peut être estimée la fonction f.

Représentation de la fonction lisse par des fonctions de base

Il est question ici de représenter la fonction lisse f dans une base appropriée. f pourra être écrite alors comme une combinaison linéaire de **fonctions de bases** issues d'une base préalablement choisie, ces fonctions de base étant supposées connues a priori. f peut alors être représentée de la manière suivante :

$$f(x) = \sum_{j=1}^{k} b_j(x)\beta_j \tag{4.3}$$

Avec:

- $b_j(x)$ la j^{ime} fonction de base

- β_j le paramètre (à estimer) associé à la j^{ime} fonction de base
- k la dimension de la base choisie, correspond également au nombre de fonctions de base.

En remplaçant l'expression de l'Équation 4.3 dans celle de l'Équation 4.2 , on se ramène à un modèle linéaire classique. En optant par exemple pour une base polynomiale de degré 3, on obtient:

$$f(x) = \beta_1 + \beta_2 x + \beta_3 x^2 + \beta_4 x^3$$

Ainsi:

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \beta_4 x_i^3 + \epsilon_i \tag{4.4}$$

Le problème de spécification de la fonction lisse f se ramène donc au choix de la base (et donc des fonctions de base) dans laquelle elle sera représentée.

Les bases polynomiales s'avèrent peu appropriées en pratique pour la représentation de fonctions lisses. Les polynômes ont de bonnes propriétés (en terme de régularité) dans des voisinages restreints de points, mais pour la représentation d'une fonction inconnue sur un domaine relativement étendu, ils deviennent très peu pertinents et ont tendance à sur-apprendre des données comme le montre la Figure 4.1 .

L'estimation polynomiale (courbe verte) pour peu qu'on lui accorde suffisamment de **degrés de liberté**[1] interpole parfaitement l'ensemble des points d'observations (points bleus). Cependant, cette estimation polynomiale présente des fluctuations très abruptes, et est en réalité assez éloignée de la fonction inconnue (courbe en traits interrompus courts) qu'elle est sensée approcher.

Une alternative aux bases polynomiales qui tend à régler ce problème de sur-apprentissage de ces dernières est la famille des bases de fonctions **linéaires par morceaux**. Nous allons présenter la plus simple de ces bases par la suite.

Bases de fonctions linéaires par morceaux

Le principe ici est de diviser la plage d'observation de la covariable x en plusieurs morceaux délimités par des points qu'on appellera nœuds, et de définir une fonction de base **linéaire** sur chacun des morceaux ainsi constitués. Pour une base de dimension k , on considère l'ensemble des nœuds ci-après : $(x_j^* : j = 1, ..., k)$, avec $x_j^* > x_{j-1}^*$, les fonctions de base sont définies par :

$\forall j = 2, ..., k-1$

[1] Le degré de liberté ici renvoie à l'ordre de la base polynomiale choisie

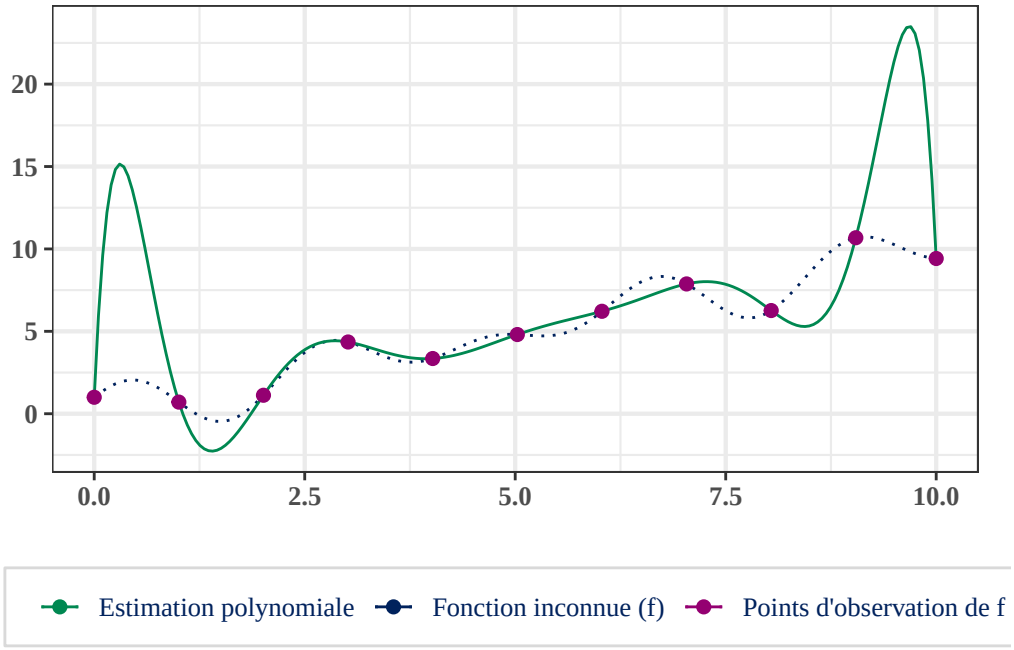

Figure 4.1 : Estimation d'une fonction par une base de fonctions polynomiales

$$b_j(x) = \begin{cases} (x - x_j^*)/(x_j^* - x_{j-1}^*) & \text{si } x_{j-1}^* < x \leq x_j^* \\ (x_{j+1}^* - x)/(x_{j+1}^* - x_j^*) & \text{si } x_j^* < x < x_{j+1}^* \\ 0 & \text{sinon} \end{cases} \tag{4.5}$$

$$b_1(x) = \begin{cases} (x_2^* - x)/(x_2^* - x_1^*) & \text{si } x < x_2^* \\ 0 & \text{sinon} \end{cases}$$

$$b_k(x) = \begin{cases} (x - x_{k-1}^*)/(x_k^* - x_{k-1}^*) & \text{si } x > x_{k-1}^* \\ 0 & \text{sinon} \end{cases}$$

Ainsi, la fonction de base $b_j(x)$ est de valeur nulle partout sauf sur le morceau de domaine délimité par les nœuds x_{j-1}^* et x_{j+1}^* .

A noter que l'Équation 4.2 peut se réécrire sous la forme matricielle suivante:

$$Y = X\beta + \epsilon \tag{4.6}$$

Avec X la matrice associée aux fonctions de base b_j, elle est définie par $X_{ij} = b_j(x_i)$ et $\beta = (\beta_1, ..., \beta_k)$

La Figure 4.2 présente les résultats obtenus en utilisant cette base de fonctions linéaires par morceaux avec les données de l'illustration précédente.

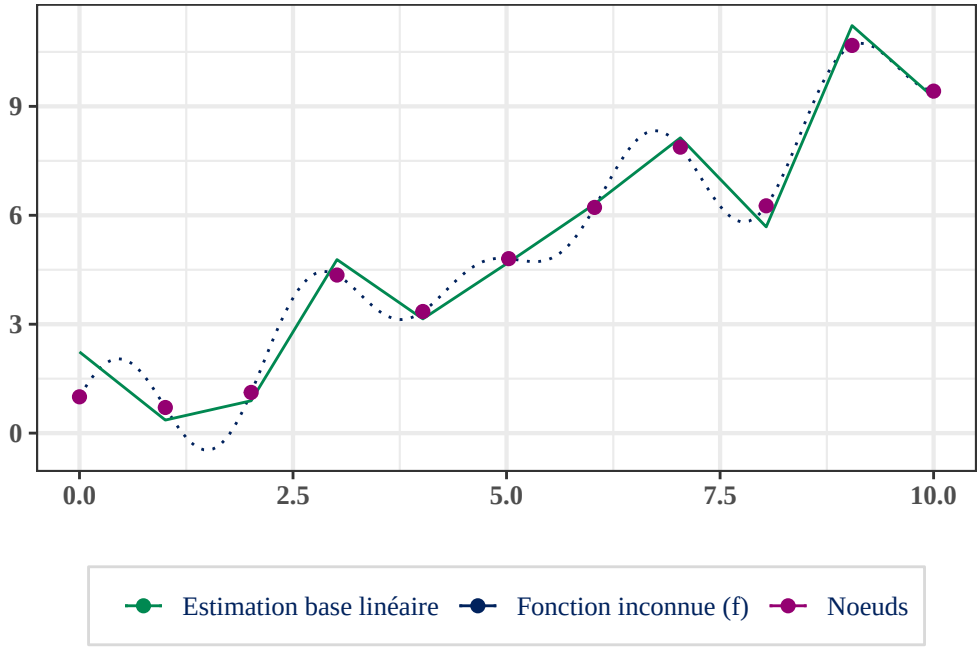

Figure 4.2 : Estimation d'une fonction par une base de fonctions linéaires par morceaux

L'estimation de la fonction inconnue f dans la base de fonctions linéaires par morceaux (courbe verte) semble mieux approcher la fonction f que ce que l'on pouvait observer avec la base polynomiale précédente, il y a moins de fluctuations et moins de risques de sur-apprentissage ici. Cependant, on est encore très loin de ce qu'on est en droit d'attendre d'une fonction **lisse**, en effet la représentation obtenue dans cette nouvelle base manque encore de **régularité**. Nous nous intéresserons donc par la suite à la façon dont il est possible de contrôler la **régularité** ou encore l'aspect **lisse** de l'estimation faite pour la fonction f.

Contrôle du niveau de lissage par pénalisation des irrégularités

Une approche intuitive pour contrôler la régularité est de faire varier le nombre de nœuds (la dimension de la base). En faisant diminuer le nombre de nœuds, la représentation de la fonction f obtenue est de plus en plus lisse. Cette approche a néanmoins quelques inconvénients :

- Elle conduit à des nœuds irrégulièrement espacés (une fois certains nœuds supprimés), ce qui peut détériorer la qualité du modèle.

- L'ajustement de modèles par des fonctions de base linéaires par morceaux est très lié à l'emplacement des nœuds, modifier ces emplacements au cours du processus d'estimation aboutirait à des modèles difficilement comparables.

Une alternative à la réduction de la dimension de la base des fonctions linéaires par morceaux est la suivante : maintenir le nombre de nœuds à un niveau légèrement supérieur à ce qui serait nécessaire. Et dès lors, contrôler la régularité du rendu en ajoutant une contrainte sur les coefficients (β_j) des fonctions de bases de sorte à pénaliser les irrégularités.

Ainsi, au lieu de minimiser $||Y - X\beta||^2$ pour obtenir le paramètre β, il sera plutôt minimisé l'objectif suivant :

$$||Y - X\beta||^2 + \lambda \sum_{j=2}^{k-1} \Big(f(x^*_{j-1}) - 2f(x^*_j) + f(x^*_{j+1}) \Big)^2 \tag{4.7}$$

Ici, le terme de sommation mesure l'irrégularité comme la somme des carrés des différenciations d'ordre 2 de la fonction lisse aux différents nœuds. Le terme de pénalité sera d'autant plus grand que l'estimation de f présentera de fortes irrégularités.

Le paramètre λ est appelé **paramètre de lissage**, il contrôle l'arbitrage entre la régularité de l'estimation de f et sa fidélité aux observations initiales. Plus ce paramètre est élevé et plus le rendu final est lisse, la Figure 4.3 illustre cet effet.

Estimation du vecteur de paramètres pour un niveau fixé du paramètre de lissage

Considérons la fonction objectif de l'Équation 4.7, le terme de pénalité de la fonction lisse peut se réécrire de la façon suivante :

$$\lambda \sum_{j=2}^{k-1} \Big(f(x^*_{j-1}) - 2f(x^*_j) + f(x^*_{j+1}) \Big)^2 = \lambda \sum_{j=2}^{k-1} \Big(\beta_{j-1} - 2\beta_j + \beta_{j+1} \Big)^2 \tag{4.8}$$

On montre en effet que pour les bases de fonctions linéaires par morceaux définies à l'Équation 4.5, on a $\beta_j = f(x^*_j)$.

En se ramenant ensuite à une écriture matricielle et en faisant le constat suivant :

$$\begin{pmatrix} \beta_1 - 2\beta_2 + \beta_3 \\ \beta_2 - 2\beta_3 + \beta_4 \\ \beta_3 - 2\beta_4 + \beta_5 \\ . \\ . \end{pmatrix} = \begin{pmatrix} 1 & -2 & 1 & 0 & . & . & . \\ 0 & 1 & -2 & 1 & 0 & . & . \\ 0 & 0 & 1 & -2 & 1 & 0 & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ . \\ . \end{pmatrix} = D\beta$$

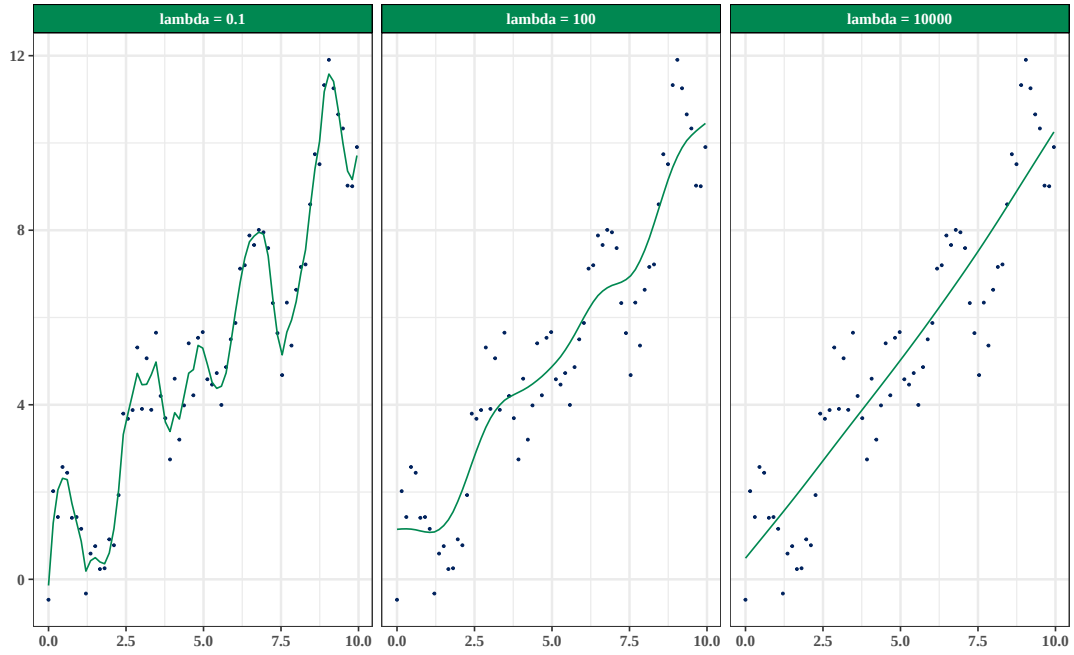

Figure 4.3 : Effet du paramètre de lissage

Le terme de pénalité s'écrit alors de la manière suivante :

$$\lambda \sum_{j=2}^{k-1} \Big(f(x^*_{j-1}) - 2f(x^*_j) + f(x^*_{j+1}) \Big)^2 = \lambda \beta^T D^T D \beta = \lambda \beta^T S \beta \tag{4.9}$$

Avec : $S = D^T D$, la matrice de **pénalité**.

Le vecteur de paramètres β est alors obtenu en minimisant la fonction objectif suivante :

$$||Y - X\beta||^2 + \lambda \beta^T S \beta \tag{4.10}$$

A noter qu'il s'agit ici d'un problème de **régression pénalisée**, il est possible d'obtenir une solution explicite pour β, elle est obtenue de la manière suivante :

$$\text{On remarque que : } ||Y - X\beta||^2 + \lambda\beta^T S\beta = \left\| \begin{pmatrix} Y \\ 0 \end{pmatrix} - \begin{pmatrix} X \\ \sqrt{\lambda}D \end{pmatrix} \beta \right\|^2$$

$$\text{On a alors : } \hat{\beta} = \left(\begin{pmatrix} X \\ \sqrt{\lambda}D \end{pmatrix}^T \begin{pmatrix} X \\ \sqrt{\lambda}D \end{pmatrix} \right)^{-1} \begin{pmatrix} X \\ \sqrt{\lambda}D \end{pmatrix}^T \begin{pmatrix} Y \\ 0 \end{pmatrix} \tag{4.11}$$

$$\text{Et donc : } \hat{\beta} = (X^T X + \lambda S)^{-1} X^T Y$$

Pour une valeur fixée du paramètre λ, on est en mesure d'estimer les paramètres du modèle et donc d'estimer la fonction lisse correspondante. Il est donc maintenant légitime de se demander comment trouver la valeur optimale du paramètre λ de sorte à ce que l'estimation de la fonction lisse soit la plus proche possible de la valeur vraie de la fonction f recherchée.

Détermination du paramètre de lissage optimal par validation croisée

Un paramètre de lissage λ trop élevé aboutirait à une estimation $\hat{f}$ de f trop lisse, quand une valeur trop faible de ce même paramètre entraînerait une estimation pas assez lisse sujette au surraprentissage. Dans les deux cas on aboutit à une estimation $\hat{f}$ qui s'éloigne de la vraie fonction f recherchée. Une condition idéale serait de choisir λ de sorte à ce qu'il minimise le critère suivant:

$$M = \frac{1}{n} \sum_{i=1}^{n} (\hat{f}_i - f_i)^2 \tag{4.12}$$

Avec $\hat{f}_i \equiv \hat{f}(x_i)$ et $f_i \equiv f(x_i)$. Ce critère ne peut cependant pas être utilisé directement vu que f est inconnue, on montre que minimiser M revient à minimiser le critère suivant appelé critère de **validation croisée généralisée** (**GCV**) :

$$\mu_g = \frac{n \sum_{i=1}^{n} (y_i - \hat{f}_i)^2}{[n - tr(A)]^2} \tag{4.13}$$

Avec:

- A la **matrice chapeau** associée au modèle, et tr la trace.

Le paramètre λ optimal sera celui qui minimisera ce critère μ_g.

En définitive, on est donc en mesure de calibrer un modèle **additif simple univarié** (estimation de la fonction lisse et détermination du paramètre de lissage optimal). Nous allons par la suite discuter des modèles **additifs multivariés** avant de passer un cran au dessus avec les **modèles additifs généralisés**.

4.2.2 Modèles additifs multivariés

Approche simplifiée

On considère un cas simple avec 2 variables explicatives x et v, considérons un modèle additif avec la structure suivante :

$$y_i = \alpha + f(x_i, v_i) + \epsilon_i = \alpha + f_1(x_i) + f_2(v_i) + \epsilon_i \tag{4.14}$$

Avec :

- α une constante à estimer
- f_1 , f_2 les fonctions lisses à estimer
- ϵ_i une variable aléatoire de loi $\mathcal{N}(0, \sigma^2)$

Cette représentation pour un modèle additif **multivarié** apporte une contrainte supplémentaire que l'on n'avait pas avec le modèle à une variable, il s'agit d'un problème d'**identifiabilité** du modèle. Ce problème s'illustre par le constat suivant :

$$\hat{y}_i = \alpha + \hat{f}_1(x_i) + \hat{f}_2(v_i) = \alpha + (\hat{f}_1(x_i) + c) + (\hat{f}_2(v_i) - c) = \alpha + \hat{f}_1^c(x_i) + \hat{f}_2^c(v_i)$$

Avec c un réel quelconque. Il y a donc une infinité d'estimations possibles de f_1 et f_2 (vu qu'il faut juste ajouter et retrancher la même constante aux 2 fonctions estimées) qui aboutissent à la même prédiction $\hat{y}$. Il faut donc imposer des contraintes d'identifiabilité au modèle avant de pouvoir l'estimer. Chacune des deux fonctions peut être représentée dans la base de fonctions linéaires par morceaux définie au niveau de l'Équation 4.5.

$$f_1(x) = \sum_{j=1}^{k_1} a_j(x)\delta_j \ ; \ f_2(v) = \sum_{j=1}^{k_2} b_j(v)\gamma_j \tag{4.15}$$

Avec :

- δ_j et γ_j les coefficients inconnus à estimer
- $a_j(x)$ et $b_j(v)$ les fonctions de bases linéaires par morceaux associées aux variables x et v.

La contrainte qui sera retenue ici pour régler le problème d'identifiabilité du modèle est la suivante :

$$\sum_{i=1}^{n} f_1(x_i) = 0 \tag{4.16}$$

Cette contrainte est prise en compte en apportant quelques modifications à la **matrice de Design** du modèle. une fois ces modifications apportées (Voir le livre de Wood pour plus de détails), le modèle à 2 variables s'écrit sous la forme matricielle suivante :

$$Y = X\beta + \epsilon$$

Avec : $X = (1, X_1, X_2)$ et $\beta^T = (\alpha, \delta^T, \gamma^T)$. X_1 (pour la variable x) et X_2 (pour la variable v) étant les matrices associées aux fonctions de bases a_j et b_j, avec quelques modifications sur X_1 pour tenir compte de la contrainte d'identifiabilité. Le coefficient β est alors obtenu de façon analogue au cas univarié (via une **régression pénalisée**) par minimisation de la fonction objectif suivante:

$$||Y - X\beta||^2 + \lambda_1 \beta^T S_1 \beta + \lambda_2 \beta^T S_2 \beta \tag{4.17}$$

Avec:

- λ_1 et λ_2 les paramètres de lissage associés aux variables x et v
- S_1 et S_2 les matrices de pénalité, définies de façon analogue à S dans le cas univarié

La solution β peut être calculée explicitement de façon analogue à l'Équation 4.11 , on obtient alors :

$$\hat{\beta} = (X^T X + \lambda_1 S_1 + \lambda_2 S_2)^{-1} X^T Y \tag{4.18}$$

La détermination des paramètres optimaux de lissage λ_1 et λ_2 se fait également par minimisation du critère de validation croisée généralisé (**GCV**).

