EURIA — EURo Institut d'Actuariat

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 22 Septembre 2017

Par : Elisée KABORE
Titre : **Estimation des améliorations futures de la mortalité des sous-groupes au sein d'une population hétérogène**

Confidentialité : Non

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

Membre présent du jury de l'Institut des Actuaires :
Yann QUERE
Signature :
Sophie BOURDET
Signature :

Entreprise :
GIE AXA
Signature :

Membres présents du jury de l'EURIA :
Franck VERMET
Signature :

Directeur de mémoire en entreprise :
Jérémy GORIS
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 :

Secrétariat :

Bibliothèque :

EURIA
EURo Institut
d'Actuariat

6, avenue le Gorgeu
CS 93837
29238 Brest Cedex 3

T +33 (0)2 98 01 66 55
F +33 (0)2 98 01 66 57
euria@univ-brest.fr

Remerciements

La réalisation de ce mémoire est le fruit de la participation de plusieurs personnes, qui ont bien voulu m'accompagner et me soutenir en m'apportant leur collaboration.

Tout d'abord, je tiens à témoigner ma reconnaissance et ma gratitude à mon Directeur de mémoire et tuteur en entreprise, Jérémy GORIS, pour son accompagnement et son implication active à la réalisation de ce mémoire. Je suis très reconnaissant pour ses précieux conseils, sa disponibilité ainsi que pour sa patience à mon égard.

Je tiens aussi à témoigner mes remerciements à ma tutrice à l'EURIA, Isabelle DEVINE, pour son suivi, ses conseils et son implication active à l'élaboration de ce mémoire.

J'adresse mes remerciements à Marine HABART, Responsable « Modèle Interne vie et Retraite & Fonds de Pension » ainsi qu'à Mohamed BACCOUCHE Responsable du « Group Risk Management Life » au sein du Groupe AXA de m'avoir donné l'opportunité de réaliser ce mémoire au sein de leurs équipes. J'ai pu bénéficier de leurs conseils avisés tout au long de mon étude.

Je remercie toute l'équipe qui m'a accompagné durant l'élaboration de cette étude. Je remercie particulièrement Tom POPA pour son implication active à sa construction.

Ce mémoire traduit l'aboutissement de trois années d'étude à l'EURIA. Je remercie grandement tout le corps professoral de l'EURIA et tout particulièrement Franck VERMET pour son engagement professionnel de haut niveau, dès le commencement, à la réussite de mes années à l'EURIA.

Enfin, je remercie ma très grande famille qui m'a soutenu jusqu'au bout. Je remercie mes parents pour leur implication entière. Je remercie vivement toutes les personnes qui se sont impliquées dans mon projet et m'ont apporté leur soutien précieux.

Résumé

Le *basis risk* en longévité est le risque que certains sous-groupes de population vivent plus longtemps que d'autres, et que cet écart évolue dans le temps. Il peut être subdivisé en deux principales composantes : le niveau et la tendance. La problématique de l'estimation du *basis risk* prend une ampleur considérable en raison de son impact grandissant sur les compagnies d'assurance.

Nous présentons dans cette étude une approche de modélisation de la composante tendancielle du *basis risk*. La méthode utilisée constitue une extension du modèle de Cox avec pour particularité l'imposition d'une fonction de tendance aux coefficients de régression. Nous appliquons la méthode à l'estimation et à la validation du *trend* de la population des rentiers en France en comparaison avec celui de la population nationale, puis à l'étude du *basis risk* géographique en Allemagne. Nous proposons une approche de modélisation par la simulation des populations à partir des tables de mortalité puis par une décomposition des individus permettant l'évolution de leurs caractéristiques au cours du temps.

Dans la première application, cela a consisté à la simulation d'une population nationale française puis d'une population de rentiers incluse dans celle-ci. Les résultats montrent l'ajustement d'un *trend* linéaire, confirmant ainsi l'absence de risque de tendance entre les données de mortalité des tables des rentiers et la population nationale en France.

Dans la seconde application, cela a consisté à la simulation des sous-populations d'Allemagne de l'Est et de l'Ouest au sein de l'Allemagne. Le modèle a permis la détection de *trend* différents en Allemagne : une tendance linéaire à l'ouest et une tendance quadratique, puis logarithmique aux années plus récentes à l'est. L'Allemagne de l'Est rattrape son retard, mais les évolutions ralentissent ces dernières années.

La démarche présentée nous a permis de détecter les différents *trends* au sein de ces populations, confirmant donc sa validité, ainsi que sa capacité à détecter et à corriger le *basis risk*.

Mots clefs: Espérance de vie, longévité, *basis risk*, risque de niveau, risque de tendance, *trend*, modèle de Cox, fonction de tendance, rentiers

Abstract

The basis risk in longevity is the risk that some subpopulations will live longer than others, and that this gap will change over time. It can be subdivided into two main components : level and trend. The estimation of the basic risk takes on a considerable scale due to its growing impact on insurance companies.

In this study, we present an approach to modeling the trend component of the basis risk. The method used is an extension of the Cox model, with the peculiarity of imposing a trend function on regression coefficients. We apply the method to the estimation and validation of the trend of the annuitant population in France in comparison with that of the national population, and then to the study of the geographic base risk in Germany. We propose an approach of modeling by simulating populations using mortality tables and then decomposing individuals to allow the evolution of their characteristics over time.

In the first application, this consisted of simulating a French national population and then an annuitants population included in it. The results show the adjustment of a linear trend, thus confirming the absence of trend risk between mortality data from the annuitant tables and the national population in France.

In the second application this consisted in the simulation of the subpopulations of East and West Germany. The model has allowed the detection of different trends in Germany : a linear trend to the west and a quadratic then logarithmic trend for the more recent years in the east. East Germany is catching up, but developments have slowed down in recent years.

This approach allowed us to detect the different trends within these populations, thus confirming his validity, as well as its ability to detect and correct the basis risk.

Keywords: Life expectancy, longevity, basis risk, level risk, trend risk, Cox model, trend function

Synthèse

Estimation des améliorations futures de la mortalité des sous-groupes au sein d'une population hétérogène

Elisée KABORE (EURIA, GIE AXA)
Septembre 2017

Nous présentons dans cette étude une approche d'estimation du basis risk en longévité et plus particulièrement du risque de tendance du basis risk au sein d'une population hétérogène. Le modèle développé est une extension du modèle de Cox avec pour particularité l'imposition d'une fonction de tendance aux coefficients de régression. Nous avons appliqué la méthode à l'estimation et à la validation du trend de la population des rentiers en France puis à l'étude du basis risk géographique en Allemagne. Nous avons montré par cette démarche que le modèle proposé permet de détecter et de corriger le risque de tendance au sein d'une population hétérogène.

Mots clefs: Longévité, *basis risk*, risque de tendance, *trend* , modèle de Cox

Problématique

Depuis la deuxième moitié du XX^e siècle, l'espérance de vie a augmenté d'environ 3 mois par an dans la plupart des pays industrialisés (Oeppen, W.Vaupel, 2002). Cependant, ce gain n'est pas uniformément réparti au sein des populations. Certains sous-groupes voient leur espérance de vie évoluer fortement, tandis que d'autres constatent une décélération de leur *trend* d'espérance de vie. Ces améliorations différentes de mortalité au sein d'une même population introduisent un risque appelé le *basis risk*[1].

Le *basis risk* en longévité est le risque que certains sous-groupes de population vivent plus longtemps que d'autres, et que cet écart évolue dans le temps. Il peut donc être décomposé en deux composantes : le niveau et la tendance. La couverture du *basis risk* a pris une importance particulière en raison de son impact sur les compagnies d'assurance et les fonds de pension. L'une des grandes problématiques sous-jacentes au *basis risk* est donc sa détection et sa correction. Nous proposons dans cette étude un approche

1. ou risque de base

d'estimation du risque de tendance du *basis risk* en modélisant l'hétérogénéité observée au sein des populations.

Modélisation du risque de tendance du *basis risk*

Deux grandes approches peuvent être utilisées pour la modélisation du *basis risk*.

Approche par discrimination

Elle consiste à effectuer une modélisation pour chaque sous-groupe homogène au sein d'une population. Ces sous-groupes homogènes représentent des individus ayant des caractéristiques similaires. Les modèles mathématiques couramment utilisés sont les modèles de mortalité à deux ou trois facteurs. L'approche par discrimination a l'avantage d'être intuitive et complète. Elle permet de modéliser finement la mortalité des sous-groupes. Cependant l'utilisation de cette méthode induit une autre problématique qui est celle du choix de la segmentation optimale.

Approche par modèle de survie

L'analyse de survie, aussi appelée analyse des durées de survie, constitue l'étude des données pour lesquelles le temps avant la réalisation d'un événement (décès) relève d'une importance particulière. Nous nous sommes intéressés dans notre étude aux méthodes qui, en plus d'estimer les distributions de survie, permettent aussi d'évaluer les effets que peuvent avoir des variables explicatives caractéristiques des sous-groupes, sur la survie. Il s'agit des modèles de régression. La littérature utilise couramment deux modèles. Il s'agit du modèle de Cox et des modèles à temps de vie accélérée [2]. Ces méthodes permettent l'estimation de l'impact de chaque covariable sur la mortalité, sans subdivision de la population.

Nous avons choisi dans notre étude de proposer une méthode d'estimation du risque de tendance du *basis risk* basée sur le modèle de Cox. Le tableau 1 explique les raisons de ce choix.

Le modèle de Cox : adaptation à la modélisation de la composante tendancielle du *basis risk* en longévité

Notations et formulations mathématiques

Soient :

- Z un vecteur de covariables désignant les caractéristiques des individus. On suppose que $Z = (Z_1, ..., Z_p)$; p étant le nombre de caractéristiques
- Z^i désignant les caractéristiques d'un individu i.

2. Accelerated Failure Time Models(AFT)

	Discrimination	**AFT**	**Cox**	**Commentaires**
Segmentation de la base	✓	✗	✗	*L'approche par discrimination impose une segmentation de la base.*
Formatage de la base	✗	✓	✓	*Pour Cox et AFT, la base doit être bien ordonnée, de façon à faire apparaître les covariables.*
Choix d'une distribution de probabilité	✓	✓	✗	*AFT nécessite le choix d'une loi de probabilité pour la modélisation des temps de survie. Les méthodes de discrimination utilisent une distribution normale pour la modélisation des séries temporelles.*
Facilités d'interprétation	✗	✓	✓	*L'interprétation des résultats dans l'approche par discrimination peut être complexe car elle dépend fortement de la segmentation et des modèles choisis.*
Simplicité dans la mise en œuvre	✗	✗	✓	*Cox ne nécessite pas d'étude pour le choix d'une distribution de probabilité ou d'étude pour choix d'une segmentation optimale de la population.*
Ressources importantes d'analyse et de calcul	✓	✓	✗	*Les ressources nécessaires en terme d'études sont importantes dans les méthodes par discrimination et AFT.*
Flexibilité	✓	✓	✓	Les modèles sont flexibles, permettent la prise en compte des effets d'interaction par exemple.
Facilité de déploiement	✗	✗	✓	*Le choix de la distribution de probabilité et de la segmentation optimale rendent plus complexe le déploiement opérationnel des modèles AFT et de l'approche par discrimination.*
Modélisation fine des sous-groupes	✓	✗	✗	*L'approche par discrimination permet une modélisation fine des sous-groupes ; possibilité d'utiliser le modèle adapté pour chaque groupe.*
Choix du modèle	✗	✗	✓	***Le modèle de Cox est le meilleur choix pour notre étude.***

TABLE 1 – Choix du modèle

- $\beta = (\beta_1, ..., \beta_p)$ les coefficients de régression du modèle
- x l'âge des individus et t l'année d'observation.

Le modèle de Cox s'écrit :

$$h(x|Z) = h_0(x) \exp(\beta^\top Z) \tag{1}$$

Avec

- $h(x|Z)$ le risque instantané de décès à l'âge x sachant les caractéristiques Z.
- $h_0(x)$ le *baseline* ou mortalité de référence. Il correspond à la mortalité d'un individu dont les caractéristiques sont $Z = (0, ..., 0)$.

Le modèle de Cox avec tendance

Pour modéliser le *trend*, nous proposons d'adapter le modèle de Cox en introduisant une fonction de tendance aux coefficients de régression. Le modèle proposé est le suivant :

$$h(x, t|Z) = h_0(x) \exp\left(\beta(t)^\top Z\right) = h_0(x) \exp\left(\sum_{j=1}^{p} \beta_j(t) Z_j\right) \tag{2}$$

$$\exists\, f : \mathbb{R} \to \mathbb{R} \;/\; \beta(t) = \beta^{le} + \beta^{tr} f(t) \tag{3}$$

Les paramètres $\beta^{le} = (\beta_1^{le}, ..., \beta_p^{le})$ et $\beta^{tr} = (\beta_1^{tr}, ..., \beta_p^{tr})$ modélisent respectivement les effets liés au niveau et au *trend*. f représente des scénarios d'évolution future de la mortalité et $h(x, t|Z)$ le risque instantané de décès à l'âge x à l'année t sachant Z.

Cadre d'application et résultats

Notre étude répond à une problématique actuarielle qui est l'estimation du risque de tendance du *basis risk*. Nous proposons ici d'en faire deux applications :

- Une application à l'étude du *basis risk* entre population de rentiers et population nationale.
- Une application à l'étude d'un *basis risk* géographique : celui entre l'Allemagne de l'Est et l'Allemagne de l'Ouest.

Pour cela, nous avons simulé ces populations grâce à leurs données de mortalité par année. L'avantage de cette démarche a été l'obtention de résultats généraux, ne dépendant pas de la structure et de la composition des portefeuilles d'individus, permettant ainsi de réduire la volatilité que peut engendrer leur utilisation.

L'objectif de ces deux applications est de montrer par la première que la méthode proposée permet bien d'estimer le risque de tendance du *basis risk*, puis de proposer une estimation concrète du risque de tendance à travers la deuxième application.

Nous avons testé 7 fonctions de tendance représentant différents scénarios d'évolution de la mortalité. Il s'agit des scénarios linéaire, quadratique, logarithmique, inverse, exponentiel inverse, tangent hyperbolique et croissant cyclique. La figure 1 nous permet de visualiser ces différents scénarios.

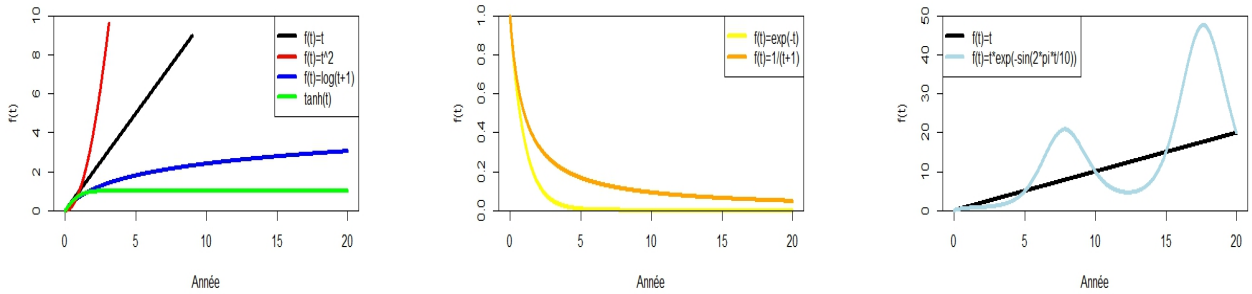

FIGURE 1 – Fonctions de tendance

Risque de tendance entre population nationale et population de rentiers

Pour l'estimation du risque de tendance entre la population de rentiers et la population nationale, nous avons simulé ces populations grâce aux tables générationelles réglementaires de rentiers en France et aux données de mortalité de la Human Mortality Database entre 1996 et 2005. Nous avons adopté une vision par période afin de visualiser réellement les améliorations à chaque année. Pour cela, nous avons calculé des tables de rentiers par période à partir des tables par cohorte. Les tables de rentiers ayant été ajustées par les données de la population nationale française lors de leur construction, nous nous attendions à retrouver le même *trend* sur la population nationale, c'est-à-dire le *trend* linéaire. Après implémentation des modèles, nous avons mesuré la qualité de l'ajustement de chaque modèle et calculé les écarts quadratiques annualisés entre les espérances de vie ajustés par chaque modèle et celles des tables de mortalité. Les résultats [3] sont regroupés dans le tableau 2.

Modèle	Hommes			Femmes		
	AIC	BIC	Écarts	AIC	BIC	Écarts
Tangente hyperbolique	4651670	4651690	0,08947672	3494599	3494619	0,3572199
Exponentiel inverse	4651661	4651681	0,07366774	3494597	3494617	0,3548081
Inverse	4651650	4651670	0,05460477	3494597	3494614	0,3517714
Croissance cyclique	4651637	4651657	0,03203398	3494591	3494611	0,3476654
Logarithmique	4651630	4651651	0,02054169	3494590	3494610	0,3463284
Quadratique	4651629	4651650	0,01992241	3494590	3494610	0,3455982
Linéaire	4651623	4651644	0,00865467	3494589	3494608	0,3441683

TABLE 2 – Écarts quadratiques annualisés, AIC et BIC

Le modèle de Cox nous permet de retrouver le résultat attendu. Le *trend* linéaire est le meilleur ajustement (voir figure 2).

3. Akaike Information Criterion(AIC) ; Bayesian Information Criterion (BIC)

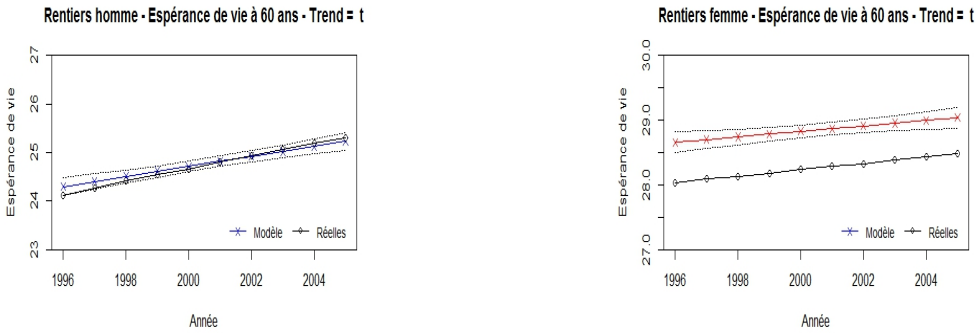

FIGURE 2 – Espérances de vies ajustées avec une tendance linéaire

Les résultats obtenus montrent que le Cox peut ainsi tout à fait détecter le *trend* d'une sous-population.

Estimation du risque de tendance : Cas du *basis risk* géographique en Allemagne

La partition de l'Allemagne en 1949 a contribué à la divergence des espérances de vie dès 1979 entre en Allemagne. L'Allemagne de l'Ouest a vu son espérance de vie fortement augmenter tandis que l'Allemagne de l'Est a connu de plus faibles améliorations. Cependant, après la réunification, la tendance s'est inversée. L'Allemagne de l'Est rattrape son retard tandis que l'Allemagne de l'Ouest connaît un essoufflement.

Pour l'étude du *basis risk* géographique en Allemagne, nous avons simulé grâce aux données de la Human Mortality Database, une population composée d'individus de l'Allemagne de l'Est et de l'ouest entre 1956 et 2015. Nous avons estimé les tendances de ces deux populations grâce au modèle de Cox.

$$h(x,t|Z) = h_0(x)\exp(\beta_E(t)Est + \beta_O(t)Ouest) \tag{4}$$

avec : $\beta_E(t) = \beta_E^{le} + \beta_E^{tr} f_e(t)$ et $\beta_O(t) = \beta_O^{le} + \beta_O^{tr} f_o(t)$; $Est \in \{0,1\}$; $Ouest \in \{0,1\}$

La modélisation nous a permis de détecter deux tendances différentes pour ces deux populations (voir tableau 3 et figure 3).

	Allemagne de l'Est	Allemagne de l'Ouest
Hommes	Quadratique	Linéaire
Femmes	Linéaire	Linéaire

TABLE 3 – Tendance des populations (1975-2015)

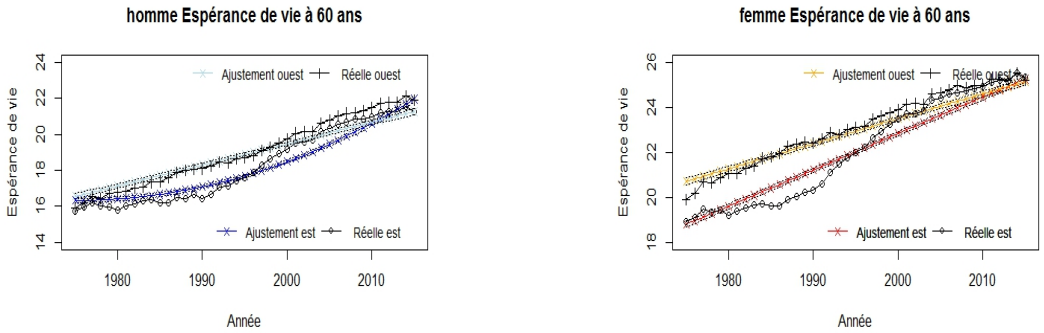

FIGURE 3 – Comparaison des espérances de vie

Nous avons, dans un deuxième temps, ajusté les tendances pour les années 1995 à 2015. Les résultats sont présentés dans le tableau 4.

	Allemagne de l'Est	Allemagne de l'Ouest
Hommes	Logarithmique	Linéaire
Femmes	Logarithmique	Linéaire

TABLE 4 – Tendance des populations (1995-2015)

Le modèle de Cox a donc pu détecter et de corriger différents *trends* au sein cette population hétérogène sans subdivision de celle-ci.

Conclusions

Nous avons développé dans cette étude une approche pertinente permettant la détection et la correction du *basis risk* au sein des populations hétérogènes. Nous avons appliqué la démarche à l'étude du *basis risk* entre la population de rentiers et la population nationale en France, puis à l'étude du *basis risk* géographique en Allemagne. Les résultats montrent que le modèle développé permet de détecter d'une part le *trend* d'espérance de vie lorsque les populations d'étude ont le même *trend* et d'autre part les *trends* hétérogènes lorsque les populations d'étude possèdent des *trends* différents. Les résultats des tests de sensibilité montrent une grande robustesse dans l'estimation du paramètre de tendance, même lorsque la taille de la sous-population étudiée est réduite.

Cette approche nouvelle peut permettre à l'assureur de modéliser plus finement la mortalité de son portefeuille ainsi que son évolution dans le temps, en tenant compte des facteurs d'hétérogénéité et en s'appuyant sur la mortalité d'une population de moyenne pour estimer celle des sous-populations de son portefeuille. Les pistes d'amélioration de l'approche développée restent nombreuses. Toutefois, cette étude peut être un cadre de construction d'une démarche plus élaborée d'estimation du *basis risk*.

Summary

Estimation of future improvements in subgroup mortality within an heterogeneous population

Elisée KABORE (EURIA, GIE AXA)
Septembre 2017

We present in this study an approach to the estimation of the basis risk in longevity and more particularly the estimation of the basis risk trend within a heterogeneous population. The model developed is an extension of the Cox model with the peculiarity of imposing a trend function on regression coefficients. We applied the method to the estimation and validation of the trend in the annuitants population in France and then to the study of geographical basis risk in Germany. We show in this study that the Cox model enables to detect and correct the basis risk in a heterogeneous population.

Mots clefs: basis risk, longevity, trend, Cox model, heterogeneous population

Problématique

Since the second half of 20th century, life expectancy has increased by about 3 months per year in most industrialized countries (Oeppen, W.Vaupel, 2002). However, this gain is not uniformly distributed among populations. Some subgroups see their life expectancy evolve with a steep slope, while others are slowing down their life expectancy increase. These different mortality improvements within the same population introduce a risk called the basic risk.

The longevity of basis risk is the risk that some subgroups of the population will live longer than others, and that this gap will change over time. It can therefore be broken down into two components : level and trend. The coverage of basis risk has become particularly important due to its impact on insurance companies and pension funds. One of the major issues underlying basis risk is therefore its detection and correction. We propose in this study an approach to estimate the risk of the trend of basis risk by modeling the heterogeneity in the populations.

Modeling the trend risk of basis risk

Two broad approaches can be used to model basis risk.

Discrimination Approach

It consists of modeling each homogeneous subgroup with similar characteristics within a population. The commonly used mathematical models are two or three factor mortality models. The discrimination approach has the advantage of being intuitive and complete. It makes it possible to finely model the mortality of each subgroups. However, the use of this method induces another problem which is the optimal segmentation.

Approach by survival model

Survival analysis, also called survival time analysis, is the study of data for which the time before the occurrence of an event (death) is the main concern. In our study, we investigated methods that, in addition to estimating survival distributions, also allow us to evaluate the effects of survivor-specific explanatory variables on sub-groups. These are the regression models. Two major models are commonly used in the literature : the Cox model and the Accelerated Failure Time Models (AFT). These methods have the advantage of allowing the estimation of the mortality for each subgroups without the subdivision of the population.

In our study, we proposed a method to estimate the trend of basis risk based on the Cox model. The reasons for this choice are presented in the table 5.

Cox model : adaptation to modeling the basis risk trend longevity

Mathematical notations and formulations

Let :

- Z a vector of covariates designating the characteristics of individuals. We suppose that $Z = (Z_1, ..., Z_p)$; p being the number of characteristics.
- Z^i designating the characteristics of an individual i.
- $\beta = (\beta_1, ..., \beta_p)$ the regression coefficients of the model.
- x the age of the individuals and t the year of observation.

The Cox model is written :

$$h(x|Z) = h_0(x) \exp(\beta^\top Z) \tag{5}$$

Avec

- $h(x|Z)$ the instantaneous risk of death at age x knowing the characteristics Z.
- $h_0(x)$ the baseline. It corresponds to the mortality of an individual whose characteristics are $Z = (0, ..., 0)$.

	Discrimination	**AFT**	**Cox**	**Comments**
Segmentation of the database	✓	✗	✗	*The discrimination approach requires segmentation of the database.*
Formatting the database	✗	✓	✓	*For Cox and AFT, the basis must be well ordered, so that the covariates appear.*
Choosing a probability distribution	✓	✓	✗	*AFT requires the choice of a probability distribution for the modeling of survival times. Discrimination methods use the normal distribution for modeling time series.*
Interpretation facilities	✗	✓	✓	*The interpretation of the results in the discrimination approach can be complex because they depend heavily on the segmentation and the models chosen.*
Simplicity in implementation	✗	✗	✓	*Cox does not require study for the choice of a probability distribution or study for choosing an optimal segmentation of the population.*
Important analysis and calculation resources	✓	✓	✗	*The necessary resources in terms of studies are important in methods by discrimination and AFT.*
Flexibility	✓	✓	✓	Models are flexible ; Allows interaction effects to be taken into account.
Ease of deployment	✗	✗	✓	*The choice of the probability distribution and the optimal segmentation make the operational deployment of the AFT models and the discrimination approach more complex.*
Fine modeling of subgroups	✓	✗	✗	*The discrimination approach allows fine modeling of subgroups ; Possibility of using the adapted model for each group.*
Choice of model	✗	✗	✓	***The Cox model is the best choice for our study***

TABLE 5 – Choice of model

The Cox model with tendance

To model the trend, we propose to adapt the Cox model by introducing a trend function in the regression coefficients. The proposed model is as follows :

$$h(x,t|Z) = h_0(x)\exp\left(\beta(t)^\top Z\right) = h_0(x)\exp\left(\sum_{j=1}^{p}\beta_j(t)Z_j\right) \tag{6}$$

$$\exists\ f:\mathbb{R}\to\mathbb{R}\ /\ \beta(t) = \beta^{le} + \beta^{tr}f(t) \tag{7}$$

The coefficients $\beta^{le} = (\beta_1^{le},...,\beta_p^{le})$ and $\beta^{tr} = (\beta_1^{tr},...,\beta_p^{tr})$ model the level and trend. f represents scenarios of future evolution of the mortality and $h(x,t|Z)$ is the instantaneous risk of death at age x and at year t knowing Z.

Application framework and results

The aim of our study is to answer an actuarial problem that is the estimation of the trend risk of basis risk. We propose here to make two applications :

- An application to the study of the basis risk between the annuitants population and the national population.
- An application to the study of a geographical basis risk : that between East Germany and West Germany.

To do this, we simulated these populations using their mortality data per year. The advantage of this approach has been the achievement of general results, not depending on the structure and composition of a portfolio, thus reducing the volatility that can result from the use of real portfolios.

The objective of these two applications is to show by the first that the proposed method enables the estimation of the risk trend and then to propose a concrete estimate of this risk through the second application.

We tested seven trend functions representing different scenarios of mortality evolution. It is the linear, quadratic, logarithmic, inverse, inverse exponential, hyperbolic tangent and cyclic ascending scenarios. The figure 4 allows us to visualize these different scenarios.

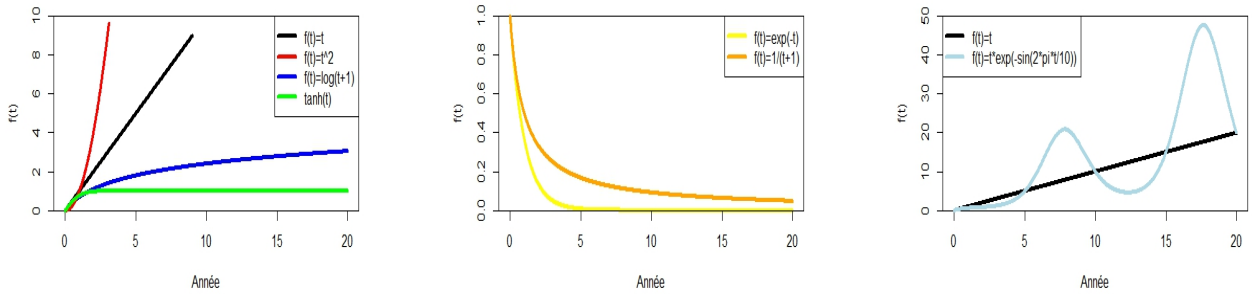

FIGURE 4 – Trend functions

Trend risk between the national population and the population of pensioners

In order to estimate the trend between the annuitants population and the national population, we simulated these populations by using the generational tables of annuitants in France and the mortality data of the Human Mortality Database between 1996 and 2005. We have adopted a period vision in order to actually visualize the improvements by year. To do this, we computed annuity tables by period from the cohort tables. Since the annuitants tables were by construction adjusted with the French national population, we expected to find the same trend on the national population. After implementing the models, we measured the quality of fit of each model and calculated the annualized quadratic deviations between the life expectancies adjusted by each model and those of the mortality tables. The results [4] are presented in Table 6.

Model	Men			Women		
	AIC	BIC	Deviations	AIC	BIC	Deviations
Hyperbolic tangent	4651670	4651690	0,08947672	3494599	3494619	0,3572199
Inverse exponential	4651661	4651681	0,07366774	3494597	3494617	0,3548081
Inverse	4651650	4651670	0,05460477	3494597	3494614	0,3517714
Cyclic ascending	4651637	4651657	0,03203398	3494591	3494611	0,3476654
Logarithmic	4651630	4651651	0,02054169	3494590	3494610	0,3463284
Quadratic	4651629	4651650	0,01992241	3494590	3494610	0,3455982
Linear	4651623	4651644	0,00865467	3494589	3494608	0,3441683

TABLE 6 – Annualized quadratic deviations, AIC et BIC

The Cox model enables us to find the expected result. The linear trend is the best fit (see the figure 5).

4. Akaike Information Criterion(AIC) ; Bayesian Information Criterion (BIC)

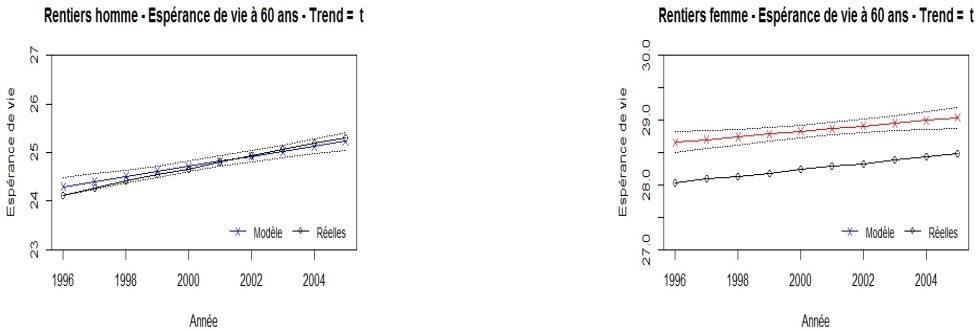

FIGURE 5 – Life expectancies adjusted with a linear trend
(On the left, the adjusted life expectancy for men, on right those of women)

The results show that the Cox can thus fully detect the trend of a subpopulation.

Estimating trend risk : Case of geographic basis risk in Germany

The partition of Germany in 1949 contributed to the divergence of life expectancy as early as 1979 within Germany. West Germany saw its life expectancy rise sharply while East Germany experienced lower improvements. However, after reunification of Germany, the trend was reversed. East Germany is catching up, while West Germany is running out of steam.

We simulated these populations using data from the Human Mortality Database, a population composed of individuals from East and West Germany between 1956 and 2015. We estimated the trends of these two populations using the Cox model.

$$h(x,t|Z) = h_0(x)\exp(\beta_E(t)Est + \beta_O(t)Ouest) \tag{8}$$

with : $\beta_E(t) = \beta_E^{le} + \beta_E^{tr} f_e(t)$ et $\beta_O(t) = \beta_O^{le} + \beta_O^{tr} f_o(t)$; $Est \in \{0,1\}$; $Ouest \in \{0,1\}$

The modeling allowed us to detect two different trends for these two populations (see table 7 and figure 6).

	East Germany	West Germany
Men	Quadratic	Linear
Women	Linear	Linear

TABLE 7 – Population Trends

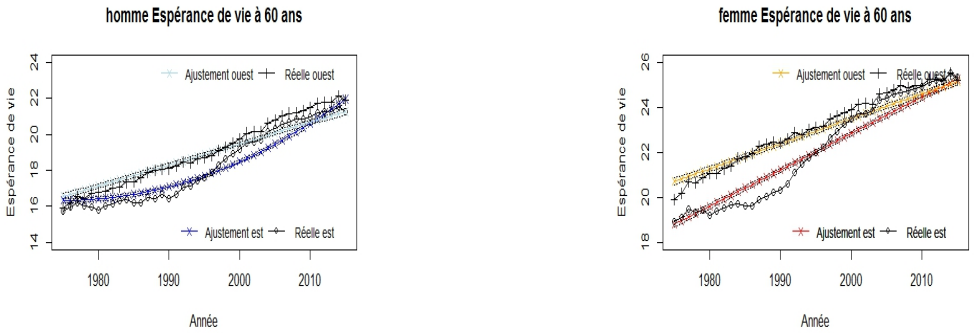

FIGURE 6 – Comparison of life expectancy
(On the left, the adjusted life expectancy for men, on right those of women)

We then adjusted the trends for the years 1995 to 2015. The results are presented in table 8.

	East Germany	West Germany
Men	Logarithmic	Linear
Women	Logarithmic	Linear

TABLE 8 – Population Trends

The Cox model was able to detect and correct the trends within this heterogeneous population without subdivision of it.

Conclusions

In this study, we have developed a pertinent approach for detecting and correcting basis risk in heterogeneous populations. We have applied the approach to the study of the basis risk between the annuitants population and the national population in France, and then to the study of the geographic basis risk in Germany. The results show that the model developed makes it possible to detect the life expectancy trend when the populations have the same trend and the heterogeneous trends when the populations have different trends. The results of the sensitivity tests show great robustness in the estimation of the trend parameter, even when the subpopulation size is reduced.

This new approach can allow the insurer to more accurately model the mortality of its portfolio and its evolution over time, taking into account heterogeneity factors and relying on the mortality of an average population to estimate the mortality of sub-populations in its portfolio. There are still many avenues for improving the approach developed. However, this study can be a framework for building a more elaborate approach to estimating the *base risk*.

Table des matières

Table des figures

Liste des tableaux

Introduction

« Nous entrons dans un nouvel Âge, un Âge où l'espérance de vie est de 90 ans et plus » [5] (Riley, 2017). Ces mots décrivent l'un des phénomènes qui fera l'objet de toutes les attentions tout au long de ce XXI^e siècle : il s'agit du phénomène de longévité.

La notion de longévité ou de longue durée de la vie était associée depuis le Moyen Âge aux souhaits de prospérité et de grandeur accordés aux rois et aux nobles. Ce souhait semble s'être réalisé pour nous aujourd'hui. L'humanité prospère au rythme des progrès de la science et des avancées technologiques. De récentes études montrent que depuis la deuxième moitié du XX^e siècle, les espérances de vie ont connu des améliorations significatives. L'Homme gagne 5 heures et demie d'espérance de vie par jour (Robin, 2013). La quête de l'immortalité est donc lancée et les anticipations d'espérance de vie se font grandissantes. En 2030 par exemple, l'espérance de vie dans bon nombre de pays développés pourrait atteindre 90 ans (Kontis et al., 2017).