Les résultats présentés ici pour un modèle avec 2 variables explicatives se généralisent aisément pour trois et plus de variables explicatives.

Dans le but de simplifier les représentations, une hypothèse forte a été faite sur la forme additive des effets lisses des variables x et v ($f(x, v) = f_1(x) + f_2(v)$), rien en effet n'impose que la fonction f soit de cette forme là. Nous allons présenter par la suite un moyen d'estimer f en s'affranchissant de cette hypothèse.

" Thin plate regression splines "

Pour estimer une fonction lisse d'une variable explicative, il est nécessaire comme nous l'avons vu jusqu'ici de disposer de fonctions de base. Il a été présenté à cet effet la base de **fonctions linéaires par morceaux**, ce type de base n'est cependant pas adapté pour l'estimation de fonctions lisses multivariées à moins de faire (comme ci dessus) une hypothèse forte sur la forme de cette fonction.

Les **thin plate spline** apportent une solution à ce problème d'estimation de fonctions lisses de plusieurs covariables. L'idée originelle présentée dans Duchon (1977) était la suivante : estimer une fonction lisse $g(x)$ à partir de n observations bruitées (y_i, x_i) avec $y_i = g(x_i) + \epsilon_i$, ici x_i étant un vecteur de covariables de dimension d quelconque et ϵ un terme d'erreur. L'estimation de g se faisait alors en recherchant la fonction f qui minimise :

$$||y - f||^2 + \lambda J_{md}(f) \tag{4.19}$$

Avec y le vecteur des y_i et $f = [f(x_1), f(x_2), ..., f(x_n)]^T$. $J_{md}(f)$ est une fonction de pénalité mesurant l'irrégularité de f , et λ le paramètre de lissage. On a :

$$J_{md} = \int_{\mathcal{R}^d} \sum_{V_1+...+V_d=m} \frac{m!}{V_1!...V_d!} \left(\frac{\partial^m f}{\partial x_1^{V_1} ... \partial x_d^{V_d}} \right)^2 dx_1 ... dx_d. \tag{4.20}$$

m est une constante entière à déterminer, de sorte que $2m > d + 1$, généralement il est choisi $E(\frac{d+1}{2}) + 1$ avec E l'opérateur partie entière.

Il est possible de montrer que la fonction f qui minimise l'Équation 4.19 est de la forme :

$$\hat{f}(x) = \sum_{i=1}^{n} \delta_i \eta_{md}(||x - x_i||) + \sum_{j=1}^{M} \alpha_j \phi_j(x) \tag{4.21}$$

Avec :

- δ et α des coefficients à estimer, δ étant soumis à une contrainte linéaire du type $T^T\delta = 0$ avec $T_{ij} = \phi_j(x_i)$.
- $M = \frac{(m+d-1)!}{d!(m-1)!}$

- Les ϕ_i des fonctions polynomiales linéairement indépendantes formant une base de l'espace des polynômes de degré inférieur à m dans $\mathcal{R}^d$. De plus , $J_{md}(\phi_i) = 0$, c'est à dire que les ϕ_i sont des fonctions complètement lisses du point de vu de la pénalité J_{md}. Par exemple, pour $m = d = 2$, on a : $\phi_1(x) = 1,\ \phi_2(x) = x_1,\ \phi_3(x) = x_2$, sachant que $x = (x_1, x_2)$.

$$\eta_{md} = \begin{cases} \frac{(-1)^{m+1+d/2}}{2^{2m-1}\pi^{d/2}(m-1)!(m-d/2)!} r^{2m-d}\log(r) & \text{si d est pair} \\ \frac{\Gamma(d/2-m)}{2^{2m}\pi^{d/2}(m-1)!} r^{2m-d} & \text{si d est impair} \end{cases}$$

Cette version originelle des **thin plate spline** nécessitait donc pour le calcul de $\hat{f}$ d'estimer autant de paramètres δ_i que d'observations (n), ce qui devient très vite rédhibitoire en terme de coût de calcul.

Pour contourner cette difficulté on a recours à des nœuds multidimensionnels $\{x_i^* : i = 1...k\}$, une approximation de f est alors obtenue par :

$$\hat{f}(x) = \sum_{i=1}^{k} \delta_i \eta_{md}(||x - x_i||) + \sum_{j=1}^{M} \alpha_j \phi_j(x) \tag{4.22}$$

On obtient ainsi un problème de dimension plus raisonnable. Les paramètres $\delta = (\delta_1, \delta_2, ..., \delta_k)^T,\ \alpha = (\alpha_1, \alpha_2, ..., \alpha_M)^T$ sont obtenus par minimisation de l'objectif suivant :

$$||y - X\beta||^2 + \lambda \beta^T S \beta \text{ sous contrainte que } C\beta = 0 \tag{4.23}$$

Avec :

- $\beta^T = (\delta^T, \alpha^T)$
- X une matrice carrée de taille $k + M$ telle que :

$$X_{ij} = \begin{cases} \eta_{md}(||x_i - x_j^*||) & j = 1, ..., k \\ \phi_{j-k}(x_i) & j = k+1, ..., k+M \end{cases}$$

- S est une matrice carrée de taille $k + M$ avec des zéros partout sauf dans le bloc supérieur gauche de taille $k \times k$ pour lequel on a : $S_{ij} = \eta_{md}(||x_i^* - x_j^*||)$
- C est une matrice de taille $M \times (k + M)$ telle que :

$$C_{ij} = \begin{cases} \phi_i(x_j^*) & j = 1, ..., k \\ 0 & j = k+1, ..., k+M \end{cases}$$

Nous sommes donc à présent capable d'estimer une fonction lisse multidimensionnelle sans faire d'hypothèse trop restrictive sur la forme de cette dernière. Cette méthode peut bien évidemment être aussi utilisée dans le cas unidimensionnel. C'est d'ailleurs elle qui est implémentée par défaut dans le package *R* `mgcv` que nous utiliserons pour calibrer les modèles additifs généralisés (**GAM**).

Cette approche a cependant une limite qui peut être rédhibitoire selon la nature des covariables en présence. En effet, lorsque les covariables ont des **ordres de grandeurs très différents**, il devient nécessaire d'avoir un niveau de lissage différent suivant chacune des dimensions (covariables) de la fonction lisse multidimensionnelle. Or, l'approche par les **thin plate regression splines** n'autorise qu'un seul niveau de lissage (paramètre unique λ) pour toutes les dimensions de la fonction. Dans ce cas de figure, il est possible de recourir à une approche via les **produits tensoriels** qui autorise un paramètre de lissage pour chaque dimension du problème. Le lecteur intéressé pourra se référer au chapitre 5 de Wood (2017) pour plus de détails concernant cette approche.

Maintenant que le cadre d'estimation des modèles **additifs simples** a été posé, nous allons passer par la suite à la présentation de leur généralisation aux modèles **additifs généralisés**.

4.3 Les modèles additifs généralisés

Les modèles **additifs généralisés** (**GAM**) sont une extension des modèles **additifs simples** à des familles de distribution de la réponse autres que la loi normale. La relation entre ces deux types de modèles est ainsi parfaitement analogue à la relation entre les modèles **linéaires** classiques et les **GLM**.

4.3.1 Représentation du modèle et estimation des paramètres

Un modèle additif généralisé (GAM) standard s'écrit sous la forme suivante :

$$g(\mu_i) = A_i\gamma + \sum_j^p f_j(x_{ij}),\ Y_i \sim EF(\mu_i, \phi) \tag{4.24}$$

Avec :

- $\mu_i \equiv \mathbb{E}(Y_i)$, Y_i étant la variable d'intérêt de loi appartenant à la famille exponentielle
- g une fonction de lien
- A_i la i^{ime} ligne de la matrice de design de la partie paramétrique du modèle , et γ le vecteur de paramètres correspondant
- f_j une fonction lisse de la covariable x_j
- p le nombre de covariables
- $EF(\mu_i, \phi)$ indique une famille de distribution exponentielle de moyenne μ_i et de paramètre d'échelle ϕ .

Par la suite, il est choisit une base de fonctions lisses ainsi que les niveaux de pénalisation pour chacune des fonctions f_j, ceci étant matérialisé par les matrices de **modèle** $X^{[j]}$ et de **pénalité** $S^{[j]}$. Si $b_{jk}(x)$ est la k^{ime} fonction de base pour f_j, alors $X^{[j]}_{ik} = b_{jk}(x_{ji})$.

Des contraintes d'identifiabilité de la forme $\sum_i f_j(x_{ji}) = 0$ sont appliquées aux estimations des fonctions lisses. L'application de ces contraintes se traduit par une **reparamétrisation** dans la base des fonctions lisses. Désignons par $X^{[j]}_*$ et $S^{[j]}_*$ les matrices de modèle et de pénalité après la reparamétrisation. La matrice A et les $X^{[j]}_*$ sont alors fusionnées par colonnes pour obtenir une matrice de design globale du modèle :

$$X = (A : X^{[1]}_* : X^{[2]}_* : \ ...)$$

Le vecteur de paramètres β correspondant à cette matrice de design globale contient γ et les coefficients associés aux différentes fonctions de bases. La pénalité globale associée au modèle peut alors s'écrire sous la forme suivante :

$$\sum_j^p \lambda_j \beta^T S_j \beta \tag{4.25}$$

Avec :

- λ_j le paramètre de lissage associé à f_j
- S_j est la matrice obtenue en intégrant $S^{[j]}_*$ comme un bloc diagonal et en complétant partout ailleurs avec des 0 de telle sorte que $\lambda_j \beta^T S_j \beta$ soit la pénalité associée à f_j.

Le modèle au final peut ainsi s'écrire comme un GLM sur-paramétré :

$$g(\mu_i) = X_i\beta,\ Y_i \sim EF(\mu_i, \phi) \tag{4.26}$$

Les paramètres du modèle sont estimés par maximisation de la fonction de **logvraisemblance pénalisée** suivante :

$$l_p(\beta) = l(\beta) - \frac{1}{2\phi}\sum_{j}^{p} \lambda_j \beta^T S_j \beta \tag{4.27}$$

Avec : $l(\beta)$ la log-vraisemblance associée à la distribution $EF(\mu, \phi)$ et aux observations $y_1, ..., y_n$ de la variable réponse Y.

Il est bien évidemment possible d'utiliser des GAM avec des fonctions lisses multivariées ($g(\mu_i) = A_i\gamma + f(x_{i1}, ..., x_{ip})$). Dans ce cas, les matrices de design X et de pénalité S seront telles que décrites dans le cadre des **thin plate regression spline** ou de l'approche par les **produits tensoriels** selon que l'on souhaite ou pas des niveaux de lissage différents pour chaque covariables.

En pratique, cette fonction de vraisemblance pénalisée est maximisée (pour des λ_j donnés) numériquement par la méthode itérative des moindres carrés pénalisés (**PIRLS**) qui se décline comme suit:

1. Initialiser $\hat{\mu}_i = y_i + \delta_i$ et $\hat{\eta}_i = g(\hat{\mu}_i)$, avec δ_i qui généralement vaut 0 mais qui peut être une constante de valeur assez faible permettant d'assurer que $\hat{\eta}_i$ soit fini.
2. Calculer des pseudo-données $z_i = g^{'}(\hat{\mu}_i)(y_i - \hat{\mu}_i)/\alpha(\hat{\mu}_i) + \hat{\eta}_i$ et des poids $w_i = \alpha(\hat{\mu}_i)/\{g^{'}(\hat{\mu}_i)^2 V(\hat{\mu}_i)\}$.
3. Trouver $\hat{\beta}$ qui minimise l'objectif suivant :

$$||z - X\beta||^2_W + \sum_{j}^{p} \lambda_j \beta^T S_j \beta$$

Mettre à jour ensuite $\hat{\eta} = X\hat{\beta}$ et $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$.

Les étapes 2 et 3 sont itérées jusqu'à la convergence. Pour plus de détails voir le chapitre 3 de Wood (2017).

Le problème de régression pénalisée de l'étape 3 est exactement le même problème résolu à l'Équation 4.17 pour les modèles additifs simples.

Avec :

- $||a||^2_W = a^T W a$, W étant la matrice diagonale telle que $W_{ii} = w_i$

- $V(\mu)$ la fonction de variance de la distribution de famille exponentielle
- $\alpha(\mu_i) = [1 + (y_i - \mu_i)\{V^{'}(\mu_i)/V(\mu_i) + g^{''}(\mu_i)/g^{'}(\mu_i)\}]$, cette grandeur vaut 1 pour la fonction de lien canonique.

Les paramètres de lissages optimaux (λ_j) peuvent ici aussi êtres obtenus par minimisation du critère de validation croisée généralisée μ_g présenté plus haut pour les modèles additifs simples.

Nous allons terminer par la présentation du cadre de calcul des intervalles de crédibilité pour les fonctions lisses estimées.

4.3.2 Intervalles de crédibilité

Dans une vision bayésienne du processus de lissage, il est possible de montrer qu'en interprétant le paramètre de lissage comme un a priori, la distribution a posteriori du vecteur de paramètres β est la suivante :

$$\beta|y \sim \mathcal{N}(\hat{\beta}, V_\beta) \tag{4.28}$$

Dans le cas d'une famille exponentielle, il est obtenu l'expression suivante pour la matrice de variance covariance du vecteur de paramètre β :

$$V_\beta = (X^T W X + S_\lambda)^{-1}\phi \tag{4.29}$$

Avec :

- $W = diag(w_i)$; $w_i = \alpha(\hat{\mu_i})/(g^{'}(\hat{\mu_i})^2 V(\hat{\mu_i}))$, celle de la dernière itération de l'algorithme **PIRLS**.
- $S_\lambda = \sum_j^p \lambda_j S_j$

Ce résultat nous permet de construire des intervalles de crédibilité pour les fonctions lisses estimées par le modèle. Considérons une fonction $f(x)$ estimée par le modèle, on souhaite alors fournir un intervalle de crédibilité pour cette estimation aux points d'observations de x. Si $\hat{f}$ est le vecteur contenant les estimations de $f(x)$ aux points d'observations de x, alors on peut écrire $\hat{f} = \tilde{X}\beta$. Soit $v = diag(\tilde{X} V_\beta \tilde{X}^T)$, alors $\hat{f}_i \pm z_{\alpha/2}\sqrt{v_i}$ est un intervalle de crédibilité approximatif à $(1-\alpha)100\%$ pour f_i. Avec $z_{\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de la loi normale centrée réduite.

Noter ici qu'il s'agit bien d'intervalles de crédibilité et non d'intervalles de confiance. En effet, il est démontré que $\mathbb{E}(\hat{f}) \neq \mathbb{E}(f)$, l'estimateur lisse $\hat{f}$ de la fonction f est donc biaisé, on parle de **biais de lissage**. Ce biais empêche ainsi la construction d'un intervalle de confiance au sens strict. La théorie bayésienne permet tout de même la construction des

intervalles de crédibilité présentés ci-dessus, ceux ci présentent tout de même de bonnes propriétés de couverture fréquentiste, voir Nychka (1988).

4.3.3 Diagnostic

La validation est une étape cruciale dans le processus de modélisation statistique. C'est à l'issue de celle-ci que la qualité du modèle en terme de capacité d'ajustement aux données est évaluée. Cette étape de validation passe par l'analyse des résidus du modèle. Pour les modèles additifs simples, cette analyse des résidus est relativement aisée car étant dans un cadre gaussien, on se contente de calculer les résidus bruts $(y_i - \widehat{\mu}_i)$ et de vérifier s'ils suivent approximativement une loi normale centrée réduite. Pour les modèles additifs généralisés, les résidus bruts ne peuvent plus êtres interprétés de la même manière car la distribution de la réponse n'est plus nécessairement normale. Il est généralement utilisé à la place des résidus bruts les **résidus de déviance** définis comme suit :

$$\widehat{\epsilon_i}^{\,d} = sign(y_i - \widehat{\mu}_i)\sqrt{dev_i} \tag{4.30}$$

Avec :

- dev_i la contribution de l'individu ou de la donnée i à la déviance du modèle, de sorte que la déviance totale soit $\sum_i dev_i$
- $sign$ la fonction signe

On montre que pour un modèle bien spécifié, on devrait avoir approximativement $dev_i/\phi \sim \mathcal{X}_1^2$, soit $\epsilon_i^d \sim \mathcal{N}(0, \phi)$. Ainsi, le modèle peut être considéré comme valide lorsque les résidus de déviance sont à peu près distribués comme des réalisations d'une loi normale centrée de variance ϕ.

Chapitre 5

Modélisation de la mortalité

Dans ce chapitre, nous allons nous atteler à la construction proprement dite de la table de mortalité en tenant compte du montant de rente. L'approche de modélisation principale se fera par les modèles additifs généralisés (GAM). Nous partirons de la vraisemblance des données et nous montrerons que sous l'hypothèse de **force de mortalité constante** avec pénalisation des paramètres, on se ramène à un GAM Poisson. Cette approche sera challengée par deux approches de modélisation secondaires par les modèles de Cox (qui eux aussi se ramènent à des GAM avec une hypothèse de hasard proportionnel), et les forêts aléatoires de survie.

Classiquement la construction des tables de mortalité se décompose en quatre étapes (Tomas et Planchet 2014) :

- Calcul des taux bruts de décès
- Lissage des taux bruts de décès
- Liaison des taux lissés aux tables réglementaires via des modèles relationnels
- Fermeture de la table aux grands âges

L'approche que nous adopterons ici permettra d'aboutir à une table prête à l'usage directement aux sorties de la modélisation par les GAM. Ceci grâce notamment au fait que la force de mortalité des tables réglementaires TGH/TGF 05 sera mise en **offset** dans le modèle.

5.1 Approche sous l'hypothèse de force de mortalité constante (GAM Poisson)

5.1.1 Spécification du modèle

Nous partons de la base de données au format agrégé tel que présenté dans la Table 3.1 . Nous avons donc des données agrégées par âge (x), sexe (s), année calendaire (y) et classe de montant de rente (m). Il a été ainsi calculé pour chacune des combinaisons

de ces variables explicatives le nombre de décès observés d et l'exposition centrale au risque ec. Deux variables supplémentaires sont rajoutées , la médiane des classes de montant de rente m_2 qui est un proxy numérique de m et la force de mortalité des tables réglementaires μ.

Vraisemblance du modèle

Nous reprenons l'expression de l'Équation 2.11 pour la vraisemblance associée aux données individuelles tronquées à gauche et censurées à droite :

$$\begin{aligned} L(t_1, ..., t_n; \theta) = L(\theta) &= c \prod_{i=1}^{n} (f_\theta(t_i))^{\delta_i} (S_\theta(t_i))^{1-\delta_i} \\ l(\theta) = \log(L(\theta)) &= \sum_{i=1}^{n} \left(\delta_i \log(f\theta(t_i)) + (1-\delta_i) \log(S_\theta(t_i)) \right) \\ &= \sum_{i=1}^{n} \left(\delta_i \log(h_\theta(t_i)) - \int_0^{t_i} h_\theta(t_i) \right) \end{aligned} \tag{5.1}$$

Les caractéristiques de chaque individu à la date de début d'observation consistent ici en un quadruplet (x_i, y_i, s_i, m_i) correspondant respectivement à l'âge, l'année calendaire, le sexe et la classe de montant de rente. A noter que le terme c a été ignoré ici car ne dépend pas de θ et n'est donc d'aucune utilité dans le processus de maximisation de la vraisemblance.

On supposera par la suite pour procéder à l'agrégation que la force de mortalité h est **constante par morceaux** entre deux âges entiers et deux années calendaires entières. Plus formellement, on écrit : $h_{x+\epsilon, y+\gamma, s, m} = h_{x,y,s,m}$ pour tout couple $(x, y) \in \mathbb{N}^2$ et tout couple $(\epsilon, \gamma) \in [0, 1[^2$. L'expression de l'Équation 5.1 se réécrit alors comme suit (Biessy 2022) :

$$
\begin{aligned}
l(\theta) &= \sum_{i=1}^{n} \left(\delta_i \log(h_{x_i+t_i,y_i+t_i,s_i,m_i}(\theta)) - \int_0^{t_i} h_{x_i+t_i,y_i+t_i,s_i,m_i}(\theta) \right) \\
&= \sum_{i=1}^{n} \left(\left[\sum_{x,y,s,m} 1_{(x\leq x_i+t_i<x+1, y\leq y_i+t_i<y+1, s_i=s, m_i=m)} \right] \delta_i \log(h_{x_i+t_i,y_i+t_i,s_i,m_i}(\theta)) \right. \\
&\left. - \int_{u=0}^{t_i} \left[\sum_{x,y,s,m} 1_{(x\leq x_i+u<x+1, y\leq y_i+u<y+1, s_i=s, m_i=m)} \right] h_{x_i+u,y_i+u,s_i,m_i}(\theta) du \right) \\
&= \sum_{x,y,s,m} \left(\log(h_{x,y,s,m}(\theta)) \sum_{(x,y,s,m} 1_{(x\leq x_i+t_i<x+1, y\leq y_i+t_i<y+1, s_i=s, m_i=m)} \delta_i \right. \\
&\left. - h_{x,y,s,m}(\theta) \sum_{i=1}^{n} \int_{u=0}^{t_i} 1_{(x\leq x_i+u<x+1, y\leq y_i+u<y+1, s_i=s, m_i=m)} du \right)
\end{aligned}
$$

Et finalement on obtient :

$$
l(\theta) = \sum_{x,y,s,m} \left(d_{x,y,s,m} \log(h_{x,y,s,m}(\theta)) - h_{x,y,s,m}(\theta) e^c_{x,y,s,m} \right) \tag{5.2}
$$

En notant, pour des individus de sexe s, de classe de montant de rente m, d'âge compris entre x et $x+1$ et pour une année calendaire comprise entre y et $y+1$:

- $d_{x,y,s,m} = \sum_{i=1}^{n} 1_{(x\leq x_i+t_i<x+1, y\leq y_i+t_i<y+1, s_i=s, m_i=m)} \times \delta_i$: le nombre de décès observés (la variable d dans notre base agrégée)
- $e^c_{x,y,s,m} = \sum_{i=1}^{n} \int_{u=0}^{t_i} 1_{x\leq x_i+u<x+1, y\leq y_i+u<y+1, s_i=s, m_i=m} du$: le nombre total d'années d'observations appelé exposition centrale au risque (la variable ec dans notre base agrégée)

On obtient donc une nouvelle expression de la logvraisemblance associée au format de données agrégées. Un dernier élément à préciser ici pour être complet est la forme paramétrique de la fonction de hasard $h(\theta)$. On supposera ici une forme log-linéaire : $h(\theta) = \exp(X\theta)$ avec X la matrice des variables explicatives encore appelée matrice de modèle ou matrice de design.

Lien avec un GAM Poisson

Il est possible d'opérer une liaison entre la logvraisemblance obtenue précédemment et le cadre des modèles additifs généralisés (GAM).

On considère D une variable aléatoire dont d (le nombre de décès) est une réalisation. On suppose que D suit une loi de Poisson dont le paramètre est proportionnel à la

force de mortalité $h(\theta)$ et à l'exposition centrale au risque e^c. Plus formellement, on a : $D|e^c \sim \mathcal{P}(h(\theta)e^c)$.

Posons $\chi = (x, y, s, m)$, la logvraisemblance associée à ce modèle s'écrit alors comme suit :

$$\begin{aligned} l(\theta) &= \log\left(\prod_{\chi} \mathbb{P}(D_\chi = d_\chi)\right) = \log\left(\prod_{\chi} \exp(-h_\chi(\theta)e^c_\chi)\frac{(h_\chi(\theta)e^c_\chi)^{d_\chi}}{d_\chi!}\right) \\ &= \sum_{\chi}\left(-h_\chi(\theta)e^c_\chi + d_\chi \log(h_\chi(\theta)) + d_\chi \log(e^c_\chi) + \log(d_\chi!)\right) \end{aligned} \tag{5.3}$$

En faisant abstraction des deux derniers termes de cette expression qui ne dépendent pas de θ et donc n'interviennent pas dans la maximisation de la logvraisemblance, on obtient la même expression de la logvraisemblance de l'Équation 5.2. Les deux estimateurs sont donc parfaitement équivalents. Pour terminer, il faudra bien évidemment rajouter aux expressions de la logvraisemblance des Équation 5.2 et Équation 5.3 le terme de pénalisation du vecteur de paramètres θ pour se ramener à la logvraisemblance pénalisée de l'Équation 4.27, et ainsi établir une équivalence parfaite avec un GAM Poisson. C'est donc cette logvraisemblance pénalisée qui sera effectivement maximisée pour estimer θ.

Sur le plan opérationnel, cette équivalence nous permettra de calibrer rapidement le modèle avec pénalisation des paramètres via un GAM poisson en utilisant le package *R* `mgcv`.

La notion d'offset dans un modèle statistique

Mettre en **offset** une variable explicative dans un modèle statistique revient à l'inclure dans le modèle tout en fixant d'emblée le coefficient qui lui est associé à la valeur 1. Ce procédé est généralement utilisé dans la modélisation des processus de comptage. Dans le cadre des études de mortalité, il est souvent observé des décès par âge. Il est logique de penser que le nombre de décès observés pour un âge donné est d'autant plus grand que le nombre total d'individus observés (exposition) à cet âge là est élevé. Dans ce cas de figure, il est plus approprié de modéliser un nombre de décès par unité d'observation (taux de décès) plutôt qu'un nombre de décès absolu qui lui dépend fortement du nombre d'individus observés. Pour modéliser un processus de comptage comme un taux, on place la variable d'exposition en offset dans le modèle (Parry 2018).

Dans le cadre de ce mémoire et de l'approche avec les modèles GAM, le nombre de décès est modélisé en supposant pour ceux-ci une distribution de Poisson. Sachant que la fonction de lien pour une distribution de Poisson est la fonction **logarithme népérien**, le modèle utilisé incluant l'exposition en offset s'écrit comme suit :

$$\log(d) = 1 \times log(ec) + \beta_0 + \beta_1 \times s + \beta_2 \times x + \beta_3 \times m \tag{5.4}$$

Ce qui est équivalent à :

$$\log(\frac{d}{ec}) = \beta_0 + \beta_1 \times s + \beta_2 \times x + \beta_3 \times m \tag{5.5}$$

En remarquant que $\frac{d}{ec}$ représente l'estimateur du maximum de vraisemblance des taux bruts de mortalité.

Mais comme précisé plus haut, en plus de l'exposition, la force de mortalité des tables réglementaires sera également mise en offset dans le modèle. Ceci afin d'imprimer la dynamique d'évolution de la mortalité des tables réglementaires dans la modélisation de celle du portefeuille. On obtient alors l'écriture suivante pour le modèle :

$$\log(d) = 1 \times log(ec) + 1 \times \log(\mu) + \beta_0 + \beta_1 \times s + \beta_2 \times x + \beta_3 \times m \tag{5.6}$$

ce qui revient à :

$$\log(\frac{d}{ec \times \mu}) = \beta_0 + \beta_1 \times s + \beta_2 \times x + \beta_3 \times m \tag{5.7}$$

ou encore :

$$\frac{d}{ec} = \mu \times \exp(\beta_0 + \beta_1 \times s + \beta_2 \times x + \beta_3 \times m) \tag{5.8}$$

Les coefficients (passés à l'exponentielle) estimés par un modèle spécifié de la sorte s'interprètent comme des écarts de la force de mortalité du portefeuille par rapport à la force de mortalité des tables réglementaires. En d'autres termes, ce sont des coefficients multiplicatifs à appliquer à la force de mortalité des tables réglementaires pour aboutir à la force de mortalité du portefeuille étudié.

5.1.2 Modélisation de la mortalité du portefeuille

A titre de rappel, les modèles additifs généralisés à la manière des GLM classiques permettent de modéliser via une fonction de lien une variable d'intérêt (d'une loi de famille exponentielle) comme fonction linéaire d'un ensemble de **fonctions lisses** des variables explicatives.

$$g(E(Y)) = \alpha + f_1(X_1) + f_2(X_2) + f_3(X_3) + ... \tag{5.9}$$

Avec Y la variable d'intérêt de loi appartenant à la famille exponentielle et g la fonction de lien associée; X_i les variables explicatives et f_i les fonctions lisses associées.

Ici, nous allons calibrer des modèles sur le nombre de décès, en supposant que ce dernier suit une loi de Poisson. Nous mettrons en offset **l'exposition centrale au risque** et la **force de mortalité** des tables réglementaires. Pour une loi de Poisson, la fonction de lien g est la fonction **logarithme népérien**.

Trois modèles différents seront calibrés :

- Un modèle avec un lissage sur l'âge et avec un traitement catégoriel du montant de rente
- Un modèle avec un lissage sur l'âge et sur les médianes des classes de montant de rente
- Un modèle avec un lissage sur l'âge, les médianes de classes de montant de rente et prise en compte de l'interaction entre ces deux variables

Ces modèles seront enfin comparés au regard du critère **AIC**.

Modèle avec le montant de rente comme variable catégorielle simple et lissage sur l'age

Le modèle calibré s'écrit comme suit :

$$\log(E(D) = d) = \alpha + S(x) + \beta_1 \times m + \beta_2 \times s + \log(ec) + \log(\mu) \tag{5.10}$$

Avec :

- $S, \alpha, \beta_1, \beta_2$: respectivement une fonction lisse de l'âge et des coefficients qui seront estimés par le modèle;
- x, d, m, s, ec, μ : l'âge, le nombre de décès, la classe de montant de rente, le sexe, l'exposition au risque et la force de mortalité des tables réglementaires.

Effet de l'âge

Sur la Figure 5.1, on observe un effet décroissant de l'âge. Ainsi l'écart entre les décès observés dans le portefeuille et ceux prédits par les tables réglementaires tend à décroître à mesure que l'âge augmente. Une interprétation de l'effet représenté sur ce graphique est la suivante : Toutes choses égales par ailleurs, pour un individu du portefeuille âgé de 60 ans, il faudrait multiplier la force de mortalité des tables réglementaires par 132% pour obtenir la force de mortalité de cet individu dans le portefeuille.

Effet de la classe de montant de rente

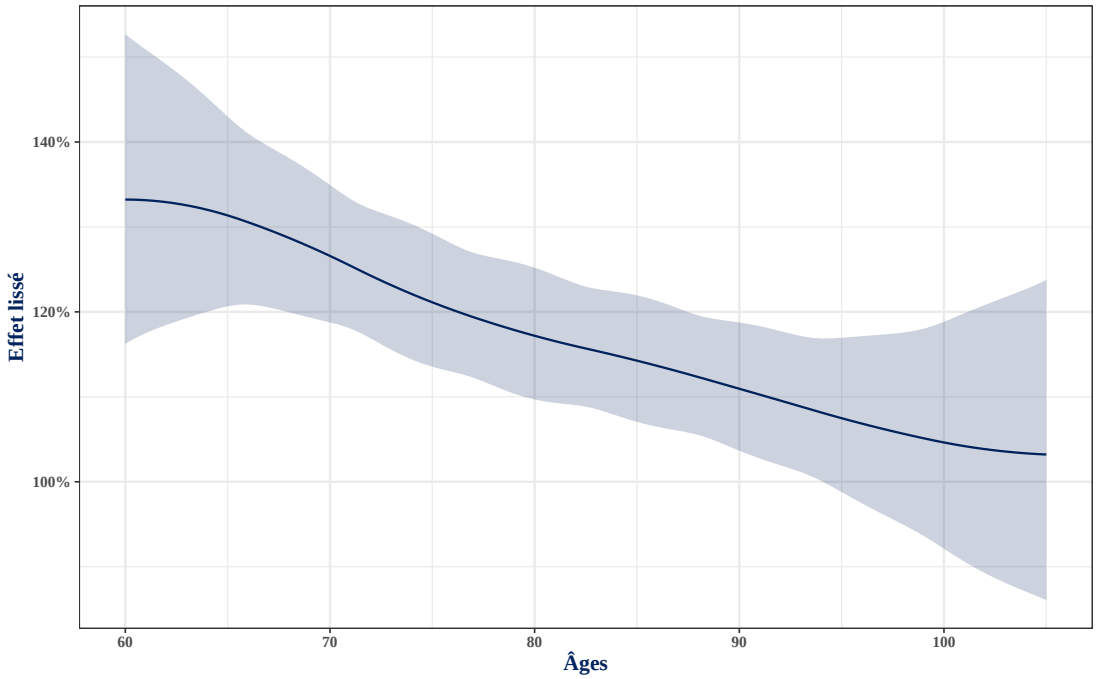

Figure 5.1 : Effet lisse de l'âge

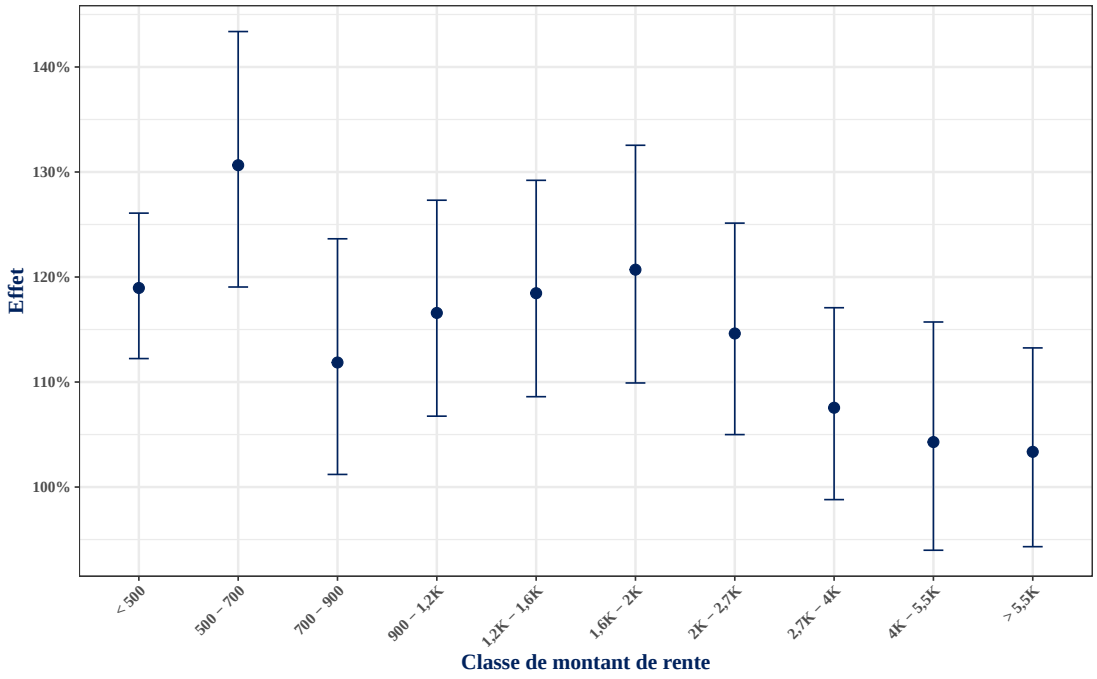

Figure 5.2 : Effet du montant de rente

Les interprétations pour la Figure 5.2 doivent là aussi se faire relativement aux tables réglementaires. Ainsi, en dehors des classes (**"500-700"** et **"1,6K-2K"**), toutes les autres classes de montant de rente semblent présenter des écarts par rapport aux tables réglementaires plus faibles que la classe **"<500"**. Il semble aussi se dégager une tendance à la décroissance de l'écart par rapport aux tables réglementaires à mesure que la classe de montant de rente augmente. Une interprétation de cet effet ici est la suivante : Toutes choses égales par ailleurs, pour un individu du portefeuille appartenant à la classe **"1,6K - 2K"**, il faudrait multiplier la force de mortalité des tables réglementaires par 120% pour obtenir la force de mortalité de cet individu dans le portefeuille.

A noter néanmoins qu'il faut être prudent avec l'interprétation de ces résultats vu la largeur des intervalles de confiances observés. Sur le plan statistique, en observant les p-valeurs renvoyées par ce modèle, seules les trois dernières classes de montant de rente se révèlent significatives au seuil de 5% . Pour mieux cerner l'effet de cette variable, il conviendrait de procéder à un lissage suivant les médianes de ces classes de montant de rente. Ce sera l'objet du modèle suivant.

Effet du sexe

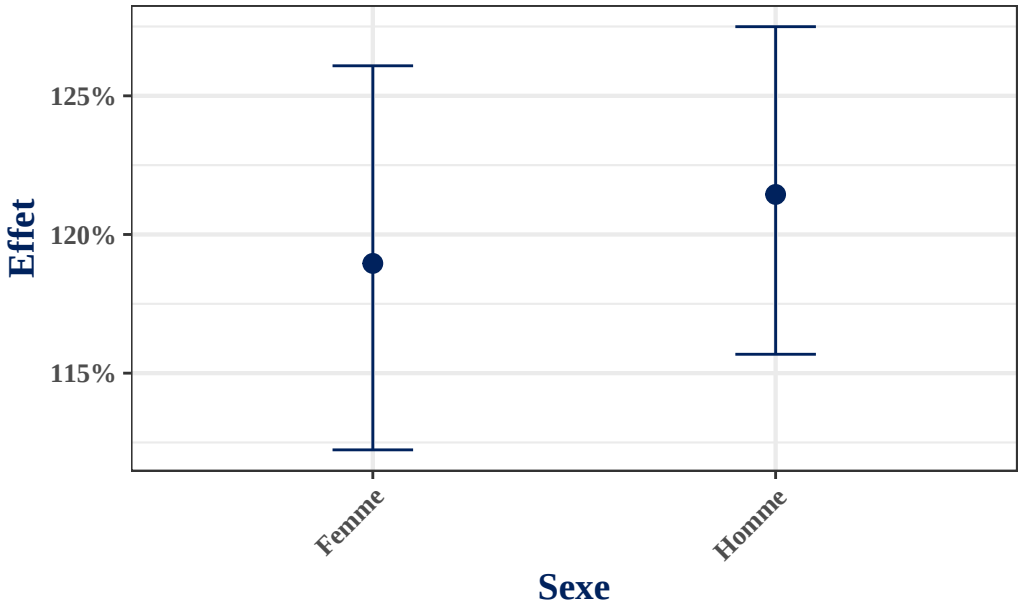

Figure 5.3 : Effet du sexe

En se référant aux p-valeurs , l'effet du sexe dans le modèle se révèle être non significatif au seuil de 5%. La Figure 5.3 annonçait déjà cette réalité vu la proximité des coefficients estimés pour les hommes et les femmes ainsi que la largeur des intervalles de confiance. Le sexe est évidement une variable très importante dans toute étude de mortalité, et il est bien connu que les femmes vivent plus longtemps en moyenne que les hommes. Il

faut bien se rappeler ici que ce n'est pas la mortalité du portefeuille de façon directe qui est modélisée, mais sa distance par rapport à la mortalité des tables réglementaires. La variable sexe apparaissant comme non significative ici signifie tout simplement que l'écart entre la mortalité du portefeuille et celle des tables réglementaires est sensiblement la même chez les hommes et chez les femmes.

Modèle avec lissage suivant les médianes des classes de montants de rente

Le modèle calibré s'écrit comme suit :

$$\log(E(D) = d) = \alpha + S_1(x) + S_2(m_2) + \beta_1 \times s + \log(ec) + \log(\mu) \tag{5.11}$$

Avec:

- S_1, S_2 des fonctions lisses à estimer pour l'âge et le montant de rente respectivement
- m_2 la médiane de la classe de montant de rente.

Effet du montant de rente

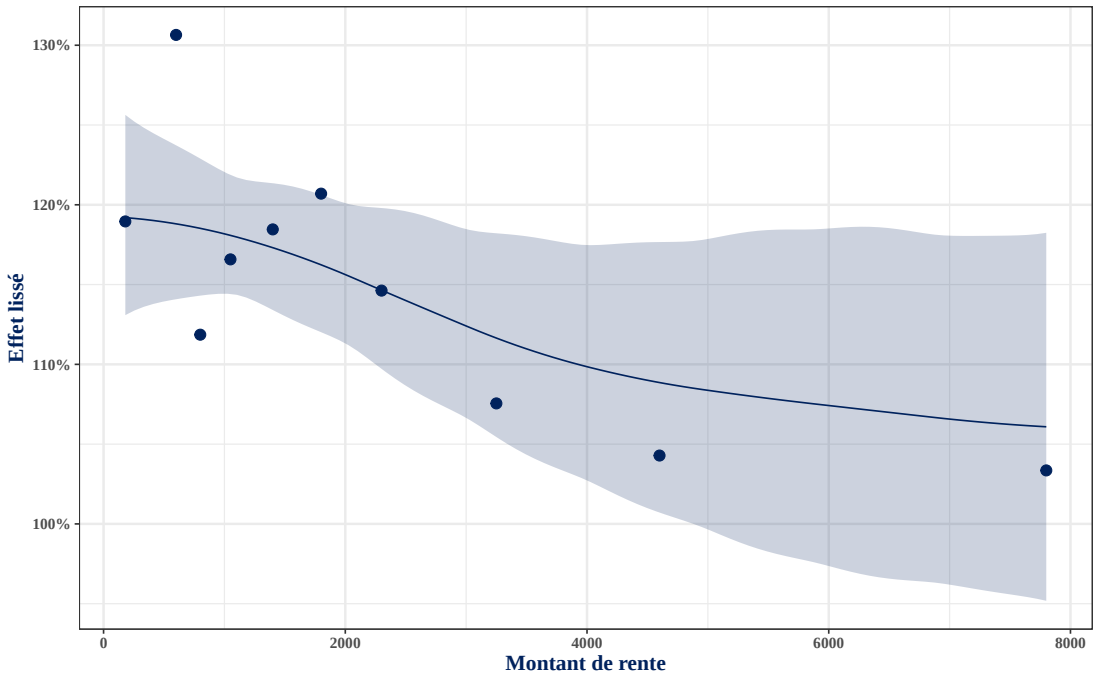

Figure 5.4 : Effet lisse du montant de rente

On constate sur la Figure 5.4 une tendance à la baisse des écarts par rapport aux tables réglementaires à mesure que la médiane de la classe de montant de rente augmente. L'effet

capturé ici est beaucoup plus net que ce que l'on observait avec le modèle précédent dans lequel la rente était traitée comme une variable catégorielle. De plus, le lissage élimine les fluctuations erratiques que l'on pouvait observer d'une classe de montant de rente à l'autre, ce qui semble plus proche de la réalité du phénomène étudié. L'effet capturé ici tend à se courber légèrement vers les montants de rente les plus élevés avec une pente de moins en moins raide. Ceci traduit une certaine atténuation de l'effet du montant de rente vers les montants élevés. Ceci est plutôt en accord avec la littérature sur le sujet qui stipule que les améliorations de mortalité liées au niveau de vie sont de plus en plus faibles à mesure que le niveau de vie augmente.

Modèle avec interaction entre l'âge et le montant de rente

Le modèle calibré s'écrit comme suit :

$$\log(E(D) = d) = \alpha + S_1(x) + S_2(m_2) + S_3(x, m_2) + \beta_1 \times s + \log(ec) + \log(\mu) \quad (5.12)$$

Avec :

- S_3 la fonction lisse à estimer pour l'interaction entre l'âge et le montant de rente, à noter que cette interaction ne comporte pas les effets principaux des variables prises individuellement.