Cependant, ces améliorations de l'espérance de vie ne profitent pas de la même façon aux populations. « *Les populations sont hétérogènes. Les plus fragiles ont tendance à mourir en premier* » (Vaupel, 2016). Les sources de l'hétérogénéité sont multiples. Elles peuvent être d'origine sociale, économique, géographique, génétique, culturelle ou même non expliquée. Les populations ne sont pas assujetties aux mêmes conditions de vie, aux mêmes niveaux de rémunération, ne vivent pas dans les mêmes zones géographiques, ne sont pas soumises au même environnement ni au même cadre d'éducation. En assurance, cela est communément appelé le *basis risk* ou le risque de base.

Le *basis risk* peut être subdivisé en deux composantes. La première a trait au niveau des espérances de vie au sein des populations et la seconde aux évolutions hétérogènes ou différentes d'une sous-population à une autre. L'un des enjeux de l'assurance précisément de l'assurance vie est aujourd'hui la détection et la correction du *basis risk*. En effet, prévoir l'évolution future de la survie de ses assurés est d'une d'importance capitale. Il en va de la bonne tarification des contrats et de la sécurité des assurés. Aucun assuré ne souhaiterait par exemple que son assureur fasse faillite et qu'ainsi son épargne se perde. Pour éviter ce cas de figure, l'assureur se doit de bien estimer la valeur et le prix des contrats proposés aux assurés. D'où la nécessite d'une bonne détection et prise en compte

5. « We're Entering a New Age, One Where the Life Expectancy Is 90+ »

du *basis risk*.

Nous nous sommes intéressés dans cette étude à la détection et à la correction du risque de tendance du *basis risk* au sein d'une population hétérogène. Une étude préalablement menée par Jérémy GORIS (Goris, 2016) présente une approche d'estimation du niveau du *basis risk*. Notre étude s'inscrit comme une suite de ses travaux sur le *basis risk*.

Trois approches ont été envisagées pour l'estimation du risque de tendance du *basis risk* :

La première approche consiste à modéliser les populations par la méthode de discrimination. Il s'agit de subdiviser la population en sous-groupes homogènes puis d'appliquer à chaque sous-groupe un modèle de mortalité afin d'en quantifier la mortalité spécifique.

La deuxième approche consiste à l'utilisation d'un modèle à temps de vie accéléré qui modélise les durées de survie à l'aide de variables explicatives définies par avance en supposant que ces durées de survie suivent une loi de probabilité donnée.

Enfin, la troisième approche consiste à l'utilisation d'un modèle à hasard proportionnel notamment le modèle de Cox qui, comme le modèle à temps de vie accéléré, modélise les durées de survie à l'aide de variables explicatives, mais sans utiliser une distribution de probabilité pour les durées de survie.

Dans la première partie, de notre étude, nous présenterons le contexte ainsi que les enjeux actuels liés à l'estimation du risque des tendance du *basis risk*. Dans une deuxième partie, nous exposerons les aspects mathématiques de la modélisation. Puis dans une troisième partie, nous présenterons les résultats de la méthode ainsi que ses limites.

Première partie

Cadre conceptuel et méthodologique

Chapitre 1

L'assurance vie et les fonds de pensions

> ***« La recherche sur l'allongement de la vie active me semble comparable aux recherches des alchimistes (...) [et] nous détourne à la fois d'une vérité inévitable et de programmes sociaux réalistes. »*** *(William Donald Hamilton)*[1]

1.1 Présentation générale

L'assurance vie est une branche de l'assurance dans laquelle les risques sous-jacents portent sur la durée de vie de l'assuré.

On distingue deux grandes formes de garantie en assurance vie :

- L'assurance en cas de décès : dans ce type de contrat, l'assureur déclenche le paiement des prestations au décès de l'assuré. Comme exemples simples de contrat d'assurance en cas de décès, il y a le contrat d'assurance temporaire décès et le contrat d'assurance vie entière. Le premier garantit le versement d'un capital ou le versement d'une rente aux bénéficiaires, si le décès de l'assuré survient durant une période de temps définie par le contrat. Si au terme de la période, l'assuré est toujours en vie, le contrat prend fin et les cotisations de l'assuré reviennent de droit à l'assureur. Le deuxième garantit le paiement d'un capital ou d'une rente aux bénéficiaires au décès de l'assuré, quelle que soit sa date de décès. Ce contrat a donc une durée de vie indéterminée car l'assuré peut décéder à priori à n'importe

1. Citation tirée de Vaupel (2016)

quel moment.

- L'assurance en cas de vie : ce contrat prévoit le versement d'un capital ou d'une rente en cas de survie de l'assuré à une date ou un âge défini par le contrat.

Les contrats les plus répandus parmi les assurances en cas de vie sont les contrats d'épargne et les contrats retraite.

Dans le contrat d'épargne, l'assuré décide de verser régulièrement des primes à l'assureur qui s'engage à lui verser un certain montant à terme, s'il est vivant. Ce contrat est en réalité un investissement pour l'assuré. Dans la pratique, l'assureur s'engage à rémunérer d'un certain pourcentage le capital versé par l'assuré. Ce type de contrat est donc très souvent utilisé dans le but de faire fructifier un patrimoine ou financer un projet futur.

Le contrat retraite fonctionne à peu près de la même manière qu'un contrat d'épargne. Dans ce contrat, l'assuré constitue une épargne en vue de préparer sa retraite. L'épargne accumulée et capitalisée lui est reversée par l'assureur à sa retraite, s'il est toujours vivant. Ce capital permettra ainsi à l'assuré de compenser la baisse de revenu liée à son départ en retraite. Un contrat retraite peut disposer de plusieurs particularités en fonction de la législation en vigueur dans le pays d'application. En France, l'épargne retraite est encadrée par l'article 83 et l'article 39 du Code des impôts. Il distingue deux modes de financement par l'assuré : les régimes à cotisations définies et ceux à prestations définies.

Dans les régimes à cotisations définies, les cotisations que l'assuré devra verser sont définies à l'avance et sont fixes au cours du temps. Le montant de cotisations de l'assuré est capitalisé à son propre compte et lui est reversé à son départ en retraite. Par contre, dans les régimes à prestations définies, le niveau de prestation est défini à l'avance. Cependant, le montant de capital accumulé n'est pas dû à l'assuré. L'épargne est perdue si l'assuré change d'entreprise par exemple. L'assuré peut aussi choisir de souscrire à un contrat d'épargne retraite individuel. Ces contrats sont des cas particuliers de contrats d'épargne dans lesquels le paiement des rentes débute obligatoirement au départ en retraite si l'assuré est toujours en vie. Il existe deux modes de gestion de l'épargne collectée : la gestion par répartition et celle par capitalisation.

Dans les systèmes de retraites par répartition, les générations actuelles financent la retraite des générations passées. Par contre, dans les systèmes de retraite par capitalisation, chaque génération épargne pour sa propre retraite, chaque assuré épargne pour sa retraite future. L'épargne constituée est très souvent gérée par les fonds de pension. En France, le système de retraites par répartition est obligatoire depuis 1941. Cependant, d'autres pays comme les Etats-Unis, le Royaume Uni ou le Chili privilégient un système de retraite par capitalisation.

Le régulateur européen, à travers la directive solvabilité (Européenne, 2016), a iden-

tifié au moins sept risques auxquels doit faire face un assureur vie. Il s'agit des risques de mortalité, de longévité, d'invalidité-morbidité, de dépense en vie, de révision, de rachat et de catastrophe en vie. Nous nous intéresserons de plus près aux risques liés à la durée de la vie humaine à savoir le risque de mortalité et de longévité :

- Risque de mortalité : il s'agit du risque de la perte de capital que peut engendrer une hausse de la mortalité. Ce risque concerne particulièrement les contrats d'assurance en cas de décès. En effet, si la mortalité augmente l'assureur devra payer plus de rentes aux assurés. Le risque sous-jacent à ces contrats est donc celui de voir la mortalité augmenter brusquement.

- Risque de longévité : il s'agit du risque de perte liée à une diminution de la mortalité des assurés. C'est le risque de voir les assurés vivre plus longtemps que prévu. Les contrats d'assurance en cas de vie sont les plus exposés à ce risque. En effet, si les assurés vivent plus longtemps que prévu, l'assureur devra alors leur verser un capital sur une plus longue période que celle anticipée. L'assureur s'expose ainsi dans ces contrats au risque ne pas disposer d'un capital suffisant pour financer la rente qui est due aux assurés.

Le risque de longévité est un risque important auquel font face les fonds de pension, surtout ceux qui sont à prestations définies en raison de leur exposition à devoir financer la retraite des assurés si ceux-ci vivent plus longtemps que prévu.

Un fonds de pension ou fonds de retraite est une structure publique ou privée dont le but est de financer la retraite des assurés. Les assurés, le plus souvent des salariés, versent leurs cotisations au fonds qui leur reversera chaque année une rente à leur départ en retraite. Ces fonds de pension fonctionnent très souvent par capitalisation et sont gérés par l'assureur.

En guise d'illustration de l'impact du risque de longévité sur la gestion des fonds de pension, supposons qu'un fonds de pension promette de payer 2 000€ par mois à 10 individus dès leur départ à la retraite fixé à 60 ans. Le taux d'actualisation est fixé à 0%. On suppose que les individus ont une espérance de vie de 80 ans. Alors, le fonds devra verser sur 20 ans en moyenne, 2 000€ par mois à chaque individu, soit un total de 4 800 000€. Supposons que l'évolution de l'espérance de vie soit conforme aux résultats publiés par le Groupe AXA (Robin, 2013) selon lesquels l'espérance de vie augmenterait de 5 heures et demie chaque jour. Alors sur une période de 20 ans, le gain d'espérance de vie serait de plus de 4 années et demi. En supposant que les individus soient à l'âge de leur retraite alors, le fonds de pension devra dépenser maintenant 5 880 000€ au lieu des 4 800 000€ initialement prévus. Au bout de 20 ans, le fond de pension aura donc dépensé 1 080 000€ de plus qu'initialement prévu. Dans la pratique, le fond de pension gère les retraites de centaines voire de milliers d'individus.

Le risque de longévité est un risque d'envergure pour les contrats de retraite (individuelle et collective) et les assureurs vie en général. L'amélioration constante de l'espérance de vie introduit de nouveaux défis et de nouvelles problématiques en matière d'organisation, de gestion des systèmes de retraite et en matière de gestion du risque pour l'assureur et le régulateur.

1.2 La longévité : un nouveau défi

Durant plusieurs siècles, une idée largement répandue était qu'il y aurait une frontière fixe à la durée de vie humaine. Pour étayer cela, Aristote dans un essai comparait la « chaleur vitale » de la vie à un feu qui se consumait (Vaupel, 2016). Ce feu peut s'éteindre prématurément lorsqu'il subit l'effet du sable ou de l'eau par exemple. Cette extinction peut correspondre à la mort d'un Homme par accident ou par maladie. Cependant, ce feu peut aussi s'éteindre naturellement après s'être entièrement consumé : c'est la mort de vieillesse. Cette mort est inéluctable. C'est pourquoi bon nombre de nos contemporains pensait que la seule cause de mortalité aux âges avancés était la vieillesse. « La vieillesse c'est le déclin, on n'y peut rien » (Guérette et al., 2006). S'il est vrai que l'Homme ne peut rien face à la détérioration de sa durée de vie avec l'âge, un constat est que cette détérioration n'est pas la même dans le temps et a même tendance à s'atténuer. Vaupel (2016), a montré que la pensée d'Aristote était vérifiée jusqu'aux années 1950. En effet, avant 1950, on peut observer que les progrès en terme d'amélioration d'espérance de vie ont été très limités. Par contre, après 1950, les améliorations ont été significatives. Le risque de décès à 80 ans des femmes suédoises par exemple a été réduit de près de 10%, de même que celui des hommes suédois à 90 ans. En ce qui concerne la population française, on constate que la probabilité de mourir à 80 ans en 2016 est la même que celle de mourir à 71 ans en 1966. Depuis 1950, dans bon nombre de pays tel que la France, l'Italie ou le Japon, l'espérance de vie a augmenté de 2 ans chaque décennie soit environ 3 mois par an (Oeppen, W.Vaupel, 2002).

Ces améliorations de la durée de vie sont encore plus prononcées aux âges avancés. En France, le nombre de centenaires a été multiplié par 20 entre 1970 et 2016. Tandis qu'en 1970 on dénombrait à peu près 1100 centenaires, on en dénombre 21 000 aujourd'hui (INSEE, 2016). Le nombre de centenaires pourrait être multiplié par 13 en 2070. Entre 15% et 48% des filles nées en 2016 pourraient atteindre l'âge de 100 ans (INSEE, 2016).

Ainsi, aujourd'hui on assiste non seulement à une augmentation de l'espérance de vie, mais les limites de la durée de vie humaine sont sans cesse repoussées. Le record de longévité jamais enregistré demeure toujours celui de la française Jeanne Calment qui a vécu 122 ans et 164 jours, puis vient en seconde position l'Américaine Sarah Knauss qui a vécu 119 ans et 97 jours [2].

2. https://www.supercentenarian.com/records.html

Le phénomène de longévité prend une ampleur considérable tant sur le plan politique, social, qu'économique. Sur le plan politique, ce phénomène a conduit à de nouvelles politiques de gestion des fonds de pension et des systèmes de retraite dans la plupart des pays industrialisés. L'une des mesures phares en France a été l'allongement de l'âge légal de départ à la retraite afin de permettre aux personnes en activité de cotiser plus longtemps. D'autres mesures phares ont déjà été énoncées lors de la campagne présidentielles en 2017 comme l'uniformisation des méthodes de calcul de la pension retraite en France. On pourrait donc s'acheminer vers un système de retraite par capitalisation. En 2010, le déficit économique par assuré au niveau des régimes obligatoires était estimé à près de 2 000 € (Rougerie, 2010). Le risque de voir le système de retraite s'effondrer est donc bien réel d'autant plus qu'on observe une diminution de la population active au fil des années. Au-delà de l'enjeu économique, ces réformes visent aussi à rétablir la confiance des assurés et retraités vis-à-vis du système en vigueur et lever le climat d'incertitude autour de la retraite future des actifs.

Une des particularités du risque de longévité est qu'il est un risque de long terme. Sa gestion requiert le provisionnement d'un capital important. La nécessité de bien anticiper l'évolution future de la longévité est donc un enjeu crucial pour les compagnies d'assurance. Face à cette problématique, de plus en plus d'assureurs préfèrent céder ce risque à d'autres acteurs tels que les réassureurs ou aux marchés financiers. Toutefois, la grande problématique du transfert du risque de longévité demeure toujours celle de l'anticipation des évolutions futures de mortalité.

Le risque de longévité peut être décomposé en deux composantes : le niveau [3] et la tendance [4]. Le niveau désigne la valeur actuelle de la mortalité, tandis que la tendance désigne l'évolution de cette valeur dans le temps. Une question fondamentale est quelle sera la tendance d'évolution de la mortalité dans le temps ? Observera-t-on un essoufflement ou plutôt une accélération de l'évolution des espérances de vie ? Les avis sont divergents sur ce sujet. S'il est « unanimement » reconnu que l'espérance de vie a connu une évolution linéaire par le passé, beaucoup pensent que l'on n'observera pas les mêmes évolutions à l'avenir. Il existe trois écoles de pensée : les futuristes, les optimistes et les réalistes (Olshansky, Carnes, 2009).

Les futuristes pensent que l'espérance de vie poursuivra son évolution avec une plus forte pente que par le passé c'est-à-dire que l'augmentation de la longévité sera de plus en plus marquée. En effet, les avancées de la science sur le plan biomédical et technologique qui ont contribué à l'éradication de certaines maladies et à la vulgarisation de nouveaux modes de vie, permettent à l'Homme de réduire considérablement son risque de décès. L'on pourrait donc penser que les limites de la vie humaine seront repoussées sans cesse, voire jusqu'à l'infini. Pour les futuristes, trois « ponts » mèneraient à l'immortalité humaine (Olshansky, Carnes, 2009). Le premier fait référence aux changements

3. Level
4. *trend*

de style de vie et à de meilleurs prise en charge de la santé qui, combinés, permettrait à l'Homme de vivre 20 années supplémentaires au-delà de l'espérance de vie actuelle. Le second concerne les avancées dans la technologie biomédicale comme la thérapie par cellules souches, le génie génétique [5] ou le rajeunissement, qui devraient être développés d'ici la moitié de ce siècle. Le troisième pont concerne le développement de technologies plus évoluées et prometteuses telles que les nanotechnologies. Une fois ces trois « ponts » franchis, l'immortalité serait alors atteignable.

La quête du « gène de l'immortalité » ne cesse de passionner les recherches. Les chercheurs ont identifié sept différences cellulaires et moléculaires entre les jeunes et les personnes âgées qui pourront permettre d'éliminer le vieillissement et ainsi rendre possible l'immortalité physique (GREY et al., 2002). Pour l'heure, le passé récent de la longévité permet de soutenir les arguments d'une possible « immortalité » de l'Homme dans le futur.

Les optimistes, s'appuyant sur des considérations démographiques, pensent que l'espérance de vie suivra la même tendance que celle du passé. Les récentes études montrent des *trends* d'espérance de vie de 3 mois par an dans la plupart des pays industrialisés (Oeppen, W.Vaupel, 2002). L'Asie de l'est par exemple qui avait une espérance de vie de 45 ans en 1950 a connu de fortes améliorations d'espérance de vie. D'après le ministère de la santé du Japon, Hong-Kong enregistre le plus haut niveau d'espérance de vie dans le monde avec 87,34 ans pour les femmes et 81,32 ans pour les hommes, suivi de près par le Japon avec 87,14 ans pour les femmes et 80,98 ans pour les hommes [6]. De même, dans la plupart des pays industrialisés l'espérance de vie poursuit sa progréssion. Les projections montrent qu'elle pourrait atteindre 90 ans d'ici 2030 (Kontis et al., 2017). Pour les optimistes, aucune raison ni démographique ni biologique ne pourrait conduire à penser que ces améliorations s'estomperaient à l'avenir (Wilmoth, 2001). Cette vision optimiste est soutenue par les données historiques de même que par les anticipations de bon nombre de modèles de projection de la mortalité.

Enfin, les réalistes pensent quant à eux que l'espérance de vie augmentera plus lentement que dans le passé. La thèse soutenue est qu'il y aurait une limite à la durée de vie de l'Homme, de même qu'à celle de tout organisme vivant. Cette limite est marquée par la vieillesse et est une contrainte biologique à l'immortalité. En effet, le déclin physiologique lié à la reproduction cellulaire, la perte des près de 80% de leurs capacités à 80 ans et l'apparition de pathologies aux âges élevés constituent une limite biologique à l'immortalité humaine (Carnes et al., 2003). Les récentes données de mortalité justifient cette école de pensée. En effet, les améliorations d'espérance de vie ont connu très récemment une décélération dans certains pays comme le Royaume-Uni (Coghlan, 2017) ou les États-Unis (Monde, 2016). Cette diminution serait liée à une mortalité inhabituelle du fait de l'augmentation considérable de la maladie d'Alzheimer, et de celle d'autres pathologies respiratoires, cardio-vasculaires et attaques cérébrales.

5. ou ingénierie génétique
6. https://mainichi.jp/english/articles/20170728/p2g/00m/0dm/002000c

Par ailleurs, l'âge maximal jamais enregistré demeure toujours 122 ans et 164 jours.

Quoi qu'il en soit, l'amélioration de l'espérance de vie est un constat bien réel et un nouveau défi pour les sociétés. Sur le plan médical, le traitement et la gestion des maladies liées à la vieillesse devient une problématique de plus en plus importante. En effet, les années d'espérance de vie gagnées le sont en général durant les années de bonne santé. Ainsi, la santé des personnes âgées déterminera certainement le *trend* futur de longévité. L'enjeu majeur serait alors d'éliminer toutes ces maladies de la vieillesse qui ont fortement progressé ces dernières années.

La longévité est un phénomène réel qui prend de l'ampleur dans le monde entier. Il suscite de nos jours des débats et des recherches de plus en plus passionnants. Le dernier en date est lié à la limite actuelle de la durée de vie humaine qui serait de 115 ans, selon des chercheurs de la faculté de médecine Albert Einstein de New York aux Etats-Unis [7]. Pour les uns, il serait de 125 ans ou 120 ans, pour les autres il n'y aurait aucune limite à la durée de vie de l'Homme. Loin d'être résolu, bon nombre de recherches s'orientent désormais vers l'hétérogénéité au sein des populations. En effet, les gains d'espérances de vie ne sont pas repartis uniformément au sein des populations. Comme simple exemple, l'espérance de vie à la naissance, en 2015, en France a été de 85 ans pour les femmes tandis que celle des hommes a été de 78 ans. Cependant, au cours de ces 20 dernières années, l'espérance de vie des femmes a connu une amélioration de 3,1 années tandis que celle des hommes s'est améliorée de 5,1 années. Ces améliorations différentes sont constatées sur toute la population française, réparties selon des considérations économiques, sociales et géographique. On voit donc émerger une nouvelle problématique qui est celle des améliorations différentes de mortalité au sein d'une population hétérogène. Communément appelé *basis risk*, le *basis risk* en longévité est un risque émergent qui a un impact significatif sur les compagnies d'assurance et fonds de pension.

Dans le chapitre suivant, nous introduirons le *basis risk* et ses différentes formes. Nous étudierons les différentes causes du basis et l'impact financier engendré sur les compagnies d'assurance.

7. http://www.nature.com/nature/journal/v538/n7624/full/nature19793.html

Chapitre 2

Le *basis risk*

« The research highlights that "one size does not fit all" and that using standard actuarial projections may lead to misleading funding assumptions » *(Graham Vidler, PLSA director of external affairs)*

2.1 Comprendre le *basis risk* et ses enjeux

Depuis la deuxième moitié du XX^e siècle, l'espérance de vie a augmenté d'environ 3 mois par an dans la plupart des pays industrialisés. Cependant, ce gain n'est pas uniformément réparti au sein des populations. Certains sous-groupes au sein de la même population présentent une espérance de vie évoluant avec une forte pente, tandis que d'autres voient leur *trend* d'espérance de vie s'atténuer. Ces améliorations différentes de mortalité au sein d'une même population introduit un risque appelé le *basis risk*. Le *basis risk*, peut être décomposé en deux composantes :

— Le niveau : c'est le niveau actuel du *basis risk*, c'est-à-dire, les différences d'espérances de vie au sein d'une population hétérogène.
— La tendance : il s'agit de l'évolution de ces différences au fil du temps.

2.1.1 Illustration de l'impact financier du *basis risk*

La couverture du *basis risk* a pris une importance particulière en raison de son impact sur les compagnies d'assurance et les fonds de pension. En guise d'illustration de l'impact du *basis risk*, considérons une population avec un niveau d'espérance de vie de 80 ans. On suppose de plus que son *trend* d'espérance de vie est de 3 mois par an. Une sous-population « A » réalise quant à elle un gain d'espérance de vie de 5 mois par an.

« Assurance Vie » un assureur vie de la place dispose d'un portefeuille retraite d'âge moyen de 50 ans, constitué majoritairement d'individus de la sous-population « A ». « Assurance Vie » s'est engagé à verser une rente de 1 000 € par mois à chaque assuré à son départ en retraite (à 60 ans). Pour plusieurs raisons « Assurance Vie » ne dispose pas de données précises et fiables lui permettant d'estimer non seulement le niveau actuel de la mortalité mais aussi le *trend* sous-jacent. Il décide donc de baser ses estimations de mortalité sur les données de populations nationales. Pour des questions de simplicité, on suppose que l'espérance de vie reste figée après 30 ans et que le taux d'actualisation est de 0%.

Le *trend* étant de 3 mois par an sur toute la population, « Assurance Vie » considérera que dans 30 ans, le niveau d'espérance de vie serait de $80 + \left(30 \times \frac{3}{12}\right)$ années soit 87,5 ans. Ainsi, pour un portefeuille de 1000 individus, l'assureur mettra en provision une somme de $1000€ \times 12 \times 1000 \times (87,5 - 60)$, soit 330 M€. Cependant, dans la réalité, l'espérance de vie de la population de son portefeuille atteindra $80 + \left(30 \times \frac{5}{12}\right)$ soit 92,5 ans dans 30 ans et l'assureur dépensera en réalité $1000€ \times 12 \times 1000 \times (92,5 - 60)$ soit 390 M€. Ainsi, celui-ci constatera une perte de 60 M€ dans 30 ans. Le *basis risk* peut donc avoir un impact financier important.

2.1.2 Les facteurs de source du *basis risk*

Au sein d'une population hétérogène, plusieurs facteurs peuvent être source de *basis risk*.

Une source évidente de *basis risk* est liée au genre c'est-à-dire aux différences de mortalité entre hommes et femmes. Les femmes vivent plus longtemps que les hommes. En se référant aux données de l'Institut national de la statistique et des études économiques [1] (Voir figure 7), on constate que l'écart d'espérance de vie [2] des hommes et femmes a été de 6,1 années en 2015 et en 2016. Plusieurs raisons peuvent expliquer ce fait. En effet, les femmes ont un avantage biologique qui leur assure un meilleur système immunitaire. Celles-ci souffrent plus de maladies chroniques moins mortelles que celles auxquelles sont exposés les hommes. « Les hommes et les femmes ne sont pas égaux devant la maladie » (tdg, 2016). Aussi, la longévité des femmes peut s'expliquer par leur attitude vis-à-vis de leur santé et à leur comportement moins risqué. L'évolution des espérances de vie montre cependant que les écarts se resserrent. De 8,2 années en 1994, le gap entre homme et femme est désormais de 6,1 années, soit une diminution de 2,1 ans. On assiste donc à une convergence de la longévité hommes et femmes. La figure 7 présente l'évolution des espérances de vie des hommes et des femmes. Aux âges plus élevés par contre, les chiffres sont beaucoup plus faibles. L'écart d'espérance de vie à 60 ans a régressé de 5,3 ans à 4,4 soit une baisse de moins d'un an (voir figure 7). Les hommes rattrapent timidement

1. Insee
2. Espérance de vie à la naissance sauf indication contraire

leur retard aux grands âges.

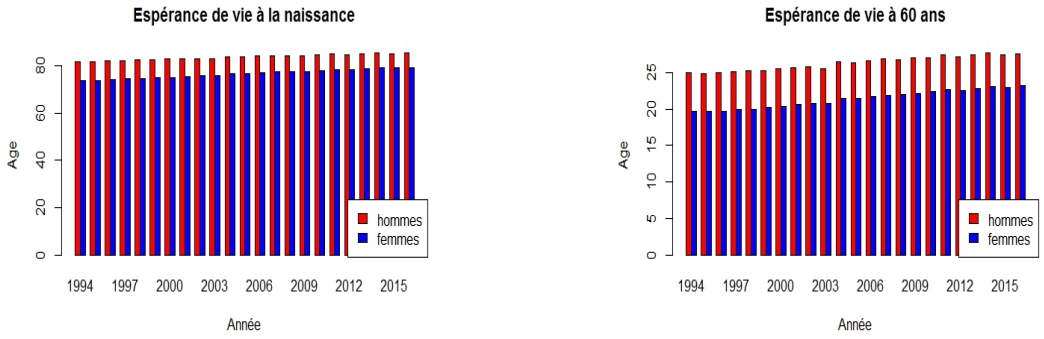

FIGURE 7 – Espérance de vie hommes et femmes entre 1994 et 2016
Source : Insee

Une autre source de *basis risk* est liée aux conditions socio-économiques. En effet, l'espérance de vie varie en fonction du revenu, de l'éducation et de la catégorie socio-professionnelle. Les plus riches vivent plus vieux que les moins riches. L'une des explications est intuitive. Les riches bénéficient de meilleures conditions de vie et ont un meilleur accès aux soins que les catégories sociales défavorisées. La PLSA [3], organisme au Royaume-Uni aidant à la gestion des fonds de pension, a réalisé une étude (PLSA, 2017) portant sur l'évolution de l'espérance de vie en fonction des facteurs socio-economiques. Elle a étudié une population de pensionnaires subdivisée en trois groupes selon leur revenus et le niveau de vie dans la zone géographique de résidence. Le premier groupe est constitué des individus percevant une pension mensuelle de retraite de plus de 8 300€ ou ceux percevant une pension de 5 500€ et habitant dans une zone avec un niveau de vie moyen ou faible. Le deuxième groupe est constitué des individus percevant entre 5 500€ et 8 300€ ou moins de 5 500€ et habitant en zone avec un faible niveau de vie. Le troisième groupe est constitué des individus ayant une pension inférieure à 5500€. Les résultats (pour les hommes) montrent que les individus du premier groupe ont vu leur espérance de vie à 65 ans évoluer de 18,1 ans à 20,3 ans entre 2000 et 2015, soit une amélioration de 2,2 années. Les individus du deuxième groupe qui avait une espérance de vie à 65 ans de 16,6 ans ont quant à eux connu une évolution 2,1 années entre 2000 et 2015 jusqu'à 18,7 ans. L'espérance de vie à 65 ans des individus du troisième groupe est quant à elle passée de 14,4 ans à 17 ans, soit une évolution de 2,6 années entre 2000 et 2015. Bien qu'ayant une espérance de vie plus faible, les individus du troisième groupe ont connu la plus forte amélioration sur toute la période. Le constat est encore plus surprenant lorsque l'on observe plus finement le *trend* de longévité sur les intervalles de temps donnés (voir tableau 9). Entre 2000 et 2005, les individus du troisième groupe ont connu la plus forte amélioration d'espérance de vie. Entre 2005 et 2010, ceux-ci ont cédé

3. Pensions and Lifetime Savings Association

leur place à ceux du deuxième groupe. Par contre entre 2010 et 2015, ceux du deuxième groupe ont connu le plus faible *trend* d'espérance de vie et ceux du premier groupe ont connu la plus forte progression. Contrairement à ce que l'on s'attendait à voir, entre 2010 et 2015, l'écart s'est creusé entre les individus les plus riches et les autres. Cependant, les individus ayant le plus faible niveau de pension tendent à rattraper leur retard.

	2000-2005	2005-2010	2010-2015
Troisième groupe	1 année	1,2 années	0,4 année
Deuxième groupe	0,6 année	1,4 années	0,1 année
Premier groupe	0,7 année	0,9 année	0,6 année
Population étudiée	0,9 année	1,3 années	0,5 année

TABLE 9 – Gain d'espérance de vie par groupe.
Source : Club Vita analysis for PLSA longevity trends report. 2017. Premier groupe : "comfortable" group, having retirement income of over £7,500 per year, or £5,000 if they live in the least deprived parts of the UK. Deuxième groupe : "making-do" group of people with modest retirement income (<£5,000), living in areas of average to low levels of deprivation. Troisième groupe : "hard-pressed" group of people living in deprived areas on low income.

Ainsi, le *trend* d'espérance de vie varie selon la catégorie sociale des individus. En France, d'après les données de l'Insee, les cadres vivent plus longtemps que les ouvriers et les inactifs (voir figure 8) et la tendance n'est pas à la baisse. La différence d'espérance de vie à 60 ans entre cadres et inactifs est passée de 6,9 années à 8,4 années entre 1976 et 2013. Du côté des femmes les écarts sont plus faibles mais ont tout de même légèrement évolué à la hausse (voir figure 8). L'écart entre cadres et inactives a évolué de 2,3 années à 2,8 années durant la même période. Les hommes cadres vivent de nos jours 8,1 années de plus que les inactifs et 4,4 années de plus que les ouvriers. Les femmes cadres quant à elles vivent 2,8 années de plus que les inactives et 2,1 années de plus que les ouvrières.

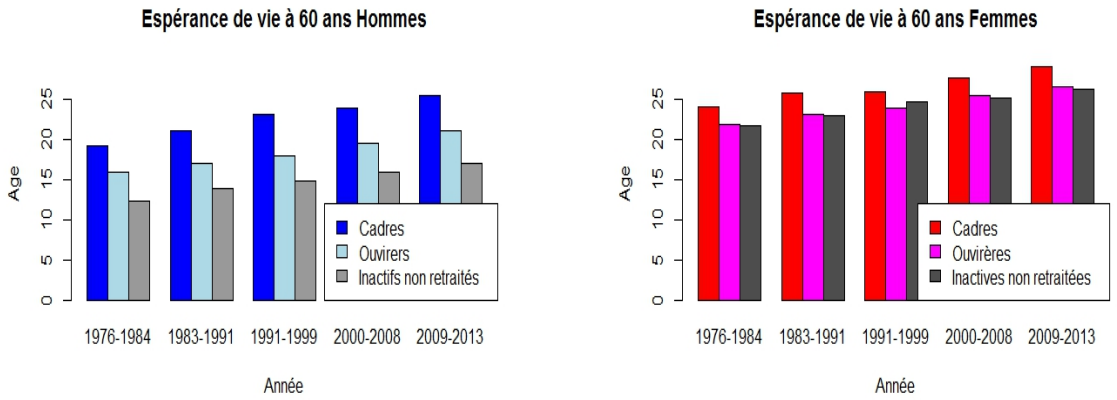

FIGURE 8 – Espérance de vie à 60 ans en fonction de la catégorie socio-professionnelle

Source : Étude Insee

La tendance est cette fois-ci à la convergence lorsque l'on considère l'espérance de vie à 60 ans en fonction du niveau d'étude (voir figure 9).

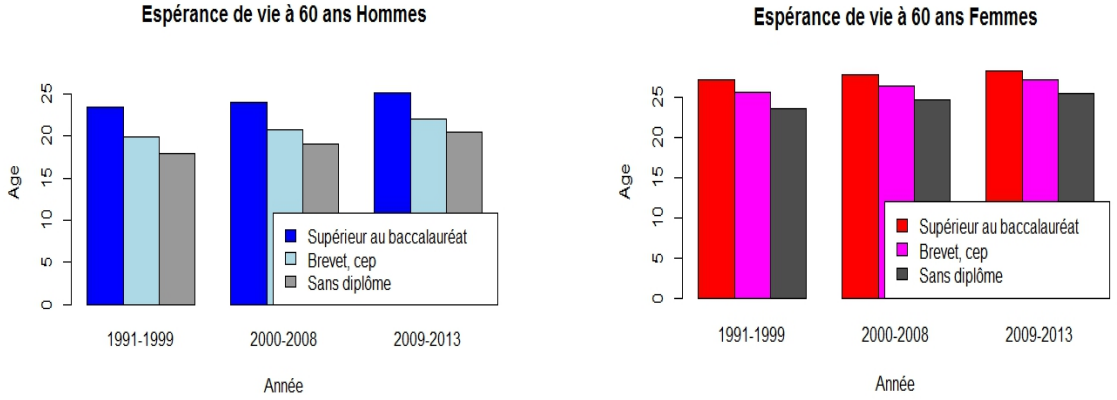

FIGURE 9 – Espérance de vie à 60 ans en fonction du niveau d'étude

Source : Etude Insee

Toutes ces différences peuvent s'expliquer par le fait que le niveau d'étude, la catégorie socio-professionnelle et le revenu déterminent le niveau de vie des individus. Avec la volonté et les efforts de l'état Français et partout ailleurs d'assurer un niveau de vie acceptable pour tous, il est tout à fait normal de constater que les populations les plus démunies rattrapent. Cependant, Des efforts restent encore à faire pour assurer de meilleurs conditions de vie aux populations démunies.

La position géographique peut aussi être source de *basis risk*. Les populations de certaines zones géographiques vivent plus longtemps que d'autres. Une explication rationnelle peut être que les facteurs socio-économiques déterminent en partie les lieux de résidence des individus. Il serait donc cohérent de constater que les populations des zones résidentielles vivent plus longtemps que celles des zones défavorisées par exemple. En France, on observe une grande disparité de l'espérance de vie selon les départements (voir figure 10). L'écart maximal demeure toujours celui entre le Pas-de-Calais et les Hauts-de-Seine. En 1977, cet écart était de 5,9 ans pour les hommes et 4,2 ans pour les femmes tandis qu'en 2007, il était de 6 ans pour les hommes et de 3,4 ans pour les femmes (Barbieri, 2013). En 2015, les écarts ont connu une légère baisse et sont de 5,7 années pour les hommes et 2,9 années pour les femmes. Aux âges avancés, les écarts sont beaucoup plus faibles soit 3,9 années pour les hommes et 2,2 années pour les femmes à 60 ans. En 2008, une étude publiée par la World Health Organisation (WHO) [4] présentait un effet surnommé l'« effet Glasgow ». L'étude souligne que l'espérance de vie d'un enfant en zone défavorisée de Glasgow en Écosse était de 28 ans inférieure à celle des zones prospères d'East Dunbartonshire, juste à 13 kms de Glasgow. Comme explications, certains pointent l'état de santé de ces populations, pourtant cela ne permet d'expliquer un tel écart n'existant nulle par ailleurs au Royaume-Uni. Il n'existe pas toujours des explications évidentes de ces disparités selon les zones géographiques. Le facteur socio-économique ne permet pas de justifier tout le *basis risk* géographique et plusieurs contradictions subsistent. L'une des plus passionnantes provient du japon, précisément de l'île d'Okinawa, surnommée l'« île des centenaires » en raison de la longévité exceptionnelle que connaissent ses habitants. Le paradoxe est que la préfecture d'Okinawa est l'une des régions les plus pauvres du Japon. Une explication de cette longévité exceptionnelle serait le mode de vie traditionnel conservé par ceux-ci en termes d'alimentation et d'habitudes. Loin du stress et de l'effervescence des grandes métropoles, ses habitants disposent d'un cadre favorable à la longévité. Ainsi, l'environnement dans lequel grandissent les individus et leurs attitudes ont un impact sur leur mortalité.