Effet d'interaction

La Figure 5.5 présente les effets combinés des variables âge et montant de rente prises individuellement ainsi que l'effet de leur interaction. Il apparaît ici que plus l'âge augmente et plus un effet éventuel du montant de rente tend à disparaître (de moins en moins d'effets verticaux vers les grands âges sur le graphique), ce qui semble logique car au-delà d'un certain âge la mortalité est déterminée exclusivement par des facteurs biologiques liés au vieillissement.

5.1.3 Diagnostic des modèles et comparaison

Il est question ici d'évaluer la qualité de l'ajustement des trois modèles calibrés précédemment. Pour cela nous allons examiner leurs résidus de déviance. Nous partons de l'Équation 4.30 et on l'adapte au cas d'un GAM Poisson, on obtient alors l'expression suivante pour les résidus de déviance :

$$\hat{\epsilon_i}^d = sign(d_i - \hat{d}_i)\sqrt{2\left(d\log(\frac{d_i}{\hat{d}_i}) - (d - \hat{d}_i)\right)} \quad (5.13)$$

Avec :

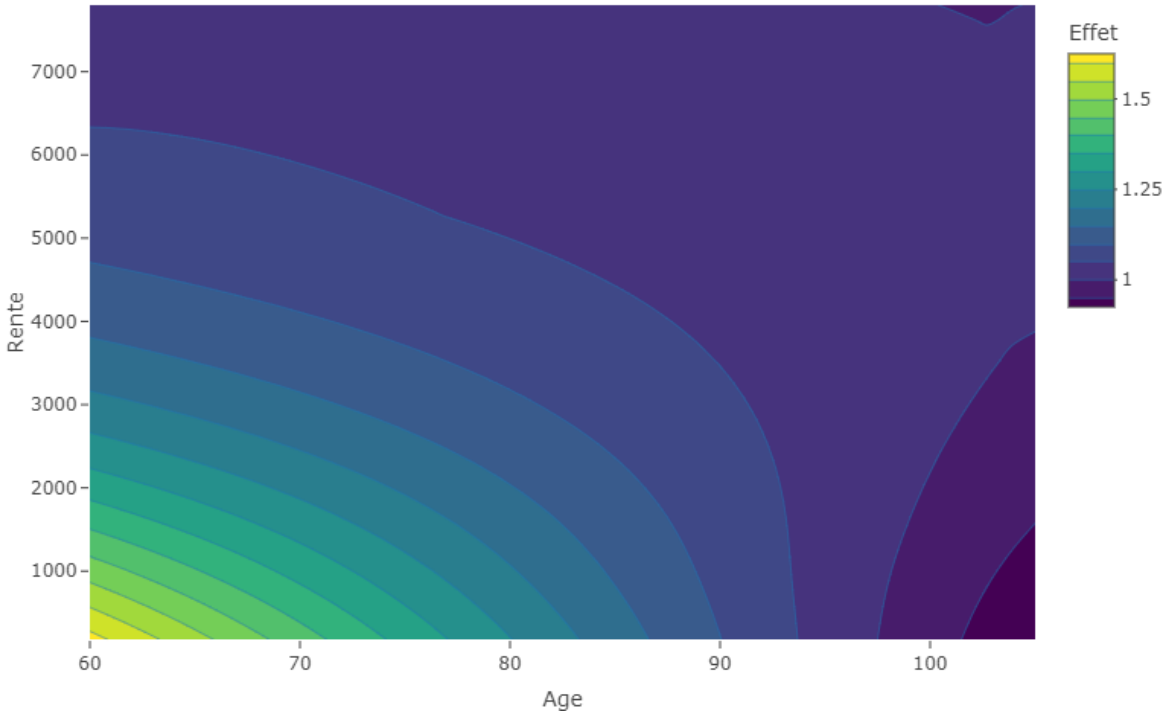

Figure 5.5 : Effet d'interaction entre âge et montant de rente

- d_i le nombre de décès observé pour la cellule i
- $\hat{d}_i$ le nombre de décès prédit par le modèle pour la cellule i
- Une cellule i donnée représente une combinaison des variables explicatives x, y, s, m

Si le modèle est bien spécifié ces résidus doivent suivre une distribution normale centrée réduite (pour une loi de Poisson, $\phi = 1$) .

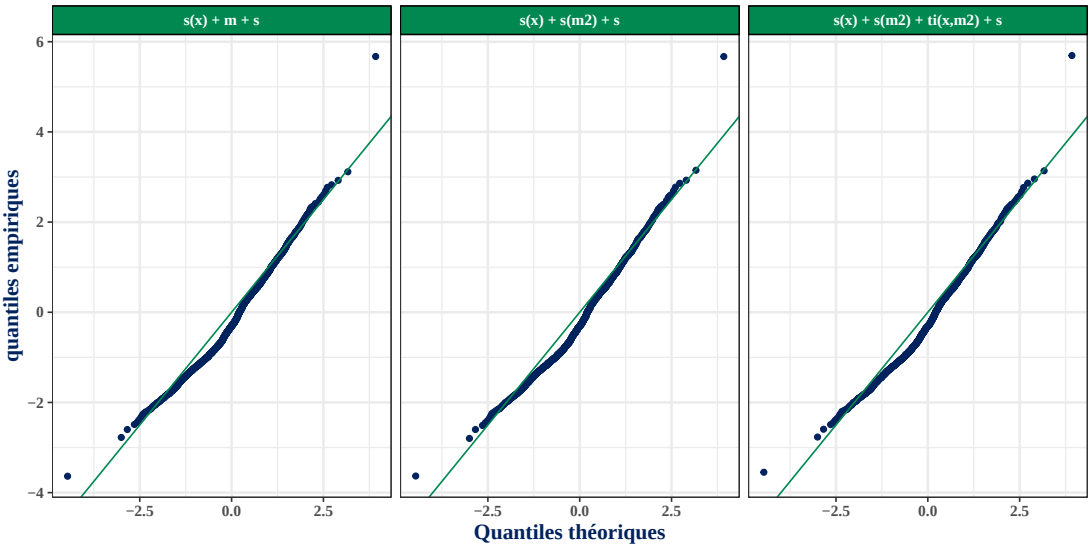

Figure 5.6 : Diagrammes quantile-quantile des résidus

Les diagrammes quantiles quantiles (Figure 5.6) des résidus des trois modèles sont assez satisfaisants et semblent accréditer l'hypothèse d'une distribution normale des résidus. A noter néanmoins pour tous les modèles une légère courbure de la distribution des résidus par rapport à la distribution normale entre -1 et 0, ainsi qu'une observation a priori aberrante.

La Figure 5.7 présente pour l'ensemble des modèles la distribution des résidus en fonction des prédicteurs linéaires. Le rendu est assez classique pour une distribution de Poisson, avec la majorité des résidus compris entre 2 et -2 . On n'observe aucune dépendance a priori entre ces résidus et les prédicteurs linéaires.

Les trois modèles présentent donc ainsi une assez bonne adéquation aux données, pour les départager nous allons les comparer suivant le critère **AIC**.

Au regard du critère **AIC**, (Figure 5.8) le modèle avec interaction semble être le meilleur, suivi du modèle avec lissage suivant l'âge et le montant de rente. Le moins bon de ces modèles est celui dans lequel la rente est traitée comme une variable catégorielle. A noter que `s()` correspond au lissage unidimensionnel et `ti()` à un effet d'interaction lisse.

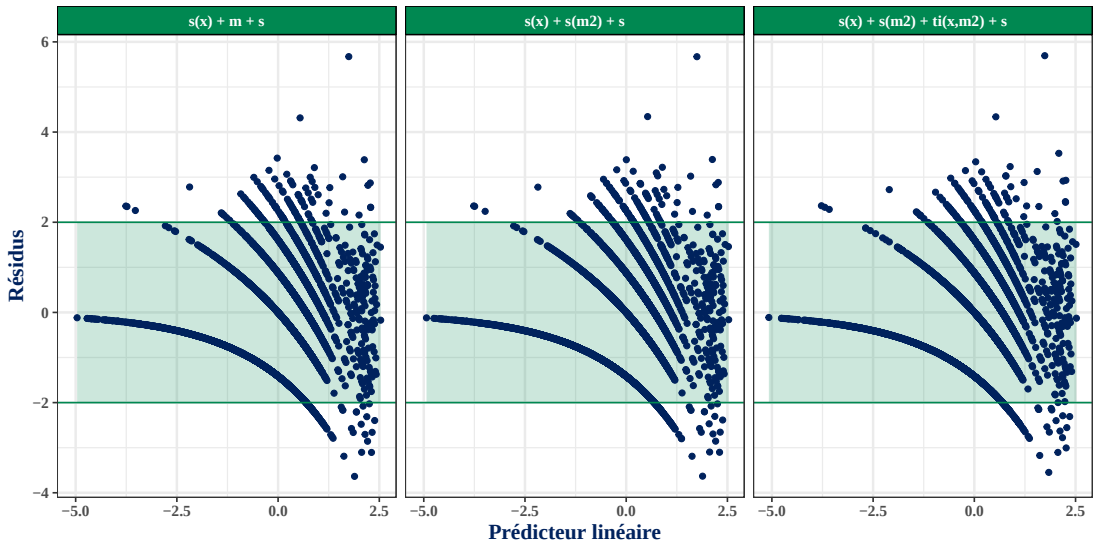

Figure 5.7 : Résidus en fonction des prédicteurs linéaires

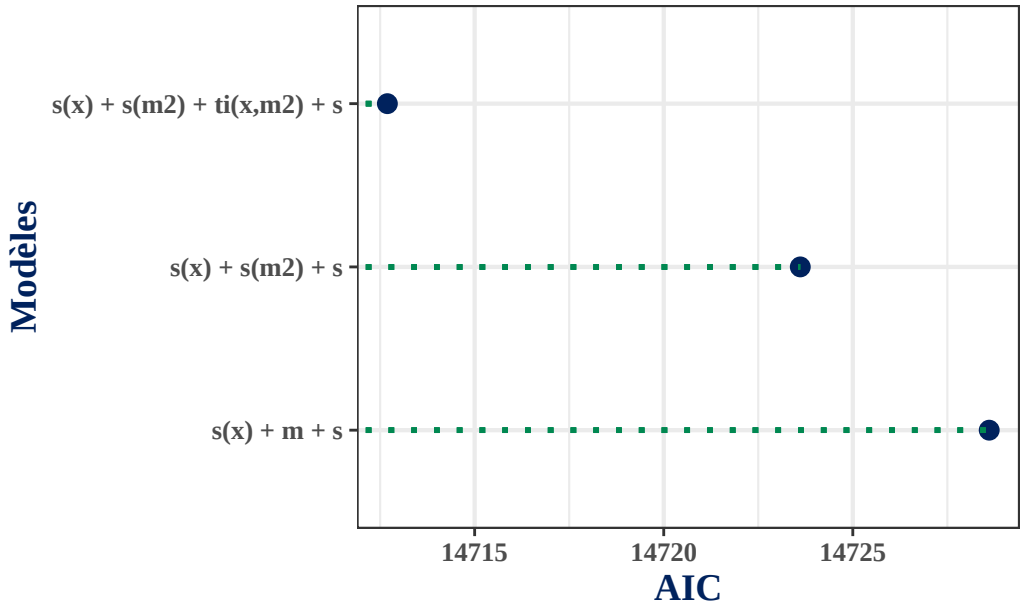

Figure 5.8 : Comparaison des modèles suivant l'AIC

5.1.4 Extrapolation aux grands âges

Comme mentionné au tout début de cette partie, l'utilisation des modèles additifs généralisés permet de court-circuiter le schéma classique de construction des tables d'expérience :

- L'aspect calcul des taux bruts de mortalité est couvert puisque le modèle explique un nombre de décès avec une variable d'exposition en offset, tout se passe comme si c'était un taux de mortalité qui était modélisé en réalité.
- L'aspect lissage est aussi pris en compte ici, vu que ces modèles nous permettent d'effectuer directement des lissages suivant l'âge et suivant le montant de rente.
- L'aspect liaison aux tables réglementaires est également traité avec le fait de mettre en offset la force de mortalité des tables réglementaires.

Nous allons aborder ici la question de la fermeture de la table aux grands ages. Lors de la construction de tables d'expériences, il n y a généralement pas assez d'observations aux grands ages pour avoir des résultats fiables sur le plan statistique. Dans l'approche classique, il est souvent utilisé des techniques telles que la méthode de *Coale & Kisker* ou encore la méthode de *Kannisto* pour extrapoler la mortalité aux grands âges, ce sujet est étudié en détail dans Quashie et Denuit (2005). L'utilisation des GAM assure en ce qui concerne les effets lisses des variables explicatives une convergence vers une limite déterminée pour les domaines où il y a un manque d'observations. Dans notre cas en ce qui concerne l'effet de l'âge, au delà de 105 ans on ne dispose clairement plus d'assez d'observations. Le prolongement de l'effet aux âges inobservés est tout de même assuré par le modèle grâce aux propriétés des fonctions lisses de base. Ce prolongement est représenté sur la Figure 5.9.

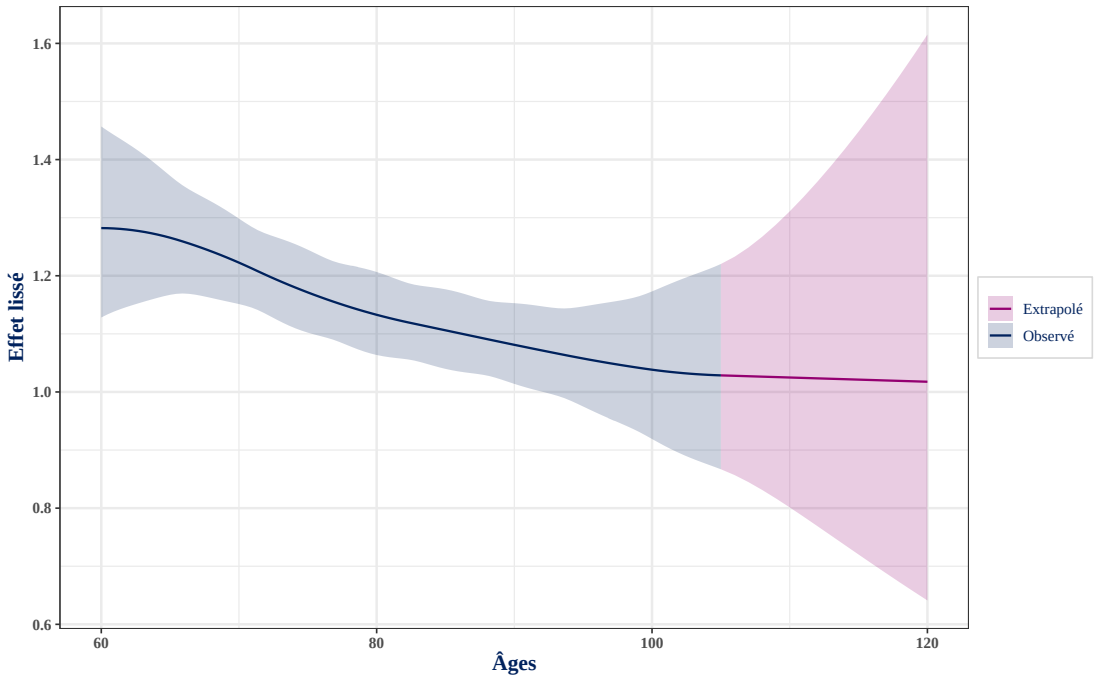

Figure 5.9 : Extrapolation de l'effet lisse de l'âge

5.2 Approche par les modèles de Cox

Nous allons à présent modéliser la mortalité de notre portefeuille au moyen d'un modèle de Cox. Il s'agit du modèle de durée de référence lorsqu'il est question d'évaluer l'influence de variables explicatives sur des durées de survie. L'idée ici est de s'assurer qu'on a des résultats équivalents avec les modèles de la partie précédente au niveau des effets des variables explicatives, notamment pour le montant de rente.

5.2.1 Présentation du modèle de Cox

Le modèle de Cox fait partie de la famille des modèles de durées **composites**, il s'agit d'une classe de modèles qui permet la prise en compte de l'hétérogénéité des populations en terme de durée de survie. Il s'agit aussi d'un modèle semi-paramétrique à **hasard proportionnel**. La fonction de hasard de ce modèle s'écrit sous la forme suivante :

$$h_X(x|z;\theta) = \exp(z^t\theta)h_0(x) \tag{5.14}$$

Avec:

- $z^t = (z_1, ..., z_p)$ un vecteur de covariables

- $\theta^t = (\theta_1, ..., \theta_p)$ un vecteur de paramètres inconnus à estimer
- $h_0(x)$ une fonction de hasard de base.

La fonction de hasard de base h_0 est inconnue ici, il faudra donc l'estimer au même titre que le vecteur de paramètres θ.

Les observations sont faites sous censure, on dispose alors de 2 échantillons indépendants $(X_1, ..., X_n)$ et $C_1, ..., C_n$, puis :

$$T_i = inf(X_i, C_i),\ D_i = 1_{(X_i \leq C_i)}$$

Avec :

- X_i la durée de survie de l'individu i
- C_i la censure pour l'individu i (Dans notre cas elle sera la même pour tous les individus)
- T_i la durée de survie effectivement observée pour l'individu i
- D_i une indicatrice qui vaut 1 si le décès de l'individu i a été effectivement observé, et 0 sinon.

L'estimation du vecteur de paramètres θ se fait par le maximum de vraisemblance en maximisant la fonction suivante :

$$L(\theta|z, h_0) = \prod_{i=1}^{n} \left(h_0(t_i) e^{\theta^t z_i} \exp\left(-H_0(t_i) e^{\theta^t z_i}\right) \right)^{d_i} \left(\exp\left(-H_0(t_i) e^{\theta^t z_i}\right) \right)^{1-d_i} \tag{5.15}$$

Avec:

- d_i la réalisation de la variable aléatoire D_i pour l'individu i
- H_0 la fonction de hasard cumulée associée à h_0
- t_i le temps de fin d'observation de l'individu i soit pour cause de décès, soit de censure.

Cette fonction cependant ne peut être maximisée lorsque h_0 et H_0 sont inconnus. Cox (1972) a proposé une fonction de vraisemblance partielle qui ne dépend plus de h_0. Cette fonction de vraisemblance partielle s'écrit alors de la manière suivante :

$$L_{Cox}(\theta|z, h_0) = \prod_{i=1}^{n} \left(\frac{e^{\theta^t z_i}}{\sum_{j=1}^{n} 1_{(t_i \leq t_j)} e^{\theta^t z_j}} \right)^{d_i} \tag{5.16}$$

La logvraisemblance $l(\theta)$ (le logarithme de la quantité ci-dessus) peut ainsi être maximisée par des méthodes numériques pour obtenir un estimateur $\hat{\theta}$ du vecteur de paramètre θ . Cet estimateur satisfait les propriétés suivantes :

- $\hat{\theta}\ _{p.s} \rightarrow \theta$: convergence
- $\sqrt{n}(\hat{\theta} - \theta)\ _{\mathcal{L}} \rightarrow \mathcal{N}(0, \Sigma)$: normalité asymptotique

Un estimateur consistant de la matrice de variance covariance Σ est donné par l'inverse de la matrice d'information de Fisher $I(\theta)$ qui est définie par :

$$I(\theta)_{ij} = \frac{\partial^2 l(\theta)}{\partial\theta_i \partial\theta_j} \tag{5.17}$$

Ces résultats servent de base à la construction d'intervalles de confiances pour l'estimateur $\hat{\theta}$.

La fonction de hasard cumulée de base est donnée par l'estimateur suivant :

$$\hat{H}_0(t) = \sum_{i:t_i^* \leq t} \frac{m_i}{\sum_{j\ \epsilon\ A_i} \exp\left(\theta^t z_j\right)} \tag{5.18}$$

Avec :

- $0 \leq t_0^* \leq t_1^* \leq\ ...\ \leq t_k^*,\ (k \leq n)$ les instants de survenance des décès
- m_i le nombre de décès observés à l'instant t_i
- A_i la population sous risque juste avant l'instant de décès t_i^*.

On déduit alors l'estimateur de Cox de la fonction de survie de X :

$$\hat{S}_X(x|z;\theta) = \exp\left(-e^{z^t\theta}\hat{H}_0(x)\right) \tag{5.19}$$

Dans le contexte des modèles additifs, il est possible de pénaliser la vraisemblance partielle de l'Équation 5.16, et donc d'intégrer à la forme semi-paramétrique de la fonction de hasard proposée par Cox des effets lisses des variables explicatives. La fonction de hasard s'écrit alors :

$$h_X(x|z;\theta) = \exp\left(z^t\theta + S_1(X_1) + S_2(X_2) + ... + S_l(X_l)\right)h_0(x) \tag{5.20}$$

Avec :

- X_i une variable explicative quantitative
- S_i une fonction lisse de la variable X_i.

L'estimation de ces fonctions lisses se fait dans le même cadre que celui présenté pour les modèles additifs généralisés. Le package `mgcv` offre la possibilité de calibrer directement un tel modèle de Cox avec pénalisation des paramètres et intégration d'effets lissés de variables explicatives.

5.2.2 Adéquation du modèle de Cox

Le modèle de Cox suppose que les risques sont proportionnels. Cette hypothèse de proportionnalité des risques stipule que l'influence des variables explicatives sur la probabilité instantanée de décès ne varie pas au cours du temps (la période d'observation). Il est nécessaire de s'assurer que cette hypothèse est vérifiée pour l'ensemble des variables explicatives lorsqu'on calibre un modèle de Cox (Vermet 2022). Une approche pour vérifier cette hypothèse est le test de **Schoenfeld,** basé sur le calcul des résidus de Schoenfeld donnés par :

$$r_{ik} = d_i \left(z_{ik} - \frac{\sum_{j \in R_i} z_{jk} \exp\left(\theta^t z_j\right)}{\sum_{j \in R_i} \exp\left(\theta^t z_j\right)} \right);\ 1 \leq i \leq n,\ 1 \leq k \leq p \tag{5.21}$$

Avec:

- R_i l'ensemble des individus sous risques au moment du décès de l'individu i ;
- z_{ik} la valeur de la k^{ime} covariable associée à l'individu i
- d_i la réalisation de l'indicatrice de décès (D_i) pour l'individu i
- p le nombre de variables explicatives.