4. http://www.who.int/bulletin/volumes/89/10/11-021011/en/

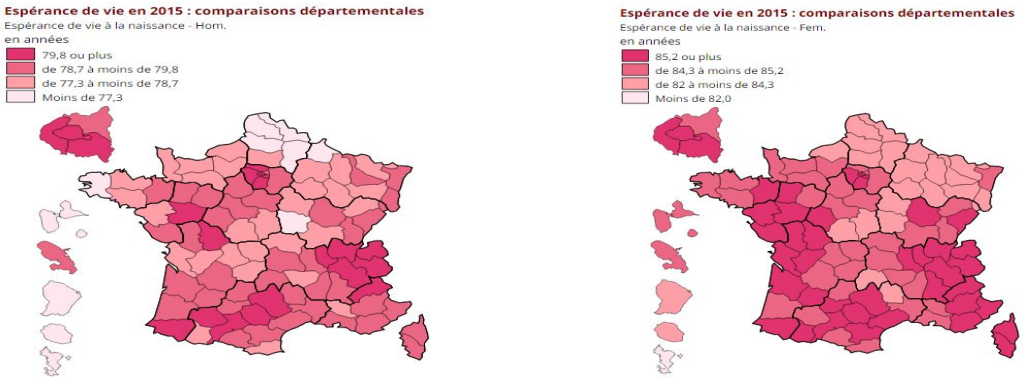

FIGURE 10 – Espérance de vie par département en 2015

Source : Insee, État civil, Estimations de population.

De récentes études sur les jumeaux tendent à prouver qu'une source de *basis risk* serait génétique. Elles montrent que l'hérédité de l'espérance de vie serait de 25%. Aux âges avancées, l'hérédité serait même de 33% pour les femmes et 48% pour les hommes. Le chromosome 3 serait porteur de la longévité (Brooks-Wilson, 2013). Une longévité exceptionnelle est expliquée donc en partie par un facteur génétique hérité des parents. On pourrait donc penser que les régions ayant subi une forte immigration verront donc leur espérance de vie changer du fait des changements génétiques.

Pour certains auteurs comme Tom Kirkwood, il y aurait une part de chance qui serait plus significative que facteur génétique dans la longévité. Toutefois, au-delà du facteur génétique et de la chance, les écarts d'espérance de vie semblent se dessiner déjà dès la naissance. En effet, le mois de naissance a une influence sur la durée de vie après 50 ans (Doblhamme, Vaupel, 2001). Les personnes nées en automne (Octobre-Décembre) par exemple vivraient plus longtemps que celles nées au printemps (Avril-Juin) [5], mais ces différences se réduiraient au fil du temps. Une explication pourrait être que les enfants nés au printemps font face très vite à la chaleur de l'été, favorable au développement des bactéries ce qui pourrait les rendre plus vulnérables.

Les sources du *basis risk* sont multiples. Sa prise en compte dans l'assurance peut donc s'avérer très complexe. En effet, prétendre éliminer ce risque impliquerait d'étudier la mortalité de chaque individu en fonction de ses caractéristiques propres, sa génétique, son passé et bien d'autres facteurs. Dans la partie suivante, nous aborderons la prise en compte du *basis risk* dans l'assurance du point de vue de l'assureur et de la réglementation.

5. Données de l'Australie et le Dannemark

2.2 Le *basis risk* dans l'assurance

Dans les chapitres précédents nous avons présenté le *basis risk*, son impact financier, ainsi que les facteurs expliquant son origine. Sa prise en compte parfaite est très complexe, mais l'assureur doit trouver des stratégies pour se couvrir contre ce risque.

La gestion des risques en assurance est encadrée par la réglementation Solvabilité 2 qui est en vigueur en Europe depuis l'année 2016. Concernant le risque de longévité, Solvabilité 2 impose aux entreprises d'assurance le provisionnement d'un montant leur permettant de faire face « à la perte de fonds propres de base des entreprises d'assurance et de réassurance résultant de la baisse soudaine permanente de 20 % des taux de mortalité utilisés pour le calcul des provisions techniques » (Européenne, 2016).

Concernant le *basis risk*, aucune mesure n'est proposée par le régulateur pour couvrir ce risque. En effet, la prise en compte du risque de longévité par la réglementation n'inclut pas le *basis risk* lié à l'hétérogénéité au sein des portefeuilles d'assurance.

La réglementation encadre par ailleurs le choix des tables de mortalité à utiliser par les assureurs. Depuis le 1^{er} janvier 2007, les tables réglementaires pour les rentiers en France sont les tables réglementaires TGH 05 pour les hommes et TGF 05 pour les femmes. Ces tables générationnelles prospectives des rentiers ont été construites à partir de 19 portefeuilles, soit 2 000 000 d'individus, sur une période de 1993 à 2005 (Planchet, 2006). La réglementation interdit aux assureurs la distinction entre hommes et femmes lors de tarification des contrats. Cependant, pour le provisionnement, ceux-ci peuvent effectuer une distinction entre hommes et femmes. Une première approche dans la prise en compte du *basis risk* pourrait donc être la différenciation entre hommes et femmes dans le cadre du provisionnement. Les tables TGH 05 et TGF 05 permettent à l'assureur de prendre en compte les écarts de mortalité liés au genre.

Au-delà de la distinction homme et femme, l'assureur peut considérer tous les facteurs de *basis risk* pour son estimation, à condition de toujours se conformer à la réglementation en vigueur dans le pays. En effet, l'utilisation de certaines données sur les assurés est proscrite par la réglementation dans le but de les protéger. Une des contraintes à cette démarche est aussi le problème de données. Les données disponibles n'ont pas toujours un historique suffisamment long permettant la modélisation des sous-populations.

Une autre forme du *basis risk* en assurance est celle entre population nationale et population assurée. En effet, la mortalité de la population nationale est supérieure à celle des assurés. Une explication pourrait être que la population assurée bénéficie de meilleures conditions de vie que la population nationale. Par ailleurs, souscrire à un contrat retraite nécessite de disposer d'un minimum de revenu. Il existe donc des différences significatives entre l'espérance de vie de la population nationale et celle de la population assurée. En

2005, l'espérance de vie des hommes de la population nationale était de 76 ans [6], tandis que celle des rentiers était de 81 ans [7] soit une différence de 5 années. La couverture du *basis risk* entre population nationale et population assurée est donc indispensable pour l'assureur.

Le régulateur impose l'utilisation des tables générationnelles prospectives de la population des rentiers par l'assureur pour la tarification et le provisionnement. L'assureur se couvre donc nécessairement contre le *basis risk* de niveau. Cependant, la couverture contre le risque de tendance du *basis risk* demeure jusqu'à présent très obscure. En effet, l'utilisation des tables prospectives de mortalité réglementaire de rentiers n'assure pas une bonne estimation de la tendance de longévité de la population assurée. Ces tables ont été extrapolées à partir des tables de la population nationale, grâce à une régression des logits [8] des taux de mortalité (Planchet, 2006). Par conséquent, les améliorations de mortalité observées au sein de ces tables sont du même ordre que celles observées sur la population nationale. Cependant, aucune étude ne permet de vérifier que la population nationale et assurée observe le même *trend* de longévité. Les compagnies d'assurance ne peuvent pas estimer rigoureusement le risque de tendance du *basis risk* entre la population nationale et celle assurée. Il existe donc un « flou » quant à l'estimation du risque de tendance du *basis risk*. L'étude de la tendance propre aux assurés est d'un enjeu capital et fera l'objet de l'application de notre étude en dernière partie. Dans son mémoire intitulé « *Risque de longévité : la référence à la tendance de la population nationale est-elle justifiée ?* », Anne-Sophie Musset montre que la population assurée possède un *trend* propre à elle-même (Musset, 2016). L'estimation rigoureuse du risque de tendance du *basis risk* entre population nationale et population assurée est donc importante pour les compagnies d'assurance.

L'Autorité Européenne des Assurances et des Pensions Professionnelles (EIOPA) a été critiquée ces dernières années en raison de la moindre prise en compte du *basis risk* sous Solvabilié 2. Certains militent d'ores et déjà en faveur de la création d'un sous module *basis risk*. En effet, il peut être extrêmement complexe pour l'assureur de calibrer lui-même ce risque car l'hétérogénéité est propre au portefeuille. Les assureurs s'exposeraient en l'absence de données fiables et d'une grande profondeur d'historique à une mauvaise estimation du risque et à une forte volatilité autour de l'estimation de la tendance du *basis risk*. La nécessité d'harmoniser la prise en compte de ce risque est donc réelle.

Dans les parties suivantes, nous nous intéresserons à l'estimation du risque de tendance du *basis risk*. Nous proposerons une approche simple et intuitive permettant d'es-

6. Données de la Human Mortality Database(HMD), *created to provide detailed mortality and population data to researchers, students, journalists, policy analysts, and others interested in the history of human longevity.*

7. Calculé à partir de la table TGH par période

8. *Rappel :* $\text{logit}(x) = \ln\left(\frac{x}{1-x}\right) \quad x \in]0\ ,\ 1[$

timer le *basis risk* suivant les facteurs d'hétérogénéité au sein d'un portefeuille. L'objectif est de montrer que le modèle proposé permet d'estimer de façon cohérente ce risque en proposant une application à l'estimation du *trend* au sein de populations hétérogènes.

Chapitre 3

Modélisation de la composante tendancielle du *basis risk*

> « ***On distingue la longévité moyenne ou espérance de vie et la longévité potentielle. Celle-ci exprime l'âge maximum qu'une espèce peut atteindre même si elle ne bénéficie qu'à un seul individu.*** » *(Lemaire Envir. 1975)*

Dans le chapitre précédent, nous avons étudié le *basis risk* sous ses différents aspects en soulignant que plusieurs facteurs pouvaient avoir une influence significative sur celui-ci. Dans ce chapitre, nous aborderons la question de l'estimation du *basis risk* d'une population hétérogène.

L'estimation de la mortalité d'une population passe nécessairement par l'estimation des lois de survie. Trois grandes approches ont été utilisées dans la littérature pour l'estimation des lois de survie d'une population hétérogène (Kamega, Planchet, 2011). La première approche consiste à modéliser de manière indépendante les sous-groupes de population en fonction de leurs caractéristiques. La deuxième approche repose sur l'utilisation de modèles intégrant directement, grâce à des variables explicatives, les facteurs d'hétérogénéité observables. Enfin, la troisième approche modélise la survie en intégrant les facteurs d'hétérogénéité inobservables ou résiduels (Horny, 2008).

Nous nous intéressons dans notre étude aux deux premières approches. Nous allons dans une première partie présenter la première approche qui correspond à une approche par méthodes de discrimination. Dans une seconde partie, nous présenterons les méthodes d'analyse de survie intégrant les facteurs d'hétérogénéité, puis, nous conclurons dans une troisième partie par le choix de la méthode qui nous semble adaptée à notre étude.

3.1 Les méthodes de discrimination

L'approche par discrimination est de prime abord relativement simple et intuitive. Cependant, sa mise en oeuvre pourrait s'avérer complexe. En effet, cette approche consiste à effectuer une modélisation pour chaque sous-groupe homogène au sein d'une population. Bien que la population globale soit hétérogène, nous pouvons constituer des sous-groupes présentant des caractéristiques similaires : on parle alors de sous-groupes homogènes. Par exemple, on pourrait avoir au sein d'une population, des cadres, des ouvriers, des chômeurs. L'approche par discrimination consisterait dans ce cas à modéliser la survie de chacun des sous-groupes c'est-à-dire de celui des cadres, des ouvriers et des chômeurs.

La méthode exhaustive (Goris, 2016) consiste à affiner la modélisation à tous les sous-groupes au sein de la population. Considérons une population $\mathcal{P}$ composée de N individus. On suppose que les individus peuvent être caractérisés par n variables $x_1, x_2, ..., x_n$. Par exemple, x_i peut représenter tant le sexe, la catégorie socio-professionnelle, la position géographique que le niveau d'étude. De plus, supposons que chaque caractéristique x_i prenne p_i valeurs distinctes. Si x_i représente la catégorie socio-professionnelle « Ouvrier » ou « Cadre » par exemple, alors p_i sera égal à 2. La méthode exhaustive appliquée à la population $\mathcal{P}$ conduirait à l'estimation de la mortalité de chaque sous-groupe au sein de cette population, c'est-à-dire pour $\prod_{i=1}^{n} p_i$ groupes. Pour illustrer cela, supposons une population où chaque individu peut être caractérisé par :

- la catégorie socio-professionnelle « *Cadre, Ouvrier ou Professions Intermédiaires* »,
- le sexe « *homme ou femme* »,
- le niveau d'étude « *Brevet, Baccalauréat, Licence, Master ou Doctorat* »,
- la position géographique « *Ville ou Campagne* ».

Un individu choisi aléatoirement au sein de cette population pourrait avoir les caractéristiques :« *Femme-Cadre-Doctorat-Ville* » et un autre : « *Homme-Professions Intermédiaires-Licence-Ville* ». Le nombre de sous-groupes que l'on pourrait constituer au sein de cette population est de : $3 * 2 * 5 * 2$, soit 60 sous-groupes. La méthode exhaustive appliquée à cette population conduirait donc à modéliser la mortalité de 60 sous-groupes. La méthode exhaustive peut donc s'avérer très coûteuse en ressources. Sa mise en place nécessite de disposer de ressources d'exécution et d'analyses considérables. En effet, il faudrait non seulement calibrer les modèles pour chacun des sous-groupes, analyser chaque résultat, mais aussi agréger chaque sous-groupe afin de tirer des conclusions sur toute la population. Aussi, il se pose le problème de cohérence entre les modèles et méthodes utilisés pour la modélisation de chaque sous-groupe. La question de la significativité des sous-groupes apparaît donc dans cette méthode très importante. En effet, le sous-groupe « *Homme-Cadre-Baccalauréat-Campagne* » par exemple contiendra probablement un faible nombre

d'individus, ce qui rendrait non significative l'estimation de sa mortalité. On pourrait donc obtenir des sous-groupes avec un nombre insuffisant de données pour l'estimation de la mortalité. Ainsi, la méthode de discrimination introduit une nouvelle problématique importante qui est celle du choix de la segmentation optimale. Pour répondre à cette problématique, on pourrait s'intéresser aux méthodes de classification non supervisée telles que les K-moyennes ou le regroupement hiérarchique (Hartigan, Wong, 1979; Joe Ward, 1963). Ces problématiques ne constituant pas l'objet de notre étude, nous ne les développerons pas dans la suite.

Une fois les sous-groupes constitués, il est alors possible d'estimer leurs mortalités et surtout d'effectuer des projections de mortalité dans le temps. Plusieurs approches ont été utilisées par les actuaires et les démographes pour la projection de la mortalité d'une population (Pitacco et al., 2009). Ces méthodes peuvent être classées en trois grandes approches (Booth, Tickle, 2008) :

- **Approche subjective** [1] : cette approche est fondée sur le jugement d'expert. Dans la pratique, plusieurs hypothèses d'évolution de la mortalité peuvent être formulées. Typiquement, on pourrait avoir un scénario central autour duquel sont construits un ou plusieurs scénarios optimistes ou pessimistes. Une autre hypothèse pourrait consister en la définition d'une valeur cible du niveau de la mortalité à un horizon de temps fixé, puis en la construction de plusieurs scénarios d'évolution de la mortalité dans le temps pour atteindre cette valeur cible. Comme exemple, on pourrait supposer que l'espérance de vie des cadres atteindra 90 ans dans 50 ans et celle des ouvriers 88 ans dans 50 ans. Des scénarios peuvent donc être construits autour de ces chiffres clés. Cette approche a l'avantage de prendre en compte dans les projections certains facteurs qui ne peuvent être pris en compte par les modèles comme l'influence des connaissances démographiques, les avancées scientifiques, ainsi que l'influence de tous les facteurs connus pouvant impacter l'évolution de la mortalité dans le temps. Cependant, elle a pour inconvénient d'être subjective et potentiellement biaisée. Le jugement d'expert pouvant s'avérer peu correct et éloigné de la réalité.

- **Approche par extrapolation** [2] : l'approche par extrapolation est l'approche la plus utilisée pour la projection de la mortalité. Elle suppose que l'évolution future de la mortalité suivra la même tendance que celle du passé. Plusieurs hypothèses de tendance sont faites dans la pratique. On pourrait supposer par exemple une tendance *linéaire*, *logarithmique*, *exponentielle* ou *logistique* de la mortalité. L'un des modèles les plus utilisés dans cette approche, est celui de Gompertz-Makeham qui suppose une tendance exponentielle.

$$\mu_x = A + Bc^x \tag{3.1}$$

1. Expectation approach
2. Extrapolation approach

μ_x le taux instantané de mortalité, x l'âge, A, B, c des constantes à déterminer.

Ce modèle est un modèle de mortalité à un facteur. En effet, la modélisation de la mortalité ne dépend que de l'âge.

Plusieurs améliorations ont donc été proposées dans la littérature afin d'étendre les modèles à un facteur, à des modèles à deux ou trois facteurs en intégrant les facteurs de temps et de cohortes. Très souvent, les modèles à deux facteurs prennent en compte dans la modélisation l'âge et le temps, tandis que ceux à trois facteurs prennent en compte l'âge, le temps et la cohorte. Le modèle le plus connu parmi les modèles à deux facteurs est le modèle de Lee-Carter (Lee, Carter, 1992) :

$$\ln \mu_{x,t} = \beta_x^{(1)} + \beta_x^{(2)} \kappa_t^{(2)} + \xi_{x,t} \tag{3.2}$$

$\mu_{x,t}$ étant le taux instantané de mortalité, x l'âge, t l'année. $\beta_x^{(i)}$ [3], $\kappa_t^{(2)}$ représentant respectivement les effets liés à l'âge et au temps et $\xi_{x,t}$ le terme d'erreur.

Ce modèle propose une décomposition de la mortalité en deux composantes, une liée au temps et l'autre à l'âge. La composante liée au temps $\kappa_t^{(2)}$ est modélisée par une série temporelle. Le modèle de Lee-Carter fait partie de la classe des modèles stochastiques de mortalité qui supposent ainsi des taux de décès aléatoires. Le tableau 10 résume les principaux modèles à deux et trois facteurs couramment utilisés dans la littérature pour la modélisation et la projection de la mortalité.

Lee, Carter (1992)	$\ln \mu_{x,t} = \beta_x^{(1)} + \beta_x^{(2)} \kappa_t^{(2)}$
Renshaw, Haberman (2006)	$\ln \mu_{x,t} = \beta_x^{(1)} + \beta_x^{(2)} \kappa_t^{(2)} + \beta_x^{(3)} \gamma_{t-x}^{(3)}$
Currie (2006)	$\ln \mu_{x,t} = \beta_x^{(1)} + \kappa_t^{(2)} + \gamma_{t-x}^{(3)}$
Cairns, Blake and Dowd (Cairns et al., 2007)	$\operatorname{logit} q(t,x) = \kappa_t^{(1)} + \kappa_t^{(2)}(x - \bar{x})$ $\operatorname{logit} q(t,x) = \kappa_t^{(1)} + \kappa_t^{(2)}(x - \bar{x}) + \gamma_{t-x}^{(3)}$ $\operatorname{logit} q(t,x) = \kappa_t^{(1)} + \kappa_t^{(2)}(x - \bar{x}) + \kappa_t^{(3)}\left((x - \bar{x})^2 - \hat{\sigma}_x^2\right) + \gamma_{t-x}^{(4)}$ $\operatorname{logit} q(t,x) = \kappa_t^{(1)} + \kappa_t^{(2)}(x - \bar{x}) + \gamma_{t-x}^{(3)}(x_c - x)$

TABLE 10 – Modèles à deux et trois facteurs

3. $i = 1$ ou $i = 2$

- **Approche explicative** [4] : cette approche est fondée sur l'utilisation des méthodes d'analyse causale. La causalité peut être définie simplement comme la relation entre la cause et l'effet. Intrinsèquement on suppose donc que tout effet a une cause. Les méthodes d'analyse causale ont connu ces dernières années des avancées notables en biométrie, économétrie et sociologie. Les modèles les plus utilisés dans cette approche sont les modèles structurels ou les modèles de cause, intégrant les causes spécifiques de la mortalité (Booth, Tickle, 2008; O'Hare, 2012). Le principal avantage de ces approches est qu'elles permettent d'identifier les facteurs de risques ayant une importance significative sur la mortalité future et de ce fait permettent d'affiner la modélisation du risque. Ces modèles sont fondés sur l'utilisation de la régression et s'inscrivent donc dans le cadre global des modèles GLM [5]. Cependant, ils diffèrent des modèles de régression classiques couramment utilisés dans l'approche par extrapolation en permettant d'intégrer des facteurs de risques ayant des effets avec un décalage temporel (O'Hare, 2012). On peut par exemple supposer qu'une variable du modèle aurait un effet significatif sur la mortalité au bout de 10 ans et non avant.

3.2 Les méthodes d'analyse de survie

L'analyse de survie aussi appelée analyse des durées de survie, est l'étude des données pour lesquelles le temps avant la réalisation d'un événement relève d'une importance particulière. Ce temps, communément appelé durée de survie, représente le temps écoulé jusqu'à la survenance d'un événement appelé « décès ». Cet événement peut être différent selon l'étude menée. Il peut représenter l'apparition d'une maladie, la réalisation d'un sinistre, la panne d'un appareil électronique ou la mort. Tout au long de notre étude, l'événement considéré sera la mort.

La durée de survie d'un individu est une grandeur inconnue. Supposons T une variable aléatoire continue, positive ou nulle mesurant la durée de survie d'un individu. La fonction de survie au temps t, représentant la probabilité que l'événement [6] survienne après t peut alors s'écrire :

$$S(t) = \mathbb{P}(T > t) \tag{3.3}$$

On définit aussi le risque instantané de décès ou taux de hasard qui représente la probabilité que le décès survienne dans un petit intervalle de temps après t :

$$h(t) = \lim_{\mu \to 0} \frac{\mathbb{P}(t \leq T < t + \mu | T \geq t)}{\mu} \tag{3.4}$$

4. Explanatory approach
5. Generalized Linear Model
6. Nous utiliserons le terme décès dans toute la suite.

De nombreuses méthodes ont été développées dans la littérature pour l'analyse des durées de survie selon qu'elles portent sur la modélisation de la fonction de survie, la modélisation du temps de survie ou la modélisation de la fonction de hasard. Parmi ces approches, il existe des méthodes non paramétriques d'estimation de la survie (Kaplan, Meier, 1958; Planchet, 2016a), des méthodes d'estimation semi-paramétriques et des méthodes paramétriques d'estimation de la survie (Planchet, 2016b).

Nous nous sommes intéressés dans notre étude aux méthodes qui, en plus d'estimer les distributions de survie, permettent aussi d'évaluer les effets que peuvent avoir des variables explicatives sur la survie. Il s'agit des modèles de régression. En effet, au sein d'un groupe, il se pourrait que certaines covariables aient un effet significatif sur la survie. Dans ce cas, il y aurait des sous-groupes ayant des survies différentes au sein du groupe. Il serait donc intéressant d'analyser à la fois les distributions de survie, les différences de survie au sein du groupe et leurs effets sur la survie du groupe.
Nous nous sommes donc intéressés aux modèles de régression les plus couramment utilisés en analyse de survie. Il s'agit du modèle de Cox et des modèles à temps de vie accélérée [7].

3.2.1 Approche par le modèle de Cox

Le modèle de Cox, introduit en 1972 par David COX (Cox, 1972) est un modèle à hasards proportionnels utilisé en analyse de données de survie pour la modélisation de l'effet que peuvent avoir diverses covariables sur les durées de survie. Le modèle de Cox est l'un des modèles les plus utilisés en biostatitique et dans le domaine médical. Cependant, ce modèle demeure encore peu utilisé dans le monde de l'actuariat. Le modèle de Cox ne s'intéresse pas tant à la mortalité elle-même mais aux effets des différentes sous-populations sur la population.

Considérons un groupe de n individus. On suppose de plus que chaque individu soit identifiable par p caractéristiques ou covariables. Les covariables peuvent être le sexe, l'âge, la zone géographique, l'état de santé, etc. Soit $X = (X_1, ..., X_p)$ les différentes covariables du groupe et $\beta = (\beta_1, ..., \beta_p)$ les effets [8] à estimer de chaque covariable sur la survie. Les modèles à hasards proportionnels se caractérisent par la relation suivante :

$$h(t|X) = h_0(t)g(\beta^\top X) \tag{3.5}$$

- g étant une fonction positive.
- $h(t|X)$ le risque instantané de décès à t sachant que l'individu a les caractéristiques X.
- $h_0(t)$ est une fonction positive inconnue appelée *baseline* [9]
- $\forall\, k \in \{1, ..., p\}$, $\beta_k \in \mathbb{R}$

7. Accelerated Failure Time Models(AFT)
8. Ou coefficients de régression
9. ou risque de base ou mortalité de référence

L'appellation modèle à hasards proportionnels vient du fait que le rapport des fonctions de hasard pour deux individus i et j ne varie pas au cours du temps. Les fonctions de hasard de i et j sont donc proportionnels. En effet :

$$\frac{h_i(t|X)}{h_j(t|X)} = \frac{g_i(\beta^\top X)}{g_j(\beta^\top X)} \tag{3.6}$$

Le modèle de Cox est un cas particulier des modèles à hasard proportionnel. Il suppose que la fonction g est une fonction exponentielle. Le modèle de Cox s'écrit donc :

$$h(t|X) = h_0(t)\exp(\beta^\top X) = h_0(t)\exp(\sum_{k=1}^{p}\beta_k X_k) \tag{3.7}$$

Avec

- $h(t|X)$ le risque instantané de décès à t
- $h_0(t)$ le *baseline*
- $\exp(\sum_{k=1}^{p}\beta_k X_k)$ représente le risque relatif

De par sa définition, le modèle de Cox doit vérifier les deux hypothèses suivantes :

- L'**hypothèse de risques proportionnels** définie par l'équation 3.6
- L'**hypothèse de log-linéarité**. En effet le logarithme du risque instantané est une fonction linéaire des covariables selon la relation suivante :

$$\log(h(t|X)) = \log(h_0(t)) + \beta^\top X \tag{3.8}$$

L'hypothèse de risques proportionnels est une hypothèse forte du modèle. Celle-ci n'est pas toujours vérifiée dans la pratique. Dans ce cas, l'effet des covariables sur la mortalité peut être mal estimé par le modèle. Plusieurs extensions du modèle de Cox ont donc été développées afin de permettre des meilleures estimations de ces effets dans le cas où l'hypothèse de risques proportionnels n'est pas vérifiée.

- **Le modèle à risques non proportionnels**

  Le modèle à risques non proportionnels est une extension du modèle de Cox pour lequel les coefficients de régression varient avec le temps [10] (Chauvel, 2014). La fonction de hasard peut alors s'écrire :

$$h(t|X) = h_0(t)\exp(\beta(t)^\top X(t)) = h_0(t)\exp(\sum_{k=1}^{p}\beta_k(t)X_k(t)) \tag{3.9}$$

  La forme de la fonction $\beta(t)$ peut être donnée en entrée du modèle ou estimée pour chaque temps.

10. Les covariables aussi peuvent être dépendante du temps

Une première approche pourrait être d'écrire : $\beta(t) = \theta f(t)$, f étant une fonction déterministe par exemple (Chauvel, 2014). Cox dans son article (Cox, 1972) à proposé une modélisation avec une fonction f linéaire. Grambsch, Therneau (1994) ont proposé une méthode d'estimation du modèle à risque non proportionnel avec $\beta(t) = \beta_1 + \beta_2 g(t)$. Anderson, Senthilselvan (1982) ont introduit le modèle de rupture $\beta(t) = \beta_0 \mathbb{I}_{\{t \leq \gamma_0\}} + \beta_1 \mathbb{I}_{\{t > \gamma_0\}}$, avec une estimation par la vraisemblance partielle. Duroux, O'Quigley (2016) ont proposé une estimation en utilisant le processus de score standardisé sur un modèle étendu $\beta(t) = \beta_0 \mathbb{I}_{\{t \leq \gamma_0\}} + \beta_1 \mathbb{I}_{\{t > \gamma_1\}} + ... + \beta_k \mathbb{I}_{\{t > \gamma_k\}}$. Ces méthodes ont l'avantage d'être simples à interpréter et s'appuient sur des techniques d'estimation classiques. D'autres approches stochastiques ont aussi été proposées dans la littérature. West (1991) a proposé une approche bayesienne avec des coefficients $\beta(t)$ étant modélisés par une série temporelle : $\beta_t = \beta_{t-1} + \xi_t$, avec $\xi_t \sim \mathcal{N}(0, \sigma_t^2)$.

- Le modèle stratifié

Le modèle stratifié permet de se passer de la nécessité de vérifier l'hypothèse de proportionnalité pour tout le groupe et de le conserver juste pour des sous-groupes. Cette extension consiste à considérer un *baseline* différent pour les sous-groupes conditionnellement aux covariables (Kalbfleisch, Prentice, 1973). Concrètement, on subdivise notre groupe en M sous-groupes puis on estime la fonction $h_0(t)$ pour chaque sous-groupe. Les fonctions de hasard pour deux individus a et b appartenant aux sous-groupes respectif A et B s'écrivent donc :

$$\begin{cases} h_a(t|X) = h_0^A(t) \exp(\beta^\top X_a) \\ h_b(t|X) = h_0^B(t) \exp(\beta^\top X_b) \end{cases} \tag{3.10}$$

Ainsi, l'hypothèse de proportionnalité est conservée en local pour chaque sous-groupe et plus en global sur tout le groupe. En effet, pour ces deux individus, le rapport des fonctions de hasard n'est pas constant par rapport au temps.

$$\frac{h_a(t|X)}{h_b(t|X)} = \frac{h_0^A(t) \exp(\beta^\top X_a)}{h_0^B(t) \exp(\beta^\top X_b)}$$

La stratification peut aussi se faire selon les valeurs des covariables qui ne vérifient pas l'hypothèse de risque proportionnel (Therneau, Grambsch, 2000).

D'autres alternatives au modèle à risque proportionnel existent. Nous pouvons citer le modèle additif d'Aalen (Hosmer, Royston, 2002) avec pour fonction de hasard

$$h(t|X) = h_0(t) + \beta(t)^\top X(t).$$

Lin, Ying (1995) ont proposé le modèle à risques additifs et multiplicatifs suivant :

$$h(t|Z) = g\Big(\beta(t)^\top W(t)\Big) + h_0(t) f\Big(\gamma(t)^\top X(t)\Big),$$

les fonctions f et g étant généralement des fonctions du type $f(x) = \exp(x)$, $f(x) = 1+x$, $g(x) = \exp(x)$, $g(x) = x$ et $Z = \left(W^\top, X^\top\right)^\top$ le vecteur des covariables.

3.2.2 Approche par les modèles à temps de vie accélérée

Les modèles à temps de vie accélérée sont des modèles de régression utilisés en analyse de survie pour la modélisation et l'analyse des distributions de survie. Ils se distinguent des modèles à hasard proportionnel dans leur formulation mathématique par l'introduction d'une loi de probabilité. Ces modèles peuvent être des alternatives utiles au modèle à risque proportionnel de Cox (Wei, 1992). Tout comme le modèle de Cox, ils permettent d'intégrer l'effet des différentes covariables sur la fonction de survie.

Soit T une variable aléatoire continue, positive ou nulle mesurant la durée de survie d'un individu. Considérons toujours un groupe de n individus ayant p covariables $X = (X_1, ..., X_p)$; $\beta = (\beta_1, ..., \beta_p)$ les coefficients de régression. Les modèles à temps de vie accélérée s'écrivent :

$$\log(T) = \beta_1 X_1 + ... + \beta_p X_p + \xi \tag{3.11}$$

ξ étant une variable aléatoire.

Contrairement au modèle de Cox qui explique le logarithme de la fonction de hasard par une régression linéaire, les modèles à temps de vie accélérée expliquent par une régression linéaire sur les covariables le logarithme des temps de survie. La particulairté est qu'ils supposent de plus que les temps de survie suivent une distribution dictée par la variable aléatoire ξ. Les lois de probabilité les plus utilisées dans la modélisation sont la loi Exponentielle, la loi de Weibull, la loi Log-normal, la loi Gamma, la loi gamma généralisé, la loi log-logistique, la loi de Fisher (Medeiros et al., 2014).

L'équation 3.11 peut être réécrite de la façon suivante :

$$T = T_0 \exp(\beta^\top X) \tag{3.12}$$

Avec $T_0 = \exp(\xi)$

Le calcul de la fonction de survie au temps t donne :

$$\begin{aligned} S(t|X) &= \mathbb{P}\Big(T > t|X\Big) && (3.13)\\ &= \mathbb{P}\Big(T_0 \exp(\beta^\top X) > t\Big) && (3.14)\\ &= \mathbb{P}\Big(T_0 > t\exp(-\beta^\top X)\Big) && (3.15)\\ &= S_0\Big(t\exp(-\beta^\top X)\Big) && (3.16) \end{aligned}$$

La fonction de survie est donc de la forme :

$$S(t|X) = S_0\Big(t\exp(-\beta^\top X)\Big) \tag{3.17}$$

Avec

$$S_0(t) = \mathbb{P}\Big(T_0 > t\Big)$$

Nous pouvons donc déduire l'expression de la fonction de hasard $h(t|X)$:

$$\begin{aligned} h(t|X) &= -\frac{S'(t|X)}{S(t|X)} && (3.18)\\ &= -\frac{\Big[S_0\Big(t\exp(-\beta^\top X)\Big)\Big]'}{S_0\Big(t\exp(-\beta^\top X)\Big)} && (3.19)\\ &= -\frac{S_0'\Big(t\exp(-\beta^\top X)\Big)}{S_0\Big(t\exp(-\beta^\top X)\Big)}\exp(-\beta^\top X) && (3.20)\\ &= h_0(t\exp(-\beta^\top X))\exp(-\beta^\top X) && (3.21) \end{aligned}$$

La fonction de hasard peut donc s'écrire sous la forme :

$$h(t|X) = h_0(t\exp(-\beta^\top X))\exp(-\beta^\top X) \tag{3.22}$$

La forme de h_0 étant conditionnée par la variable aléatoire ξ.

3.3 Choix du modèle

Dans les parties précédentes, nous avons présenté deux grandes approches permettant la prise en compte du *basis risk*.

L'approche par discrimination présente l'avantage d'être intuitive et complète. Elle permet de modéliser finement la mortalité des sous-groupes. Cependant, l'utilisation de cette méthode induit une autre problématique qui est celle du choix de la segmentation optimale. De plus, la significativité des sous-groupes constitués, la cohérence des méthodes et modèles utilisés, et l'interprétation des résultats obtenus peuvent s'avérer complexes. La projection de sous-groupes de façon indépendante n'est sans doute pas adéquate pour notre étude. En effet, il se peut qu'il y ait des effets d'interaction et de corrélation entre les sous-groupes, ce qui peut biaiser l'estimation du risque de tendance. De plus, il se pose le problème de la perte d'information liée à l'agrégation de résultat. Aussi, d'un point de vue opérationnel, cette approche est peu adaptée. En effet, si la méthode devait être appliquée à plusieurs bases de données, cela nécessiterait plusieurs études de segmentation de chacune des bases. Ainsi, les résultats obtenus pour chacune

des bases ne seraient pas comparables. Cette approche sera donc très peu privilégiée pour une entreprise ayant par exemple plusieurs entités et souhaitant mener une étude de *trend* pour l'ensemble de ces entités.

Toutes ces raisons nous ont poussés à nous intéresser de plus près aux méthodes d'analyse de survie, notamment aux méthodes de régression pour l'estimation du risque de tendance du *basis risk*. En effet, ces méthodes ont l'avantage de permettre l'estimation, sans subdivision de la population, de l'impact de chaque covariable sur la mortalité. Implicitement, elles permettent donc de comprendre l'effet de chaque sous-groupe sur la mortalité du groupe et donc d'estimer le *basis risk*. La problématique du choix de la segmentation optimale ne se pose donc plus dans l'utilisation de ces méthodes, ce qui est très avantageux d'un point de vue opérationnel. En effet, appliquées à plusieurs bases, il faudrait que celles-ci aient la bonne structure, c'est-à-dire que les données soient bien ordonnées en faisant apparaître les variables explicatives du modèle. Aussi, ces méthodes sont flexibles dans l'estimation des interactions entre les sous-groupes. En effet, elles permettent de prendre en compte intuitivement les effets d'interactions au sein du groupe en introduisant uniquement des variables d'interactions dans les modèles.

Les méthodes de régression en analyse de survie peuvent être classées en trois groupes. Il s'agit des méthodes **paramétriques**, des méthodes **non paramétriques** et des méthodes **semi-paramétriques**.

Rappelons que les méthodes paramétriques supposent que les temps de survie suivent une loi de distribution choisie. Elles ont ainsi l'avantage de proposer des estimateurs de paramètres convergeants (Kamega, Planchet, 2011), en utilisant des techniques d'estimation relativement simples et couramment utilisées [11]. Cependant, l'inconvénient principal de ces méthodes demeure l'utilisation d'hypothèses de lois de distribution qui ne s'ajustent pas toujours aux données utilisées. On peut donc faire face à un risque de mauvaise estimation du comportement des données.

Les modèles non paramétriques quant à eux n'utilisent pas d'hypothèses de loi de distribution particulière. Cependant, ils peuvent induire une problématique particulière dans l'estimation des paramètres qui est le « fléau de dimension » (Lopez, 2007), encore appelé « malédiction de la dimension [12] ». Ce terme, inventé par Richard Bellman en 1961 est relatif aux difficultés d'optimisation dues à l'augmentation du nombre de données lorsqu'on ajoute des dimensions supplémentaires dans un espace mathématique. Les estimateurs de paramètres dans ce cas peuvent donc présenter des difficultés de convergence lorsque le nombre de covariables devient important [13](Viallon, 2006; Lopez, 2007).

Enfin, les modèles semi-paramétriques se présentent comme une synthèse des mo-

11. Maximum de vraisemblance, Moindres carrés
12. Curse of dimensionality
13. Supérieur à 3

dèles paramétriques et non paramétriques. Elles ont l'avantage de permettre une grande souplesse dans l'estimation des paramètres sans imposer une distribution particulière aux données. De plus, ces modèles permettent de limiter la problématique du fléau de dimension, entraînant ainsi une meilleure prise en compte de l'hétérogénéité au sein de la population. Les modèles semi-paramétriques se posent donc comme alternatives adéquates aux modèles paramétriques et non paramétriques.

Dans la cadre de notre étude, nous avons donc choisi d'estimer le risque de tendance du *basis risk* par l'utilisation de l'approche semi-paramétrique, en l'occurrence par l'utilisation du modèle de Cox. Le modèle de Cox a l'avantage de ne pas imposer une segmentation de la base initiale de données. Il nécessite par contre un formatage correct de celle-ci en entrée. Il n'impose pas le choix d'une distribution de probabilité contrairement aux modèles à temps de vie accélérée par exemple. Ainsi, ce modèle ne requiert pas de ressources importantes en terme d'implémentation et surtout en terme de déploiement opérationnel. Le modèle de Cox présente une grande simplicité en terme d'interprétation des résultats et possède des propriétés particulièrement intéressantes notamment en terme de flexibilité pour l'ajout du *trend* et pour la prise en compte d'effets d'interaction. Cependant, le modèle de Cox ne permet pas une modélisation fine des sous-groupes. En effet, il suppose un *baseline* identique pour tous les individus de la base de données. Pour résoudre ce problème, un modèle de Cox stratifié pourrait être utilisé, ce qui introduirait implicitement la problématique de segmentation optimale. Nous avons présenté dans le tableau 11 une comparaison entre le modèle de Cox, le modèle à temps de vie accélérée et les méthodes de discrimination, suivant des critères pertinents du point de vue des objectifs de notre étude.

Le modèle de Cox, malgré ses propriétés intéressantes n'est pas très présent dans la littérature actuarielle et dans le monde de l'assurance. Ce modèle a été utilisé par Jérémy Goris (Goris, 2016) pour la modélisation du risque de niveau du *basis risk*. Nous proposons donc d'étendre et d'adapter ce modèle pour la modélisation du risque de tendance du *basis risk*. Dans le chapitre suivant, nous nous intéresserons à l'utilisation du modèle de Cox pour l'estimation du *basis risk*. Nous introduirons plusieurs approches permettant l'estimation du risque de tendance du *basis risk*.

	Discrimination	**AFT**	**Cox**	**Commentaires**
Segmentation de la base	✓	✗	✗	*L'approche par discrimination impose une segmentation de la base.*
Formatage de la base	✗	✓	✓	*Pour Cox et AFT, la base doit être bien ordonnée, de façon à faire apparaître les covariables.*
Choix d'une distribution de probabilité	✓	✓	✗	*AFT nécessite le choix d'une loi de probabilité pour la modélisation des temps de survie. Les méthodes de discrimination utilisent une distribution normale pour la modélisation des séries temporelles.*
Facilités d'interprétation	✗	✓	✓	*L'interprétation des résultats dans l'approche par discrimination peut être complexe car elle dépend fortement de la segmentation et des modèles choisis.*
Simplicité dans la mise en œuvre	✗	✗	✓	*Cox ne nécessite pas d'étude pour le choix d'une distribution de probabilité ou d'étude pour choix d'une segmentation optimale de la population.*
Ressources importantes d'analyse et de calcul	✓	✓	✗	*Les ressources nécessaires en terme d'études sont importantes dans les méthodes par discrimination et AFT.*
Flexibilité	✓	✓	✓	Les modèles sont flexibles, permettent la prise en compte des effets d'interaction par exemple.
Facilité de déploiement	✗	✗	✓	*Le choix de la distribution de probabilité et de la segmentation optimale rendent plus complexe le déploiement opérationnel des modèles AFT et de l'approche par discrimination.*
Modélisation fine des sous-groupes	✓	✗	✗	*L'approche par discrimination permet une modélisation fine des sous-groupes ; possibilité d'utiliser le modèle adapté pour chaque groupe.*
Choix du modèle	✗	✗	✓	***Le modèle de Cox est le meilleur choix pour notre étude.***

TABLE 11 – Choix du modèle

Deuxième partie

Le modèle de Cox : adaptation à la modélisation de la composante tendancielle du *basis risk* en longévité

Chapitre 1

Adaptation du modèle de Cox

Dans ce chapitre, nous proposons une adaptation du modèle de Cox à la modélisation du risque de tendance.

1.1 Notations et formulations mathématiques

Soient :

— $\mathcal{P}$ une population.

— Z un vecteur de covariables désignant les caractéristiques des individus dans $\mathcal{P}$. On suppose que $Z = (Z_1, ..., Z_p)$; p étant le nombre de caractéristiques.

— Z^i désignant les caractéristiques d'un individu i appartenant à $\mathcal{P}$.
$\forall\, i \in \mathcal{P}$, $Z^i = (Z_1^i, ..., Z_p^i)$.

— $\beta = (\beta_1, ..., \beta_p)$ les coefficients de régression du modèle.

— D le nombre de décès dans $\mathcal{P}$ sur toute la période d'observation.

— X une variable aléatoire modélisant les temps ou âges de décès des individus de $\mathcal{P}$; x une réalisation de X.

— C une variable aléatoire modélisant les temps ou âges de censure des individus de $\mathcal{P}$; c une réalisation de C.

— d le nombre de temps distincts de décès.

— d' le nombre de temps distincts de censure.

— $x_1 < x_2 < ... < x_d$ les temps ou âges distincts de décès des individus de $\mathcal{P}$. On pose $x_0 := 0$.

— $c_1 < c_2 < ... < c_{d'}$ les temps de censures distincts. On pose $c_0 := 0$.

— $d(x)$ le nombre de décès à l'âge x ; $\sum_{i=1}^{d} d(x_i) = D$

— $\mathcal{R}(x)$ l'ensemble des individus à risque c'est-à-dire l'ensemble des individus dont le temps de décès ou de censure est au moins égal à x [1].

— $\mathcal{D}(x)$ l'ensemble des individus décédés à l'âge x.

— T une variable désignant l'année d'observation de la population $\mathcal{P}$; t une réalisation de T.

Le modèle de Cox appliqué à la modélisation de notre population peut s'écrire de la manière suivante :

$$h(x|Z) = h_0(x) \exp(\beta^\top Z) \tag{1.1}$$

Avec
- $h(x|Z)$ le risque instantané à l'âge x sachant les caractéristiques Z.
- $h_0(x)$ le *baseline* ou mortalité de référence. Il correspond à la mortalité d'un individu dont les caractéristiques sont $Z = (0, ..., 0)$.

La fonction de hasard pour un individu i dans $\mathcal{P}$ s'écrit donc :

$$h_i(x|Z^i) = h_0(x) \exp(\beta^\top Z^i) = h_0(x) \exp\left(\sum_{j=1}^{p} \beta_j Z_j^i\right) \tag{1.2}$$

Le modèle ci-dessus permet d'estimer le risque de niveau [2] du *basis risk* mais ne permet pas l'estimation du risque de tendance. Dans la section suivante, nous proposerons plusieurs extensions pour l'estimation du risque de tendance du *basis risk*.

1.2 Le modèle de Cox avec tendance

Considérons une population $\mathcal{P}$ et $Z = (Z_1, ..., Z_p)$ le vecteur des covariables. Soit N le nombre de sous-groupes que l'on pourrait constituer au sein de cette population (Voir 3.1). L'objectif est d'estimer l'évolution dans le temps de la mortalité de chaque sous-groupe. En d'autres termes, nous voulons quantifier les impacts futurs de l'effet des sous-groupes sur la mortalité dans $\mathcal{P}$.

En guise d'illustration, reconsidérons l'exemple introduit en 3.1. On suppose que dans $\mathcal{P}$, chaque individu peut être caractérisé par la catégorie socio-professionnelle « *Cadre,*

1. Individus encore à risque
2. Level

Ouvrier ou Professions Intermédiaires », le sexe « *Homme ou Femme* », le niveau d'étude « *Brevet, Baccalauréat, Licence, Master ou Doctorat* » et la position géographique « *Ville ou Campagne* ». Ces variables étant qualitatives, le vecteur de covariables Z doit transformé en une variable numérique. Deux grandes méthodes peuvent être adoptées :

- **Codage dichotomique :** le vecteur de covariables Z s'écrit ici de la manière suivante [3] :

  $$Z = (Sexe,\, CSP,\, Etudes,\, PG)$$

  Avec par exemple :
  — $Sexe \in \Big\{0, 1\Big\}$, $Homme = 1$ et $Femme = 0$.
  — $CSP \in \Big\{1, 2, 3\Big\}$, $Cadre = 1$, $Ouvrier = 2$, $Professions\ Intermédiaires = 3$.
  — $Etudes \in \Big\{1, 2, 3, 4, 5\Big\}$, $Brevet = 1$,..., $Doctorat = 5$.
  — $PG \in \Big\{1, 2\Big\}$, $Ville = 1$ et $Campagne = 2$.

  Ainsi, le vecteur de covariables d'un individu ayant les caractéristiques « *Femme-Cadre-Doctorat-Ville* » s'écrira :

  $$Z = (0,\, 1,\, 5,\, 1)$$

- **Codage binaire** : le vecteur de covariables s'écrit :

  $$Z = (Homme,\, Femme,\, Cadre,\, Ouvrier,\, Professions\ Interme^{'}diaires,\, Brevet,\, Baccalaureat,\, Licence,\, Master,\, Doctorat,\, Ville,\, Campagne)$$

  Avec : $\forall\, i,\, Z_i \in \Big\{0, 1\Big\}$.

  Un individu ayant les caractéristiques « *Femme-Cadre-Doctorat-Ville* » aura comme vecteur de covariables :

  $$Z = (0,\, 1,\, 1,\, 0,\, 0,\, 0,\, 0,\, 0,\, 0,\, 1,\, 1,\, 0)$$

  Nous pouvons réduire la dimension du vecteur de covariable en supprimant l'une des caractéristiques. Pour la caractéristique sexe « *homme ou femme* » par exemple, on pourrait garder l'une des caractéristiques « *homme* » ou « *femme* » dans le vecteur de covariables.

  $$Z = (Homme,\, Cadre,\, Ouvrier,\, Professions\ Interme^{'}diaires,\, Brevet,\, Baccalaureat,\, Licence,\, Master,\, Doctorat,\, Ville,\, Campagne)$$

  ou

3. CSP : catégorie socio-professionnelle ; PG : position géographique

$$Z = (Femme, Cadre, Ouvrier, Professions\ Interme'diaires, Brevet, Baccalaureat, Licence, Master, Doctorat, Ville, Campagne)$$

L'individu précédent aura donc respectivement comme vecteur de covariable :

$$Z = (0,\ 1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1,\ 1,\ 0)$$

ou

$$Z = (1,\ 1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1,\ 1,\ 0).$$

On peut de même enlever une modalité pour chaque variable explicative prise en compte dans le modèle.

Nous avons choisi dans notre étude d'utiliser le codage binaire. En effet, il traduit clairement l'appartenance à un groupe. Le codage dichotomique quant à lui introduit une relation d'ordre dans l'attribution d'une valeur plus élevée à une caractéristique plutôt qu'à une autre. Cela ne traduit donc pas l'appartenance au groupe. De plus, le codage binaire a l'avantage d'estimer directement les effets de chacun des sous-groupes sur la mortalité, ce que ne permet pas le codage dichotomique. En utilisant le codage dichotomique, on obtiendrait :

$$h(x|Z) = h_0(x) \exp\Big(\beta_s Sexe + \beta_{csp} CSP + \beta_e Etude + \beta_{pg} PG\Big) \tag{1.3}$$

β_s traduisant l'impact de la caractéristique $Sexe$ sur la mortalité, β_{csp} celui de la caractéristique CSP sur la mortalité. Dans un tel cas, en considérant toujours que $Etudes \in \Big\{1, 2, 3, 4, 5\Big\}$ avec $Brevet = 1$,..., $Doctorat = 5$, l'effet du fait de disposer d'un doctorat sur la mortalité est traduit par $\exp(\beta_e * 5)$ par contre celui du fait de disposer du brevet est de : $\exp(\beta_e * 1)$. Cela introduit donc une relation d'ordre entre les différentes caractéristiques. Ainsi, cette approche nécessite de connaître l'ordre des valeurs des caractéristiques et ensuite de savoir quelles valeurs attribuer à celles-ci. Une de ses limites pourrait être aussi le fait que l'on ne sache pas par exemple si posséder un doctorat a un effet sur la mortalité équivalent au fait posséder un master, puisqu'une relation est déjà établie. Par contre grâce au codage binaire, on obtient :

$$h(x|Z) = h_0(x) \exp\Big(\beta_h Homme + \beta_{ca} Cadre + ... + \beta_c Campagne\Big) \tag{1.4}$$

Les coefficients $(\beta_h, ..., \beta_c)$ traduisant l'impact des sous-groupes sur la mortalité.

Si β_h est strictement positif, alors une conclusion pourrait être que être un $Homme$ impacterait la mortalité à la hausse. De même, si β_c est négatif, alors vivre en $Campagne$

impacterait la mortalité à la baisse. L'interprétation peut aussi se faire en terme de risque. En effet, le rapport des risques $RR_{h/f}$ des sous-groupes $Homme$ et $Femme$ donne :

$$RR_{h/f} = \frac{h_0(x)\exp\Big(\beta_h * 1 + \beta_{ca}Cadre + ... + \beta_cCampagne\Big)}{h_0(x)\exp\Big(\beta_h * 0 + \beta_{ca}Cadre + ... + \beta_cCampagne\Big)} \tag{1.5}$$

$$= \frac{\exp\Big(\beta_h + \beta_{ca}Cadre + ... + \beta_cCampagne\Big)}{\exp\Big(\beta_{ca}Cadre + ... + \beta_cCampagne\Big)} \tag{1.6}$$

$$= \exp(\beta_h) \tag{1.7}$$

D'où

$$RR_{h/f} = \exp(\beta_h)$$

$RR_{f/f} = RR_{h/h} = 1$: c'est le « risque zéro », c'est-à-dire le risque en absence d'hétérogénéité.

- Si $\beta_h = 0$, alors $RR_{h/f} = RR_{f/f} = RR_{h/h} = 1$. Le fait d'être $Homme$ n'a pas d'impact sur la mortalité.

- Si β_h strictement positif, alors $RR_{h/f}$ est strictement supérieur à 1. Le risque a ainsi augmenté de $\Big[\exp(\beta_h) - 1\Big]$. Par exemple, si $\beta_h = 0,5$, alors $RR_{h/f} = 1.64$. Ainsi, le fait d'être $Homme$ augmenterait donc le risque instantané de décès de $0,64$ soit 64%.

- Si β_h strictement négatif, alors $RR_{h/f}$ est strictement inférieur à 1. Le risque a ainsi diminué de $\Big[1 - \exp(\beta_h)\Big]$. Si $\beta_h = -0,5$, alors aussi $RR_{h/f} = 0,60$; être $Homme$ diminuerait le risque instantané de décès de 40%.

Les même analyses peuvent être faites pour tous les N sous-groupes dans $\mathcal{P}$. Le codage binaire permet donc de répondre correctement à l'objectif d'estimation de l'impact des sous-groupes sur la mortalité.

Grâce aux coefficients de régression calibrés dans le modèle, nous pouvons donc quantifier l'impact de chaque sous-groupe sur la mortalité. L'objectif étant d'estimer l'évolution de ces impacts dans le temps, il est alors évident de supposer que les coefficients de régression peuvent évoluer dans le temps. Nous proposons ainsi le modèle suivant pour estimer le risque de tendance :

$$h(x,t|Z) = h_0(x)exp\Big(\beta(t)^\top Z\Big) = h_0(x)\exp\Big(\sum_{j=1}^{p}\beta_j(t)Z_j\Big) \tag{1.8}$$

Avec

- $h(x,t|Z)$ le risque instantané de décès à l'âge x à l'année t sachant les caractéristiques Z.
- $h_0(x)$ le *baseline* ou la mortalité de référence.

On pourrait complexifier le modèle en supposant un *baseline* dépendant du temps.

$$h(x,t|Z) = h_0(x,t)exp(\beta(t)^\top Z) = h_0(x,t)\exp\Big(\sum_{j=1}^{p}\beta_j(t)Z_j\Big) \tag{1.9}$$

Cette approche qui consiste en réalité à la stratification du *baseline*, correspond au modèle de Cox stratifié présenté en 3.2.1. Elle a l'avantage de permettre une déformation de la fonction de mortalité dans le temps. Cependant, nous n'aborderons pas dans le cadre de notre étude la stratification par modèle de Cox.

Une autre approche peut être de supposer des covariables dépendantes du temps.

$$h(x,t|Z) = h_0(x)exp\Big(\beta(t)^\top Z(t)\Big) = h_0(x)\exp\Big(\sum_{j=1}^{p}\beta_j(t)Z_j(t)\Big) \tag{1.10}$$

Cette approche est cohérente dans la mesure où l'on suppose une population dynamique dans laquelle l'état des individus pourrait évoluer d'une année à l'autre. En effet, un individu ayant un niveau de salaire s € à t_0 pourrait obtenir une revalorisation salariale à un niveau de salaire de $s(1+r)^{(t-t_0)}$ € à l'année t années ; r étant le taux de revalorisation [4] salariale. De même, un individu habitant en campagne pourrait être amené à habiter en ville dans le futur. On pourrait donc de façon cohérente supposer que les covariables évolueraient avec le temps. Cependant, il demeure difficile de prédire le comportement des individus d'une population. Pourquoi un individu habitant en ville pourrait-il décider d'habiter en campagne et vice versa ? Plusieurs raisons peuvent expliquer le comportement des individus, ce qui rend complexe toute tentative de prédiction. Cette branche de l'étude des comportements des individus est prise en compte dans l'assurance notamment dans la modélisation des lois de rachat. Cependant, cela nécessite plusieurs autres prérequis pour être appliqué à notre étude. Nous ne l'étudierons donc pas, car elle est complexe et surtout non nécessaire à notre étude. En effet, les évolutions temporelles des covariables peuvent aussi être captées par les coefficients de régression. Supposons par exemple que les salaires soient revalorisés chaque année avec un taux r. Dans ce cas :

$$\begin{aligned}\beta(t) * Salaire(t) &= \beta(t) * Salaire_{initial}(1+r)^{(t-t_0)} &(1.11)\\ &= \beta(t)(1+r)^{(t-t_0)} * Salaire_{initial} &(1.12)\\ &= \kappa(t) * Salaire_{initial} &(1.13)\end{aligned}$$

4. dans l'hypothèse d'une revalorisation fixe

Avec $\kappa(t) = \beta(t)(1+r)^{(t-t_0)}$.

Cela revient donc à considérer un coefficient de régression de la forme $\beta(t)(1+r)^{(t-t_0)}$ dans le modèle.

Pour toutes ces raisons, nous avons donc retenu le modèle avec une évolution temporelle des coefficients de régression pour l'estimation du risque de tendance du *basis risk* :

$$h(x,t|Z) = h_0(x)exp(\beta(t)^\top Z) = h_0(x)\exp\Big(\sum_{j=1}^{p}\beta_j(t)Z_j\Big)$$

Ayant pour but d'effectuer des projections dans le futur, plusieurs tendances d'évolution des coefficients $\beta(t)$ peuvent être considérées en s'inspirant des méthodes développées dans les modèles à risques non proportionnels (voir 3.2.1).

Une première approche pourrait être de supposer que :

$$\exists\ f : \mathbb{R} \to \mathbb{R}\ /\ \beta(t) = \beta^0 f(t) \tag{1.14}$$

Avec $\beta^0 = (\beta_1^0, ..., \beta_p^0)$.

Cette approche suppose une tendance dictée par une fonction de tendance f. Ici, $\beta(t)$ devra forcément dépendre soit du temps soit être nul. Cependant, il se peut que l'effet des covariables soit constant et non nul. Ce modèle ne permet pas de capter de tels effets. Une deuxième approche pourrait être donc :

$$\exists\ f : \mathbb{R} \to \mathbb{R}\ /\ \beta(t) = \beta^{le} + \beta^{tr} f(t) \tag{1.15}$$

Les paramètres $\beta^{le} = (\beta_1^{le}, ..., \beta_p^{le})$ et $\beta^{tr} = (\beta_1^{tr}, ..., \beta_p^{tr})$ modélisant respectivement les effets liés au niveau et à la tendance. Si l'effet des covariables est constant avec le temps, l'estimation des paramètres conduirait alors à trouver β^{tr} très proche de 0.

Le choix de la fonction de tendance f a une très grande importance sur la projection de la mortalité. En effet, la fonction f conditionne l'évolution future des coefficients de régression $\beta(t)$ et donc l'évolution de la mortalité. Il faudrait donc s'assurer de la bonne adéquation de la fonction choisie. Choisir par exemple la fonction identité $f : \left|\begin{array}{ccc}\mathbb{R} & \longrightarrow & \mathbb{R}\\ t & \longmapsto & t\end{array}\right.$, permettrait de tester si l'effet des covariables évolue linéairement avec au temps ainsi que son sens d'évolution grâce au signe de β^{tr}. De même, si l'on choisit la fonction logarithme $f : \left|\begin{array}{ccc}\mathbb{R} & \longrightarrow & \mathbb{R}\\ t & \longmapsto & \ln(t)\end{array}\right.$, alors $\beta(t) = \beta^{le} + \beta^{tr}\ln(t)$, on pourrait ainsi tester scénario d'évolution de pente plus faible de la mortalité.

Une approche simple pour le choix de la fonction f peut être de supposer que :

$$\beta(t) = \sum_{k=0}^{n}\beta^k \mathbb{I}_{]t_k, t_{k+1}]}(t) \tag{1.16}$$

En effet, on calibre un paramètre β^k à chaque année t_k et on ajuste la meilleure fonction $f : \mathbb{R} \to \mathbb{R}$ tel que : $\forall\, k \in \{0, ..., n\},\ f(t_k) = \beta^k$.

On peut aussi étudier finement chaque covariable. Dans un tel cas, on suppose que chaque covariable évolue suivant une tendance propre à elle. Dans ce cas :

$$\forall\, i \in \{1, ..., p\},\ \exists\, f_i : \mathbb{R} \to \mathbb{R},\quad f : \left| \begin{array}{ccc} \mathbb{R} & \longrightarrow & \mathbb{R}^p \\ t & \longmapsto & \Big(f_1(t), ..., f_p(t)\Big) \end{array} \right. /\beta(t) = \beta^{le} + \beta^{tr} \times f(t) \tag{1.17}$$

Les méthodes proposées jusqu'à présent sont des méthodes déterministes. On peut aussi supposer une évolution stochastique des coefficients de régression. L'utilisation des séries temporelles pourrait dans ce cas nous permettre de modéliser l'évolution aléatoire des coefficients de régression dans le temps.

Nous n'aborderons pas dans cette étude la modélisation des coefficients par l'utilisation des séries temporelles pour des raisons de simplicité et d'interprétabilité. Cela requiert des études supplémentaires qui relèvent du domaine des méthodes de séries temporelles. Nous nous concentrerons dans cette étude à l'estimation du risque de tendance en supposant une tendance déterministe des coefficients de régression.

Chapitre 2

Calibrage, intervalles de confiance, adéquation

> ***« Les actuaires savent que s'il est très important de choisir le bon modèle, il est encore plus important de le calibrer correctement. »***
> *(W. James MacGinnitie, FSA, MAAA, FCAS)*

L'estimation des paramètres est une étape importante dans l'implémentation de notre modèle. Rappelons que le modèle proposé s'écrit :

$$h(x,t|Z) = h_0(x)\exp\Big(\beta(t)^\top Z\Big) \tag{2.1}$$

avec : $\beta(t) = \beta^{le} + \beta^{tr} f(t)$; f étant une fonction de $\mathbb{R}$ dans $\mathbb{R}$.

L'estimation des paramètres peut être réalisée grâce aux méthodes classiques d'estimation des paramètres du modèle de Cox. L'idée est d'utiliser le temps comme une covariable du modèle. En effet, on peut écrire :

$$\begin{aligned}
\beta(t)Z &= \Big(\beta^{le} + \beta^{tr} f(t)\Big)Z && (2.2)\\
&= \beta^{le}Z + \beta^{tr} f(t)Z && (2.3)\\
&= \beta^{le}Z + \beta^{tr}\Big(f(t)Z\Big) && (2.4)\\
&= \beta^{le}Z + \beta^{tr} Z^* && (2.5)
\end{aligned}$$

Avec $Z^* = f(t)Z$.

On retrouve donc la même forme que modèle de Cox classique avec Z^* comme covariable du modèle. Cependant pour ce faire, il convient de spécifier clairement comment cette covariable doit être définie et ainsi adapter la forme des données à la modélisation.

2.1 La variable de temps dans la base de données

Un individu dans une base de données est observé en général durant une période de temps qui s'étend de sa date de naissance ou date d'entrée jusqu'à sa mort ou sa sortie du portefeuille. Comme exemple, considérons la base de données suivante constituée de 5 individus observés sur des périodes de temps différentes :

N^o	**Année_E**	**Année_S**	**Sexe**	**Sortie**	**Année_N**	**Âge_E**	**Âge_S**
1	1992	2013	H	DÉCÈS	1942	50	71
2	1994	2017	F		1959	35	58
3	1995	2013	F		1954	41	59
4	1975	2014	H	DÉCÈS	1935	40	79
5	1996	2015	F		1952	44	63

TABLE 12 – Exemple de base de données

Avec :
— N^o : Le numéro de contrat ou le numéro police (N)
— Année_E : L'année d'entrée dans le portefeuille
— Sexe : Homme (H) ou Femme (F)
— Année_S : L'année de sortie du portefeuille
— Sortie : La cause de sortie (DÉCÈS ou Censure)
— Année_N : L'année de naissance
— Âge_E : L'âge d'entrée
— Âge_S : L'âge de sortie

Globalement, les individus dans cette base de données ont été observés sur une plage d'années s'étendant de 1996 à 2017. Avant 1996, seuls les individus 1, 2, 3 et 4 avaient été observés, et après 2013 seuls les individus 2, 4 et 5 l'étaient encore.
Supposons par exemple que l'on veuille estimer la mortalité au cours de l'année 2013. On pourrait penser qu'au cours de l'année 2013, on aurait enregistré qu'un décès, celui de l'individu N^o1, et une sortie en censure, celle de l'individu N^o3. Cependant, au cours de l'année 2013, on a enregistré 1 décès et 4 sorties en censures puisque les individus 2, 3, 4, 5 ont survécu au cours de l'année 2013. Il est donc important de disposer d'une base de données permettant d'observer chaque individu à chaque année afin de disposer des bonnes informations sur les décès ou censures pour le calibrage des coefficients $\beta(t)$. Il s'agit donc de spécifier clairement le temps ou l'année d'observation dans la base de données.

Pour ce faire, considérons un individu I observé du temps t_d au temps t_f. On suppose qu'il décède en t_f. Notons x_t l'âge de cet individu au temps t. On peut supposer que l'individu I correspond en réalité n individus, observés à chaque période de temps c'est-à-dire entre t_i et t_{i+1} $d \leq i < f$. La figure 11 nous en donne une illustration.

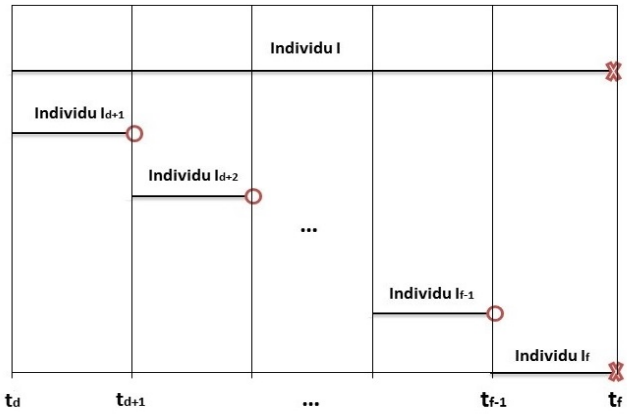

FIGURE 11 – Décomposition d'un individu

Les individus I_{d+1},..., I_f correspondent à l'individu I observé durant les temps t_d, ..., t_{f-1}. I_{d+1} représente un individu d'âge x_{t_d}, observé entre t_d et t_{d+1}, qui sort en censure à t_{d+1} du fait que l'individu I ait survécu durant l'année t_d. I_{d+1} possède les mêmes caractéristiques que I en t_d. L'individu I_f quant à lui décède en t_f tout comme I.

Cette décomposition nous permet ainsi de créer une nouvelle base de données avec le temps comme covariable du modèle.

Concrètement, si l'on applique cette approche aux individus N^o1 et N^o2 en les observant à chaque temps, on obtiendrait la nouvelle base de données suivante :

N^o	**Année_E**	**Année_S**	**Sexe**	**Sortie**	**Année_N**	**Âge_E**	**Âge_S**
1.1	1992	1993	H		1942	50	51
1.2	1993	1994	H		1942	51	52
1.3	1994	1995	H		1942	52	53
	...	...	...	...	...	...	...
1.19	2010	2011	H		1942	68	69
1.20	2011	2012	H	DÉCÈS	1942	69	70
2.1	2013	1994	H		1959	54	35
2.2	1994	1995	F		1959	35	36
2.3	1995	1996	F		1959	36	37
	...	...	...	...	...	...	...
2.22	2015	2016	F		1959	56	57
2.23	2016	2017	F	DÉCÈS	1959	57	58

TABLE 13 – Base de données des individus N^o1 et N^o2 du tableau 12 observés chaque année

L'individu N^o1 a été décomposé en 20 sous-individus. De même, l'individu N^o2 a

été décomposé en 23 sous-individus. Remarquons que les individus 1.1, ..., 1.20 ont la même année de naissance que l'individu N^o1 et que les individus 2.2, ..., 2.23 ont aussi la même année de naissance que l'individu N^o2. Cependant, leurs âges sont différents et correspondent à l'âge des individus N^o1 et N^o2 durant une année d'observation bien précise.

La décomposition ne modifie pas l'information initiale sur chaque individu. C'est simplement une manière subtile de réécrire l'information déjà contenue dans la base de données initiale, pour prendre réellement en compte le temps comme covariable. Dans la pratique, on pourrait avoir dans la base de données d'autres covariables telles que la zone géographique et la catégorie socio-professionnelle, qui pourraient être amenées à évoluer selon l'année d'observation.

Grâce à cette nouvelle base de données, il est donc possible de considérer l'année d'observation comme une covariable. Dans l'exemple précédent, cette covariable correspond la variable « Année_E » qui désigne l'année d'entrée dans le portefeuille. Le modèle peut alors s'écrire :

$$h(x|Z) = h_0(x) \exp\left(\beta^{le\top} Z + \beta^{tr\top} Z^*\right) \tag{2.6}$$

Avec $Z^* = f(t)Z = f(Année\_E)Z$.

Ainsi, on peut estimer les paramètres β^{le} et β^{tr} grâce aux méthodes classiques d'estimation des paramètres du modèle de Cox.

Deux groupes de paramètres doivent en réalité être estimés dans le modèle de Cox : les coefficients de régression et le *baseline*. Dans la section suivante, nous présenterons les différentes approches utilisées pour le calibrage des paramètres du modèle de Cox.

2.2 Estimation des coefficients de régression

Nous avons montré grâce à l'équation 2.6 que nous pouvions nous ramener, grâce à l'introduction du temps dans les données, au modèle de Cox classique pour l'estimation des paramètres. Considérons ainsi le modèle de Cox :

$$h(x|Z) = h_0(x) \exp\left(\beta^\top Z\right)$$

Avec : $\beta = \beta^{le} + \beta^{tr} f(t)$; $\beta^{le} = (\beta_1^{le}, ..., \beta_p^{le})$; $\beta^{tr} = (\beta_1^{tr}, ..., \beta_p^{tr})$.

Cox (1972) a proposé une approche fondée sur le maximum de vraisemblance, précisément sur la vraisemblance partielle, pour l'estimation des coefficients de régression. On se place dans un cadre continu dans lequel il n'y a qu'un seul décès à chaque temps de décès [1]. Deux individus ne peuvent donc pas décéder au même moment. Rappelons que la probabilité qu'un individu i décède entre x_k et $x_k + \Delta x$ pour Δx suffisamment petit,

1. Cela induit que $D = d$

est donnée par le risque instantané de décès $h(x_k|Z^i) = h_0(x_k)\exp\left(\beta^\top Z^i\right)$.

Supposons qu'il y ait un décès à l'âge x_k. La probabilité que ce soit l'individu i qui décède s'écrit :

$$\frac{h(x_k|Z^i)}{\sum_{j\in\mathcal{R}(x_k)} h(x_k|Z^j)} = \frac{\exp\left(\beta^\top Z^i\right)}{\sum_{j\in\mathcal{R}(x_k)} \exp\left(\beta^\top Z^j\right)} \tag{2.7}$$

Il s'agit de la contribution de l'individu i au risque de décès de l'ensemble des individus à risque. Il est important de remarquer que le *baseline* n'intervient pas dans cette probabilité. L'objectif est de trouver les paramètres β maximisant cette quantité pour tous les D décès observés dans la base. On cherche à estimer les coefficients β de telle sorte que les probabilités données par le modèle soient les plus vraisemblables, d'où la notion de vraisemblance partielle. Supposons que chaque individu i décède à l'âge x_i. La vraisemblance partielle sur l'ensemble des décès est définie par :

$$L(\beta) = \prod_{i=1}^{D} \frac{\exp\left(\beta^\top Z^i\right)}{\sum_{j\in\mathcal{R}(x_i)} \exp\left(\beta^\top Z^j\right)} \tag{2.8}$$

L'hypothèse de continuité introduite ci-dessus, n'est pas toujours vérifiée par les données dont on dispose. En effet, il pourrait arriver que deux ou plusieurs individus décèdent au même âge. Pour assurer la continuité, il faudrait alors disposer d'informations plus fines sur l'instant de décès comme le jour ou l'heure de décès. Ainsi, l'hypothèse de continuité n'est pas vérifiée dans un cadre réel. Il est très probable que dans notre base de données plusieurs individus décèdent au même âge. Trois grandes approches ont donc été adoptées dans ce cas pour le calcul de la vraisemblance partielle.

- **La méthode exacte** (Therneau, Grambsch, 2000) : dans le cas où plusieurs individus décéderaient au même temps, la méthode exacte consiste à supposer que ceux-ci décèdent les uns après les autres. Supposons par exemple deux individus i_1 et i_2 décédant au même temps x_k. Si l'information sur le temps de décès avait été plus complète, on aurait pu savoir avec certitude, lequel des deux individus était décédé avant l'autre. On pourra toujours considérer deux possibilités :

  — soit c'est l'individu i_1 qui est décédé avant i_2 et dans un tel cas la contribution à la vraisemblance partielle s'écrirait :

$$\frac{h(x_k|Z^{i_1})}{\sum_{j\in\mathcal{R}(x_k)} h(x_k|Z^j)} \frac{h(x_k|Z^{i_2})}{\sum_{j\in\mathcal{R}(x_k)\setminus i_1} h(x_k|Z^j)} = \frac{\exp\left(\beta^\top Z^{i_1}\right)\times\exp\left(\beta^\top Z^{i_2}\right)}{\sum_{j\in\mathcal{R}(x_k)}\exp\left(\beta^\top Z^j\right)\times\sum_{j\in\mathcal{R}(x_k)\setminus i_1}\exp\left(\beta^\top Z^j\right)} \tag{2.9}$$

— soit c'est l'individu i_2 qui est décédé avant i_1. La contribution à la vraisemblance s'écrirait alors :

$$\frac{h(x_k|Z^{i_1})}{\sum_{j\in\mathcal{R}(x_k)} h(x_k|Z^j)} \frac{h(x_k|Z^{i_2})}{\sum_{j\in\mathcal{R}(x_k)\setminus i_2} h(x_k|Z^j)} = \frac{\exp\left(\beta^\top Z^{i_1}\right) \times \exp\left(\beta^\top Z^{i_2}\right)}{\sum_{j\in\mathcal{R}(x_k)} \exp\left(\beta^\top Z^j\right) \times \sum_{j\in\mathcal{R}(x_k)\setminus i_2} \exp\left(\beta^\top Z^j\right)} \tag{2.10}$$

Ainsi, la contribution globale à la vraisemblance peut être calculée en sommant les deux contributions précédentes.