On définit ensuite les résidus standardisés r^*_{ik} comme les résidus r_{ik} divisés par leur écart-type. Si l'hypothèse des risques proportionnels est vérifiée, alors ces résidus ne doivent présenter aucune tendance d'évolution au cours du temps. Le test de Schoenfeld consiste alors à tester la nullité du coefficient de corrélation entre les r^*_{ik} et les t_i (temps de fin d'observation de l'individu i).

5.2.3 Implémentation du modèle

Modèle de Cox original

Il sera calibré ici le modèle de Cox original (sans lissage sur les variables) avec pour variables explicatives l'âge (x), le sexe (s) et la classe de montant de rente (m) ou le montant de rente (m_2).

Modèle avec le montant de rente

La fonction de hasard modélisée ici s'écrit de la manière suivante :

$$h_X(t|(x, s, m_2); \theta) = \exp(\alpha \times x + \beta \times s + \gamma \times m_2) h_0(t) \tag{5.22}$$

Avec :

- $\theta = (\alpha, \beta, \gamma)$ le vecteur de paramètres à estimer
- $h_0(t)$ la fonction de hasard de base à estimer également.
- **Diagnostic**

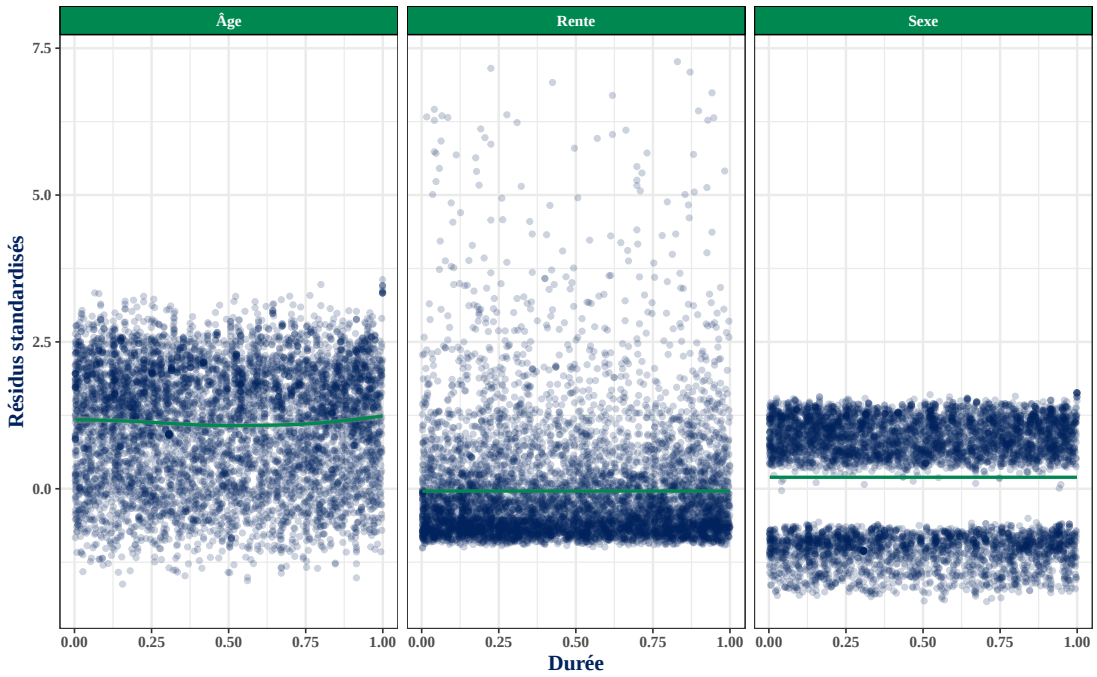

Figure 5.10 : Résidus de Schoenfeld (1)

En examinant la Figure 5.10 il ne semble pas se dégager une quelconque tendance d'évolution des résidus au cours du temps. Ce constat graphique est confirmé par un test statistique plus formel qui établit que l'hypothèse de hasard proportionnel est vérifiée pour le montant de rente le sexe, et l'âge.

- **Effets des variables explicatives**

Les 3 variables explicatives du modèle s'avèrent être significatives ici au seuil de 5%. Les effets présentés par la Figure 5.11 s'interprètent de la manière suivante :

- **Toutes choses égales par ailleurs**, une augmentation du montant de rente de 1 000 € entraîne une diminution de la probabilité instantanée de décès de $1,6\%$;

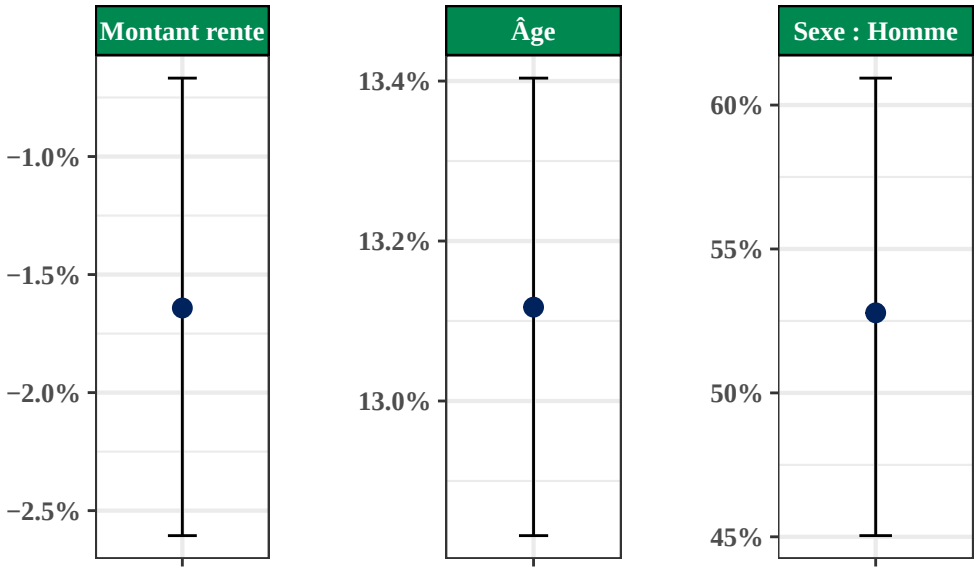

Figure 5.11 : Effets des variables explicatives

- **Toutes choses égales par ailleurs**, un individu qui vieillit d'un an voit sa probabilité instantanée de décès augmenter de $13,16\%$;
- **Toutes choses égales par ailleurs**, la probabilité instantanée de décès d'un homme est $52,5\%$ plus élevée que celle d'une femme.

Modèle avec la classe de montant de rente

La fonction de hasard modélisée ici s'écrit de la manière suivante :

$$h_X(t|(x,s,m);\theta) = \exp(\alpha \times x + \beta \times s + \gamma \times m)h_0(t) \tag{5.23}$$

- **Diagnostic**

Sur la Figure 5.12 , on observe globalement les mêmes résultats que pour le modèle précédent, l'hypothèse de hasard proportionnel est vérifiée pour la classe de montant de rente, pour le sexe, et pour l'âge.

- **Effets des variables explicatives**

Une fois ce modèle calibré, seules les trois dernières classes de montant de rente se sont révélées significatives. La Figure 5.13 s'interprète **toutes choses égales par ailleurs** en terme de variation à la baisse de la probabilité instantanée de décès lorsqu'on passe de la classe de référence qui est moins de 500 € ($<$ **500**) à une autre classe. Ainsi, toutes

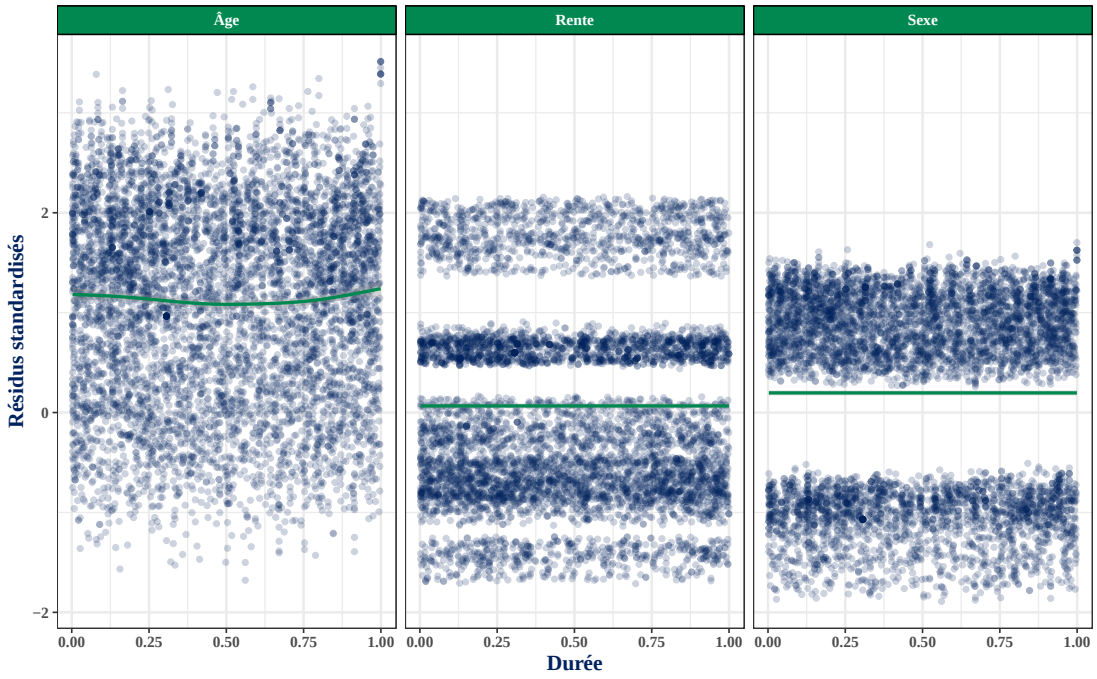

Figure 5.12 : Résidus de schoenfeld (2)

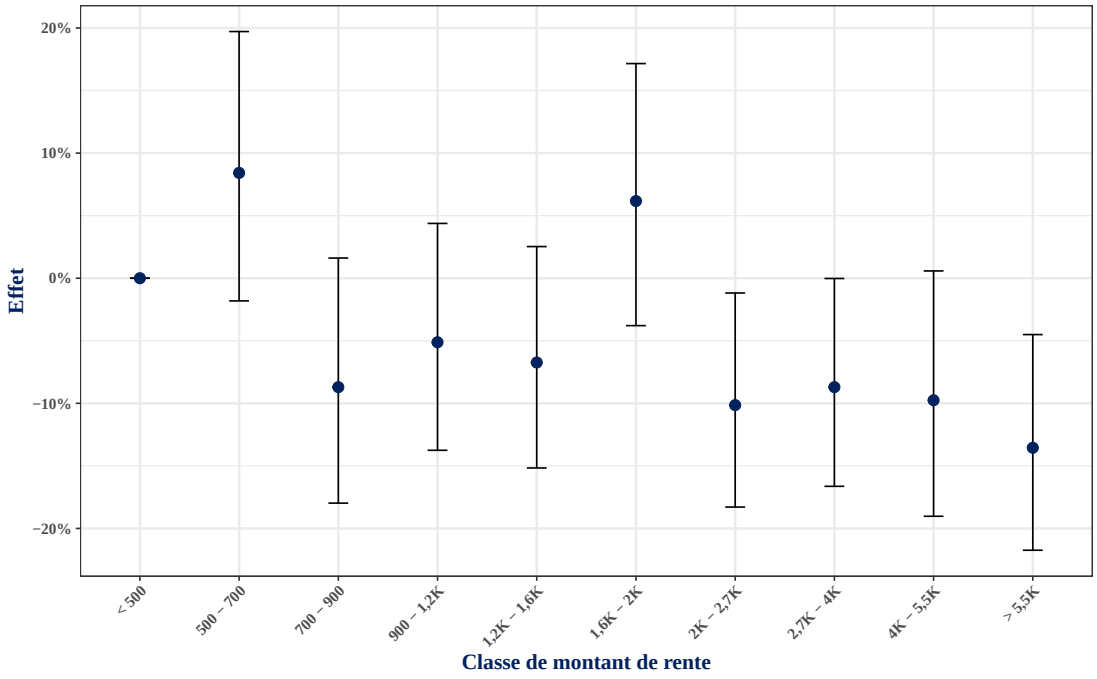

Figure 5.13 : Effet de la classe de montant de rente

choses égales par ailleurs, un individu avec plus de 5 500 € de montant de rente (**>5,5K**) a une probabilité instantanée de décès inférieure de $13,33\%$ à celle d'un individu avec moins de 500 € de montant de rente. Les effets de l'âge et du sexe restent quasiment inchangés ici par rapport au modèle précédent. A noter ici comme dans le cas des modèles GAM, un lissage suivant les médianes de classes de montant de rente pourrait clairement améliorer la prise en compte de cette variable , et de passer outre le fait que la plupart des classes ici semblent non significatives.

Modèle de Cox avec intégration de fonctions lisses des variables explicatives.

Le modèle calibré ici est un modèle de Cox avec pour variable explicative le sexe auquel il est intégré un effet lissé de l'âge et un effet lissé du montant de rente. La fonction de hasard modélisée ici s'écrit de la manière suivante :

$$h_X(t|(x,s,m_2);\theta) = \exp(S_1(x) + \beta \times s + S_2(m_2))h_0(t) \tag{5.24}$$

Avec S_1 et S_2 des fonctions lisses à estimer pour l'âge et le montant de rente respectivement.

- **Effet de l'âge**

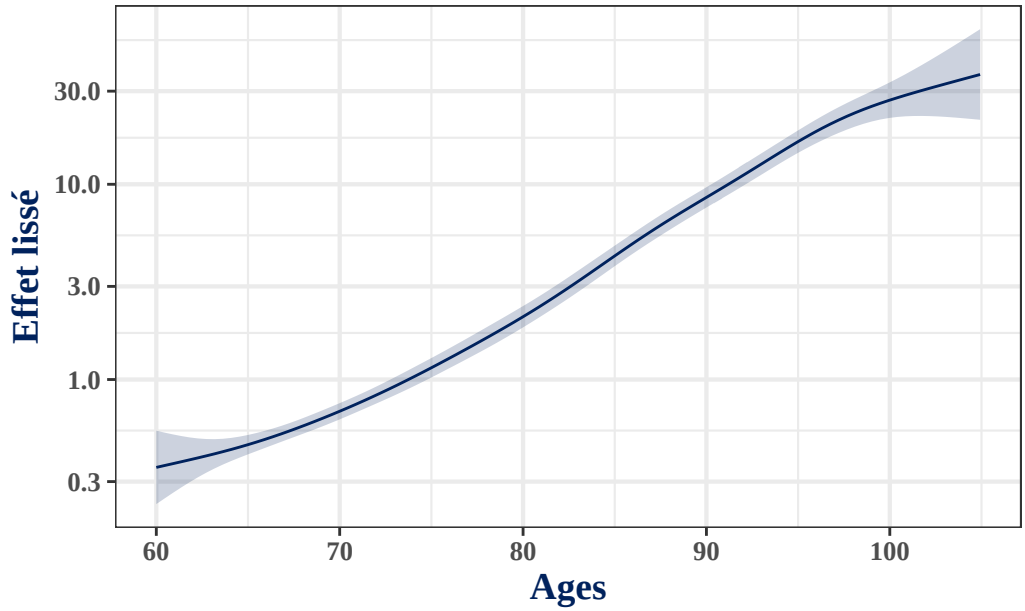

Figure 5.14 : Effet lissé de l'âge

On observe sur la Figure 5.14 un effet croissant assez classique de l'âge pour ce type de modèle. Cet effet s'interprète comme un coefficient multiplicatif à appliquer à la fonction

de hasard de base pour obtenir l'intensité instantanée de décès pour un individu à un âge donné, toutes choses égales par ailleurs. On observe des effets jusqu'à $\times 30$ pour les âges les plus élevés.

- **Effet du montant de rente**

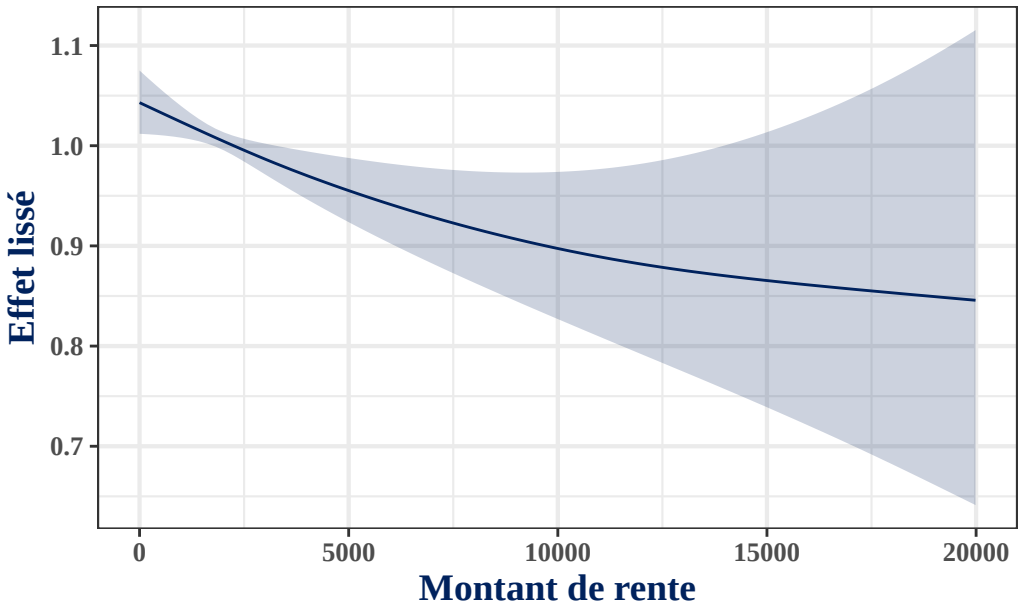

Figure 5.15 : Effet lissé du montant de rente

Il est observé sur la Figure 5.15 un effet décroissant du montant de rente sur l'intensité instantanée de décès. Là encore cet effet s'interprète comme un coefficient à appliquer à la fonction de hasard de base pour obtenir l'intensité de décès d'un individu ayant un montant de rente donné, toutes choses égales par ailleurs. On constate cependant une incertitude qui devient très grande au delà de 10 000 € de montant de rente, ceci du fait qu'il n y a pas énormément d'information au delà de ce montant de rente (3% des individus).

5.2.3.1 Diagnostic des modèles et comparaison

Il est question ici d'évaluer la qualité d'ajustement des trois modèles calibrés ici. Ceci se fera à travers l'examen des résidus de déviance. Pour un modèle de Cox, ces résidus s'écrivent comme suit :

$$r_i = sign(d_i - \hat{H}_i)\sqrt{-2(d_i - \hat{H}_i + d_i \log(\hat{H}_i))} \tag{5.25}$$

Avec :

- $\hat{H}_i$ la valeur de la fonction de hasard cumulée pour l'individu i à son temps de fin d'observation t_i

A noter que contrairement au cas d'un GAM Poisson, dans le cas du modèle de Cox, il n y a pas de littérature formelle indiquant que ces résidus doivent suivre une distribution normale. Ces derniers doivent cependant être assez faibles en valeur absolue lorsque le modèle est bien spécifié.

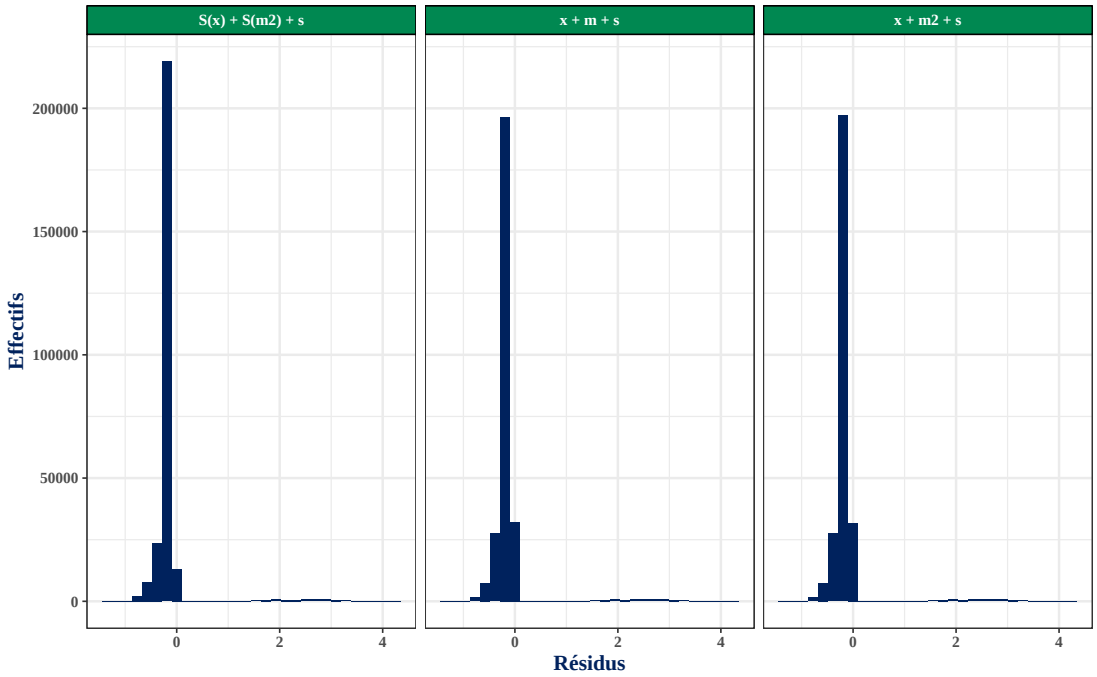

Figure 5.16 : Résidus des modèles

Il est observé des résidus assez faibles en valeur absolue (Figure 5.16), les différents modèles semblent bien s'ajuster aux données. Nous allons par la suite les comparer suivant le critère **AIC**.