On peut facilement constater que lorsqu'il y a un grand nombre d'individus qui décèdent en même temps, cette méthode peut donc s'avérer complexe. En effet, il faudrait considérer tous les cas possibles d'ordre de décès des individus.

- **La méthode de Breslow** : cette méthode est en réalité une approximation de la méthode exacte. Elle suppose que :

$$\sum_{j\in\mathcal{R}(x_k)\setminus m} \exp\left(\beta^\top Z^j\right) \approx \sum_{j\in\mathcal{R}(x_k)} \exp\left(\beta^\top Z^j\right) \tag{2.11}$$

Cette approximation est respectée lorsque le nombre de décès simultanés à l'âge x_k est faible ou lorsque le nombre d'individus à risques à cet âge est élevé. Dans ce cas, la vraisemblance à l'âge x_k s'écrit :

$$L(\beta; x_k) = \frac{\exp\left[\beta^\top \sum_{i\in\mathcal{D}(x_k)} Z^i\right]}{\left[\sum_{j\in\mathcal{R}(x_k)} \exp(\beta^\top Z^j)\right]^{d(x_k)}} \tag{2.12}$$

et la vraisemblance partielle totale :

$$L(\beta) = \prod_{k=1}^{d} L(\beta; x_k) \tag{2.13}$$

- **La méthode d'Efron** (Efron, 1975) : dans ce cas la vraisemblance partielle est approchée par : :

$$L(\beta) = \prod_{k=1}^{d} \frac{\exp\left[\beta^\top \sum_{i\in\mathcal{D}(x_k)} Z^i\right]}{\prod_{m=1}^{d(x_k)} \left[\sum_{j\in\mathcal{R}(x_k)} \exp(\beta^\top Z^j) - \frac{m-1}{d(x_k)} \sum_{j\in\mathcal{D}(x_k)} \exp(\beta^\top Z^j)\right]} \tag{2.14}$$

Pour mieux comprendre cette relation, considérons l'exemple précédent de deux individus i_1 et i_2 qui décèdent à x_k. Supposons que pour des raisons de clarté et de simplicité, l'on ait 4 individus dans la population initiale. Alors, l'approximation proposée est la suivante :

$$L(\beta; x_k) = \left(\frac{E_1}{E_1 + E_2 + E_3 + E_4}\right)\left(\frac{E_2}{\frac{1}{2}E_1 + \frac{1}{2}E_2 + E_3 + E_4}\right) \tag{2.15}$$

Avec $E_j = \exp(\beta^\top Z^{i_j})$
Si l'on avait observé trois décès, celui de i_1, i_2 et i_3, on aurait :

$$L(\beta; x_k) = \left(\frac{E_1}{E_1 + E_2 + E_3 + E_4}\right)\left(\frac{E_2}{\frac{2}{3}E_1 + \frac{2}{3}E_2 + \frac{2}{3}E_3 + E_4}\right)\left(\frac{E_3}{\frac{1}{3}E_1 + \frac{1}{3}E_2 + \frac{1}{3}E_3 + E_4}\right) \tag{2.16}$$

Avec $E_j = \exp(\beta^\top Z^{i_j})$.

Intuitivement, l'interprétation est la suivante : chaque individu peut décéder en première, deuxième ou troisième position. Ainsi, la vraisemblance peut être décomposée en trois contributions : celle du premier décès, celle du deuxième décès et celle du troisième décès. On est certain que tous les individus sont pris en compte dans la contribution du premier décès à la vraisemblance (quel que soit l'individu qui décède). En effet, cette contribution s'écrit :

$$\left(\frac{E_j}{E_1 + E_2 + E_3 + E_4}\right) \tag{2.17}$$

j=1, 2 ou 3
Ensuite, chaque individu à $\frac{2}{3}$ de chance de décéder deuxièmement, donc de participer à la contribution du deuxième décès à la vraisemblance puis enfin $\frac{1}{3}$ de probabilité de décéder en troisième position (Therneau, Grambsch, 2000).

Plusieurs études ont montré que l'utilisation de la méthode de Breslow créait un biais dans l'estimation des paramètres β, surtout lorsque le rapport du nombre d'individus décédés à un temps donné sur le rapport du nombre d'individus à risque est important, alors que la méthode exacte et celle d'Efron fournissent de bien meilleurs résultats (Borucka, 2014). L'une des difficultés liées à l'utilisation de la méthode exacte est le temps de calcul. La méthode d'Efron peut aussi être lourde en terme de temps de calcul, mais reste tout de même un bon compromis pour l'estimation des paramètres. Nous utiliserons donc dans notre étude la méthode d'Efron pour l'estimation des coefficients de régression.

Comme souligné précédemment, l'objectif est de trouver les paramètres β qui maximisent la vraisemblance. Dans la pratique, pour des raisons de simplicité, on maximise la log-vraisemblance.

Soit $\mathcal{L}$ la log-vraisemblance. On a :

$$\mathcal{L}(\beta) = \log\big(L(\beta)\big) \tag{2.18}$$

En utilisant l'expression de la vraisemblance partielle selon la méthode d'Efron, on trouve :

$$\mathcal{L}(\beta) = \sum_{k=1}^{d}\left[\sum_{i\in\mathcal{D}(x_k)} \beta^\top Z^i - \sum_{m=1}^{d(x_k)} \log\left(\sum_{j\in\mathcal{R}(x_k)} \exp(\beta^\top Z^j) - \frac{m-1}{d(x_k)}\sum_{j\in\mathcal{D}(x_k)} \exp(\beta^\top Z^j)\right)\right] \tag{2.19}$$

On définit alors la fonction score par :

$$U(\beta) = \frac{\partial\mathcal{L}(\beta)}{\partial\beta} = (\frac{\partial\mathcal{L}(\beta)}{\partial\beta_1}, ..., \frac{\partial\mathcal{L}(\beta)}{\partial\beta_n}) \tag{2.20}$$

Après calcul on trouve :

$$U(\beta) = \sum_{k=1}^{d}\left[\sum_{i\in\mathcal{D}(x_k)} Z^k - \sum_{m=1}^{d(x_k)} \frac{\sum_{j\in\mathcal{R}(x_k)} Z^j \exp(\beta^\top Z^j) - \frac{m-1}{d(x_k)}\sum_{j\in\mathcal{D}(x_k)} Z^j \exp(\beta^\top Z^j)}{\sum_{j\in\mathcal{R}(x_k)} \exp(\beta^\top Z^j) - \frac{m-1}{d(x_k)}\sum_{j\in\mathcal{D}(x_k)} \exp(\beta^\top Z^j)}\right] \tag{2.21}$$

On trouve enfin les coefficients optimaux en résolvant l'équation :

$$U(\beta) = 0 \tag{2.22}$$

Cette équation n'admet malheureusement pas de solution évidente. Une alternative serait d'utiliser un algorithme d'optimisation afin de trouver numériquement la solution. L'algorithme le plus souvent utilisé est celui de Newton-Raphson qui recherche itérativement les coefficients β optimaux selon l'itération :

$$\hat{\beta}^{(n+1)} = \hat{\beta}^{(n)} + \mathcal{I}^{-1}(\hat{\beta}^{(n)})U(\hat{\beta}^{(n)}) \tag{2.23}$$

Avec $\mathcal{I}$ la matrice d'information de Fisher. $\Big(\mathcal{I}(\beta)\Big)_{1\leq i,j\leq p,} = -\frac{\partial^2\mathcal{L}(\beta)}{\partial\beta_i\beta_j}$

Après avoir estimé les coefficients de régression, il est alors possible d'estimer le *baseline*.

2.3 Estimation du *baseline*

Le *baseline* h_0 a été considéré dans la section 2.2 comme paramètre de nuisance en vue de l'estimation des coefficients de régression. Nous allons, dans cette partie, proposer une méthode d'estimation de celui-ci. Dans la pratique, on s'intéresse plus à l'estimation

du *baseline* cumulé $H_0(x) = \int_0^x h_0(u)du$. D'après l'équation 3.4, on peut écrire la relation suivante entre la fonction de hasard et la fonction de survie.

$$h(x,t|Z) = \frac{f(x,t|Z)}{S(x,t|Z)} = -\frac{\partial}{\partial x}\ln\Big(S(x,t|Z)\Big) \tag{2.24}$$

La fonction de survie s'écrit alors :

$$S(x,t|Z) = \exp\left(-H(x,t|Z)\right) \tag{2.25}$$

Avec $H(x,t|Z) = \int_0^x h(u,t|Z)du$, la fonction de hasard cumulée.

Dans le cas où on a un modèle qui ne dépend pas du temps t, on peut écrire :

$$\begin{aligned} H(x|Z) &= \int_0^x h(u|Z)du && (2.26)\\ &= \int_0^x h_0(u)\exp\Big(\beta^\top Z\Big)du && (2.27)\\ &= \exp\Big(\beta^\top Z\Big)\int_0^x h_0(u)du && (2.28)\\ &= H_0(x)\exp\Big(\beta^\top Z\Big) && (2.29) \end{aligned}$$

Ces relations ne sont plus valables dans le cas où il y aurait une dépendance au temps. En effet, il existe une relation intuitive entre l'âge et le temps du fait que l'on considère toujours l'âge à un temps t. Un individu né au cours de l'année t_0 aura un âge x au cours de l'année t avec $t = t_0 + x$. Plus rigoureusement, l'équation s'écrirait :

$$t = t_0 + x_t \tag{2.30}$$

Avec :
- t_0 = Année de naissance
- t = Année d'observation
- x_t = Âge à l'année t

Ainsi, la fonction de hasard pourrait s'écrire :

$$\begin{aligned} h(x_t,t|Z) &= h_0(x_t)\exp\Big(\beta(t)^\top Z\Big) && (2.31)\\ &= h_0(x_t)\exp\Big(\beta(t_0+x_t)^\top Z\Big) && (2.32)\\ & && (2.33) \end{aligned}$$

On déduit donc que :

$$H(x_t,t|Z) = \int_0^{x_t} h_0(u)\exp\Big(\beta(t_0+u)^\top Z\Big)du \tag{2.34}$$

Un estimateur classique du *baseline* cumulé est l'estimateur de Breslow, obtenu par maximisation de la vraisemblance complète.

Notons $x_{[1]}, ..., x_{[n]}$ les temps de décès observés dans la base de données et $c_{[1]}, ..., c_{[n]}$ les temps de censure observés, n étant le nombre d'individus.

Pour l'individu i on observe soit un temps de décès $x_{[i]}$, soit un temps de censure $c_{[i]}$. Notons respectivement X et C les variables représentant respectivement l'instant de décès et celui de la censure. Posons $\Delta_i = I_{\{X_i<C_i\}}$. $\delta_i = I_{\{x_{[i]}<c_{[i]}\}}$ une variable indiquant l'observation ou pas du décès d'un individu i. Si $\delta_i = 1$, alors le décès a été observé. Notons $Y_i = \min(X_i, C_i)$.

La vraisemblance complète de l'ensemble des individus en cas de censure déterministe s'écrit :

$$L(\beta) = \prod_{i=1}^{n} \Big[f(y_i|Z^i)\Big]^{\delta_i} \Big[S(y_i|Z^i)\Big]^{1-\delta_i} \tag{2.35}$$

La démonstration de cette formule vient du fait que :

$$L(\beta) = \prod_{i=1}^{n} \mathbb{P}\Big(Y_i \in [y_i, y_i + dy_i], \Delta_i = \delta_i\Big) \tag{2.36}$$

Intuitivement, c'est la probabilité que pour chaque individu, les lois de censure et de survie soient conformes aux observations dans la base. On montre que (Planchet, 2016b) :

$$\mathbb{P}\Big(Y_i \in [y_i, y_i + dy_i], \Delta_i = \delta_i\Big) = \Big[f(y_i|Z^i)\Big]^{\delta_i} \Big[S(y_i|Z^i)\Big]^{1-\delta_i} \tag{2.37}$$

Donc,

$$L(\beta) = \prod_{i=1}^{n} \Big[f(y_i|Z^i)\Big]^{\delta_i} \Big[S(y_i|Z^i)\Big]^{1-\delta_i} \tag{2.38}$$

$$= \prod_{i=1}^{n} \Big[h(y_i|Z^i)\Big]^{\delta_i} S(y_i|Z^i) \tag{2.39}$$

$$\tag{2.40}$$

L'approche proposée par Breslow pour l'estimation du *baseline* consiste à décaler les temps de censure survenant dans l'intervalle $\Big[x_{k-1}\ ;\ x_k\Big[$ à $x_{k-1}\ \forall\ k = 1, ..., d$ et à supposer que le *baseline* est constant par morceau sur les intervalles de décès appartenant aux intervalles $\Big]x_{k-1}\ ;\ x_k\Big]\ \forall\ k = 1, ..., d$ (Augustin, 2002).

On se place dans un cadre réel où plusieurs individus peuvent décéder en même temps. L'idée est de maximiser la vraisemblance complète en supposant que les coefficients de

régression sont égaux à leur valeur estimée $\hat{\beta}$ obtenue par la méthode de vraisemblance partielle. Ainsi, au lieu de maximiser $L(\beta)$, on s'intéresse à la maximisation de $L\Big(\hat{\beta}, h_0\Big)$ avec $\hat{\beta}$ connu.

$$L\Big(\hat{\beta}, h_0\Big) = \prod_{i=1}^{n} \Big[h_0(x_{[i]}) \exp\Big(\hat{\beta}(t)^\top Z^i\Big)\Big]^{\delta_i} \exp\left(- \int_0^{x_{[i]}} h_0(u) \exp\Big(\hat{\beta}(t)^\top Z^i\Big) du \right) \tag{2.41}$$

Comme pour l'estimation des coefficients de régression, l'on revient grâce à l'introduction du temps dans les données au modèle suivant :

$$h(x|Z) = h_0(x) \exp\Big(\beta^{0\top} Z + \beta^{1\top} Z^*\Big)$$

Considérons pour des raisons de simplicité, le modèle général $h(x|Z) = h_0(x) \exp\Big(\beta^\top Z\Big)$. La vraisemblance complète s'écrit alors :

$$L\Big(\hat{\beta}, h_0(x_{[i]})\Big) = \prod_{i=1}^{n} \Big[h_0(x_{[i]}) \exp\Big(\hat{\beta}^\top Z^i\Big)\Big]^{\delta_i} \exp\left(- \int_0^{x_{[i]}} h_0(u) \exp\Big(\hat{\beta}^\top Z^i\Big) du \right) \tag{2.42}$$

$$= \prod_{k=1}^{d} \left[\prod_{i \in \mathcal{D}(x_k)} h_0(x_k) \exp\Big(\hat{\beta}^\top Z^i\Big) \right] \left[\exp\left(- \sum_{i \in \mathcal{R}(x_k)} \int_{x_{k-1}}^{x_k} h_0(u) \exp\Big(\hat{\beta}^\top Z^i\Big) du \right) \right] \tag{2.43}$$

$$= \prod_{k=1}^{d} \left[\prod_{i \in \mathcal{D}(x_k)} h_0(x_k) \exp\Big(\hat{\beta}^\top Z^i\Big) \right] \left[\exp\left(- \sum_{i \in \mathcal{R}(x_k)} (x_k - x_{k-1}) h_0(x_k) \exp\Big(\hat{\beta}^\top Z^i\Big) du \right) \right] \tag{2.44}$$

Avec $x_k - x_{k-1} = 1$.

La fonction de log-vraisemblance s'écrit alors :

$$\mathcal{L}\Big(\hat{\beta}, h_0\Big) = \sum_{k=1}^{d} \left[\sum_{i \in \mathcal{D}(x_k)} \Big[\log\Big(h_0(x_k)\Big) + \hat{\beta}^\top Z^i \Big] - \sum_{i \in \mathcal{R}(x_k)} \exp\Big(\hat{\beta}^\top Z^i\Big) h_0(x_k) \right] \tag{2.45}$$

$$= \sum_{k=1}^{d} \left[d(x_k) \log\Big(h_0(x_k)\Big) + \sum_{i \in \mathcal{D}(x_k)} \hat{\beta}^\top Z^i - h_0(x_k) \sum_{i \in \mathcal{R}(x_k)} \exp\Big(\hat{\beta}^\top Z^i\Big) \right] \tag{2.46}$$

On estime alors le *baseline* à chaque temps en résolvant l'équation suivante :

$$\frac{\partial \mathcal{L}\Big(\hat{\beta}, h_0(x_k)\Big)}{\partial h_0(x_k)} = 0 \ \forall\, k \in [\![1 \ ; \ d]\!] \tag{2.47}$$

On trouve alors :

$$\hat{h}_0(x_k) = \frac{d(x_k)}{\sum_{i \in \mathcal{R}(x_k)} \exp\left(\hat{\beta}^\top Z^i\right)} \quad \forall\, k \in [\![1\ ;\ d]\!] \tag{2.48}$$

On déduit alors un estimateur du *baseline* cumulé :

$$\hat{H}_0(x) = \sum_{k=1}^{d} \frac{\mathcal{I}_{\{x_k \leq x\}} d(x_k)}{\sum_{i \in \mathcal{R}(x_k)} \exp\left(\hat{\beta}^\top Z^i\right)} \tag{2.49}$$

On peut aussi écrire l'estimation de la fonction de hasard cumulée et de la fonction de survie.

$$\hat{H}(x|Z) = \hat{H}_0(x) \exp\left(\hat{\beta}^\top Z\right) = \sum_{k=1}^{d} \frac{\mathcal{I}_{\{x_k \leq x\}} d(x_k)}{\sum_{i \in \mathcal{R}(x_k)} \exp\left(\hat{\beta}^\top Z^i\right)} \exp\left(\hat{\beta}^\top Z\right) \tag{2.50}$$

$$\hat{S}(x|Z) = \exp\left(- \hat{H}(x|Z)\right) \tag{2.51}$$

2.4 Intervalles de confiance et tests statistiques

Dans cette partie, nous allons nous intéresser particulièrement à la construction d'intervalles de confiance et de tests statistiques pour les coefficients de régression.

2.4.1 Intervalles de confiance

Tsiatis (1981), Andersen, Gill (1982) ont démontré deux propriétés importantes des estimateurs des coefficients de régression que sont : la consistance et la convergence en lois vers la loi normale. On dit qu'un estimateur $\hat{\theta}_n$ de θ est consistant si et seulement si $\hat{\theta}_n$ converge en probabilité vers θ c'est-à-dire si :

$$\forall\, \xi > 0, \quad \lim_{n \to +\infty} \mathbb{P}\Big(|\hat{\theta}_n - \theta| \geq \xi\Big) = 0 \tag{2.52}$$

$\hat{\theta}_n$ converge presque sûrement vers θ si :

$$\mathbb{P}\Big(\lim_{n \to +\infty} \hat{\theta}_n = \theta\Big) = 1 \tag{2.53}$$

On dit qu'une suite $\hat{\theta}_n$ converge en loi vers θ si :

$$\lim_{n \to +\infty} F_{\hat{\theta}_n}(\alpha) = F_\theta(\alpha) \ \ \forall\, \alpha \in \mathbb{R} : F \textit{ continue} \tag{2.54}$$

F étant la fonction de répartition de θ.

Tsiatis (1981), Andersen, Gill (1982) ont montré que l'estimateur $\hat{\beta}$ des coefficients de régression, estimé par la méthode de vraisemblance partielle, converge presque sûrement vers sa vraie valeur β lorsque la taille des données augmente. De plus $\hat{\beta} - \beta$ converge en loi vers un vecteur gaussien de moyenne 0 et de matrice de variance-covariance $\mathcal{I}^{-1}(\hat{\beta})$.

$$(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} \mathcal{N}\Big(0, \mathcal{I}^{-1}(\hat{\beta})\Big) \tag{2.55}$$

Grâce à ces deux résultats, nous allons calculer un intervalle de confiance de niveau α pour les coefficients $\beta(t)$.

L'estimateur $\hat{\beta}$ présenté précédemment peut s'écrire sous la forme : $\hat{\beta} = \Big(\hat{\beta}_1^{le}, ..., \hat{\beta}_p^{le}, \hat{\beta}_1^{tr}, ..., \hat{\beta}_p^{tr}\Big)$.
On peut alors écrire : $\hat{\beta}(t) = \Big(\hat{\beta}_1(t), ..., \hat{\beta}_p(t)\Big)$ avec : $\forall\ j \leq p,\ \ \hat{\beta}_j(t) = \hat{\beta}_j^{le} + \hat{\beta}_j^{tr} f(t)$.
De l'équation 2.4.1, on déduit que : $\beta \xrightarrow{\mathcal{L}} \mathcal{N}\Big(\hat{\beta}, \mathcal{I}^{-1}(\hat{\beta})\Big)$.

Comme β est un vecteur gaussien, $\beta(t) = \beta^{le} + \beta^{tr} f(t)$ est aussi un vecteur gaussien de moyenne $\hat{\beta}(t)$ et de matrice de variance-covariance que nous noterons $C\Big(\hat{\beta}(t)\Big)$.

$$C\Big(\hat{\beta}(t)\Big) = \left(\Big[C\Big(\hat{\beta}(t)\Big)\Big]_{i,j}\right)_{i,j \leq p}$$

Avec :

$$\begin{aligned}
\Big[C\Big(\hat{\beta}(t)\Big)\Big]_{i,j} &= Cov\Big(\hat{\beta}_i(t), \hat{\beta}_j(t)\Big) \\
&= Cov\Big(\hat{\beta}_i^{le} + \hat{\beta}_i^{tr} f(t), \hat{\beta}_j^{le} + \hat{\beta}_j^{tr} f(t)\Big) \\
&= Cov\Big(\hat{\beta}_i^{le}, \hat{\beta}_j^{le}\Big) + Cov\Big(\hat{\beta}_i^{le}, \hat{\beta}_j^{tr} f(t)\Big) + Cov\Big(\hat{\beta}_i^{tr} f(t), \hat{\beta}_j^{le}\Big) \\
&\quad + Cov\Big(\hat{\beta}_i^{tr} f(t), \hat{\beta}_j^{tr} f(t)\Big) \\
&= Cov\Big(\hat{\beta}_i^{le}, \hat{\beta}_j^{le}\Big) + f(t)\left[Cov\Big(\hat{\beta}_i^{le}, \hat{\beta}_j^{tr}\Big) + Cov\Big(\hat{\beta}_i^{tr}, \hat{\beta}_j^{le}\Big)\right] \\
&\quad + \Big[f(t)\Big]^2 Cov\Big(\hat{\beta}_i^{tr}, \hat{\beta}_j^{tr}\Big)
\end{aligned}$$

La matrice de variance-covariance $\mathcal{I}^{-1}(\hat{\beta})$ peut s'écrire :

$$\mathcal{I}^{-1}(\hat{\beta}) =: \begin{pmatrix} \mathcal{I}_{le}^{-1} & \mathcal{I}_{le \times tr}^{-1} \\ \mathcal{I}_{tr \times le}^{-1} & \mathcal{I}_{tr}^{-1} \end{pmatrix} \tag{2.56}$$

Avec : $\left(\mathcal{I}_{le}^{-1}\right)_{i,j} = Cov\left(\hat{\beta}_i^{le}, \hat{\beta}_j^{le}\right)$; $\left(\mathcal{I}_{tr}^{-1}\right)_{i,j} = Cov\left(\hat{\beta}_i^{tr}, \hat{\beta}_j^{tr}\right)$; $\left(\mathcal{I}_{le\times tr}^{-1}\right)_{i,j} = Cov\left(\hat{\beta}_i^{le}, \hat{\beta}_j^{tr}\right)$ et $\left(\mathcal{I}_{tr\times le}^{-1}\right)_{i,j} = Cov\left(\hat{\beta}_i^{tr}, \hat{\beta}_j^{le}\right)$. Avec : $\mathcal{I}_{le\times tr}^{-1} = \left(\mathcal{I}_{tr\times le}^{-1}\right)^\top$.

On déduit donc que :

$$\left[C\left(\hat{\beta}(t)\right)\right]_{i,j} = \left(\mathcal{I}_{le}^{-1}\right)_{i,j} + f(t)\left[\left(\mathcal{I}_{le\times tr}^{-1}\right)_{i,j} + \left(\mathcal{I}_{le\times tr}^{-1}\right)_{j,i}\right] + \left[f(t)\right]^2\left(\mathcal{I}_{tr}^{-1}\right)_{i,j} \tag{2.57}$$

D'où :

$$\left(\hat{\beta}(t) - \beta(t)\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, C\left(\hat{\beta}(t)\right)\right) \tag{2.58}$$

Nous pouvons ainsi construire des intervalles de confiance pour $\beta(t) = \left(\beta_1(t), ..., \beta_p(t)\right)$. Cependant, les coefficients $\left(\beta_1(t), ..., \beta_p(t)\right)$ n'étant pas indépendants, il est alors plus rigoureux de construire une région de confiance au lieu d'intervalles de confiance afin de prendre en compte les effets liés à la corrélation entre les différents paramètres. Vous trouverez en annexes, un exemple d'illustration.

Nous avons montré que :

$$(\hat{\beta}(t) - \beta(t)) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, C\left(\hat{\beta}(t)\right)\right)$$

Ainsi, on déduit qu'à la limite :

$$\left(\hat{\beta}(t) - \beta(t)\right)^\top \left[C\left(\hat{\beta}(t)\right)\right]^{-1}\left(\hat{\beta}(t) - \beta(t)\right) \sim \chi_p^2$$

χ_p^2 étant une loi du « khi-deux à p degrés de liberté ». La région de confiance des coefficients $\beta(t)$ s'écrira donc :

$$\mathcal{A} = \left\{x, y \;\mid\; z_\alpha \leq \left(\hat{\beta}(t) - \beta(t)\right)^\top \left[C\left(\hat{\beta}(t)\right)\right]^{-1}\left(\hat{\beta}(t) - \beta(t)\right) \leq z_{1-\alpha}\right\}$$

Avec

$$\mathbb{P}\left(z_\alpha \leq \chi_p^2 \leq z_{1-\alpha}\right) = 1 - \alpha$$

Il est aussi possible de calculer une région de confiance pour le rique de base. Tsiatis, Andersen, Gill ont montré que l'estimateur du *baseline* présenté précédemment converge en loi vers un processus gaussien. Nous ne présenterons pas dans cette études les résultats du calcul des intervalles de confiance pour le *baseline*. Le lecteur intéressé pourra ce référer à la littérature proposée.

2.4.2 Tests statistiques

Plusieurs classes de tests statistiques sont utilisées pour effectuer des tests d'hypothèses sur les coefficients de régression β. Ils peuvent être utilisés dans notre cas pour tester la significativité d'une variable du modèle, pour ne garder que les variables significatives. Aussi, ces tests peuvent nous permettre de juger de la significativité des effets liés au niveau ou au temps. En effet, il se peut par exemple qu'il n'y ait pas de tendance liée à une des variables, ou que la valeur de l'effet lié au temps soit proche de zéro. Le test statistique peut nous permettre d'établir un jugement rigoureux. Les tests les plus utilisés dans le cadre du modèle de Cox sont le test du rapport de vraisemblance, le test de Wald et le test du Score. On teste l'hypothèse :

$$H_0 : \hat{\beta} = \beta_0$$

— Test du rapport de vraisemblance : le principe est de mesurer l'écart entre les valeurs de log-vraisemblance partielles en $\hat{\beta}$ et β_0. La statistique du test s'écrit :

$$T_R = 2\Big(\mathcal{L}(\hat{\beta}) - \mathcal{L}(\beta_0)\Big) \tag{2.59}$$

On montre sous H_0 que T_R tend vers une loi du chi-deux.

— Test de Wald : il propose de mesurer l'écart entre $\hat{\beta}$ et β_0. La statistique du test est la suivante :

$$T_W = \Big(\hat{\beta} - \beta_0\Big)^\top \mathcal{I}(\hat{\beta})\Big(\hat{\beta} - \beta_0\Big) \tag{2.60}$$

$\mathcal{I}$ étant la matrice d'information de Fisher. On montre de même que T_W tand vers une loi du chi-deux sous H_0.

— Test du Score : le principe est de calculer la pente $U(\beta) = \frac{\partial \mathcal{L}(\beta)}{\partial \beta}$ de la log-vraisemblance en β_0. Cette pente doit être la plus petite possible sous H_0. La statistique du test s'écrit :

$$T_S = U(\beta_0)^\top \mathcal{I}^{-1}(\beta_0) U(\beta_0) \tag{2.61}$$

T_S tend vers une loi du chi-deux sous H_0.

Ces tests sont asymptotiquement équivalents lorsque le nombre de données ou de décès tant vers l'infini. Cependant, le test du rapport de vraisemblance est considéré comme plus performant que les deux autres (Therneau, Grambsch, 2000). En effet, il ne nécessite pas le calcul de la fonction score et de la matrice de Fisher.
Dans notre étude, nous avons choisi d'utiliser le test de Wald pour la simple raison que la méthode est cohérente avec la méthode de calcul des régions de confiance. Il est important de préciser que choisir une autre classe de test n'aurait pas changé les conclusions de notre étude.

Comme exemple de test, supposons que l'on veuille tester la significativité de l'effet de la première variable à un temps t_0 donné. On teste :

$$H_0 : \Big(\hat{\beta}_1(t_0), ..., \hat{\beta}_p(t_0)\Big) = \Big(0, \hat{\beta}_2(t_0), ..., \hat{\beta}_p(t_0)\Big)$$

ou écrit plus simplement :

$$H_0 : \hat{\beta}_1(t_0) = 0$$

Dans ce cas la statistique du test de Wald s'écrirait :

$$T_W = \Big(\hat{\beta}(t_0) - \beta_0(t_0)\Big)^\top \Big[C\Big(\hat{\beta}(t_0)\Big)\Big]^{-1} \Big(\hat{\beta}(t_0) - \beta_0(t_0)\Big) \overset{H_0}{\sim} \chi^2_1 \tag{2.62}$$

avec : $\hat{\beta}(t_0) = \Big(\hat{\beta}_1(t_0), \hat{\beta}_2(t_0), ..., \hat{\beta}_p(t_0)\Big)$ et $\beta_0(t_0) = \Big(0, \beta_2(t_0), ..., \beta_p(t_0)\Big)$.

On peut alors s'intéresser à la p-value du test afin d'émettre une conclusion sur la valeur de la statistique de test. Mathématiquement, la p-value est la probabilité, sous H_0, d'obtenir une statistique encore plus grande que celle observée. Pour un seuil que l'on aura choisi, on pourra considérer qu'une p-value inférieure à ce seuil indique que l'hypothèse H_0 pourra être rejetée. Dans notre cas, la p-value peut indiquer la probabilité de trouver un meilleur modèle sous H_0 (Goris, 2016). De façon générale, elle pourra s'interpréter comme une mesure de la présomption à rejeter l'hypothèse H_0 (Wasserman, 2004). Ainsi, plus la p-value est petite, plus la présomption de rejeter H_0 est forte. . Le lecteur intéressé par plus de détails sur la notion de p-value peut aussi se référer à l'article écrit par Gibbons, Pratt (1975).

Soit p_{value} la valeur de la p-value du test du Wald.

$$H_0 : \hat{\beta}(t) = \beta_0(t)$$

La statistique du test s'écrit [2] :

$$T_W(t) = \Big(\hat{\beta}(t) - \beta_0(t)\Big)^\top \Big[C\Big(\hat{\beta}(t)\Big)\Big]^{-1} \Big(\hat{\beta}(t) - \beta_0(t)\Big) \overset{H_0}{\sim} \chi^2_k \tag{2.63}$$

La p-value s'écrit alors :

$$p_{value}(t) = \mathbb{P}\Big(\chi^2_k \geq T_W(t)\Big) \tag{2.64}$$

Les seuils de la p-value sont en général choisi de manière aléatoire. Cependant, certains seuils ont été admis et largement utilisés dans la littérature.

Pour notre étude, nous prendrons comme seuil de p-value, 5%. Si la p-value est inférieure à 5%, on rejette H_0, et si elle est supérieure à 5%, on accepte H_0.

2. $k \leq p$

p-value	Interprétation
$[0\ ;\ 0,01]$	Très forte présomption contre H_0
$[0,01\ ;\ 0,05]$	Forte présomption contre H_0
$[0,05\ ;\ 0,1]$	Faible présomption contre H_0
$[0,1\ ;\ 1]$	Peu ou pas de présomption contre H_0

TABLE 14 – Interprétations de la p-value

2.5 Adéquation du modèle

Plusieurs mesures sont couramment utilisées dans la littérature pour mesurer la qualité d'ajustement d'un modèle. Classiquement, on utilise la vraisemblance. En effet, plus le modèle ajuste bien les données, plus la valeur de la vraisemblance est élevée. Ainsi, on peut comparer deux modèles en utilisant le critère de la vraisemblance. Cependant, l'un des inconvénients de l'utilisation de la vraisemblance comme critère de discrimination est qu'elle est d'autant plus grande que le nombre de variables dans le modèle augmente. Une alternative peut être d'utiliser d'autres critères comme l'AIC [3] ou le BIC [4]. L'AIC permet de pénaliser l'augmentation du nombre de paramètres du modèle.

$$AIC = 2k - 2\ln(L) \tag{2.65}$$

avec k le nombre de paramètres et L la vraisemblance.

Le BIC quant à lui introduit une plus forte pénalisation des paramètres en introduisant le nombre d'observations.

$$BIC = 2\ln(n)k - 2\ln(L) \tag{2.66}$$

n le nombre d'observations.

Le meilleur modèle est celui qui possède le plus faible AIC ou BIC.

Il existe encore d'autres méthodes permettant de mesurer la qualité d'ajustement du modèle de Cox. Le lecteur intéressé pourra consulter la thèse rédigée par Chauvel (2014) qui propose une large littérature sur le sujet. Dans notre étude, nous avons choisi d'utiliser l'AIC et le BIC pour des raisons de simplicité.

3. Akaike Information Criterion
4. Bayesian Information Criterion

Troisième partie

Application du modèle : analyses et résultats

Chapitre 1

Cadre applicatif et simulation des données

> « ***....l'homme est mal à l'aise et la femme plus en confort : le surplus de longévité est pour cette dernière*** » *Guy Tassin*[1]

Dans cette partie, nous présenterons les résultats de l'implémentation du modèle de Cox sur plusieurs populations.

1.1 Cadre d'application

Le but de notre étude est de répondre à une problématique actuarielle qui est celle de l'estimation du risque de tendance du *basis risk*. Nous avons énuméré en première partie de cette étude des facteurs pouvant avoir une influence significative sur le niveau et la tendance du *basis risk* tels que le facteur social, le facteur économique, le facteur géographique, le facteur génétique et le facteur chance. Nous proposons ici d'en faire deux applications :

— Une application à l'étude du *basis risk* entre population de rentiers et population nationale.

— Une application à l'étude d'un *basis risk* géographique : celui entre l'Allemagne de l'Est et l'Allemagne de l'Ouest.

L'objectif de ces deux applications est de montrer par la première que la méthode proposée permet bien d'estimer le risque de tendance du *basis risk*, puis de proposer une

1. Sex-ratios : Le genre féminin à Haveluy de 1701 à 1870, p.172, L'Harmattan, 2005

estimation concrète du risque de tendance à travers la deuxième application.

1.2 Risque de tendance entre population nationale et population de rentiers

Nous proposons ici une approche générale d'estimation du risque de tendance entre la population de rentiers et la population nationale. Il s'agit dans cette approche de simuler une population de rentiers grâce aux tables générationelles réglementaires de rentiers en France et une population nationale grâce aux données de mortalité de la Human Mortality Database.

Cette démarche à l'avantage d'être générale et de ne pas dépendre de la structure ni de la composition d'un portefeuille d'assurance. En effet, pour l'estimation du risque de tendance entre la population de rentiers et la population nationale, il serait possible d'utiliser des données de rentiers avec un historique suffisamment long pour détecter une tendance. Dans l'idéal, il faudrait disposer de données de portefeuille permettant d'observer les facteurs d'hétérogénéité par individu, sur une période assez longue (au moins 15 ans), pour espérer détecter un *trend*. L'avantage de l'utilisation d'un portefeuille réel de rentiers est celui de pouvoir tenir compte de tous les facteurs d'hétérogénéité du portefeuille.