La Figure 5.17 indique que le modèle avec les lissages est celui qui fournit les meilleurs résultats en terme de AIC.

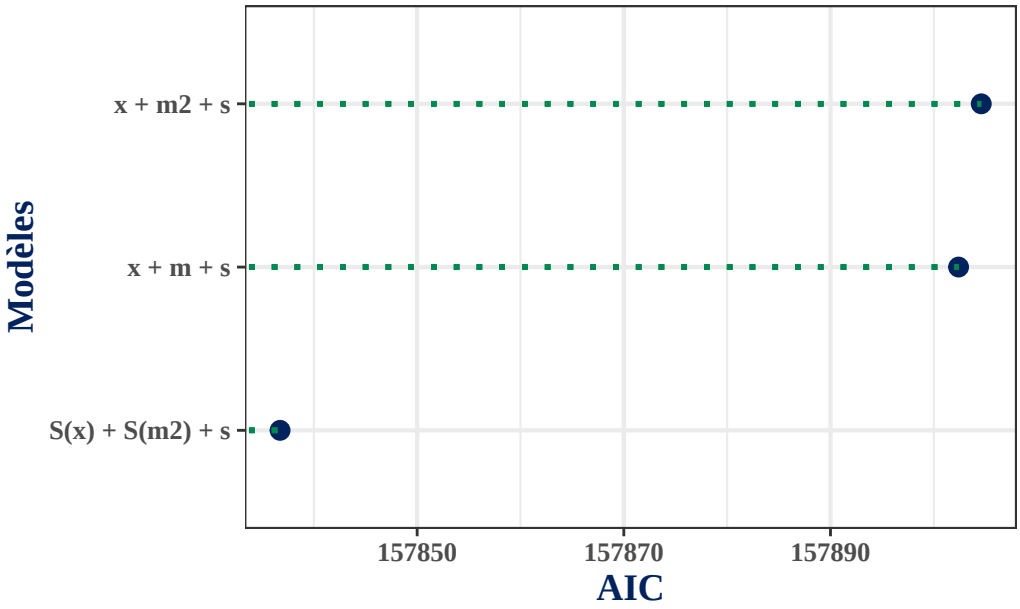

Figure 5.17 : Comparaison des modèles suivant l'AIC

5.3 Approche par les forêts aléatoires de survie

Nous allons tester ici une dernière approche par les forêts aléatoires de survie pour modéliser la mortalité du portefeuille. L'ensemble des éléments théoriques qui seront présentés ici sont tirés de Ishwaran et al. (2008) et Ehrlinger et Blackstone (2019).

5.3.1 Les forêts aléatoires

Les forêts aléatoires font partie des méthodes ensemblistes (agrégation de modèles). Elles consistent en l'agrégation de plusieurs modèles d'arbres de décision en vu d'améliorer la précision en prédiction par rapport à des modèles d'arbre pris individuellement. La mise en place de ce type de modèle ne nécessite pas d'a priori sur d'éventuelles liaisons paramétriques entre les variables explicatives et la variable réponse (linéaire ou non) ou encore des interactions entre les variables explicatives elles mêmes. Le principe de construction des forets aléatoires est résumé ainsi qu'il suit:

- Construction par Bootstrap de **B** sous échantillons de la population initiale;
- Entrainement d'un arbre de décision sur chacun des sous échantillons ainsi constitués. Chaque arbre est construit en partitionnant de manière récursive le sous échantillon par optimisation d'une règle de division dans l'espace des variables explicatives (de dimension **p**). Au moment de chaque partition (on

parlera de nœuds pour désigner le point de partitionnement) , un sous échantillon des variables explicatives (de dimension $\mathbf{m < p}$) est tiré, et c'est suivant ces variables explicatives qu'est effectuée la partition. Chaque nœud est ainsi séparé en 2 nœuds filles en maximisant la différence entre les sous populations créées au regard de la règle de division retenue. Le processus de partition est itéré avec les nœuds filles créés jusqu'à ce qu'une condition de fin soit enclenchée (généralement un nombre minimal d'individus dans chaque nœud). Au final chaque individu se retrouve dans un unique nœud terminal pour un arbre donné.

- Agrégation des différents arbres construits précédemment pour constituer la foret aléatoire. L'estimation de la réponse pour chaque individu est calculée en faisant une moyenne (dans le cas de la régression) ou un vote (dans le cas d'une classification) sur l'ensemble des résultats fournis par les **nœuds terminaux** des **B** arbres de décision qui constituent la forêt aléatoire.

Lors de la construction par Bootsrap des sous échantillons sur lesquels sont construits les arbres de décisions, chaque sous échantillon est composé d'environ **63 %** de la population initiale en moyenne. Les **37 %** d'individus restants sont qualifiés d'individus **OOB** (**O**ut **O**f **B**ag) et sont utilisés pour évaluer la précision du modèle entre autre.

5.3.2 Les forêts aléatoires de survie

Les forêts aléatoires de survie sont une extension pour les données de survie avec censure à droite des forêts aléatoires classiques présentées plus haut. L'approche a été proposée et décrite par Ishwaran et al. (2008). Le principe repose sur la segmentation de la population initiale en sous groupes (nœuds terminaux de chaque arbre) ayant des comportements de survie homogènes. Ensuite pour chacun de ces sous-groupes, un estimateur non paramétrique (estimateur de **Nelson-Aalen**) de la fonction de hasard cumulée est calculé. Cet aspect non paramétrique permet de capter des effets non linéaires et des interactions complexes des différentes variables explicatives. La moyenne sur plusieurs arbres de décisions, et l'aléa dans le processus de sélection des variables permet d'approcher des fonctions de survie assez complexes en s'affranchissant de certaines hypothèses comme celle du hasard proportionnel (modèle de Cox).

Un résumé des étapes de l'approche est fourni ci après :

- Reéchantillonnage par Bootstrap (**B** fois) de la base de données originale. Il n'est conservé dans chaque sous échantillon que à peu près 63% des données originales, le reste étant des données **OOB**;
- Construction d'un arbre de survie sur chacun des **B** échantillons Bootstrap constitués. A chaque nœud d'un arbre, une sélection aléatoire de variables explicatives est faite parmi l'ensemble des variables explicatives. La séparation des groupes au niveau des nœuds est ensuite effectuée en utilisant parmi les variables tirées aléatoirement celle qui maximise la différence de survie entre les groupes;

- Développement de l'arbre jusqu'à sa taille maximale sous contrainte d'avoir un nombre minimal de décès d_0 (strictement positif) dans chaque nœuds terminaux;
- Calcul de la fonction de hasard cumulée pour chaque arbre ainsi construit, la moyenne de ces fonctions pour chaque arbre permet d'obtenir la fonction de hasard cumulée globale;

Fonction de hasard cumulée globale

La fonction de hasard cumulée globale est un élément central de l'algorithme des forêts aléatoires de survie. Il sera détaillé par la suite le principe de construction de cette fonction de hasard cumulée.

Arbre de survie binaire

De façon analogue à l'algorithme des forêts aléatoires classiques, un arbre de survie est conçu de façon récursive par division de nœuds, en partant d'un nœud primaire qui contient l'ensemble des individus. Le critère de division ici est la maximisation de la différence de survie entre les 2 populations obtenues à l'issue de la division. Cette différence de survie est évaluée suivant la statistique du test de **Log-rank**. Le développement de l'arbre est stoppé dès lors que la contrainte sur le nombre minimal de décès par nœuds d_o n'est plus respectée.

Fonction de hasard cumulée aux nœuds terminaux

Une fois que l'arbre de survie atteint sa croissance terminale, ses nœuds les plus extrêmes sont appelés **nœuds terminaux.** On obtient ainsi dans les différents nœuds terminaux de l'arbre des sous populations homogènes en terme de survie.

Désignons par Γ l'ensemble de ces nœuds terminaux. Soient $(T_{1,h}, \delta_{1,h}), ..., (T_{n(h),h}, \delta_{n(h),h})$ les **temps de survie** et la variable de **censure** relative à chaque individu dans le nœud terminal $h \, \epsilon \, \Gamma$. Un individu i est censuré à droite au temps $T_{i,h}$ si $\delta_{i,h} = 0$; et si l'individu est décédé au temps $T_{i,h}$, alors $\delta_{i,h} = 1$.

Soient $t_{1,h} < t_{2,h} < ... < t_{n(h),h}$ les différents temps d'évènements distincts. On définit par $d_{l,h}$ et $Y_{l,h}$ respectivement le nombre de décès et d'individus exposés au temps $t_{l,h}$. La fonction de hasard cumulée pour le nœud terminal h est calculée avec l'estimateur de Nelson-Aalen et est donnée par :

$$\hat{H}_h(t) = \sum_{t_{l,h} \leq t} \frac{d_{l,h}}{Y_{l,h}} \tag{5.26}$$

Tous les individus dans un nœud terminal donné ont la même fonction de hasard cumulée.

Ainsi, pour un individu i donné, avec un vecteur de covariables x_i. Désignons par $H(t|x_i)$ la fonction de hasard cumulée de l'individu i. Pour déterminer cette fonction on recherchera parmi les nœuds terminaux de l'arbre, le nœud h qui correspond au vecteur x_i (ce nœud est unique), la fonction de hasard cumulée de l'individu i sera alors celle de ce nœud terminal.

$H(t|x_i) = \hat{H}_h(t)$, si $x_i \in h$.

La fonction de hasard cumulée globale est alors obtenue en faisant la moyenne des fonctions de hasard cumulées de tous les arbres qui constituent la foret aléatoire de survie. Il est proposé 2 versions de la fonction de hasard cumulée globale dans Ishwaran et al. (2008) : une fonction de hasard cumulée **Bootstrap** et une fonction de hasard cumulée **OOB**.

- **La fonction de hasard cumulée Bootstrap**

Elle est donnée par la formule suivante :

$$\hat{H}(t|x_i) = \frac{1}{B}\sum_{b=1}^{B} \hat{H}_b(t|x_i) \tag{5.27}$$

A remarquer ici que cette fonction de hasard cumulée est calculée sur tous arbres qui constituent la forêt aléatoire.

- **La fonction de hasard cumulée OOB**

Elle est donnée par la formule suivante :

$$\hat{H}^*(t|x_i) = \frac{\sum_{b=1}^{B} I_{i,b}\hat{H}_b(t|x_i)}{\sum_{b=1}^{B} I_{i,b}} \tag{5.28}$$

Avec : $I_{i,b} = 1$ si i est un individu **OOB** pour l'arbre **b**, sinon $I_{i,b} = 0$.

A noter ici donc qu'un individu donné ne contribue à la fonction de hasard cumulée globale qu'à travers les arbres pour lesquels il n'a pas été pris en compte (donnée **OOB**) dans le processus de construction. Il s'agit d'une moyenne sur les arbres pour lesquels l'individu i est OOB.

Conservation du nombre de décès

Les estimations issues des modèles de forêts aléatoires de survie vérifient le principe de **conservation du nombre d'évènements**. Ce principe stipule que la somme des valeurs de la fonction de hasard globale à tous les temps de survenance d'évènements (décès et censure) est égale au nombre total de décès observés. Ainsi, pour un nœud terminal $h \in \Gamma$ d'un arbre donné, on a :

$$\sum_{i=1}^{n(h)} \hat{H}_h(T_{i,h}) = \sum_{i=1}^{n(h)} \delta_{i,h} \tag{5.29}$$

En d'autres termes, il y a conservation du nombre de décès dans le nœud terminal h.

Mortalité d'ensemble

La mortalité d'ensemble se définit comme la valeur en espérance de la somme des fonctions de hasard cumulées sur tous les temps de survenance d'évènements, conditionnellement à un vecteur de caractéristiques donné. Elle s'interprète comme le nombre total de décès attendus si tous les individus avaient des comportements de survie similaires. Pour un individu i de vecteur de caractéristiques X_i, la mortalité d' ensemble est donnée par :

$$M_i = \mathbb{E}_i\left(\sum_{j=1}^{n} H(T_j|X_i)\right) \tag{5.30}$$

Dans le cadre des forets aléatoires , cette grandeur est approchée par l'estimateur suivant :

$$\hat{M}_i = \sum_{j=1}^{n} \hat{H}(T_j|X_i) \tag{5.31}$$

Cet indicateur peut également être calculé en utilisant la fonction de hasard cumulée **OOB**.

$$\hat{M}_i^* = \sum_{j=1}^{n} \hat{H}^*(T_j|X_i) \tag{5.32}$$

Erreur de prédiction

L'erreur de prédiction pour un modèle de forêts aléatoires de survie est estimé par l'indice de concordance de Harrell (**C-index**). Cet indice évalue la probabilité pour une sélection aléatoire de paires d'individus, que l'individu qui sort (décès ou censure) en premier ait une mortalité d'ensemble supérieure à celle de l'autre individu de la paire. Cette métrique a l'avantage de ne pas nécessiter de fixer une valeur précise de la durée pour être évaluée, elle permet également de tenir compte de la censure.

Le **C-index** se calcule en suivant les étapes suivantes :

- Construire toutes les paires possibles d'individus sur les données de départ;
- Exclure des paires ainsi constituées celles pour lesquelles le plus petit temps de survie est censuré. Exclure également les paires pour lesquelles les 2 durées de survie sont égales et censurées. On désigne alors par **"permis"** le nombre de paires restant;
- Pour une paire (i, j) restante donnée,
  - Si $T_i \neq T_j$, compter 1 si la plus petite durée de survie a la pire sortie prédite. Compter $\frac{1}{2}$ si les sorties prédites sont égales.
  - Si $T_i = T_j$ et les 2 individus décèdent, compter 1 si les sorties prédites sont égales, autrement compter $\frac{1}{2}$.
  - Si $T_i = T_j$ et au moins l'un des individus ne décède pas (censure), compter 1 si le décès a la pire des sorties prédites, et $\frac{1}{2}$ dans le cas contraire. Désignons par la **"concordance"** la somme de ces éléments sur toutes les paires.
- Le **C-index** est donné par : $C = \frac{Concordance}{Permis}$. Cet indice est compris entre 0 et 1, plus il est élevé et meilleure est la précision du modèle.

La notion de **sortie prédite** évoquée ci-dessus est assez similaire à celle de mortalité d'ensemble définie précédemment. Désignons par $t_1 < t_2 < ... < t_m$ l'ensemble des temps uniques de décès observés , pour deux individus i et j de vecteur de caractéristiques X_i et X_j, i a une **pire** sortie prédite que j lorsque

$$\sum_{l=1}^{m} \hat{H}^*(t_l|X_i) > \sum_{l=1}^{m} \hat{H}^*(t_l|X_j) \tag{5.33}$$

5.3.3 Application de la méthode aux données

Ici, il est question de modéliser directement une variable de durée, la base de données à disposition a donc dû être transformée en conséquence. Il est utilisé ici la base de données au format individuelle dans laquelle chaque individu est représenté par une et une seule ligne. Il a été créé une variable de durée d'observation (**ec**) ainsi qu'une variable de censure (**d**) pour indiquer si le décès de l'individu a été observé ou pas (Table 3.3).

Croissance de la forêt

Il sera calibré ici un modèle qui explique la durée de survie des individus par les variables âge (x), sexe (s) et montant de rente (m_2). L'estimation se fait grâce au package *R* `randomForestSRC` développé par *Hemant Ishwaran* et *Udaya Kogalur*.

Nous avons opté ici pour une forêt de 500 arbres, ce qui semble largement suffisant vu le faible nombre de variables explicatives dont nous disposons.

Il est à noter que en terme de précision, ce modèle affiche un **C-Index** de 74% . Une représentation schématique de l'un des arbres de la forêt construite est fournie par la Figure 5.18 .

Effets des variables explicatives

Importance des variables

L' importance d'une variable dans le processus de construction de la forêt aléatoire de survie se définit comme la différence entre la précision du modèle obtenu en bruitant la variable (permutation aléatoire de ses valeurs) et la précision du modèle obtenu avec la variable originale. Plus cette valeur est élevée et plus la variable est importante dans le processus de construction de la forêt aléatoire.

Sans beaucoup de surprise, l'âge apparaît comme la variable la plus déterminante dans la modélisation de la durée de survie (Figure 5.19). Fait un peu plus surprenant, le montant de rente semble ici être plus déterminant que le sexe dans la construction de la forêt aléatoire de survie. A noter tout de même que les conditions dans lesquelles ce modèle est appliqué ici sont loin d'être optimales, notamment car comme mentionné plus haut, nous ne disposons que de trois variables explicatives. Les forêts aléatoires en général de par leur construction sont d'autant plus performantes que le nombre de variables discriminantes pour le phénomène étudié est important.

Influence des variables

Pour quantifier de façon concrète l'influence des variables explicatives sur la durée de survie, on utilise la mortalité d'ensemble ou tout du moins son estimateur $\widehat{M}_i$ défini à

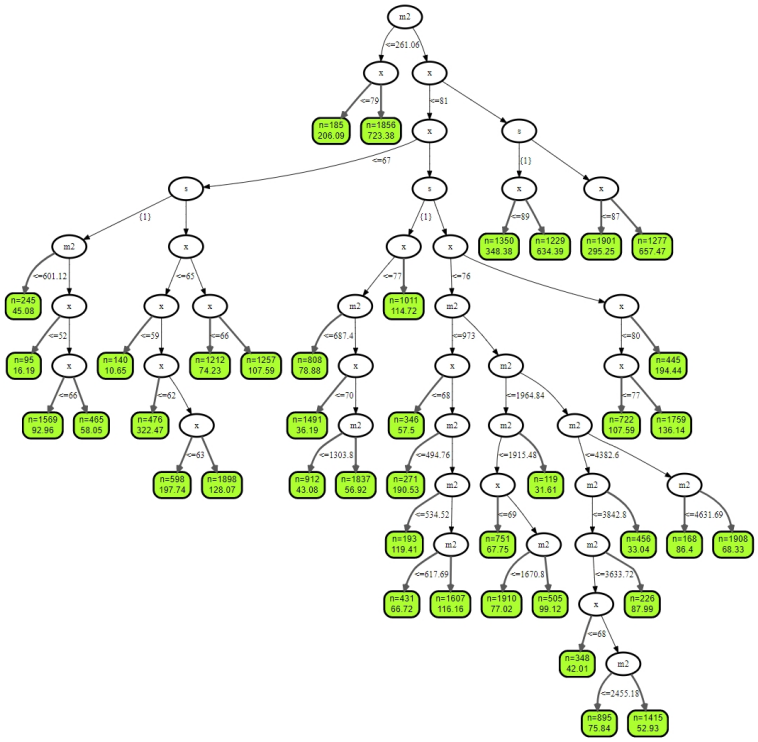

Figure 5.18 : Prototype d'arbre de la forêt aléatoire de survie

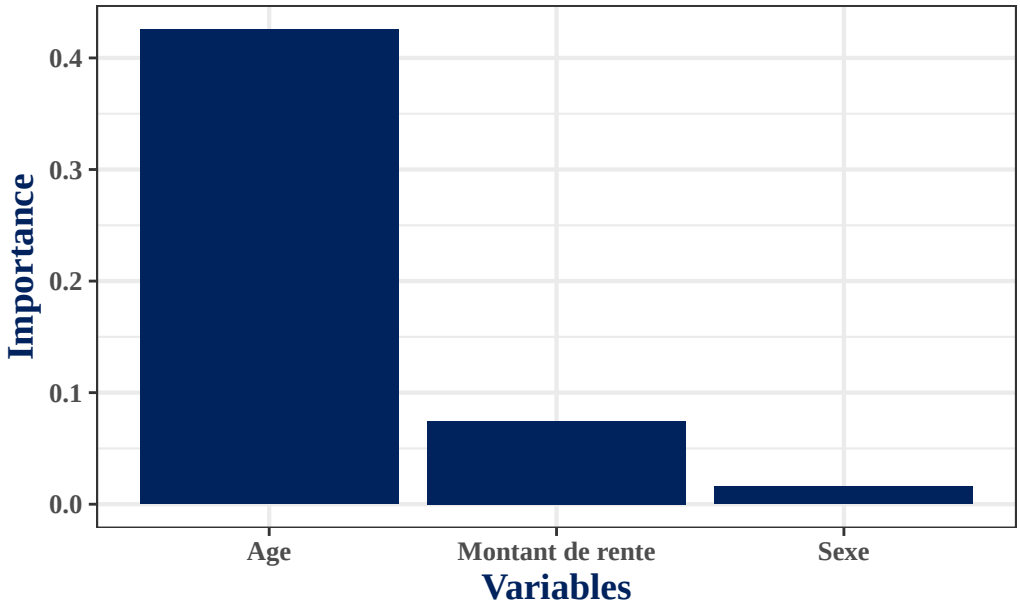

Figure 5.19 : Importance des variables explicatives

l'Équation 5.31. Pour rappel, il s'interprète comme le nombre de décès qu'on observerait si tous les individus avaient les mêmes caractéristiques que l'individu i.

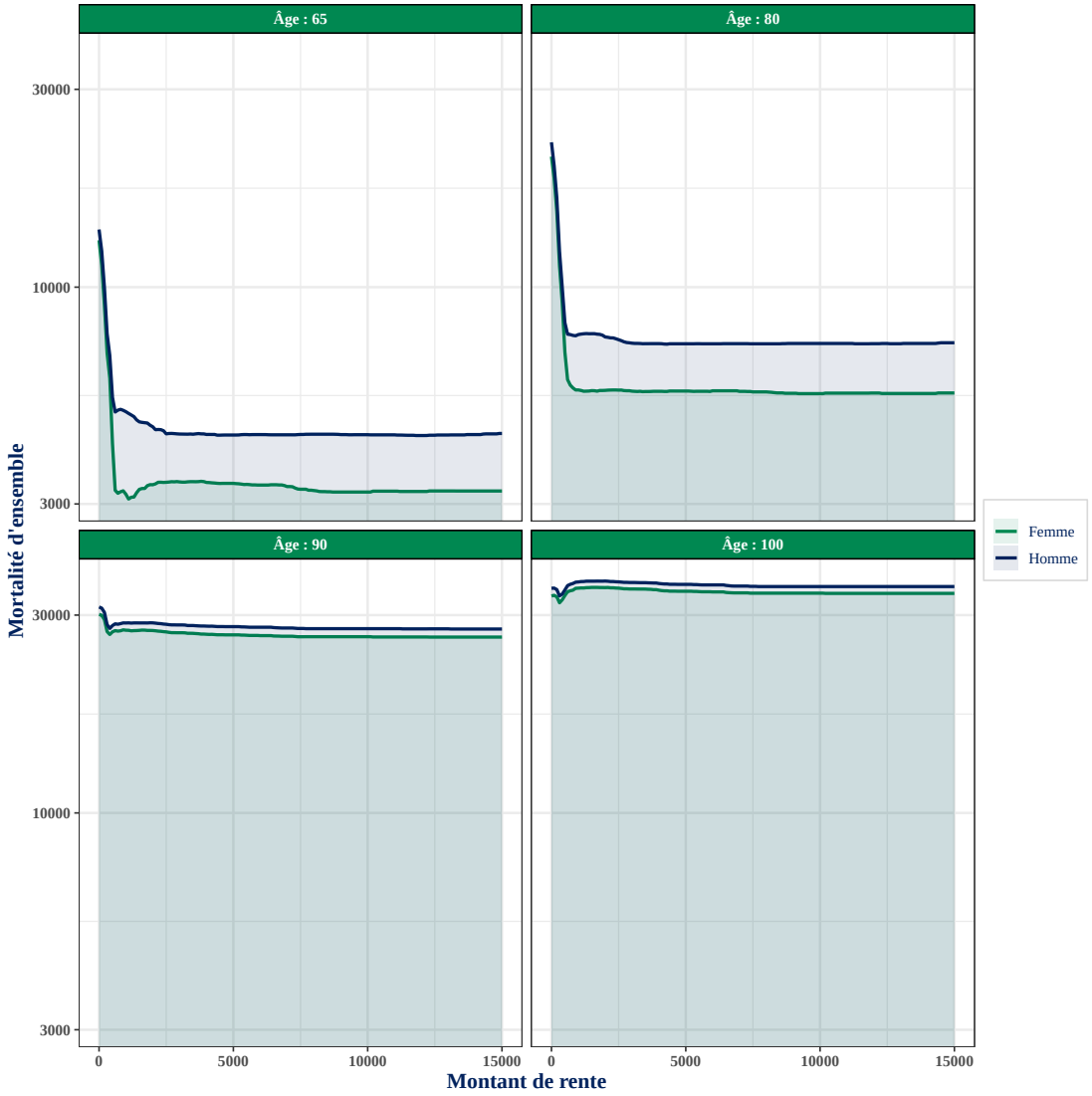

Figure 5.20 : Effet des variables explicatives

La Figure 5.20 nous apporte de nombreuses informations :

- Tout d'abord la mortalité d'ensemble diminue à mesure que le montant de rente augmente;
- La diminution de la mortalité d'ensemble avec l'augmentation du montant de rente est de moins en moins nette à mesure que l'âge avance. A 100 ans (graphique en bas à droite), les courbes d'évolution des mortalités d'ensemble pour les hommes

et pour les femmes sont quasiment plates. Un éventuel effet du montant de rente tend donc à s'estomper aux très grand âges;

- La mortalité d'ensemble des hommes est toujours supérieure à celle des femmes. Là aussi cette différence s'amenuise à mesure que l'âge augmente. A 100 ans les courbes des hommes et des femmes sont quasiment confondues pour tous les montants de rente.

Pour isoler l'effet moyen du montant de rente ici, il a été calculé une moyenne sur le sexe et l'âge de la mortalité d'ensemble pour chaque montant de rente. La Figure 5.21 en donne une illustration, on constate là encore globalement une tendance à la baisse de la mortalité d'ensemble à mesure que le montant de rente augmente.

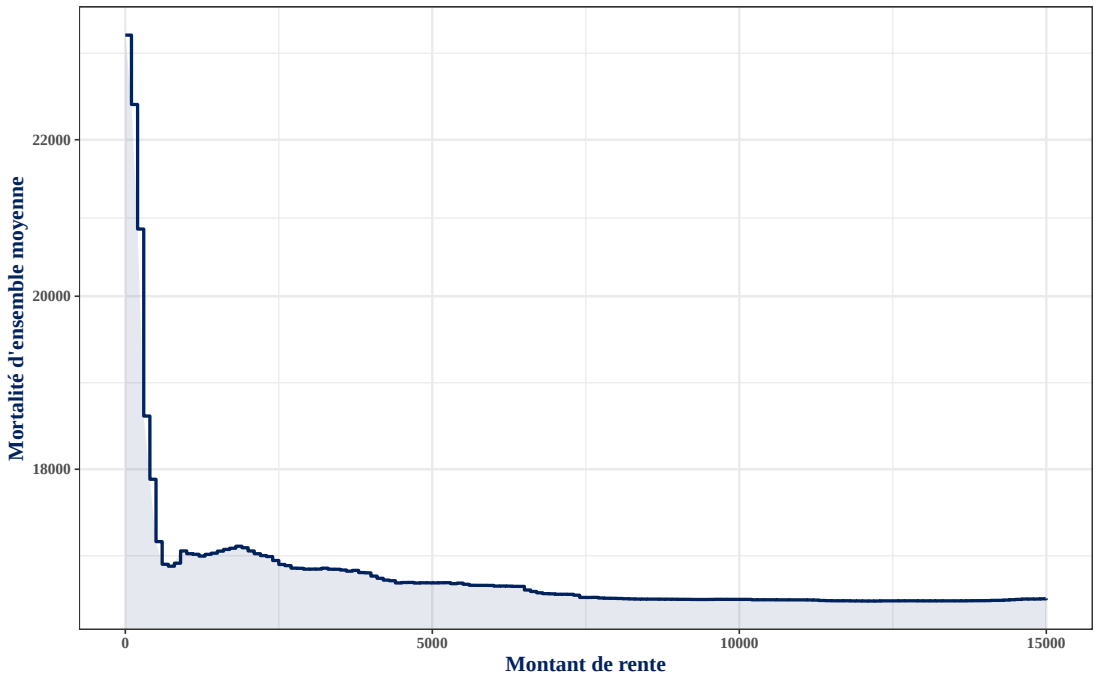

Figure 5.21 : Effet du montant de rente

Au global pour ces trois familles de modèles, en terme d'effets des variables explicatives , on obtient des résultats assez similaires. Il est ainsi constaté pour chacun de ces modèles une diminution du niveau de mortalité avec l'augmentation du montant de rente. Ceci est en accord avec le constat historique fait sur l'ensemble de la population française, selon lequel les individus les plus aisés vivent en moyenne plus longtemps que les plus modestes (INSEE 2018).

5.4 Comparaison des modèles

Un travail supplémentaire a été mené en vue de permettre une comparaison fiable des modèles sur leur capacité à prédire la mortalité du portefeuille. Ayant été calibrés sur des données de formats différents, il est nécessaire de trouver un socle commun pour comparer ces modèles. L'idée développée ici est de calculer pour chaque individu la probabilité de décéder au cours de la période d'observation. Cette probabilité peut être obtenue pour chaque modèle. Ensuite, à partir de ces probabilités de décès, une logvraisemblance binomiale est calculée par la formule suivante :

$$V = \sum_{i=1}^{n} \Big(d_i \log(p_i) + (1 - d_i) \log(1 - p_i) \Big) \tag{5.34}$$

Avec :

- d_i l'indicatrice de décès sur la période pour l'individu i
- p_i la probabilité de décès sur la période d'observation pour l'individu i, elle est fournie par les différents modèles
- n le nombre d'individus.

Pour un modèle donné, plus la valeur de V est élevée et meilleure est la capacité de ce modèle à prédire la mortalité du portefeuille sur la période d'observation. Cette métrique a l'avantage d'être applicable à tout type de modèle, sans réelle contrainte sur la manière dont il a été calibré, le seul élément requis étant de pouvoir calculer à partir de ce modèle une probabilité de décès pour chaque individu. Cette métrique sera utilisée pour comparer sept modèles :

- Le meilleur modèle de la Figure 5.8 ($S(x) + S(m_2) + ti(x, m_2) + s$) qu'on désignera sous le nom de modèle **Poisson.** Il sera décliné en deux versions :
  - Une version calibrée sur données agrégées (celle étudiée en début de chapitre) : **Poisson données agrégées**
  - Une version calibrée sur données individuelles fractionnées : **Poisson données individuelles**
- Le meilleur modèle de Cox de la Figure 5.17 ($S(x) + S(m_2) + s$)
- le modèle des forêts aléatoires de survie calibré dans la troisième partie du chapitre
- Le modèle utilisé par Swiss Life
- Un modèle à **coefficient d'abattement** : C'est un modèle à un degré de liberté (un paramètre) qui renvoie un coefficient unique à appliquer à la force de mortalité des tables réglementaires pour obtenir celle du portefeuille

- Un modèle avec **Intercept seu**l : Il s'agit là aussi d'un modèle à un degré de liberté qui calcule une probabilité de décès unique commune à tous les individus

Le modèle avec **coefficient d'abattement** est le modèle le plus simple qui puisse être calibré avec les tables réglementaires en offset. Le modèle avec **Intercept seul** quant à lui est le modèle le plus basique qui puisse être calibré. Ils sont inclus dans cette comparaison à des fins d'étalonnage.

5.4.1 La Méthode de Brass

La méthode de Brass est une méthode de positionnement par rapport à la référence réglementaire. Elle consiste en une régression des logits des taux de mortalité bruts du portefeuille sur les logits des taux de mortalité des tables réglementaires $TGH/TGF05$, cette approche est présentée dans Planchet et Thérond (2006). Le modèle est spécifié comme suit :

$$logit(q_{xt}) = a \times logit(q_{xt}^{ref}) + b + \epsilon_{xt}$$
$$\log(\frac{q_{xt}}{1-q_{xt}}) = a \times \log(\frac{q_{xt}^{ref}}{1-q_{xt}^{ref}}) + b + \epsilon_{xt}$$

Avec :

- q_{xt} : le taux de décès à l'âge x l'année calendaire t dans le portefeuille
- q_{xt}^{ref} : le taux de décès à l'âge x l'année calendaire t pour les tables de référence
- a, b : des paramètres à estimer
- ϵ_{xt} : un terme d'erreur

Les taux ajustés sont finalement obtenus en inversant la fonction logistique :

$$\hat{q}_{xt} = \frac{\exp(y_{xt})}{1+\exp(y_{xt})}, \text{avec } y_{xt} = a \times \log(\frac{q_{xt}^{ref}}{1-q_{xt}^{ref}}) + b$$

5.4.2 Comparaison

Le pire des modèles au regard de cette métrique V est sans surprise le modèle qui renvoie une probabilité de décès unique pour l'ensemble des individus (Intercept seul). La Table 5.1 présente un ensemble de métriques basées sur la vraisemblance de l'Équation 5.34. Le meilleur modèle ici est le modèle Poisson sur données individuelles, suivi du modèle Cox. Le modèle Poisson sur données agrégées se classe troisième au regard de cette métrique. L'approche sur données agrégées est privilégiée pour les

modèles Poisson malgré tout car elle présente une meilleure stabilité du point de vue des intervalles de confiance et de la possibilité de calcul de taux bruts, ainsi qu'une justification théorique plus évidente comparée à l'approche sur données individuelles. Le modèle des forêts aléatoires est ici celui qui s'en sort le moins bien au regard de cette métrique (après le pire modèle), il est précédé par la méthode de Brass.

Table 5.1 : Comparaison des modèles

Modèles	R2	R2 ajusté	AIC	Vraisemblance (V)	edf
Poisson données individuelles	16,2 %	16,19 %	42987	-21482	12,24
Cox	15,97 %	15,97 %	43097	-21540	9,61
Poisson données agrégées	15,77 %	15,76 %	43201	-21592	9,73
Tables avec coeff d'abattement	15,58 %	15,58 %	43279	-21639	1
Méthode de Brass	15,34 %	15,34 %	43404	-21701	2
Forêts aléatoires	14,67 %	14,55 %	44482	-21873	368,752
Intercept seul	0 %	0 %	51267	-25633	1

Les résultats de cette comparaison de modèles confirme dans ce cas précis la supériorité des modèles avec lissage (Poisson et Cox) par rapport aux forêts aléatoires de survie et la méthode de Brass notamment. Rappelons que le modèle conservé au final est bien celui de Poisson sur données agrégées. Ce modèle par rapport à celui de Cox a l'avantage de permettre la mise en offset des tables réglementaires, ce qui est d'un grand intérêt sur le plan opérationnel.

NB :

- $R2 = 1 - \frac{dev}{dev_{max}}$, avec $dev = -2V$, et dev_{max} la déviance maximale correspondant au pire des modèles
- $R2\ ajuste = 1 - \frac{dev}{dev_{max}} \times \frac{n-1}{n-edf}$, avec n le nombre de ligne de la base de données sur laquelle le modèle a été entraîné.
- edf correspond au nombre de degrés de liberté du modèle
- Le nombre de degré de liberté pour la forêt aléatoire est donné par le nombre moyen de feuilles des arbres qui la constituent

Chapitre 6

Calcul des provisions

Dans ce chapitre, il sera question d'évaluer concrètement l'impact de la table de mortalité construite dans le chapitre précédent sur le calcul des provisions pour les rentiers présents dans le portefeuille. Pour évaluer cet impact, les provisions seront calculées avec quatre modèles différents :

- Le modèle GAM retenu au chapitre précédent tenant compte du montant de rente
- Un modèle GAM ne prenant pas en compte le montant de rente
- Le modèle utilisé par Swiss Life
- Les tables réglementaires.

6.1 Principe des provisions techniques en assurance

Provisionner de façon générale c'est mettre des ressources de côté en vue d'évènements futurs. En théorie, toutes les entreprises quel que soit leur secteur d'activité pourraient être amenées à constituer des provisions. Mais la nécessité de provisionnement est particulièrement exacerbée pour les compagnies d'assurance, ceci du fait de la nature même de leur activité. En effet, alors que pour une entreprise normale, la vente du produit final se fait après avoir supporté (et donc quantifié précisément) l'ensemble des coûts de productions, pour une compagnie d'assurance le produit finale (le contrat d'assurance) est d'abord vendu et ce n'est qu'au moment de la survenance éventuelle d'un sinistre que ce qui pourrait s'apparenter à un coût de production est évalué. Le cycle de production est donc inversé.

Le provisionnement apparaît alors comme une question centrale pour les compagnies d'assurance, car leur rentabilité en est fortement tributaire. Le calcul de ces provisions par les compagnies d'assurance est également très encadré par le législateur [1], ceci du fait que ce sont ces provisions qui assurent la capacité des assureurs à honorer leurs engagements vis à vis de leurs assurés. C'est donc dans ce cadre que des provisions précises et les règles de calcul y afférentes ont été définies et consignées dans le Code des

[1] Le législateur ici est l'Autorité de Contrôle Prudentiel et de Résolution (ACPR)

assurances. Ces provisions portent le nom de **provisions techniques**, et elles doivent obligatoirement êtres constituées par les compagnies d'assurance selon la nature des produits dont elles disposent dans leurs portefeuilles.

Il existe une grande variété de provisions techniques spécifiques aux compagnies d'assurance Vie et Non-Vie. La provision à laquelle nous nous intéresserons ici est la **provision mathématique**, c'est la provision constituée pour que l'assureur puisse à tout moment honorer ses engagements envers ses assurés.

6.2 Formalisation mathématique

On se place dans la phase de restitution de l'épargne, c'est-à-dire la phase durant laquelle la rente est effectivement versée aux rentiers. Dans cette phase, le calcul des provisions mathématiques est basé sur la valeur actuelle des annuités viagères a_x , elle correspond à la somme actualisée des flux futurs unitaires probabilisés jusqu'au décès du bénéficiaire. Cette valeur actuelle peut s'interpréter comme une sorte d'espérance de vie résiduelle actualisée, une durée prévisionnelle durant laquelle l'assureur devra payer la rente à un assuré donné.

$$\begin{aligned} a_x &= \sum_{k=1}^{N-x} v^k \times_x P_k \\ &= \sum_{k=1}^{N-x} v^k \times \mathbb{P}(T > k+x | T > x) \\ &= \sum_{k=1}^{N-x} v^k \times \frac{\mathbb{P}(T > k+x, T > x)}{\mathbb{P}(T > x)} \\ &= \sum_{k=1}^{N-x} v^k \times \frac{\mathbb{P}(T > k+x)}{\mathbb{P}(T > x)} \end{aligned} \tag{6.1}$$

Avec :

- $v = \frac{1}{1+i}$: le facteur d'actualisation, et i le taux technique en vigueur
- x : l'âge du rentier à la date de calcul
- N : Un âge ultime de survie
- T : la variable aléatoire de durée de vie
- ${}_xP_k$: la probabilité pour le rentier de survivre k années supplémentaires sachant qu'il a déjà vécu x années.

La provision mathématique pour un individu d'âge x percevant une rente d'un montant m est alors donnée par :

$$PM = m \times a_x \tag{6.2}$$

En toute rigueur, les provisions associées aux rentes des conjoints (têtes secondaires) devraient également être calculées, mais elles seront ignorées ici pour simplifier les analyses.

6.3 Évaluation de l'impact de la prise en compte du montant de rente

Ici, les provisions seront calculées concrètement en utilisant trois modèles différents comme précisé en début de chapitre. La différence dans les calculs pour chacun des modèles se situera au niveau du ${}_xP_k$ dans l'Équation 6.1 qui aura une évaluation propre à chaque modèle.

A noter ici que les provisions sont calculées au taux technique de $1,5$ % , en date du 1^{er} Janvier 2023 pour les rentiers du portefeuille encore en vie à cette date.

6.3.1 Ecarts globaux

la Figure 6.1 présente les provisions totales évaluées par les différents modèles en pourcentage des provisions évaluées par le modèle utilisé par Swiss Life. Le premier constat ici est que les provisions évaluées suivant les deux modèles GAM et le modèle Swiss Life sont plus prudentes que les tables réglementaires. Ensuite, le modèle Swiss Life se révèle moins prudent que le modèle GAM avec montant de rente , mais plus prudent que le modèle GAM sans montant de rente. L'ensemble de ces différences entre les provisions des différents modèles seront étudiées plus précisément par la suite.

6.3.2 Modèle Swiss Life et modèle GAM sans prise en compte de la rente

On s'intéresse ici au modèle GAM sans le montant de rente et au modèle Swiss Life pour comprendre de façon plus précise les écarts de provision entre ces modèles vu que les deux n'incluent pas la dimension montant de rente. La différence fondamentale entre ces deux modèles réside dans le fait que le modèle GAM tient compte de façon explicite de l'âge dans l'évaluation des écarts entre la mortalité du portefeuille et celle des tables réglementaires, alors que le modèle Swiss Life est un peu moins flexible sur ce point. La Figure 6.2 présente les taux de mortalité par âge en 2023 fournis par le modèle GAM sans montant de rente et le modèle Swiss Life en pourcentage des taux de

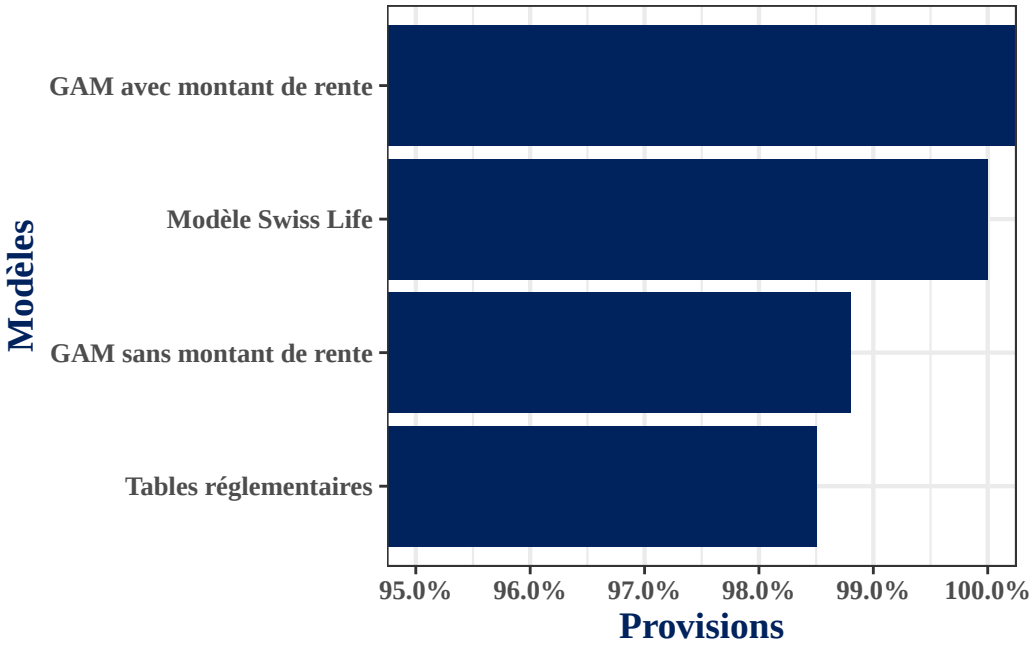

Figure 6.1 : Provisions globales par modèle exprimées en terme de pourcentage des provisions du modèle interne

mortalité réglementaires. Le premier constat ici est que les taux de mortalité du modèle GAM sont toujours supérieurs à ceux du modèle Swiss Life. Ceci est probablement dû à la plus grande flexibilité du modèle GAM. La Figure 6.3 présente les espérances de vie résiduelles des modèles GAM et de Swiss Life exprimées en pourcentage des espérances de vie résiduelles des tables réglementaires. Il vient confirmer le constat fait précédemment, avec des espérances de vie résiduelles par âge pour le modèle Swiss Life qui sont systématiquement supérieures à celles du modèle GAM, et qui s'écartent de plus en plus des tables réglementaires avec l'âge. Ces éléments expliquent la prudence des provisions du modèle Swiss Life par rapport à celles du modèle GAM ne prenant pas en compte le montant de rente. On constate également sur la Figure 6.2 que les deux modèles à partir d'un certain âge fournissent des taux de mortalité inférieurs aux taux réglementaires (le modèle Swiss Life plongeant plus rapidement que le modèle GAM), ceci explique in fine la prudence des provisions des deux modèles par rapport aux tables réglementaires.