Cependant, les portefeuilles d'assurance peuvent avoir une structure différente selon les assureurs. Chaque assureur, en fonction de son positionnement sur le marché, peut disposer d'un portefeuille de rentiers de composition très différente d'un autre. L'un attirera par exemple des individus disposant d'un niveau de rémunération élevé et un autre ceux qui disposent d'un niveau de rémunération moyen. Ainsi, les résultats obtenus peuvent être très volatils d'un assureur à un autre.

Nous proposons une approche rigoureuse par une simulation de la population des rentiers grâce aux tables de mortalité réglementaires, ce qui nous conduirait à des résultats généraux.

Les tables générationelles TGH-TGF ont été construites grâce à des données de rentiers sur une période allant de 1994 à 2004. Disposant d'un faible historique, les taux bruts récoltés ont été ajustés lors de son élaboration sur des données de population nationale française afin de les extrapoler à d'autres années. Le modèle d'extrapolation utilisé est le suivant :

$$\ln\left(\frac{q_{xt}^{rentiers}}{1-q_{xt}^{rentiers}}\right) = a_x \ln\left(\frac{q_{xt}^{Nationale}}{1-q_{xt}^{Nationale}}\right) + b_x + \xi_{xt} \tag{1.1}$$

a_x et b_x étant des paramètres à estimer et ξ_{xt} le bruit.

Le résultat a été la construction de tables prospectives pour les générations 1900 à 2005, sur la tranche d'âge de 40 ans à 95 ans, prolongée à 120 ans grâce à une clôture de

table.

Plusieurs raisons nous contraignent d'utiliser ces tables ainsi construites pour modéliser le risque de tendance. D'abord, le fait d'avoir utilisé des données de la population nationale pour leur ajustement a créé un biais dans l'estimation du *trend*. En effet, les tendances des améliorations de mortalité sur la population nationale ont été transmises à la population de rentiers. De ce fait, on devrait retrouver le même *trend* sur ces tables TG.

Aussi, ces tables sont des tables par cohorte. Il existe deux visions pour la construction une table de mortalité : la vision par cohorte et celle par période. La vision par cohorte consiste à calculer la mortalité sur les générations. Comme exemple, pour la génération 1950 cela consiste à considérer à chaque âge, les individus nés en 1950. A l'âge 1 an, on ne considérera que les individus ayant 1 an en 1951 ; de même à l'âge 50 ans, on considérera uniquement les individus qui ont 50 ans en 2000.

La vision par période en revanche, consiste à observer les individus d'une année, quelles que soient leurs années de naissance.

La vision par cohorte, contrairement à la vision par période, prend en compte les améliorations futures dans l'estimation de la mortalité puisque l'on observe les individus sur différentes années dans le futur. Ainsi, modéliser le *trend* par des tables par cohorte conduirait inévitablement à prendre en compte les anticipations futures de la mortalité donc à une surestimation du niveau et du *trend*.

Il paraît donc évident que l'utilisation de ces tables devrait nous conduire à un résultat connu d'avance, à savoir à retrouver le même *trend* que celui de la population nationale. Nous pouvons cependant exploiter ce résultat pour tester la validité de l'approche proposée. En effet, le modèle n'ayant pas encore été utilisé pour modéliser le *trend*, il est alors important d'évaluer sa capacité à bien estimer le risque de tendance à travers un résultat à priori connu d'avance. Nous allons montrer par cette première application que l'approche proposée est donc valide, et cohérente.

Par la suite, nous adopterons la démarche décrite dans le figure 12.

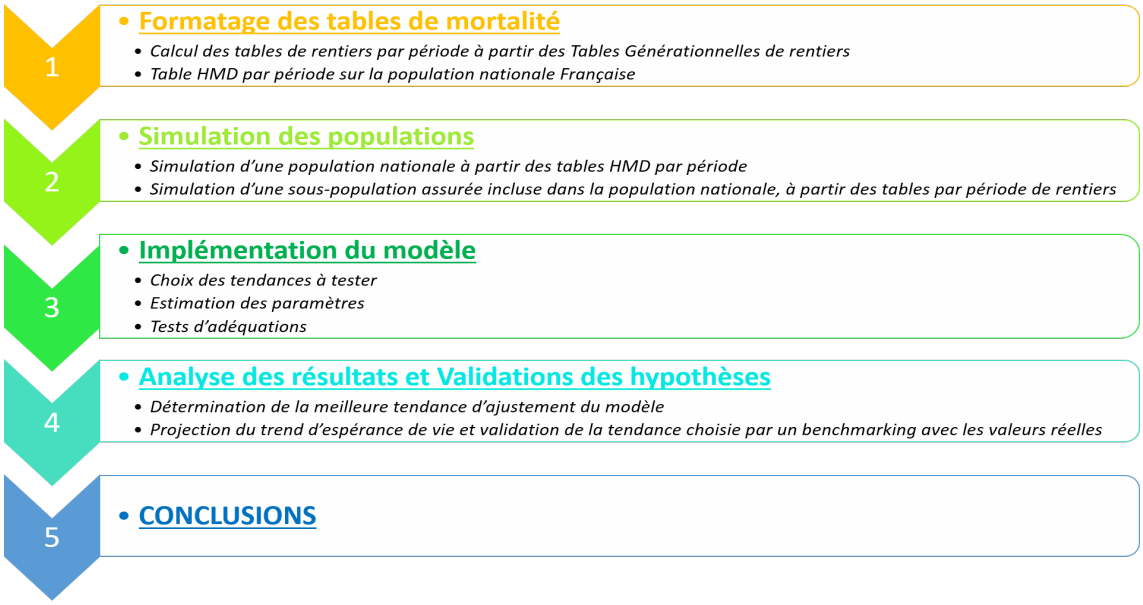

FIGURE 12 – Étapes de la modélisation

1.2.1 Formatage des tables de mortalité

Dans cette partie, nous présenterons l'approche adoptée pour le formatage des tables de mortalité. Il s'agira principalement de recalculer les tables de rentiers par période à partir des tables par cohorte.

Description de la méthode

Le principe de la transformation des tables par cohorte en tables par période peut être visualisé grâce au triangle de Lexis.

Le triangle de Lexis ou diagramme de Lexis est un représentation graphique utilisée par les démographes pour représenter des évènements ou individus par âge et par période. On y retrouve très souvent en abscisse, l'année d'observation et en ordonnée l'âge des individus (voir 13).

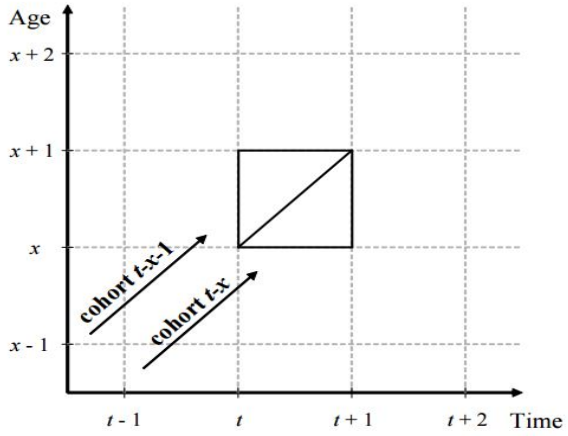

FIGURE 13 – Diagramme de Lexis
Source : www.mortality.org

Le diagramme de Lexis ainsi représenté nous fournit une représentation des individus par période et par cohorte. Par exemple, un individu observé durant l'année t [2] à l'âge x [3] sera représenté dans le carré en trait plein.

Une période sur le diagramme s'observe selon la verticale tandis qu'une cohorte s'observe suivant la diagonale. Pour en donner une illustration plus claire, étudions la représentation de la figure 14.

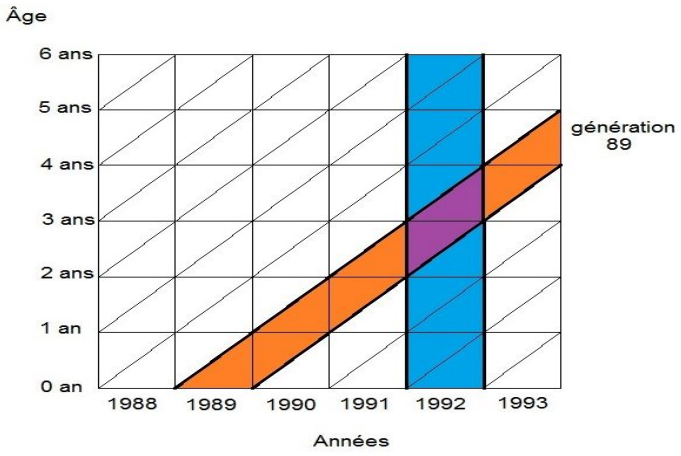

FIGURE 14 – Visualisation des périodes et cohortes
Source : quanti.hypotheses.org

En bleu, on observe les individus sur l'année 1992 tandis qu'en orange on observe les

2. c'est-à-dire dans l'intervalle $[t\ ,\ t+1[$
3. c'est-à-dire dans l'intervalle $[x\ ,\ x+1[$

individus de la cohorte 1989, c'est-à-dire tous ceux qui sont nés durant l'année 1989. Les individus d'une génération sont observés à chaque période jusqu'à extinction de celle-ci. On peut observer que dans une période il y a deux cohortes, de même une cohorte est composée de deux périodes. Par exemple, la cohorte 1989 à l'âge 3 ans est composée des individus de la période 1992 (en violet) et de ceux de la période 1993 (en orange). De même, la période 1992 est composée à l'âge 3 ans des individus de la cohorte 1989 (en violet) et de ceux de la cohorte 1988 (en bleu). La mortalité à l'année t et à l'âge x dépend donc de celle des générations $g = t - x$ et $g = t - x - 1$.

Soit Tm_p une table de mortalité par période et Tm_c une table de mortalité par cohorte. Notons x l'âge, t l'année et $g = t - x$ la génération. $Tm_c(x, g)$ représente le taux de mortalité à l'âge x pour la cohorte g et $Tm_p(x, t)$ représente la mortalité à l'âge x durant l'année t. Disposant des données de Tm_c, on peut retrouver la table par période Tm_p grâce à l'approximation suivante :

$$Tm_c(x, g = t - x) = Tm_p(x, t) + o\Big(Tm_p(x, t)\Big) \tag{1.2}$$

Cette approximation est justifiée dans la pratique par les données. Par exemple, les données de la Human Mortality Database sur la mortalité française par cohorte et par période montrent que le taux de mortalité pour les hommes à l'âge $x = 80$ ans de la cohorte $g = t - x = 1924$ est de 0,06344 tandis que celui des hommes d'âge $x = 80$ ans au cours de l'année $t = x + g = 2004$ est de 0,06257, soit une différence de 0,00087. On trouve donc que :

$$\frac{Tm_c(80, 1924) - Tm_p(80, 2004)}{Tm_p(80, 2004)} = \frac{0,06344 - 0,06257}{0,06257} = 0,01390443 \tag{1.3}$$

On retrouve donc que :

$$Tm_c(80, 1924) = Tm_p(80, 2004) + o\Big(Tm_p(80, 2004)\Big) \tag{1.4}$$

Cette approximation peut aussi être justifiée si l'on considère que les écarts de mortalité entre deux générations voisines sont négligeables sauf en cas de surmortalité exceptionnelle ou de longévité exceptionnelle.

Par ailleurs, trouver une relation mathématique rigoureuse entre les tables par cohorte et celle par période n'est pas du tout évidente.

Les données

Les tables réglementaires TG sont disponibles en ligne sur ressources-actuarielles.net. A partir de ces tables, on peut alors reconstituer la nouvelles tables de mortalité par période.

Les tables ainsi construites sont disponibles sur une période allant de 1996 à 2100 et pour pour les âges 0 à 120 ans.

Pour notre modélisation, nous avons décidé d'utiliser les données des tables entre 1996 et 2005 c'est-à-dire aux périodes sur lesquelles elles ont été ajustées. Nous effectuerons des projections grâce au modèle jusqu'en 2015 puis nous comparerons les données ajustées à celle des tables TG jusqu'en 2015. Le choix de se limiter à l'année 2015 nous permet de rester cohérent avec les données HMD sur la population française qui sont, elles aussi, disponibles jusqu'en 2015. Nous n'exploiterons les données qu'entre 40 ans à 95 ans afin d'éviter tout biais que pourrait induire l'utilisation des méthodes de clôture de table.

Les résultats pour l'année 1996 des tables de mortalité par période sont représentés par la figure 15. Les résultats pour les années 2000 et 2005 sont présentés en annexes.

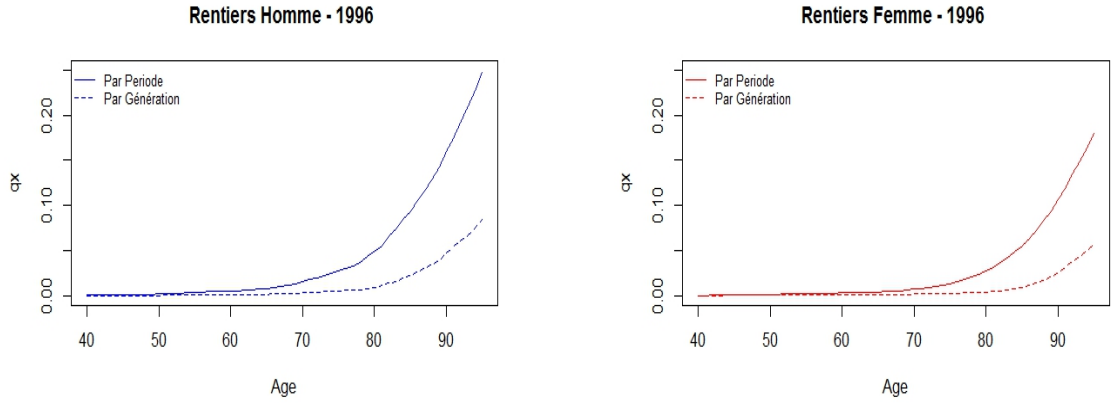

FIGURE 15 – Mortalité des rentiers pour l'année et la génération 1996

1.2.2 Simulation des populations

Nous présenterons dans cette partie la démarche et les résultats de la simulation des populations.

Description de la démarche

Nous avons défini deux grandes approches pour la simulation des populations :

— La simulation disjointe de population :

Dans cette approche, on simule indépendamment les sous-populations. On les regroupe ensuite pour constituer une population complète. La population ainsi constituée est composée de plusieurs sous-populations disjointes. Un individu ne peut donc pas à la fois appartenir à deux sous-groupes.

— La simulation inclusive de population :

Cette approche consiste à simuler des populations incluses les unes dans les autres. On simule d'abord la population la plus large, puis on choisit au sein de cette population les individus qui vont représenter les différentes sous-populations. Un individu peut alors dans ce cas appartenir à plusieurs sous-populations.

Pour l'étude du risque de tendance entre population assurée et population nationale, il paraît évident que la simulation inclusive de population est la plus adaptée. En effet, la population des rentiers est incluse dans celle nationale. On ne prendra pas en compte les phénomènes d'immigration dans cette démarche.

Simulation de la population nationale

Nous avons décrit en annexes la démarche de simulation de la population nationale sous forme d'algorithme. Il est important de préciser que l'objectif n'est pas de simuler une population ayant la même composition que la population nationale en France mais plutôt de simuler une population ayant la même mortalité que la population nationale française. En effet, il est tout à fait inutile de se lancer dans une démarche de réplication de la population nationale française en terme de composition parce que le modèle peu sensible au nombre d'individus présents par âge et par année, mais plutôt à la mortalité par âge et par année. La preuve est donnée par l'estimateur du *baseline* calculé dans la section 2.3 de la deuxième partie de notre étude. Nous avons montré que :

$$\hat{h}_0(x) = \frac{d(x)}{\sum_{i \in \mathcal{R}(x)} \exp\left(\hat{\beta}^\top Z^i\right)} \tag{1.5}$$

Ainsi pour Z=0, on trouvera par exemple que $\hat{h}_0(x) \simeq 0,011641$ si $d(x) = 1000$ et $Card(\mathcal{R}(x)) = 85903$ (c'est-à-dire 85903 individus présents dans le portefeuille à l'âge x avec 1000 décès l'âge x) ou si $d(x) = 757257$ et $Card(\mathcal{R}(x)) = 65050873$[4] (c'est-à-dire 65050873 individus présents dans le portefeuille à l'âge x avec 757257 décès l'âge x). En terme d'estimation du *baseline*, un plus gros portefeuille n'apporte aucune information supplémentaire si ce n'est plus de précisions.

Cependant si l'estimation du *baseline* est peu sensible aux données, l'estimation des paramètres par contre y est plus sensible. Nous présenterons une illustration de la sensibilité du modèle à la taille des données dans le chapitre suivant.

Les tables de mortalité sur la population nationale sont issues des données par période de mortalité de la Human Mortality Database.

4. $Card$ = Cardinal

Afin de mieux répliquer la mortalité à chaque âge surtout aux âges avancés, nous avons décidé de simuler la population nationale par tranche d'âge en repartant de la population initiale $N_{x_0,t}$. Nous avons donc effectué la simulation pour chaque tranche d'âges 40 ans - 60 ans, 61 ans - 80 ans, 81 ans - 90 ans et 91 ans - 95 ans en repartant de la population initiale de 10000 individus. Cette subdivision d'âge nous semblais être la plus optimale en terme de taille de la population simulée et de la justesse de sa mortalité. Sur la figure 16, on peut observer les statistiques pour l'année 1996 pour les hommes et les femmes.

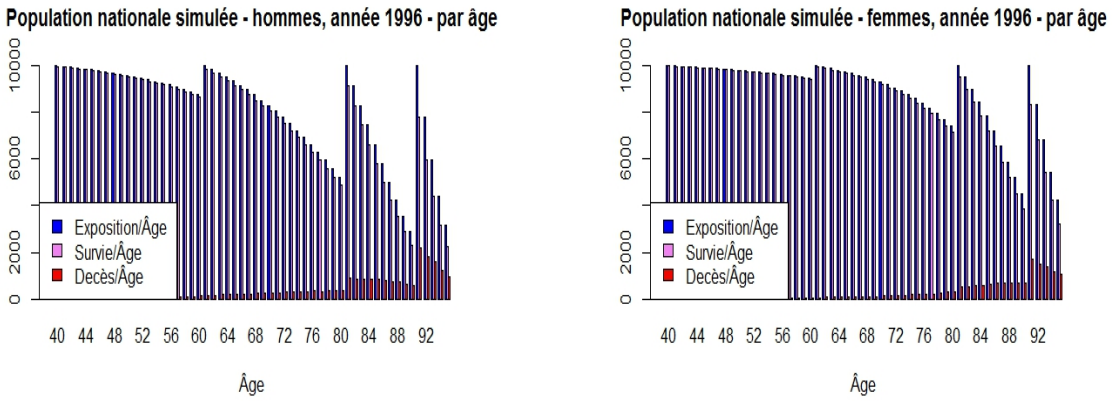

FIGURE 16 – Populations nationales simulées - 1996

Après simulation, nous avons constitué deux portefeuilles d'individus observés entre 1996 et 2005 : celui des hommes avec une population de 4,6 millions individus et celui des femmes avec une population de 4,9 millions individus.

Nous avons comparé la mortalité simulée à la mortalité réelle de la population nationale. Les résultats pour l'année 1996 sont présentés par la figure 17. Les résultats pour les années 2000 et 2005 sont présentés en annexes et conduisent à des conclusions identiques.

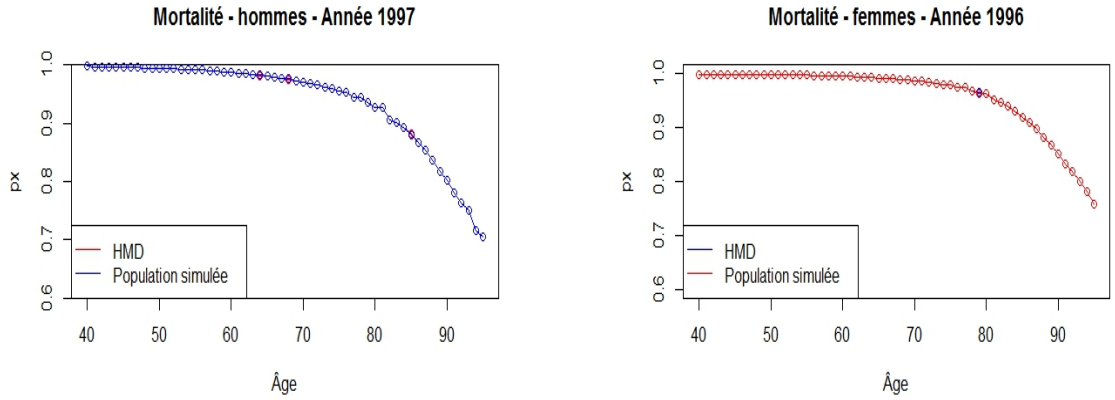

FIGURE 17 – Comparaison de la mortalité simulée et celles donnée par les tables de mortalité.

On peut voir que les graphes sont « parfaitement » superposés. En réalité, la mortalité est proprement répliquée à 10^{-4} près. Pour une réplication à 10^{-5} près comme dans les données initiales, il faudrait simuler la population avec une population initiale d'au moins 100000 individus. Cela conduirait à une trop grande de base de données ce qui rendrait difficile la modélisation du fait que la capacité des ordinateurs à disposition est limitée.

Simulation de la population de rentiers

Après avoir simulé la population nationale, nous avons simulé celle des rentiers. La démarche est celle d'une simulation inclusive de population. A partir de la population, nous avons défini grâce à une variable les individus qui feront partie de la population de rentiers de telle sorte que pour chaque année et pour chaque âge, la mortalité de celle-ci soit égale à celle observé dans les tables de rentiers par période précédemment construite. Pour cela, il faut en premier lieux définir la proportion de la population des rentiers dans la population nationale. Nous avons pris l'hypothèse qu'un tiers de la population nationale fait partie de celle des rentiers. Cette hypothèse est faite de façon arbitraire. Nous testerons dans le chapitre suivant la sensibilité des résultats obtenus à cette hypothèse. La méthode de simulation est présentée sous forme d'algorithme en annexes.

Après implémentation de la méthode nous avons comparé la mortalité de la population des rentiers simulée à celle donnée par les tables générationelles par période. Les résultats pour l'année 1996 peuvent être visualisés dans la figure 18. Les résultats pour les années 2000 et 2005, ainsi que les statistiques sont présentés en Annexe.

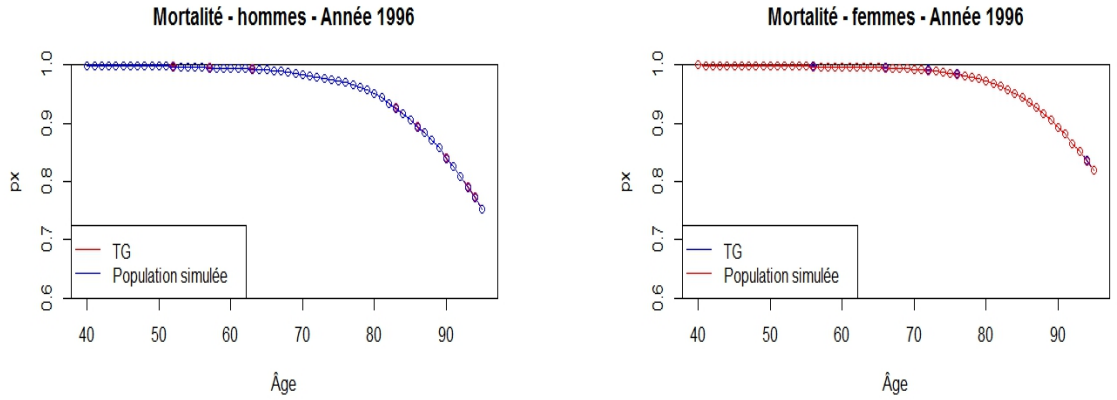

FIGURE 18 – Comparaison de la mortalité simulée des rentiers et des tables de mortalité

On peut constater que la mortalité des rentiers simulée est égale à 10^{-4} près à celle donnée par les tables de mortalité.

1.2.3 Implémentation du modèle

Le modèle implémenté s'écrit :

$$h(x,t|Z) = h_0(x)\exp(\beta(t)^\top Z) \tag{1.6}$$

avec : $\beta(t) = \beta^{le} + \beta^{tr} f(t)$; f étant la fonction de tendance.

Dans notre cas, Z représente la variable Population. Il prend la valeur 1 si l'individu appartient à la sous-population de rentiers et 0 sinon.

Choix de la fonction de *trend*

Nous avons testé 7 fonctions de tendance pouvant représenter différents scénarios d'évolution de la mortalité. Il s'agit des scénarios :

— **Linéaire** : on suppose une évolution linéaire de l'effet de la covariable sur la mortalité ; $f(t) = t$.

— **Quadratique** : on suppose un scénario extrême d'évolution de pente plus une forte que celle linéaire ; $f(t) = t^2$.

— **Logarithmique** : on suppose évolution croissance mais avec une pente plus faible, $f(t) = \ln(t)$.

— **Inverse** : on suppose une décroissance de l'effet de la covariable $f(t) = \frac{1}{t}$.

— **Exponentiel inverse** : on suppose un scénario extrême de forte décroissance, $f(t) = \exp(-t)$.

— **Tangent hyperbolique** : on suppose une évolution modérée qui s'estompe à un instant donné, $f(t) = tanh(t) = \frac{\exp(z)-\exp(-z)}{\exp(z)+\exp(-z)}$.

— **Croissant cyclique** : on suppose que la mortalité augmente mais de façon cyclique avec un période de 10 ans : $f(t) = t \times \exp\Big(-\sin(\frac{2\pi t}{10})\Big)$.

Nous pouvons visualiser ces différents scénarios sur la figure 19.

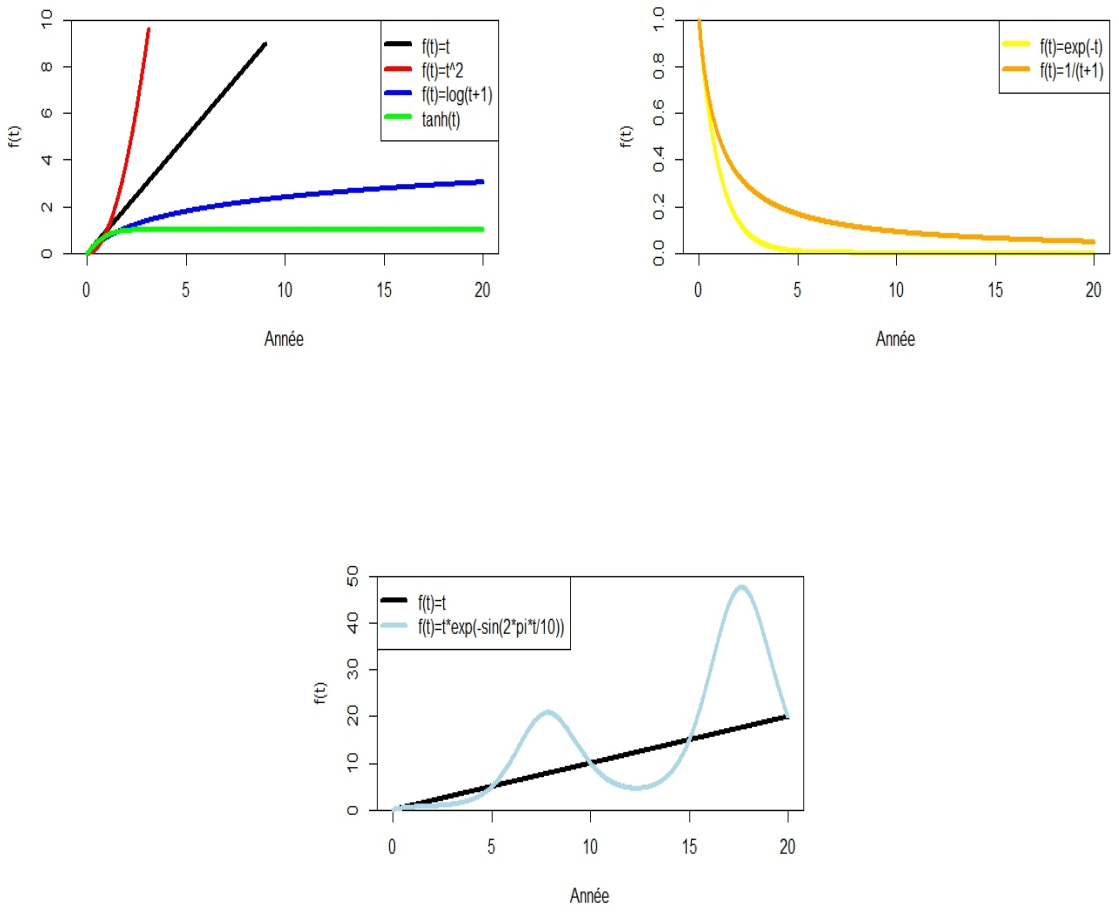

FIGURE 19 – Fonctions de tendance

Estimation des paramètres

Nous avons estimé les coefficients β^{le} et β^{tr} associés aux différents scénarios de tendance. Les résultats pour le portefeuille des hommes et des femmes sont présentés par dans le tableau 15. Nous pouvons constater que le modèle estime correctement la pente de l'évolution de la mortalité. En effet, lorsque la fonction de tendance choisie est croissante, le paramètre de tendance estimé est alors négatif tandis que lorsque la fonction de tendance choisie est décroissance, le paramètre de tendance estimé est positif. Le modèle capte donc la pente de l'évolution décroissante de la mortalité au fil du temps.

Modèle	Coefficients Estimées	
	Hommes	Femmes
1	$\beta(t) = -0,41 - 0,05 \times \ln(t+1)$	$\beta(t) = -0,42 - 0,0028 \times \ln(t+1)$
2	$\beta(t) = -0,45 - 0,001 \times t^2$	$\beta(t) = -0,45 - 0,0007 \times t^2$
3	$\beta(t) = -0,43 - 0,013 \times t$	$\beta(t) = -0,43 - 0,007 \times t$
4	$\beta(t) = -0,53 + 0,12 \times \frac{1}{t+1}$	$\beta(t) = -0,49 + 0,064 \times \frac{1}{t+1}$
5	$\beta(t) = -0,51 + 0,09 \times \exp(-t)$	$\beta(t) = -0,48 + 0,049 \times \exp(-t)$
6	$\beta(t) = -0,42 - 0,08 \times \tanh(t)$	$\beta(t) = -0,43 - 0044 \times \tanh(t)$
7	$\beta(t) = -0,46 - 0,004 \times t \times \exp\Big(-\sin(\frac{2\pi t}{10})\Big)$	$\beta(t) = -0,45 - 0,002 \times t \times \exp\Big(-\sin(\frac{2\pi t}{10})\Big)$

TABLE 15 – Coefficients de régression estimé

Après avoir estimer les coefficients de régression, nous avons calculé le risque instantané de décès puis ajusté les taux de mortalité. Les surfaces mortalité ajustées pour les tendances quadratique et croissante sinusoïdales sont représentées en annexes.

Tests statistiques

Nous avons testé la significativité des coefficients estimés grâce au test de Wald.

$$H_0 : \hat{\beta}(t) = 0$$

Nous avons obtenu des p-values très proches de 0 à chaque temps et pour chaque modèle. On rejette donc l'hypothèse H_0. Les résultats de la statistique du test ainsi que des p-values pour les hommes sont présentés en annexes.

Tests d'adéquation

Nous avons mesuré l'adéquation de chaque modèle à partir de l'AIC et le BIC. Les résultats sont consignés dans le tableau 16.

Modèle	Hommes		Femmes	
	AIC	BIC	AIC	BIC
Tangente hyperbolique	4651670	4651690	3494599	3494619
Exponentiel inverse	4651661	4651681	3494597	3494617
Inverse	4651650	4651670	3494594	3494614
Croissance cyclique	4651637	4651657	3494591	3494611
Logarithmique	4651630	4651651	3494590	3494610
Quadratique	4651629	4651650	3494590	3494610
Linéaire	4651623	4651644	3494589	3494608

TABLE 16 – AIC et BIC
(Ordonné par AIC décroissant)

On peut constater que chez les hommes et les femmes, le meilleur modèle en terme d'ajustement est le modèle avec un *trend* linéaire. C'est une première étape à la confirmation de nos attentes.

Un autre constat est que la qualité d'ajustement est globalement plus faible chez les femmes que chez les hommes. Cela devrait se matérialiser par de plus forts écarts sur le niveau des espérances de vie chez les femmes par rapport aux données empiriques.

1.2.4 Analyse des résultats et validation des hypothèses

Après avoir estimé les paramètres pour chaque modèle, nous avons ajusté les espérances de vie. Nous avons défini la mesure d'erreur suivante pour comparer les différents modèles ajustés :

$$Erreur_x^{aj} = \frac{1}{n} \sum_{t=1996}^{2005} \left(EV_{x,t}^{ajustee} - EV_{x,t}^{reelle} \right)^2 \tag{1.7}$$

Avec :

- $EV_{x,t}^{ajustee}$ l'espérance de vie ajustée par le modèle à l'âge x et l'année t
- $EV_{x,t}^{reelle}$ l'espérance de vie réelle du portefeuille à l'âge x et l'année t
- n le nombre d'année
- $Erreur_x^{aj}$ l'erreur quadratique d'ajustement annualisée.

On peut définir de même l'erreur de projection :

$$Erreur_x^{pr} = \frac{1}{n} \sum_{t=2006}^{2015} \left(EV_{x,t}^{projetee} - EV_{x,t}^{reelle} \right)^2 \tag{1.8}$$

Avec :

- $EV_{x,t}^{projetee}$ l'espérance de vie projetée par le modèle à l'âge x et l'année t
- $EV_{x,t}^{reelle}$ l'espérance de vie réelle du portefeuille à l'âge x et l'année t
- n le nombre d'année
- $Erreur_x^{pr}$ l'erreur quadratique de la projection annualisée.

Ajustement des espérances de vie

Nous nous sommes focalisés sur l'espérance de vie aux grands âges. Nous présenterons ici les résultats pour l'espérance de vie à 60 ans, sachant que les mêmes analyses peuvent être faites aux autres âges. Les résultats de l'ajustement de l'espérance de vie pour les trois meilleurs modèle en terme d'ajustement sont représentés par les figures 20, 21 et 22. On peut voir que la tendance estimée est très proche ou parallèle à la tendance réelle, ce qui traduit une bonne estimation du risque de tendance par le modèle.

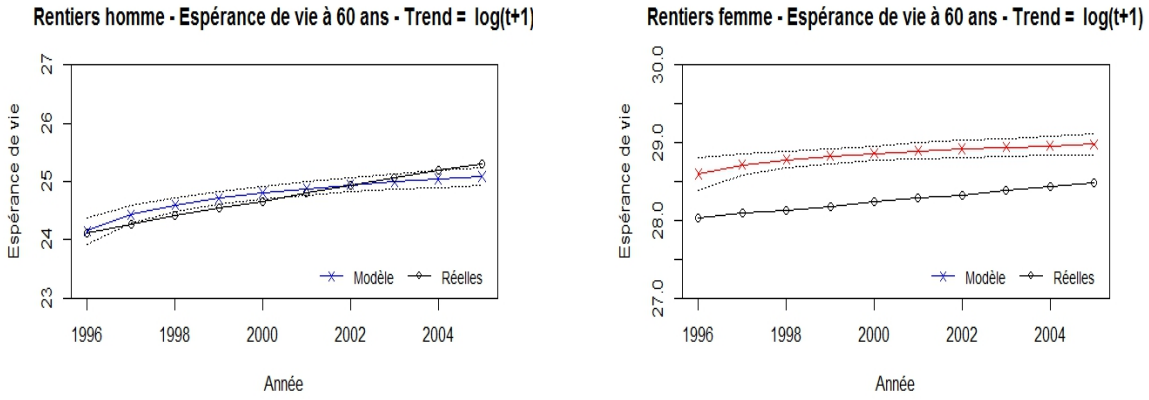

FIGURE 20 – Espérances de vies ajustées par une tendance logarithmique

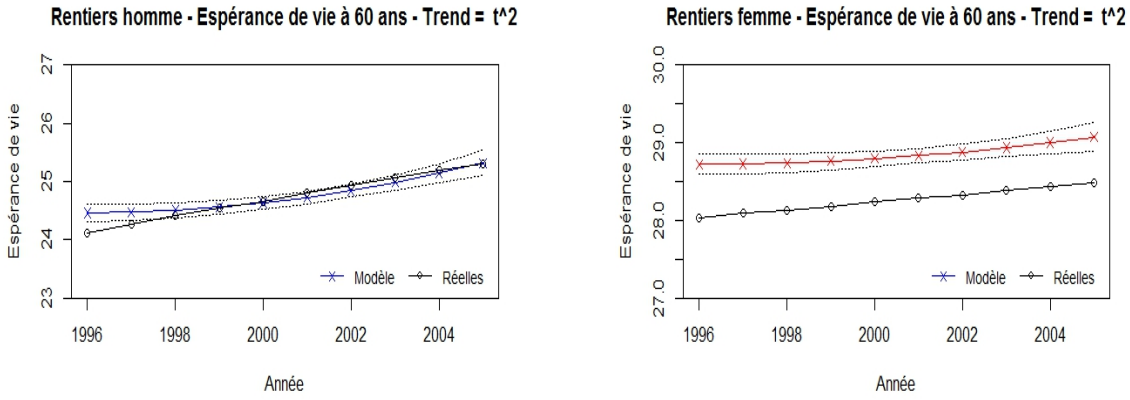

FIGURE 21 – Espérances de vies ajustées par une tendance quadratique

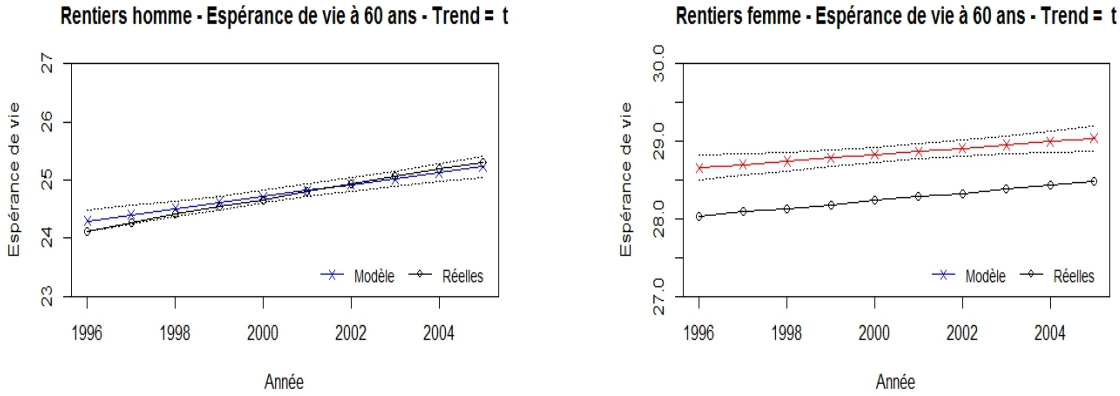

FIGURE 22 – Espérances de vies ajustées par une tendance linéaire

Les écarts sont plus marqués chez les femmes que chez les hommes même s'ils restent relativement faibles, de l'ordre de 0,2 année pour les hommes et 0,6 année pour les femmes. Les écarts quadratiques d'ajustement annualisés sont présentés par le tableau 17.