6.3.3 Modèles GAM avec et sans prise en compte du montant de rente

Pour isoler l'effet du montant de rente, nous allons comparer les provisions du modèle GAM incluant le montant de rente et celles du modèle GAM n'incluant pas le montant de rente. Ceci permet d'éliminer les interférences qui pourraient être liées aux différences

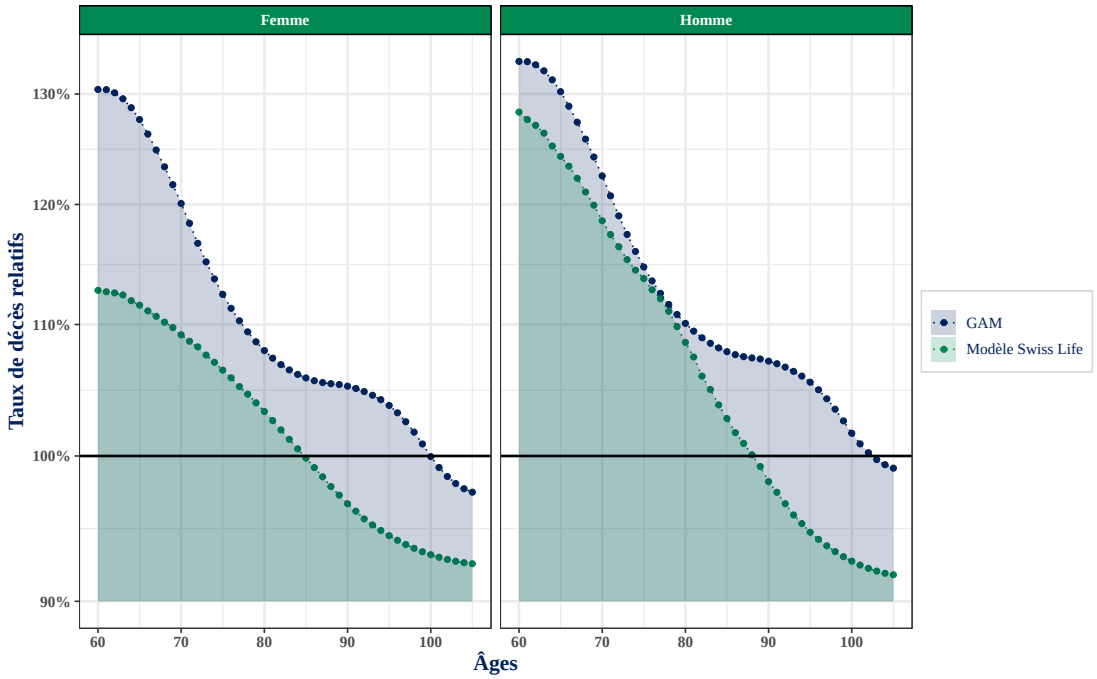

Figure 6.2 : Taux de mortalité (en 2023) par âge exprimés en terme de pourcentage des taux de mortalité réglementaires

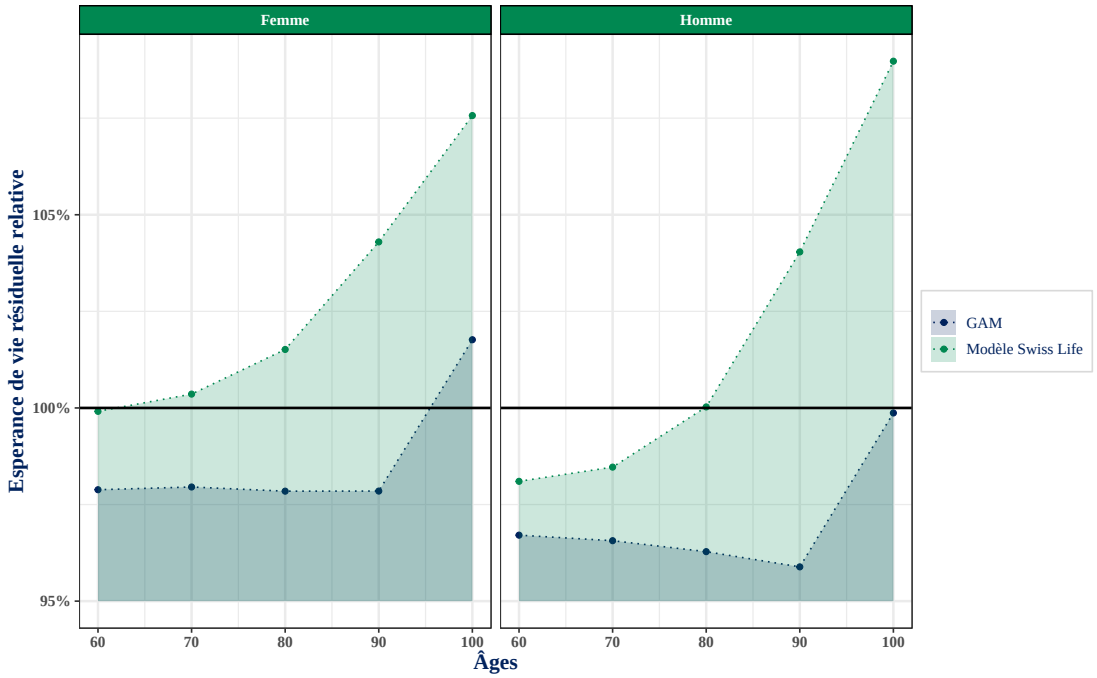

Figure 6.3 : Espérances de vie résiduelles par âge et par modèle

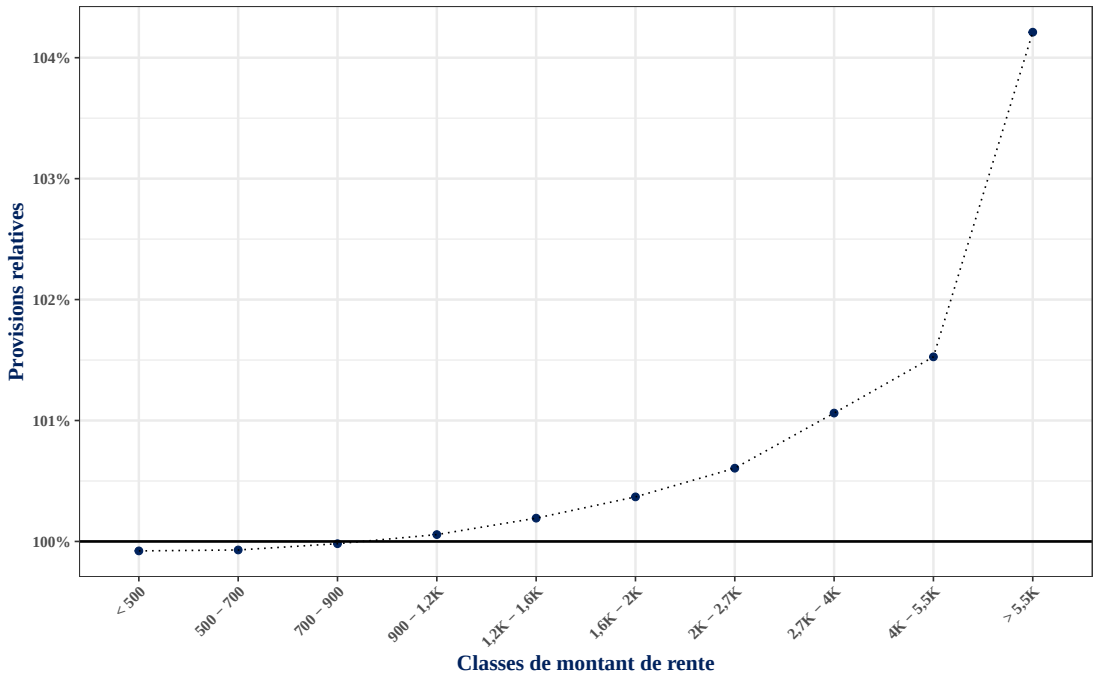

Figure 6.4 : Provisions par montant de rente exprimées en terme de pourcentage des provisions GAM sans prise en compte de la rente

intrinsèques entre les modèles GAM et le modèle Swiss Life. La Figure 6.4 présente les provisions par montant de rente (pour le modèle GAM incluant le montant de rente) en pourcentage des provisions fournies par le modèle GAM n'incluant pas le montant de rente. On constate des écarts de provisions de plus en plus élevés à mesure que le montant de rente augmente. Ceci est assez logique vu que le modèle GAM avec montant de rente comme vu dans le chapitre précédent, décrit une diminution de la mortalité avec le montant de rente. De ce fait, à âge et sexe inchangés, il fournit des espérances de vie résiduelles actualisées (a_x) et in fine des provisions d'autant plus élevées (relativement au modèle ne prenant pas en compte le montant de rente) que le montant de rente est élevé. On constate un écart maximal de plus de 4% pour la classe de montant de rente la plus élevée.

Conclusion

L'objectif principal de l'étude menée dans ce mémoire était la construction pour un portefeuille de rentiers d'une table de mortalité prospective qui tienne compte du montant de rente perçu par les individus dans l'évaluation des probabilités de décès. Pour ce faire, des modèles additifs généralisés ont été utilisés. Ces modèles par rapport à l'approche classiquement utilisée dans la construction des tables de mortalité ont l'avantage de permettre à la fois d'effectuer des lissages suivant des covariables et d'effectuer directement des liaisons avec la mortalité des tables de référence. Les résultats obtenus avec ces modèles ont été confrontés à ceux obtenus avec deux autres types de modèles : le modèle de Cox et les forêts aléatoires de survie. Les résultats obtenus avec ces modèles étaient assez concordants notamment en ce qui concerne l'impact du montant de rente sur la mortalité du portefeuille. Il a été observé pour le portefeuille étudié une décroissance du niveau de mortalité avec l'augmentation du montant de rente. C'est un résultat cohérent au vu de la littérature sur le sujet de la longévité et de l'espérance de vie, qui stipule que les individus les plus aisés vivent en moyenne plus longtemps que les individus les plus modestes. Ce phénomène est montré dans Kleinow, Cairns, et Wen (2019) pour le cas du Royaume-Uni, et dans INSEE (2018) pour le cas de la France.

Des provisions mathématiques ont été calculées avec la table prospective ainsi construite, et comparées à celles obtenues avec le modèle utilisé jusqu'ici par Swiss Life. Au global, les provisions obtenues via la table construite sont plus prudentes que celles obtenues avec le modèle de Swiss Life ou encore les tables réglementaires. Il est à noter aussi des provisions fournies par la table construite qui sont de plus en plus prudentes (par rapport aux provisions du modèle de Swiss Life) vers les rentes élevées, vu que la table construite prédit toutes choses égales par ailleurs des probabilités de décès de plus en plus faibles à mesure que le montant de rente augmente. On aboutit donc à une table plus prudente et assurément plus représentative du risque réel du portefeuille que la table qui est utilisée jusqu'ici. Cette étude présente un certains nombre de limites qu'il convient de mentionner ici.

Tout au long de cette étude, une hypothèse implicite a été faite, hypothèse selon laquelle le montant de la rente perçue était représentatif du niveau de vie des assurés. Or, il est possible par exemple que des assurés ayant des montants de rente faibles dans le portefeuille étudié ici aient d'autres contrats de rente dans d'autres compagnies avec des rentes bien plus élevées. Il faudrait alors utiliser la somme des montants sur

les différents contrats pour être parfaitement exhaustif, chose qui nous est impossible malheureusement vu que cette information est inaccessible.

Les années Covid $(2020-2021)$, il a été observé une surmortalité de 9% par rapport aux autres années d'observation. Ces années dans l'étude menée ici n'ont fait l'objet d'aucun traitement particulier. Bien que nous ayons jugé cette surmortalité relativement faible, une action de redressement sur ces années Covid ne pourrait que contribuer à améliorer la robustesse des analyses menées ici. Une piste interessante serait d'inclure dans les modèles une variable qualitative indicatrice de la période Covid.

Des données plus nombreuses et la prise en compte de la nature exacte des produits contenus dans le portefeuille étudié (contrats collectifs ou individuels) auraient pu permettre d'enrichir l'analyse menée dans ce mémoire. Les montants de rente pouvant varier selon la nature des produits auxquels ils sont associés. L'exploitation de variables comme le code postal ou encore la CSP (si elle avait été disponible) aurait pu nous renseigner d'avantage sur le niveau de richesse des individus.

Références

Bastien, Lorraine. 2020. « Utilisation de la théorie de la crédibilité pour la construction d'une table de mortalité prospective sur un portefeuille de rentes restreint ». *Institut des Actuaires*.

Biessy, Guillaume. 2022. « Cours de Modèle de durée (EURIA) ».

Côté, Steven. 2016. « Modèles additifs généralisés dans la modélisation de l'impact du kilométrage et de l'exposition au risque en assurance automobile ». *Université du Québec, Montréal*.

Cox, D. R. 1972. « Regression Models and Life-Tables ». *Journal of the Royal Statistical Society. Series B (Methodological)* 34 (2): 187-220.

Debonneuil, Edouard, Stéphane Loisel, et Frédéric Planchet. 2018. « Do actuaries believe in longevity deceleration? » *Insurance: Mathematics and Economics* 78: 325-38.

Duchon, Jean. 1977. « Splines minimizing rotation-invariant semi-norms in Sobolev spaces ». In *Constructive Theory of Functions of Several Variables: Proceedings of a Conference Held at Oberwolfach April 25–May 1, 1976*, 85-100. Springer.

Dupâquier, Jacques, et Michel Dupâquier. 1985. *Histoire de la démographie: la statistique de la population des origines à 1914*. FeniXX.

Durieux, Solène, et Gaud Samba. 2013. « Apport de la théorie des copules à l'élaboration des modèles de mortalité multivariés ». *Institut des Actuaires*.

Ehrlinger, John, et Eugene H Blackstone. 2019. « ggRandomForests: Survival with Random Forests ».

Fall, El Hadji. 2019. « Construction de table de mortalité prospective d'expérience pour les contrats de rentes viagères en périmètre retraite ». *Institut des Actuaires*.

Gourieroux, Christian, et Yang Lu. 2015. « Love and death: A Freund model with frailty ». *Insurance: Mathematics and Economics* 63: 191-203.

Guez, Ruth. 2018. « Etude du risque de longévité d'un portefeuille de rentes collectives: Impact d'une table de mortalité prospective en vision "Solvabilité 2"" ». *Institut des Actuaires*.

INSEE. 2018. « L'espérance de vie par niveau de vie ». https://www.insee.fr/fr/statistiques/fichier/3322051/F1801.pdf.

———. 2022a. « Impact de l'épidémie de Covid-19 : 95 000 décès de plus qu'attendus de Mars 2020 à Décembre 2021 ». https://www.insee.fr/fr/statistiques/6445335.

———. 2022b. « Les filles nées en 2022 pourrraient vivre en moyenne 93 ans, les garçons 90 ans ». https://www.insee.fr/fr/statistiques/6655536.

Ishwaran, Hemant, Udaya B Kogalur, Eugene H Blackstone, et Michael S Lauer. 2008. « Random survival forests ».

Kabore, Elisee. 2017. « Estimation des améliorations futures de mortalité des sous-

groupes au sein d'une population hétérogène ». *Institut des Actuaires*.

Kleinow, T, A Cairns, et J Wen. 2019. « Deprivation and life expectancy in the UK ». *The Actuary, April*.

Kontis, Vasilis, James E Bennett, Colin D Mathers, Guangquan Li, Kyle Foreman, et Majid Ezzati. 2017. « Future life expectancy in 35 industrialised countries: projections with a Bayesian model ensemble ». *The Lancet* 389 (10076): 1323-35.

Milne, Joshua. 1815. *A treatise on the valuation of annuities and assurances on lives and survivorships: on the construction of tables of mortality and on the probabilities and expectations of life*. Vol. 2. Longman, Hurst, Rees, Orme,; Brown.

Nychka, Douglas. 1988. « Bayesian confidence intervals for smoothing splines ». *Journal of the American Statistical Association* 83 (404): 1134-43.

Parry, Stephen. 2018. « To offset or not: using offsets in count models ».

Planchet, Frédéric. 2022. « Modèles de durée ». http://www.ressources-actuarielles.net/C1256F13006585B2/0/1430AD6748CE3AFFC1256F130067B88E/$FILE/Seance4.pdf?OpenElement.

Planchet, Frédéric, et Pierre Thérond. 2006. « Modèles de durée ». *Economica*.

Price, Richard. 1780. *An Essay on the Population of England: From the Revolution to the Present Time*. Cadell.

Quashie, Aki, et Michel Denuit. 2005. « Modèles d'extrapolation de la mortalité aux grands âges ». *Institut des Sciences Actuarielles et Institut de Statistique, Université Catholique de Louvain, WP*.

Tomas, Julien, et Frédéric Planchet. 2014. « Construire une table de mortalité prospective : le package ELT ».

Vermet, Franck. 2022. « Cours de Modèle de durée (EURIA) ».

Vial, Géraldine. 2012. « Qu'est ce qu'une rente viagère? »

Wood, Simon N. 2017. *Generalized additive models: an introduction with R*. CRC press.

Yikmis, Nizar. 2020. « Construction et validation de tables de mortalité prospectives best-estimate pour un portefeuille de rentes viagères dans un contexte de données peu nombreuses ». *Institut des Actuaires*.

Ziegelmeyer, Kevin. 2015. « Tarification d'un swap de longévité : Modélisation de l'impact de la rente sur la mortalité du portefeuille réassuré ». *Institut des Actuaires*.

Annexe

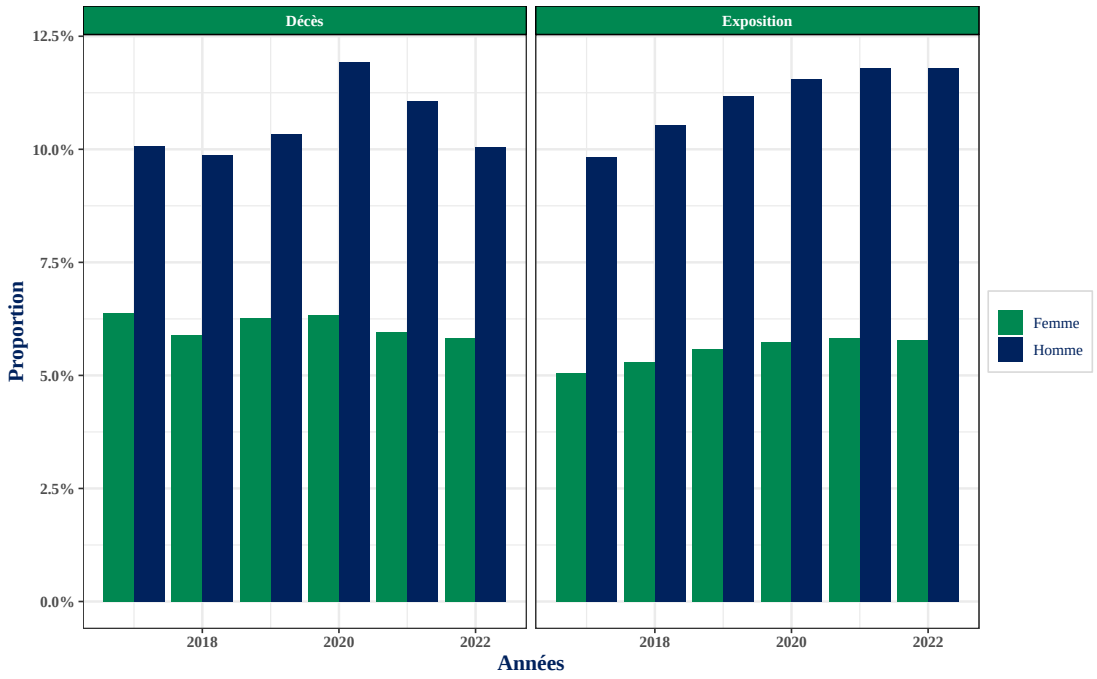

Figure 5 : Décès et exposition par année pour chaque sexe