Choix du meilleur modèle

Grâce aux données d'ajustement de l'AIC, le BIC et l'écart quadratique d'ajustement annualisé, nous pouvons conclure que le modèle linéaire est le meilleur ajustement pour ces deux populations. Cela confirme bien nos attentes. Les espérances de vie ajustées grâce au modèle linéaire sont sensiblement égales à celles projetées voir (tableaux 32 et 33). La figure 23 présente la projection des espérances de vie à l'année 2015 par la tendance linéaire.

Modèle	Hommes			Femmes		
	AIC	BIC	Écarts	AIC	BIC	Écarts
Tangente hyperbolique	4651670	4651690	0,08947672	3494599	3494619	0,3572199
Exponentiel inverse	4651661	4651681	0,07366774	3494597	3494617	0,3548081
Inverse	4651650	4651670	0,05460477	3494597	3494614	0,3517714
Croissance cyclique	4651637	4651657	0,03203398	3494591	3494611	0,3476654
Logarithmique	4651630	4651651	0,02054169	3494590	3494610	0,3463284
Quadratique	4651629	4651650	0,01992241	3494590	3494610	0,3455982
Linéaire	4651623	4651644	0,00865467	3494589	3494608	0,3441683

TABLE 17 – Écarts quadratiques annualisés, AIC et BIC

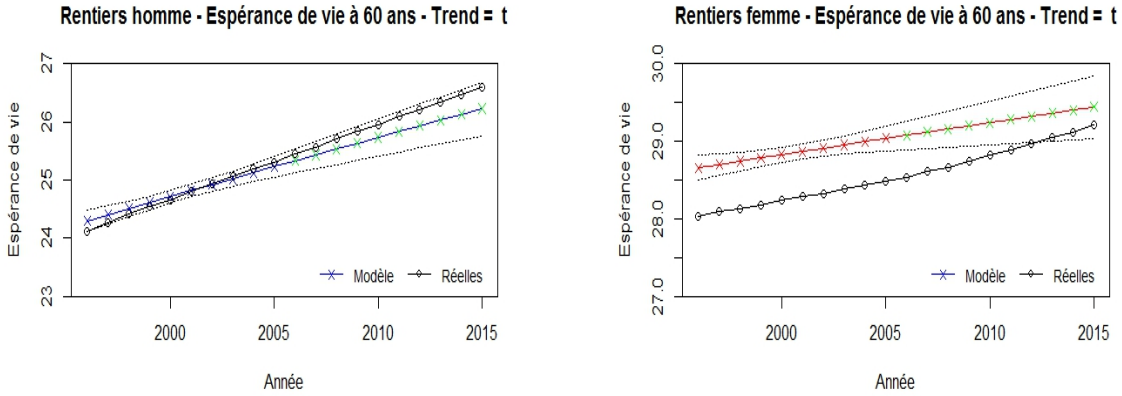

FIGURE 23 – Espérances de vies ajustées et projetées par une tendance linéaire

Nous n'avons pas d'explication rigoureuse des écarts constatés chez les femmes. Mais une piste d'explication pourrait être liée aux données ainsi que leur ajustement après 2005. Le choix d'un *baseline* qui ne dépend que de l'âge pourrait aussi être une piste d'explication.

1.2.5 Conclusion

L'objectif de cette première partie était de vérifier que le modèle proposé nous permettait de retrouver un résultat a priori connu d'avance. Nous nous attendions à retrouver que la tendance linéaire était le meilleur ajustement pour la population des rentiers. Les résultats confirment nos attentes. Cela démontre ainsi la capacité du modèle à détecter la tendance spécifique à une sous-population. Toutefois, la faible profondeur d'historique ne nous permettait pas d'obtenir des résultats rigoureux sur le niveau des espérances de vie.

Une démarche similaire peut être appliquée pour l'étude du *basis risk* spécifique à un

portefeuille d'assurance, en utilisant cette fois-ci la population du portefeuille à la place de la population de rentiers simulée.

Par cet exemple, nous avons montré la validité de cette démarche proposée. Nous avons montré que le modèle permet de détecter le *trend* dans le cas où les populations posséderaient le même *trend*. Dans la prochaine partie, nous montrerons cette fois-ci que le modèle peut détecter les *trends* lorsque les populations ont des *trends* différents. Nous appliquerons la méthode à l'étude du *basis risk* géographique entre l'Allemagne de l'Est et l'Allemagne de l'Ouest.

1.3 Estimation du risque de tendance : cas du *basis risk* géographique en Allemagne

La partition de l'Allemagne en 1949 a contribué à la divergence des espérances de vie dès 1979 au sein de l'Allemagne. D'un côté l'Allemagne de l'Ouest a vu son espérance de vie augmenter fortement tandis que de l'autre, l'Allemagne de l'Est a connu de plus faibles améliorations. Cependant, après la réunification, la tendance s'est inversée. L'Allemagne de l'Est rattrape son retard tandis que l'Allemagne de l'Ouest connaît un essoufflement. Les écarts se réduisent progressivement passant de 3,4 années pour les hommes et 2,8 années pour les femmes en 1990 à 1,3 années pour les hommes et 0,17 année pour les femmes en 2013 (Pechholdová et al., 2017). Assiste-t-on à une convergence des espérances vie ou plutôt à un retournement de situation ? Quels sont les *trends* de ces deux populations ? Quelles peuvent être les anticipations futures ?

Dans les sections suivantes, nous donnerons des éléments de réponse à ces questions en déterminant le meilleur ajustement de tendance.

1.3.1 Les données

Les données utilisées sont celles fournies par la Human Mortality Database sur la population allemande. On y trouve :

- Les données de mortalité sur la population Allemagne de l'Est entre 1956 et 2015.
- Les données de mortalité sur la population Allemagne de l'Ouest entre 1946 et 2015.
- Les données de mortalité sur la population Allemagne réunifiée entre 1990 et 2015.

1.3.2 Simulation des populations

Les populations étant disjointes, nous avons adopté la démarche de simulation disjointe présentée précédemment pour la simulation des populations de l'Allemagne de l'Est et de l'Ouest. L'algorithme de simulation est le même que celui utilisé lors de la

simulation de la population nationale dans l'exemple précédent.

Toutefois, contrairement à l'application précédente, nous allons dans cette partie projeter les deux tendances : celle de la population de l'Allemagne de l'Est et celle de la population de l'Allemagne de l'Ouest. Le choix du *baseline* est donc très important. Le modèle s'écrit alors :

$$h(x,t|Z) = h_0(x)\exp(\beta_E(t)Est + \beta_O(t)Ouest) \tag{1.9}$$

avec : $\beta_E(t) = \beta_E^{le} + \beta_E^{tr} f_e(t)$ et $\beta_O(t) = \beta_O^{le} + \beta_O^{tr} f_o(t)$; $Est \in \{0,1\}$; $Ouest \in \{0,1\}$

Rappelons que le *baseline* h_0 représente le risque instantané de décès lorsque toutes les covariables sont égales à 0. Intuitivement, h_0 représente le risque de décès instantané d'un individu n'appartient pas spécifiquement à une des deux Allemagne, mais à la population nationale, à l'Allemagne réunifiée par exemple. Cependant, les données sur l'Allemagne réunifiée n'étant disponibles qu'à partir de 1990, nous avons choisi d'utiliser les données moyennes des deux Allemagne pour la constitution du *baseline*. Cette démarche consiste donc à simuler trois populations :

— Une population de l'Allemagne de l'Ouest avec pour covariables $Z = (Est, Ouest) = (0,1)$.

— Une population de l'Allemagne de l'Est avec pour covariables $Z = (Est, Ouest) = (1,0)$.

— Une population de l'Allemagne, constituée par l'union des deux populations simulées avec pour covariables $Z = (Est, Ouest) = (0,0)$.

Comme dans l'exemple précédent, nous avons choisi de modéliser la mortalité aux grands âges. Nous utiliserons les données des tables entre les âges 60 ans à 95 ans et les années 1975 à 2015. Nous ajusterons par la suite la tendance entre 1995 à 2015 afin de parvenir à des projections plus cohérentes avec les niveaux actuels.

1.3.3 Implémentation du modèle

Nous avons choisi de tester les 7 scénarios de tendance introduits précédemment :

— **Linéaire** (Lin)
— **Quadratique** (Quad)
— **Logarithmique** (Log)
— **Inverse** (Inv)
— **Exponentiel inverse** (Expo Inv)
— **Tangente hyperbolique** (Tang Hyp)

— **Croissance cyclique** (Croiss Cycl)

Pour chacune des covariables, nous avons testé ces 07 scénarios soit un total de 49 modèles implémentés.

Tests d'adéquation

Les AIC des différents modèles après ajustement sont donnés par les tableaux 18 et 19

AIC	*hommes*	**Ouest**						
		Log	Quad	Lin	Inv	Expo Inv	Tang Hyp	Croiss Cycl
Est	Log	10805766	10805066	10804908	10807761	10808614	10808777	10807305
	Quad	10803971	10803270	10803112	10805970	10806824	10806988	10805513
	Lin	10803890	10803189	10803031	10805890	10806744	10806908	10805432
	Inverse	10808933	10808235	10808078	10810921	10811771	10811934	10810466
	Expo Inv	10810039	10809342	10809185	10812025	10812873	10813036	10811570
	Tang Hyp	10810201	10809504	10809347	10812186	10813034	10813197	10811731
	Croiss Cycl	10807632	10806933	10806776	10809622	10810473	10810636	10809167

TABLE 18 – AIC de l'ajustement sur portefeuille homme
En rouge le plus faible AIC, en vert le deuxième plus faible AIC et en bleu le troisième plus faible AIC

AIC	*femmes*	**Ouest**						
		Log	Quad	Lin	Inv	Expo Inv	Tang Hyp	Croiss Cycl
Est	Log	9623701	9623783	9623325	9625516	9626424	9626625	9625575
	Quad	9622426	9622509	9622050	9624242	9625152	9625352	9624302
	Lin	9621802	9621885	9621426	9623620	9624530	9624731	9623679
	Inverse	9627876	9627957	9627501	9629685	9630590	9630790	9629743
	Expo Inv	9629491	9629572	9629117	9631298	9632202	9632401	9631356
	Tang Hyp	9629761	9629842	9629387	9631567	9632471	9632670	9631625
	Croiss Cycl	9626734	9626815	9626359	9628544	9629450	9629650	9628602

TABLE 19 – AIC de l'ajustement sur portefeuille femmes
En rouge le plus faible AIC, en vert le deuxième plus faible AIC et en bleu le troisième plus faible AIC

Les trois meilleurs modèles du point de vue de l'AIC sont :

- Pour le portefeuille homme :
  - Modèle A : $\begin{cases} \beta_E(t) = 0,42 - 0,018 \times t \\ \beta_O(t) = 0,26 - 0,014 \times t \end{cases}$
  - Modèle B : $\begin{cases} \beta_E(t) = 0,29 - 0,00045 \times t^2 \\ \beta_O(t) = 0,26 - 0,014 \times t \end{cases}$

— Modèle C : $\begin{cases} \beta_E(t) = 0,42 - 0,018 \times t \\ \beta_O(t) = 0,15 - 0,00036 \times t^2 \end{cases}$

- Pour le portefeuille femme :
  - Modèle A : $\begin{cases} \beta_E(t) = 0,51 - 0,022 \times t \\ \beta_O(t) = 0,24 - 0,015 \times t \end{cases}$
  - Modèle B : $\begin{cases} \beta_E(t) = 0,51 - 0,022 \times t \\ \beta_O(t) = 0,13 - 0,00035 \times t^2 \end{cases}$
  - Modèle C : $\begin{cases} \beta_E(t) = 0,51 - 0,022 \times t \\ \beta_O(t) = 0,47 - 0,19 \times \log(t+1) \end{cases}$

Les résultats de l'ajustement nous permettent de valider deux scénarios de tendance pour la population d'Allemagne de l'Est à savoir la tendance linéaire et la tendance quadratique. Cela est tout à fait cohérent avec les observations réelles sur l'Allemagne de l'Est qui a connu de fortes améliorations d'espérances de vie. Pour l'Allemagne de l'Ouest même si elle aussi a connu de forte évolutions, on assiste à un essoufflement. Nous pouvons valider les trois scénarios de tendance linéaire, quadratique et logarithmique. Une modélisation fine aux années avancées nous permettrait d'affiner les résultats.

Tests statistique

Nous avons aussi testé la significativité des paramètres estimés.

$$H_0 : \beta(t) = (\beta_E, \beta_O) = 0$$

Les p-value du test sont très proches 0. On rejette donc l'hypothèse H_0. Les coefficients estimés sont donc significatifs. Les résultats pour le modèle A sont présentés en annexe.

1.3.4 Analyse des résultats et Validations des hypothèses

Nous présenterons les résultats pour les trois meilleures tendances déterminées précédemment.

Ajustement des espérances de vie

Les espérances de vie à 60 ans ajustées par le modèle sont présentées par en annexe.

Choix du meilleur modèle

Nous avons ensuite calculé l'erreur d'ajustement pour chacun de ces modèles. Les résultats obtenus sont présentés dans les tableaux 20 et 21.

Du point de vue de l'écart quadratique annualisé, le Modèle A est le meilleur pour les hommes de la population d'Allemagne de l'Ouest. La tendance linéaire est donc le meilleur ajustement pour cette population. De même pour les femmes, le Modèle A présente les écarts les plus faibles. Le *trend* linéaire alors la meilleure tendance.

Hommes	AIC	BIC	Ecart Est	Ecart Ouest
Modèle A	10803031	10803075	0,3102364	0,2625852
Modèle B	10803112	10803156	0,2647039	0,2627079
Modèle C	10803189	10803233	0,3103481	0,3890151

TABLE 20 – Écarts de quadratique d'ajustement anualisé pour les hommes

Femme	AIC	BIC	Ecart Est	Ecart Ouest
Modèle A	9621426	9621470	0,3041358	0,1479375
Modèle B	9621885	9621929	0,3044019	0,4101922
Modèle C	9621802	9621846	0,3042726	0,4281576

TABLE 21 – Écarts de quadratique d'ajustement anualisé pour les femmes

En conclusion, pour la population de l'Allemagne de l'Ouest, il apparaît donc que la tendance linéaire est celle qui ajuste le mieux les données (Figure 24).

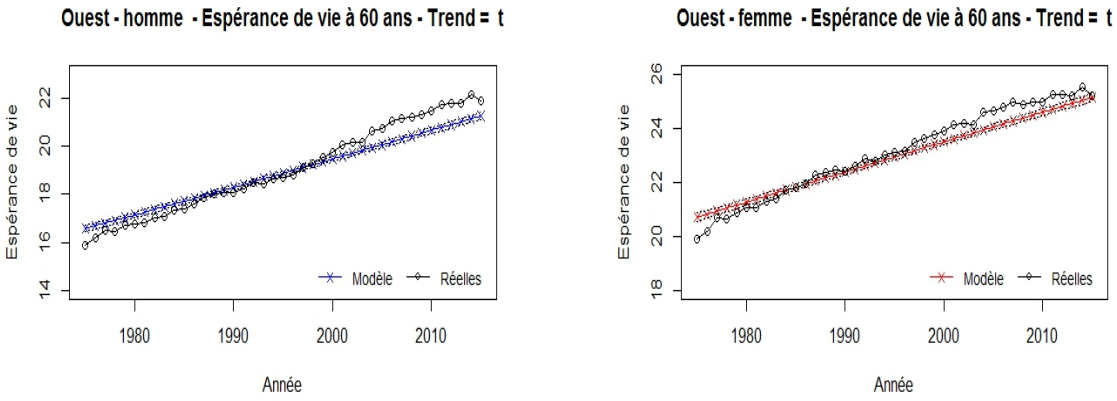

FIGURE 24 – Espérances de vies ajustées l'Allemagne de l'Ouest

On peut voir sur la figure 24 que la tendance est globalement bien ajustée sur l'ensemble de l'historique. On constate cependant une déviation aux années récentes. Un meilleur ajustement peut être réalisé aux années plus récentes.

Quant à la population de l'Allemagne de l'Est, on peut voir pour les hommes que le Modèle B est le meilleur modèle du point de vue de l'écart quadratique d'ajustement annualisé, contrairement aux résultats de l'AIC. La tendance quadratique conduit à des écarts plus faibles que la tendance linéaire mais à un AIC plus élevé. Par ailleurs, la différence d'AIC entre le Modèle B et le Modèle A est de 81 sur sur plus de 10 000 000 tandis que les écarts d'ajustement, près de 0,05, sont plus significatifs pour ces deux

modèles. Il apparaît donc que la tendance quadratique est la meilleure tendance pour la population des hommes de l'Allemagne de l'Est. Par contre, pour les femmes, les résultats montrent que la tendance linéaire est le meilleur ajustement (Figure 25).

	Allemagne de l'Est	Allemagne de l'Ouest
Hommes	Quadratique	Linéaire
Femmes	Linéaire	Linéaire

TABLE 22 – Tendance des populations

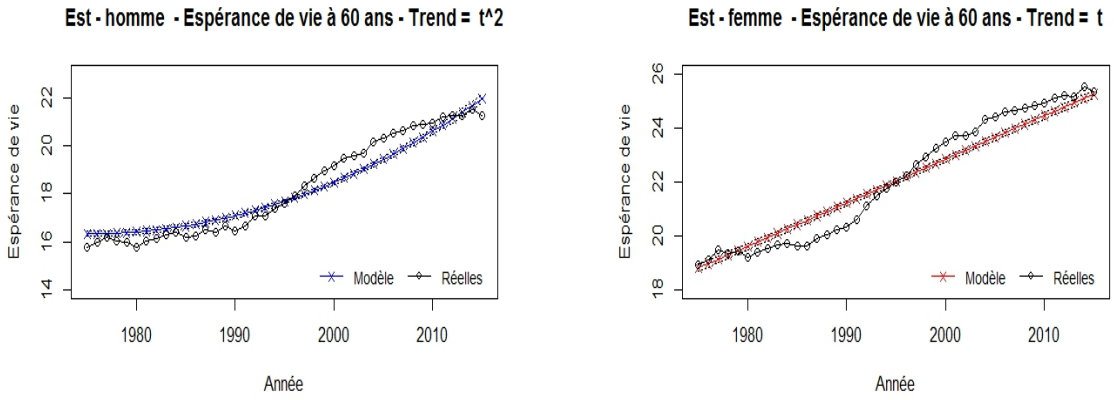

FIGURE 25 – Espérances de vies ajustées l'Allemagne de l'Est

La figure 25 montre que les espérances de vie sont globalement bien ajustées sur l'ensemble de l'historique. Cependant, chez les hommes, il semble y avoir un changement de tendance entre 1990 et 1995. Une modélisation plus fine peut être menée pour corriger la tendance.

Par ailleurs, la tendance des espérances de vie chez les femmes semble être de forme croissante cyclique. Des études supplémentaires peuvent être menées afin de déterminer si une fonction cyclique pourrait être adaptée, et définir la période associée.

La figure 26 permet de comparer les espérances de vie ajustées pour l'Allemagne de l'Est et de l'Ouest. Les résultats de l'ajustement montrent que l'espérance de vie de l'Allemagne de l'Est pourrait dépasser celle de l'Allemagne de l'Ouest. Par ailleurs, nous pouvons observer un ralentissement du *trend* d'espérance de vie de l'Allemagne de l'Ouest. Dans la section suivante, nous corrigerons les tendances ajustées pour les années 1995 à 2015.

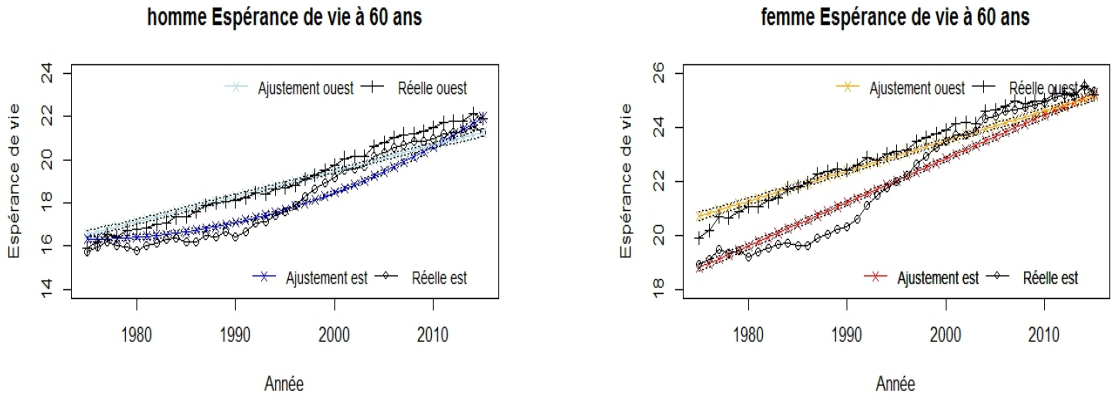

FIGURE 26 – Comparaison des espérances de vie

1.3.5 Correction des tendances

Nous avons corrigé les tendances précédemment ajustées pour les années 1995 à 2015.

L'AIC ajusté pour les trois meilleurs modèles est présenté dans les tableaux 23 et 24.

AIC	*hommes*	**Ouest**		
		Logarithmique	Quadratique	Linéaire
Est	Logarithmique	4974961	4975003	4974905
	Quadratique	4975103	4975145	4975047
	Linéaire	4974959	4975001	4974903

TABLE 23 – AIC des modèles ajustés entre 1995 et 2015 (hommes)
En rouge le plus faible AIC, en vert le deuxième plus faible AIC et en bleu le troisième plus faible AIC

AIC	*femmes*	**Ouest**		
		Logarithmique	Quadratique	Linéaire
Est	Logarithmique	4304528	4304571	4304517
	Quadratique	4304645	4304687	4304633
	Linéaire	4304545	4304588	4304534

TABLE 24 – AIC des modèles ajustés entre 1995 et 2015 (femmes)
En rouge le plus faible AIC, en vert le deuxième plus faible AIC et en bleu le troisième plus faible AIC

L'ajustement de l'AIC montre que les tendances linéaire et logarithmique sont les

meilleurs ajustements pour ces deux populations. Les écarts des AIC des deux *trends* sont relativement faibles.

Après ajustement des espérances de vie, nous avons calculé les écarts quadratiques annualisés. Les résultats sont présentés dans les tableaux 25 et 26.

hommes	**Est**		**Ouest**	
	Linéaire	Logarithmique	Linéaire	Logarithmique
Linéaire	0,16006359	0,16008783	0,09999323	0,17327629
Logarithmique	0,15551358	0,15551397	0,09999124	0,17327425

TABLE 25 – Écarts quadratiques annualisés (Hommes)

femmes	**Est**		**Ouest**	
	Linéaire	Logarithmique	Linéaire	Logarithmique
Linéaire	0,15353743	0,15353979	0,05279014	0,06605085
Logarithmique	0,14561749	0,14561984	0,05278709	0,06605085

TABLE 26 – Écarts quadratiques annualisés (Femmes)

Nous pouvons donc conclure que le *trend* logarithmique est le meilleur ajustement pour la population d'Allemagne de l'Est tandis que le *trend* linéaire demeure toujours le meilleur ajustement pour l'Allemagne de l'Ouest entre 1995 et 2015. Les figures 27 et 28 nous présentent les résultats de l'ajustement des modèles ainsi qu'une projection sur 20 années. Nous pouvons y voir que les ajustements sont encore plus proches des données empiriques.

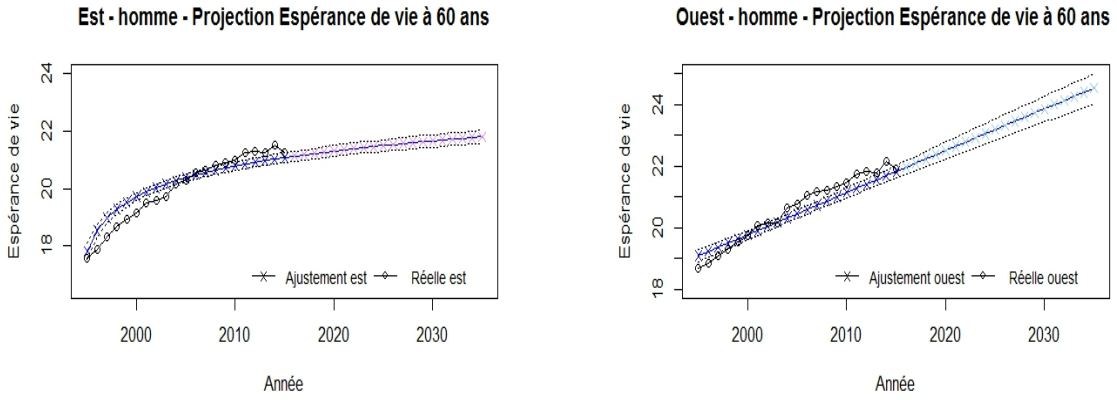

FIGURE 27 – Projection des espérances de vie (Hommes)

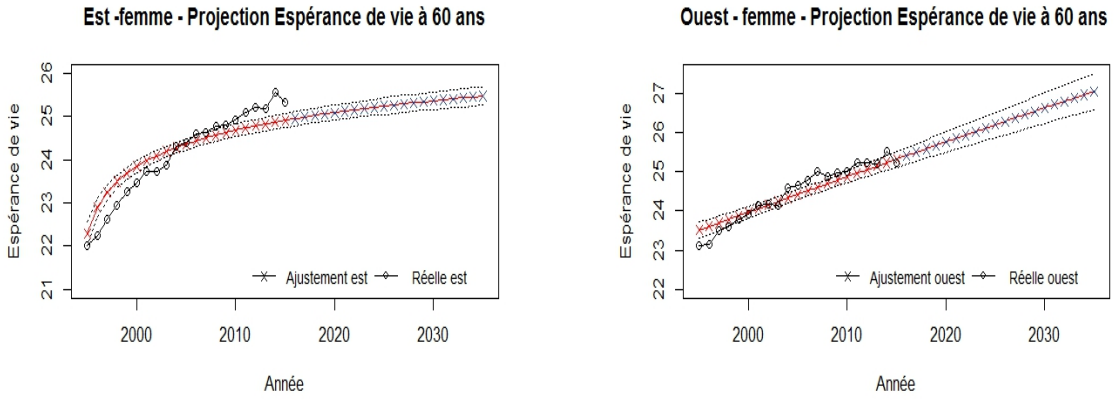

FIGURE 28 – Projection des espérances de vie (Femmes)

Les statistique du test

$$H_0 : \beta(t) = (\beta_E, \beta_O) = 0$$

1.3.6 Conclusion

Nous avons pu identifier grâce à cette extension du modèle de Cox, les différents *trends* d'espérance de vie au sein de la population allemande, sans subdivision de celle-ci. Le modèle nous a permis, en s'appuyant sur la mortalité de la population moyenne en Allemande, de retrouver les évolutions différentes de la mortalité de ses sous-populations. Après avoir ajusté des tendances entre 1965 et 2015, nous avons ensuite corrigé celles-ci entre 1995 et 2015.

Les résultats permettent de confirmer la validité de la méthode proposée. En plus de pouvoir détecter le *trend* lorsque les sous-populations ont le même *trend*, le modèle proposé permet de détecter les *trends* lorsque les sous-populations ont des *trends* différent.

Nous ne présenterons pas dans cette étude un impact économique sous-jacent à une prise en compte du *basis risk* de tendance ajustée par le modèle. L'objectif était de montrer que la méthode développée permet de détecter et de corriger les tendances des populations hétérogènes.

L'avantage de l'utilisation de ce modèle est qu'il permet de détecter puis de projeter différentes tendances sans subdivision de la base de données. Dans notre cas, nous avons constitué une population hétérogène grâce aux données de mortalité sur la population allemande, puis détecté les tendances sans modélisation fine au niveau des sous-groupes. La flexibilité du modèle permet de tester différents scénarios d'évolution. En effet, la forme de la fonction de tendance imposée en entrée du modèle impacte aussi celle présente dans ses sorties. Cela est un réel avantage lorsque l'on veut ajuster une tendance estimée. Par

exemple, la tendance logarithmique présentée ici peut encore être affinée. On pourrait encore tester des tendances de la même forme telles que les fonctions $t^{\frac{1}{2}}$, $t^{\frac{3}{5}}$, $t^{\frac{7}{10}}$ (Figure 29).

Des études similaires peuvent être menées par l'assureur sur son portefeuille de rentier pour détecter les différents *trends* au sein de son portefeuille.

L'assureur n'est donc pas conditionné dans cette approche par une estimation stochastique des tendances parfois « non maîtrisée » des évolution futures mais peut décider de modéliser différents scénarios d'évolution. Un exemple typique dans notre cas aurait pu être celui de tester un scénario d'évolution de la forme $f(t) = t^2\mathbb{I}_{\{1975\leq t<1995\}} + t^{\frac{7}{10}}\mathbb{I}_{\{1995\leq t<2005\}} + \log(t)\mathbb{I}_{\{2005\leq t<2015\}}$.

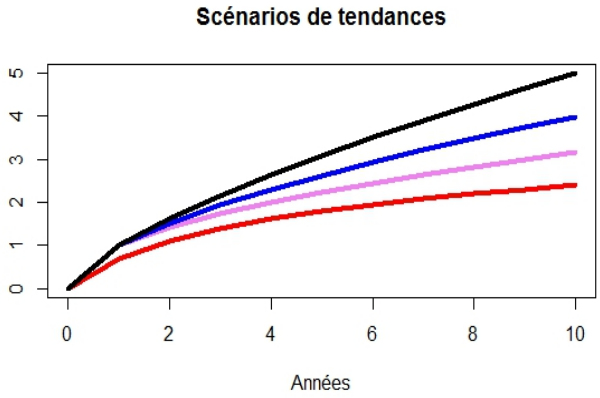

FIGURE 29 – Scénarios de tendance
En rouge $f(t) = log(t)$, *en violet* $f(t) = \sqrt{t}$, *en bleu* $f(t) = t^{\frac{3}{5}}$ *et en noir* $f(t) = t^{\frac{7}{10}}$

Dans le chapitre suivant nous présenterons des études de sensibilité du modèle.

Chapitre 2

Sensibilités, limites et perspectives

> « ***Sa barbe n'avait pas encore poussé. Ce retard annonce ... une grande longévité*** » *(Balzac, Béatrix, p. 57)*

Dans ce chapitre, nous étudierons dans une première partie la sensibilité de l'estimation du modèle aux données simulées, notamment à la taille des sous-populations. Nous nous intéresserons aussi à la sensibilité de l'estimation des paramètres aux tranches d'âges utilisées. Dans une deuxième partie, nous présenterons les limites de la méthodologie proposée ainsi que les pistes d'amélioration en troisième partie.

2.1 Études de sensibilité

L'étude des sensibilités présentée ici sera appliquée dans une première partie au portefeuille des femmes de la population des rentiers et de la population nationale. Les mêmes études peuvent être réalisées sur la population de l'Allemagne.

2.1.1 Sensibilité de l'ajustement à la taille des sous-populations

Dans le premier chapitre de cette partie, nous avons estimé la tendance de la population des rentiers en simulant une population nationale au sein de laquelle est incluse une population de rentiers. Nous avions supposé précédemment qu'un tiers de la nationale était rentiers. Selon les statistiques de l'INSEE, 36,5% de la population en France détiennent un contrat d'assurance vie [1]. On pourrait penser que la proportion de rentiers en France est beaucoup moins élevée. Il serait intéressant de tester la sensibilité des estimations précédentes lorsque la taille de la population étudiée diminue.

1. http://www.cbanque.com/actu/55552/epargne-qui-possede-une-assurance-vie-en-2015

Nous avions constaté précédemment des écarts plus marqués chez des femmes de la population des rentiers. La démarche adoptée maintenant est de reprendre l'implémentation en changeant la taille de la population assurée. Nous ferons des hypothèses de simulation de population assurée avec une proportion d'un quart et d'un cinquième, puis nous testerons la sensibilité des paramètres et de l'espérance de vie ajustée par le *trend* linéaire.

Les résultats de l'estimation des paramètres ainsi que des erreurs quadratiques d'ajustements anualisés sont présentés dans le tableau 27.

	Coefficients estimés
Proportion $=\frac{1}{3}$	$\beta(t) = -0,43 - 0,00730 \times t$
Proportion $=\frac{1}{4}$	$\beta(t) = -0,39 - 0,00719 \times t$
Proportion $=\frac{1}{5}$	$\beta(t) = -0,36 - 0,00739 \times t$

TABLE 27 – Ajustement du modèle (Proportion)

On peut remarquer que la taille de la sous-population étudiée à un impact sur les coefficients estimés. L'impact est plus marqué sur le niveau que sur la tendance. En effet, le paramètre de niveau a changé de 10% lorsque l'on est passé d'une population de rentiers de proportion un tiers à une population de rentiers de proportion un quart par rapport à la population nationale. La tendance quant à elle a changé de 1,5%. De même, on observe un changement de 8% sur le niveau et de 2,8% sur le *trend* entre les proportions un quart et un cinquième. Ainsi, entre les proportions un tiers et un cinquième, le paramètre de niveau a évolué de 20% tandis que le paramètre tendance a évolué de 1,2%.

Le paramètre de tendance est donc peu sensible à la taille des populations simulées. Ainsi, même lorsque la taille de la sous-population est réduite, le risque de tendance est globalement bien estimé par le modèle (Figure 48 en annexe).

2.1.2 Sensibilité aux tranches d'âges

Dans cette partie, nous étudierons la sensibilité des paramètres estimés aux tranches d'âges utilisées pour la calibration. Précédemment, nous avions utilisé les tranches d'âges 40 ans à 95. Nous présenterons ici les résultats de la calibration pour les âges 50 ans à 95 ans et 60 ans à 95 ans, lorsque la population de rentiers à une proportion d'un tiers par rapport à la population nationale nationale.

Les paramètres estimés sont présentés dans le tableau 28.

	Coefficients estimés
Âges : 40 ans - 95 ans	$\beta(t) = -0.4399 - 0,00730 \times t$
Âges : 50 ans - 95 ans	$\beta(t) = -0,4385 - 0,00731 \times t$
Âges : 60 ans - 95 ans	$\beta(t) = -0,4391 - 0,00740 \times t$

TABLE 28 – Ajustement du modèle (tranches d'âges)

Les coefficients de régression calibrés varient très peu en fonction des tranches d'âges. Les paramètres de niveau varient de moins de 0,3%. Nous pouvons donc conclure que le modèle est peu sensible aux différentes tranches d'âges de calibrage sur le niveau. Par contre, les paramètres de tendance connaissent une plus forte variation de 1,3% entre les calibrations pour les âges 40 ans à 95 ans et 60 ans à 95 ans. Le *trend* est donc plus sensible à la bande d'âge de calibration utilisée. Toutefois cela induit très peu de changement sur la tendance des espérances de vie (Figure 49).

2.2 Limites

Nous avons proposé dans cette étude une démarche d'estimation du risque de tendance au sein d'une population hétérogène grâce à une extension du modèle de Cox. Par deux applications, nous avons montré que le modèle peut détecter et corriger les différentes tendances au sein d'une population hétérogène. La flexibilité du modèle permet d'ajuster ces tendances et de confronter plusieurs scénarios d'évolution. Ces coefficients nous ont ainsi permis d'estimer les effets des différents sous-groupes et surtout leurs évolutions dans le temps. Nous avons pu constater que la forme des tendances imposées en entrée du modèle impacte celles des espérances de vie ajustées. Le choix de la tendance adéquate est donc déterminant. Cela peut nécessiter de tester plusieurs scénarios d'ajustement et plusieurs combinaisons de formes pour parvenir au meilleur choix. Cela peut être long à réaliser mais intéressant à analyser.

Nous avons montré précédemment que la démarche d'estimation proposée était sensible à la taille des populations surtout sur le niveau, tandis que la tranche d'âges de calibrage choisie a un impact plus marqué mais faible sur le paramètre de *trend*. Ainsi, la composition du portefeuille peut avoir un impact non négligeable sur le niveau des espérances de vie ajustées et un impact moins marqué sur le *trend*. Nous avons testé ici la sensibilité dans le cas de la projection d'un seul trend. Cependant, pour la projection

de plusieurs *trends*, les déviations peuvent être plus marquées, car il y aurait plusieurs sources d'erreurs.

La démarche de simulation des populations présentée peut aussi être améliorée. Une simulation rigoureuse devrait prendre en compte les expositions à chaque âge et à chaque temps de façon à rester cohérent avec les pyramides des âges.

2.3 Perspectives

Quatre axes d'améliorations ont ainsi pu être identifiés :

- **Le *baseline***

  Nous avons tout au long de notre étude supposé un *baseline* constant au fil du temps. Cependant, la réalité est que les populations évoluent avec des structures différentes de mortalité par âge. Considérer un *baseline* constant c'est donc supposer une stabilité de la structure de mortalité au fil du temps. Une piste d'amélioration serait d'estimer un *baseline* qui pourrait prendre une structure différente au fil du temps. Cela permettrait d'estimer des niveaux de mortalité cohérents avec les données de mortalité de l'année. De plus, un *baseline* pouvant prendre une structure différente dans le temps permettrait un meilleur ajustement de la pente des *trends* d'espérances de vie estimées. Une autre approche serait aussi la stratification du *baseline*. En effet, en présence de variable pouvant prendre plusieurs modalités, il peut être intéressant de définir un *baseline* propre à chaque modalité. Cela assurerait une meilleure estimation des paramètres. Toutefois cela nécessiterait une subdivision la population initiale.

- **Les tendances**

  Grâce aux résultats des ajustements, nous avons pu constater l'impact du choix de la fonction de tendance sur les *trends* d'espérances de vie ajustés. En revanche, cela nécessite au préalable de définir la fonction de tendance à tester. Une piste d'amélioration serait d'introduire une tendance qui pourrait s'ajuster elle-même au fil des années, c'est-à-dire une tendance stochastique. Un exemple peut être l'utilisation des séries temporelles ou des modèles stochastiques de taux. Une autre alternative serait de construire des scénarios de tendances des sous-population qui s'ajusteraient les unes en fonction de l'évolution autres. On pourrait aussi explorer une modélisation des tendances qui dépendrait à la fois du temps et de l'âge. Les champs d'explorations sont très vastes.

- **La prise en compte des effets cachés :**

Nous avons adopté dans notre démarche une vision période pour la projection des *trend* d'espérances de vie. Cela nous a permis de visualiser réellement les améliorations sur les années. Cependant, une piste d'amélioration serait de prendre en compte dans la modélisation de facteurs d'interactions entre les différentes cohortes. La prise en compte des effets cohortes peut permettre de meilleurs ajustements des tendances aux différentes années. De même une autre approche peut être d'estimer des effets d'interaction entre les sous-populations. Toutefois, cela augmenterait le nombre de paramètres du modèle à estimer.

- **La prise en compte de l'hétérogénéité non observée :**

En effet, l'hétérogénéité n'est pas toujours observée au sein d'une population. Elle peut relever d'un facteur inconnu ou mal maîtrisé potentiellement aléatoire. Modéliser les facteurs d'hétérogénéité non observée peut être une piste d'amélioration.

Conclusion

Le but de cette étude était de proposer une approche d'estimation des améliorations futures de mortalité au sein des populations hétérogènes. Après avoir étudié plusieurs des approches proposées dans la littérature, leurs avantages ainsi que leurs inconvénients, nous avons développé une approche pertinente permettant la détection et la correction du *basis risk* au sein des populations hétérogènes en partant du modèle de Cox.

La particularité de cette approche a été l'imposition d'une fonction de tendance représentant un scénario d'évolution future de la mortalité aux coefficients de régression. Grâce, à une bonne décomposition des populations d'étude, nous avons autorisé l'évolution dans le temps des effets des sous-groupes hétérogènes au sein des populations sur leur mortalité.

Nous avons appliqué la démarche à l'étude du *basis risk* entre la population de rentiers et la population nationale en France, puis à l'étude du *basis risk* géographique en Allemagne. Pour cela, nous avons simulé ces populations grâce à leurs données de mortalité par années.

L'avantage de cette démarche a été l'obtention de méthodologie générale, ne dépendant pas de la structure et de la composition d'un portefeuille, réduisant ainsi la volatilité que peut engendrer l'utilisation de portefeuilles réels.

Nous avons testé sept scénarios d'évolution future de la mortalité. Il s'agit des scénarios d'évolution linéaire, quadratique, logarithmique, inverse, exponentiel inverse, tangent hyperbolique et croissant cyclique.

Pour l'étude du *basis risk* entre la population nationale et la population de rentiers, nous avons simulé dans un premier temps une population nationale puis une population de rentiers incluse dans celle-ci. Après application de la méthode, il est ressorti que le meilleur ajustement sur la population des rentiers était celui avec une tendance linéaire. Cela était tout à fait conforme à nos attentes, du fait que les tables de mortalité des rentiers ont été ajustées, lors de leur construction, grâce aux données de mortalité de la population nationale.

L'application sur la population allemande a montré des *trends* différents entre l'Allemagne de l'Est et de l'Ouest. Après avoir simulé les sous-populations au sein de l'Allemagne, nous avons détecté grâce au modèle une tendance linéaire pour la population de l'Allemagne de l'Ouest et une tendance quadratique pour celle de l'Allemagne de l'Est. Une correction des tendances aux années plus récentes a permis de constater un ralentissement du *trend* en l'Allemagne de l'Est par la validation d'un scénario d'évolution logarithmique.

Ainsi, l'application montre que le modèle développé permet d'une part de détecter le *trend* d'espérance de vie lorsque les populations d'étude ont le même *trend* et d'autre part de détecter les *trends* hétérogènes lorsque les populations possèdent des *trends* différents et cela sans subdivision des populations d'étude mais en s'appuyant sur la mortalité d'une population moyenne pour estimer celle des sous-groupes. Les résultats des tests de sensibilité montrent une grande robustesse dans l'estimation du paramètre de tendance, même lorsque la taille de la sous-population étudiée est réduite. La méthode proposée permet donc de détecter et de corriger le risque de tendance du *basis risk* au sein des populations hétérogènes. Elle peut aussi bien s'appliquer dans un cadre général d'étude du *basis risk* entre plusieurs populations, que dans un cadre spécifique d'étude au sein d'un portefeuille d'assurance.

Cette approche nouvelle permettra à l'assureur de modéliser plus finement la mortalité de son portefeuille ainsi que son évolution dans le temps en tenant compte des facteurs d'hétérogénéité. La flexibilité de la méthode lui permettra de tester puis de valider plusieurs scénarios d'évolution future.

L'une des principales limites à l'utilisation de cette méthode réside dans le choix des tendances à tester. Cela peut s'avérer difficile mais toutefois intéressant à étudier. Une piste d'amélioration peut être celle d'utiliser une tendance stochastique. Cependant, des études supplémentaires doivent être menées pour s'assurer de la robustesse et de la cohérence d'une telle démarche.

Les pistes d'amélioration de l'approche présentée restent nombreuses. Toutefois, cette étude peut être un cadre de construction d'une démarche plus élaborée d'estimation du risque de tendance du *basis risk* en tenant compte des facteurs d'hétérogénéités au sein des populations.

Annexes

Intervalles de confiance - Régions de confiance

Pour illustrer la différence entre les intervalles de confiance et les régions de confiance, considérons un vecteur gaussien $\begin{pmatrix} x - \hat{x} \\ y - \hat{y} \end{pmatrix}$ de moyenne 0 et de matrice de variance-covariance $A = \begin{pmatrix} 1 & 0,5 \\ 0,5 & 1 \end{pmatrix}$, avec $\hat{x} = 2$, $\hat{y} = -1$. Supposons que l'on veuille calculer un intervalle de confiance avec niveau de confiance de $1 - \alpha = 99.5\%$ pour x et y. La méthode classique de calcul (en supposant l'indépendance) donnerait :

$$x \in \left[\hat{x} - z_{1-\frac{\alpha}{2}}\sigma_x, \hat{x} + z_{1-\frac{\alpha}{2}}\sigma_x\right]$$

$$y \in \left[\hat{y} - z_{1-\frac{\alpha}{2}}\sigma_y, \hat{y} + z_{1-\frac{\alpha}{2}}\sigma_y\right]$$

Avec : $z_{1-\frac{\alpha}{2}}$ le quantile d'ordre $1 - \frac{\alpha}{2}$ d'une loi normale centrée réduite. On trouve donc :

$$x \in \Big[-0,80 \ , \ 4.80\Big]$$

$$y \in \Big[-3.80 \ , \ 1.80\Big]$$

La région de confiance s'écrirait dans ce cas :

$$\mathcal{A}_1 = \Big\{x, y \ \mid \ -0,80 \leq x \leq 4.80 \ \ \& \ \ 3.80 \leq y \leq 1.80\Big\}$$

$\mathcal{A}_1$ définit la surface d'un rectangle (Voir figure 30).

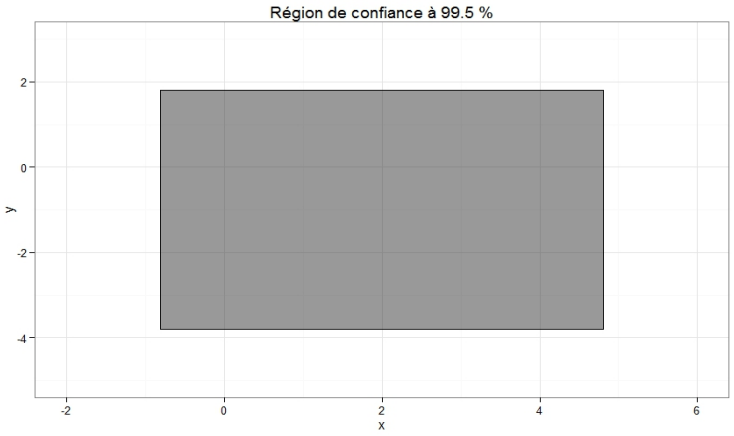

FIGURE 30 – Région de confiance $\mathcal{A}_1$

Une construction rigoureuse consisterait à utiliser l'information fournie par toute la matrice de covariance. En effet, on montre que si $\begin{pmatrix} x-\hat{x} \\ y-\hat{y} \end{pmatrix}$ est un vecteur gaussien de moyenne 0 et de matrice de variance-covariance $A = \begin{pmatrix} 1 & 0,5 \\ 0,5 & 1 \end{pmatrix}$, alors $U^\top A^{-1} U$ avec $U = \begin{pmatrix} x-\hat{x} \\ y-\hat{y} \end{pmatrix}$, suit une loi χ^2_2.

$$\mathbb{P}\Big(z_\alpha \leq \chi^2_2 \leq z_{1-\alpha}\Big) = 1-\alpha$$

On cherche l'ensemble $\mathcal{A}_2$ des x et y tels que : $z_\alpha \leq U^\top A^{-1} U \leq z_{1-\alpha}$

$$\mathcal{A}_2 = \Big\{x, y \;\mid\; z_\alpha \leq (x-\hat{x})^2 + (x-\hat{x})(y-\hat{y}) + (y-\hat{y})^2 \leq z_{1-\alpha}\Big\}$$

La surface décrite par $\mathcal{A}_2$ est celle d'une ellipse (Voir figure 31).

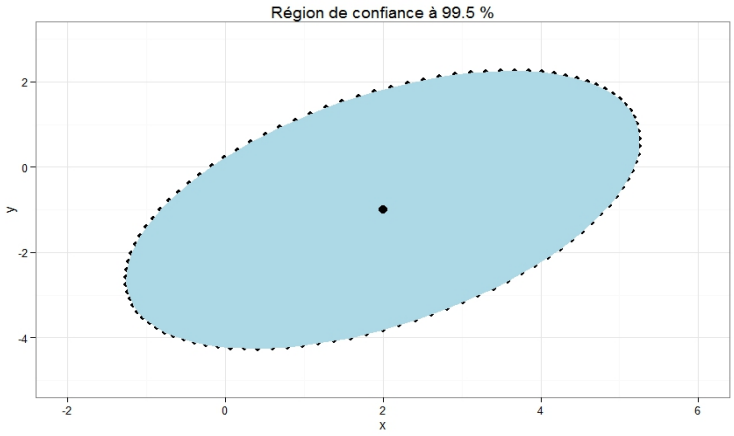

FIGURE 31 – Région de confiance $\mathcal{A}_2$

On peut comparer les régions $\mathcal{A}_1$ et $\mathcal{A}_2$ (Voir figure 32). $\mathcal{A}_1$ couvre en réalité une partie de la région de confiance $\mathcal{A}_2$. On peut constater notamment que les bornes du rectangle n'atteignent pas celles de l'ellipse. De ce fait, une partie de l'information est ignorée lorsqu'on ne prend pas en compte la corrélation. La région $\mathcal{A}_1$ peut donc sous-estimer la taille des intervalles de confiance tandis que $\mathcal{A}_2$ nous donne une information plus juste. En réalité $\mathcal{A}_1$ estime correctement les intervalles de confiance lorsque les paramètres estimés sont à leurs valeurs moyennes. $\mathcal{A}_2$ quant à lui permet de calculer les intervalles de confiance même lorsque les paramètres s'écartent de leurs moyennes.

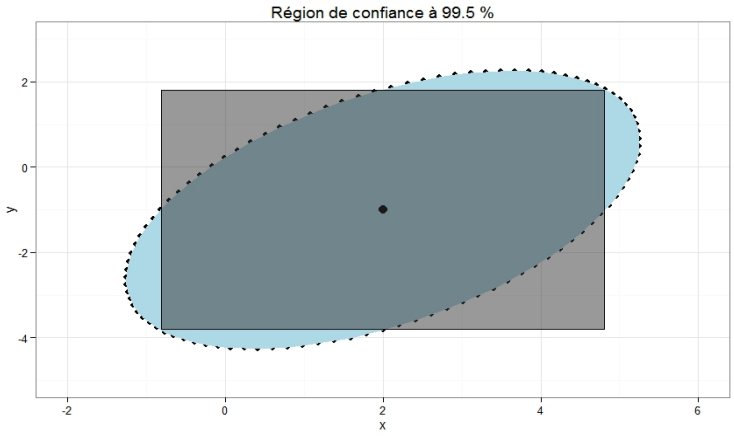

FIGURE 32 – Comparaison de $\mathcal{A}_1$ et $\mathcal{A}_2$

Risque de tendance entre population nationale et population de rentiers

Table TG par période - Femme

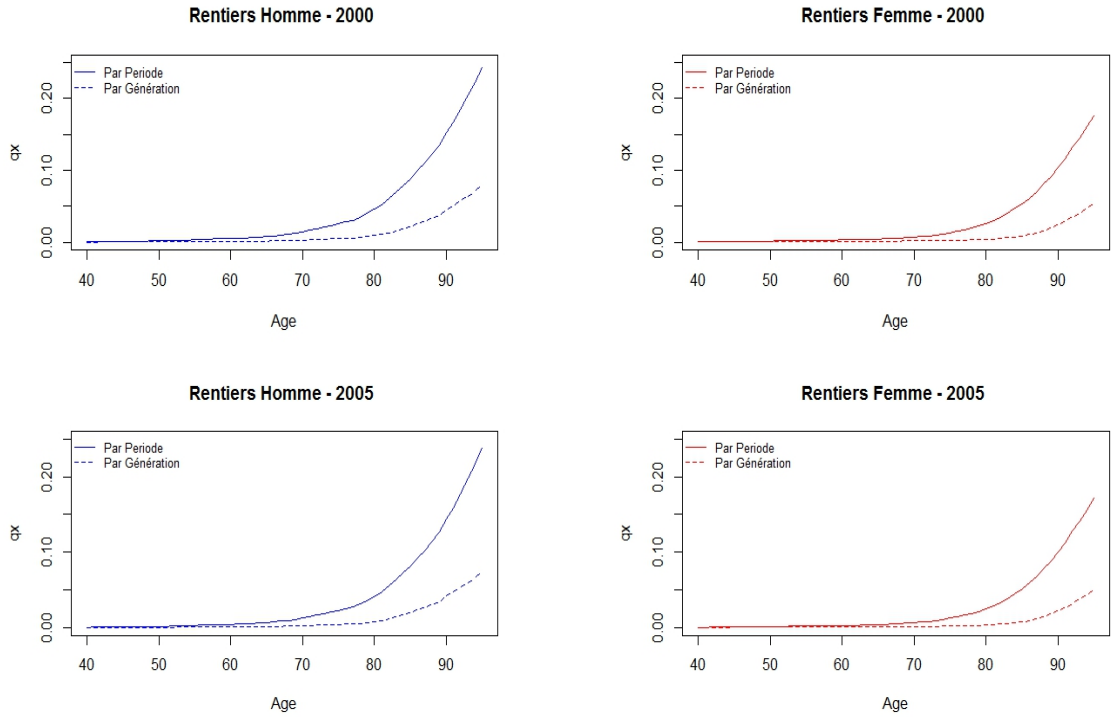

FIGURE 33 – Mortalité des rentiers pour l'année et la génération 2000 et 2005

Simulation des populations

Algorithme de simulation de la population nationale

Algorithm 1 Simulation population nationale

for $t = 1996$ to 2005 **do**
  Créer une matrice pour stocker les individus de l'année t.// *(Une matrice de 5 colonnes : Âge d'entrée, Âge de sortie, Année, Motif de sortie, Population)*
  $x_0 \leftarrow 40$ ans
  Choisir la taille $N_{x_0,t}$ initiale de la population à l'âge x_0 $\left(N_{x_0,t} \leftarrow 10000\right)$
  Créer $N_{x_0,t}$ lignes d'individus (dans la matrice) ayant pour caractéristiques :
  - Âge d'entrée $\leftarrow x_0$
  - Âge de sortie $\leftarrow (x_0 + 1)$
  - Année $\leftarrow t$
  - Motif de sortie $\leftarrow 0$ // *(Censure = 0, Décès = 1)*
  - Population $\leftarrow 0$ //*(National = 0, Rentiers = 1)*

  Calculer le nombre de décès $N_{x_0,t}^{deces}$ à l'âge x_0 ; $\left(N_{x_0,t}^{deces} \leftarrow q_{x_0,t} * N_{x_0,t}\right)$
  Choisir $N_{x_0,t}^{deces}$ lignes parmi les $N_{x_0,t}$ et modifier la caractéristique « Motif de sortie »
  - Motif de sortie $\leftarrow 1$

  for $Age = x_0 + 1$ to x_{max} **do**
    Créer $N_{Age-1,t}^{survie} \leftarrow (N_{Age-1,t} - N_{Age-1,t}^{deces})$ lignes de plus dans le tableau avec pour caractéristiques :
    - Âge d'entrée $\leftarrow Age$ ans
    - Âge de sortie $\leftarrow (Age + 1)$ ans
    - Année $\leftarrow t$
    - Motif de sortie $\leftarrow 0$
    - Population $\leftarrow 0$

    Calculer le nombre de décès $N_{Age,t}^{deces}$ à l'âge Age ; $\left(N_{Age,t}^{deces} \leftarrow q_{Age,t} * N_{Age-1,t}^{survie}\right)$
    Choisir $N_{Age,t}^{deces}$ lignes parmi les $N_{Age-1,t}^{survie}$ nouvellement créées pour l'âge Age et modifier la caractéristique « Motif de sortie »
    - Motif de sortie $\leftarrow 1$

  end for
end for
Regrouper les matrices et constituer la population totale individu par individu à chaque âge

Comparaisons de la mortalité de la population nationale simulée et la mortalité réelle

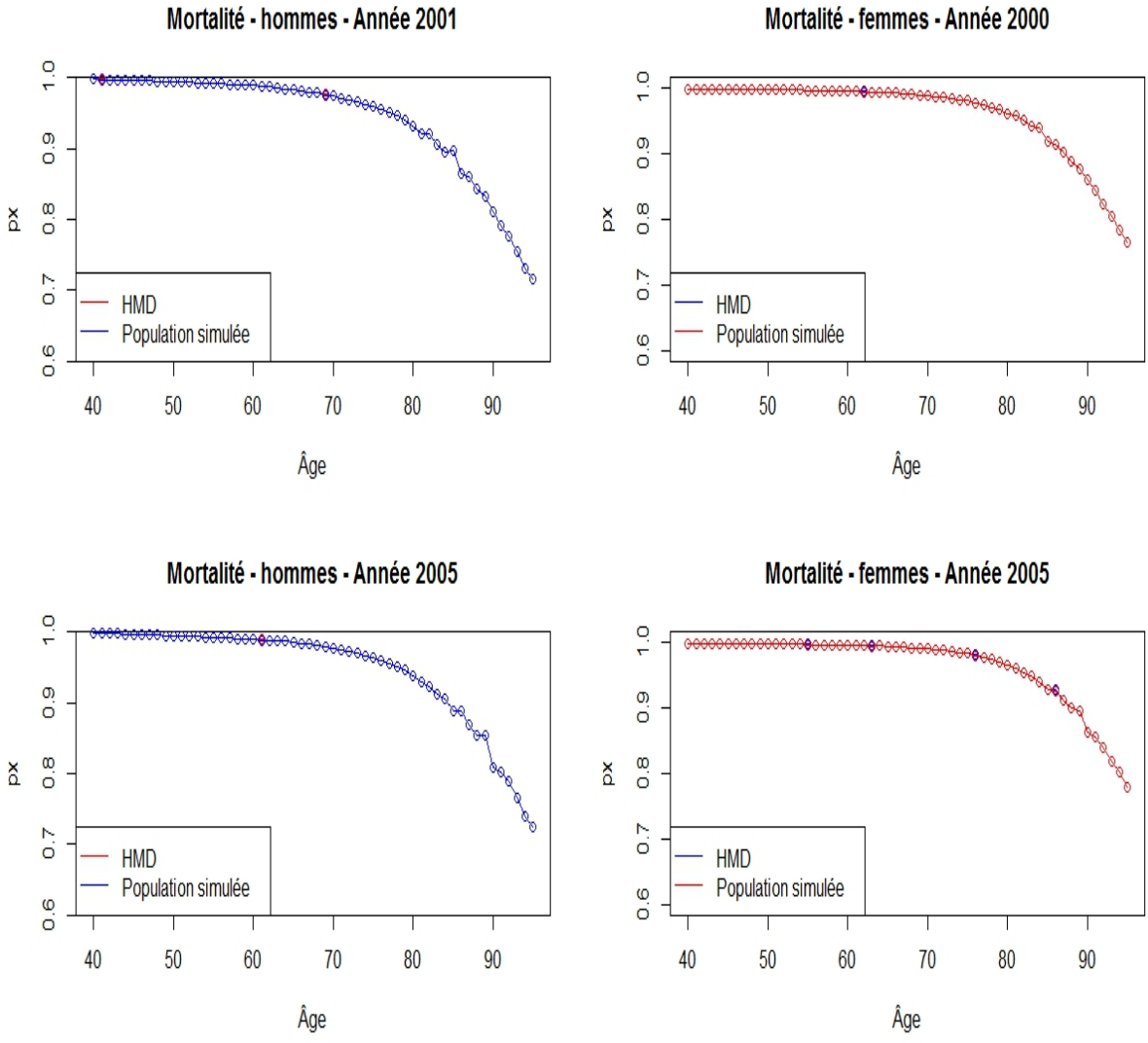

FIGURE 34 – Comparaison de la mortalité simulées et celles donnée par les tables de mortalité.

Algorithme de simulation de la population de rentiers

Algorithm 2 Simulation de la population de rentiers

1: Partir de la matrice de la population nationale simulée
2: Choisir la proportion Q de la population de rentiers par rapport à la population nationale ($Q \leftarrow \frac{1}{3}$)
3: **for** $t = 1996$ to 2005 **do**
4: **for** $Age = x_0$ to x_{max} **do**
5: Calculer l'exposition dans le portefeuille $E^{national}_{Age,t}$
6: Déduire l'exposition de la population de rentiers ($E^{rentiers}_{Age,t} \leftarrow E^{national}_{Age,t} * Q$)
7: Calculer le nombre de décès $n^{deces}_{Age,t}$ de rentiers ($n^{deces}_{Age,t} \leftarrow E^{rentiers}_{Age,t} * q^{rentiers}_{Age,t}$)
8: Choisir $n^{deces}_{Age,t}$ individus parmi les $N^{deces}_{Age,t}$ individus qui décèdent dans le portefeuille et modifier leur caractéristique « Population »
- Population $\leftarrow 1$ *//(National = 0, Rentiers = 1)*

9: Calculer le nombre de survie $n^{survie}_{Age,t}$ de rentiers ($n^{survie}_{Age,t} \leftarrow E^{rentiers}_{Age,t} - n^{deces}_{Age,t}$)
10: Choisir $n^{survie}_{Age,t}$ individus parmi les $N^{survie}_{Age,t}$ individus du le portefeuille qui ne décèdent pas à l'âge Age et l'année t et modifier leur caractéristique « Population »
- Population $\leftarrow 1$ *//(National = 0, Rentiers = 1)*

11: **end for**
12: **end for**

Comparaisons de la mortalité de la population nationale simulée et la mortalité réelle (Rentiers)

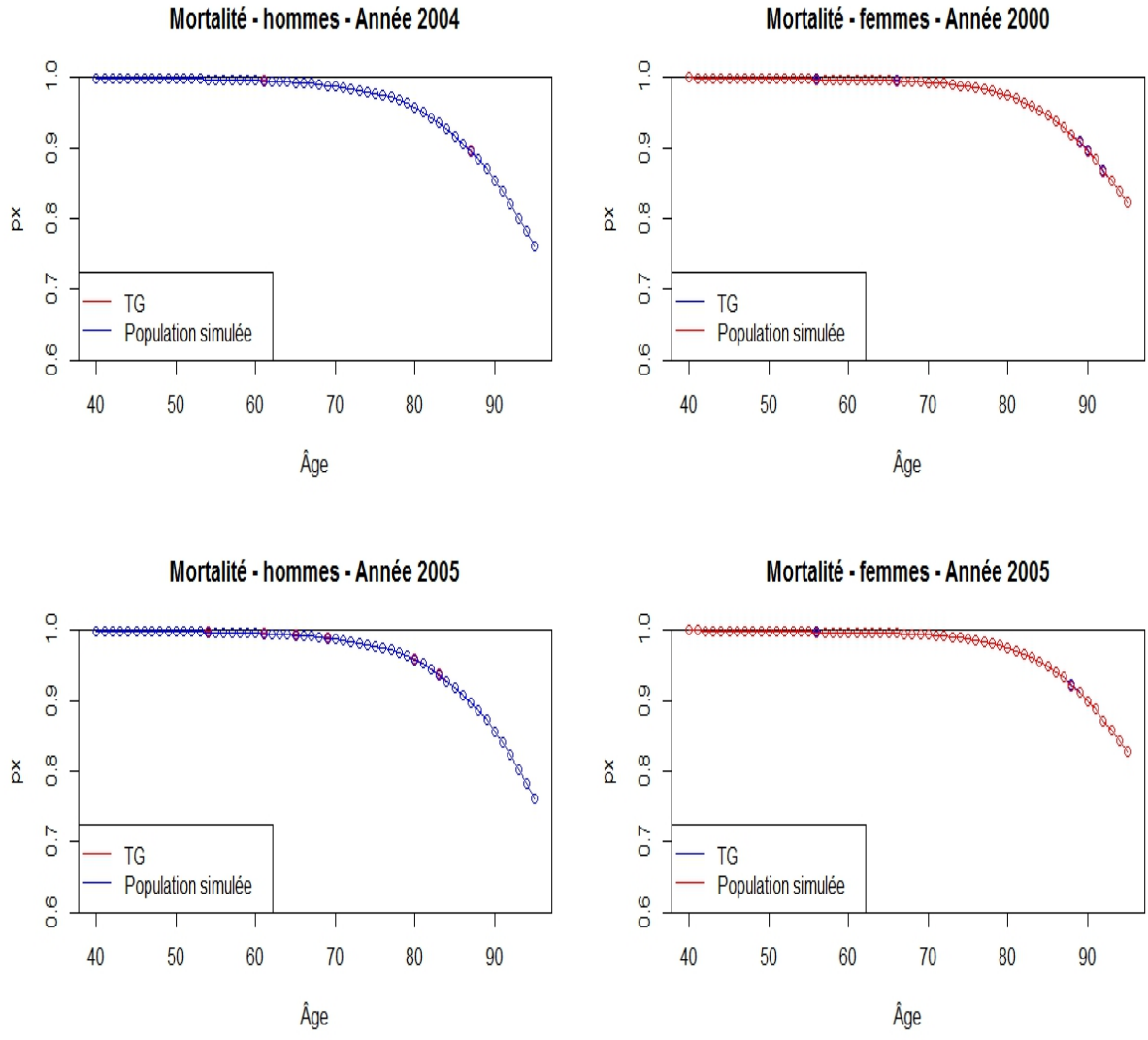

FIGURE 35 – Comparaison de la mortalité simulées et celles donnée par les tables de mortalité(Rentiers)

Ajustements des modèles

Statistiques de test de Wald et p-value (hommes)

Test	Statistique	p-value
1996	1475,72419	7×10^{-323}
1997	3915,85	0
1998	6906,1	0
1999	9006,54	0
2000	9459,11	0
2001	8848,29	0
2002	7910,48	0
2003	7000,22	0
2004	6220,48	0
2005	5577,86	0

TABLE 29 – Tests statistique - *Trend* logarithmique

Test	Statistique	p-value
1996	4314,41	0
1997	4485,06	0
1998	5032,60	0
1999	6051,82	0
2000	7586,47	0
2001	9152,49	0
2002	9219,63	0
2003	7193,51	0
2004	4802,39	0
2005	3118,61	0

TABLE 30 – Tests statistique - *Trend* quadratique

Test	Statistique	p-value
1996	2554,84	0
1997	3664,76	0
1998	5258,11	0
1999	7257,10	0
2000	9005,24	0
2001	9433,38	0
2002	8378,19	0
2003	6743,73	0
2004	5252,10	0
2005	4104,05	0

TABLE 31 – Tests statistique - *Trend* Linéaire

surfaces de mortalité ajustées

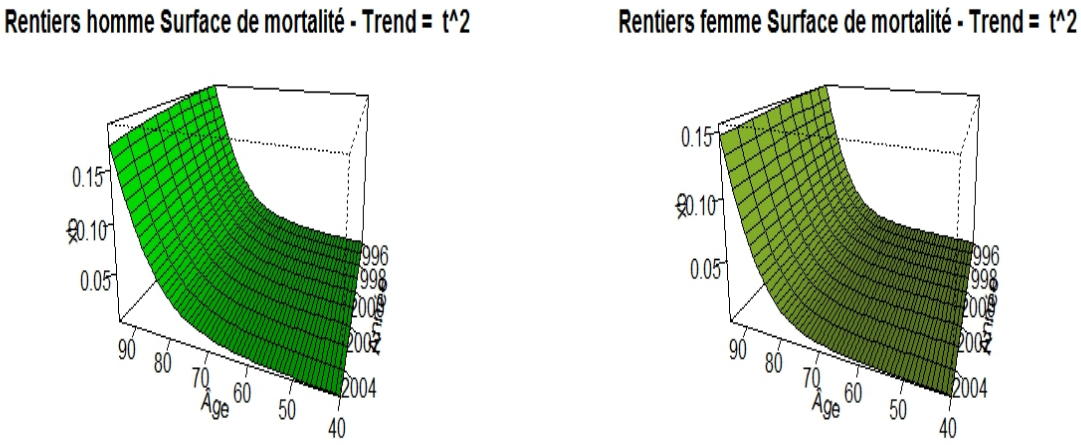

FIGURE 36 – Mortalité ajustée avec tendance quadratique

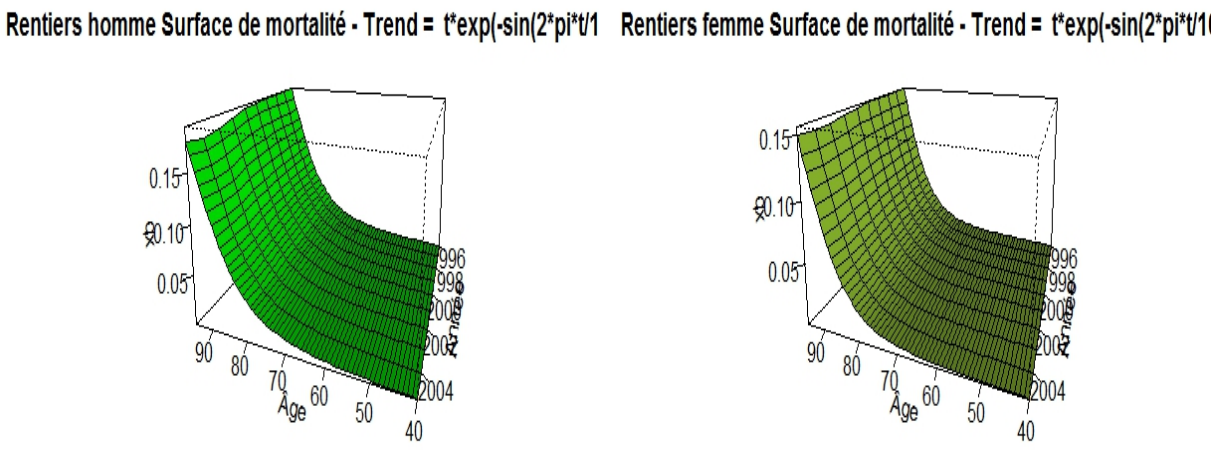

FIGURE 37 – Mortalité ajustée avec une tendance croissante sinusoïdale

Espérances de vie ajusté

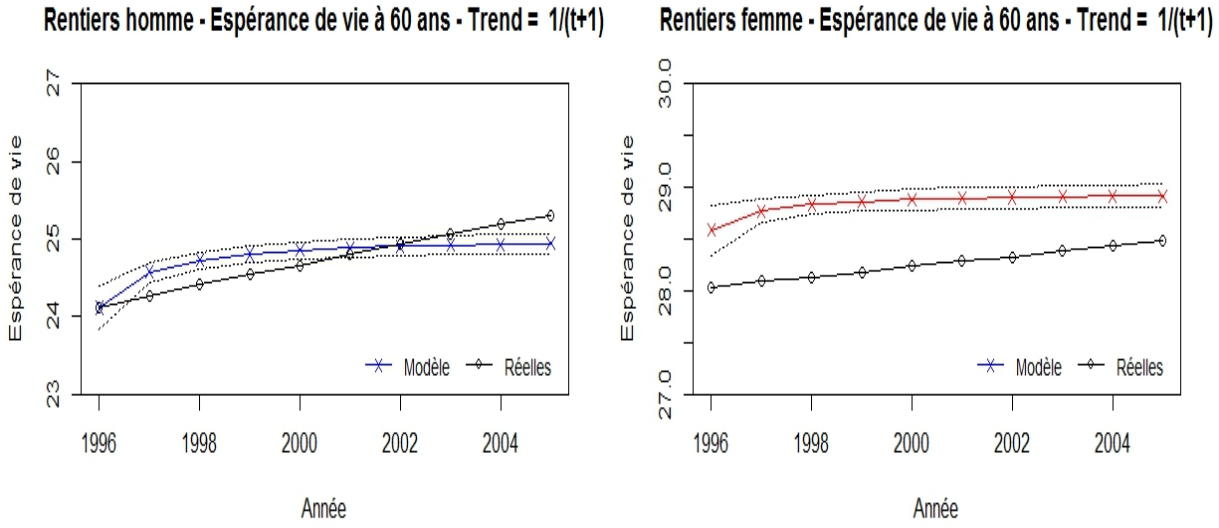

FIGURE 38 – Espérances de vies ajustées avec une tendance inverse

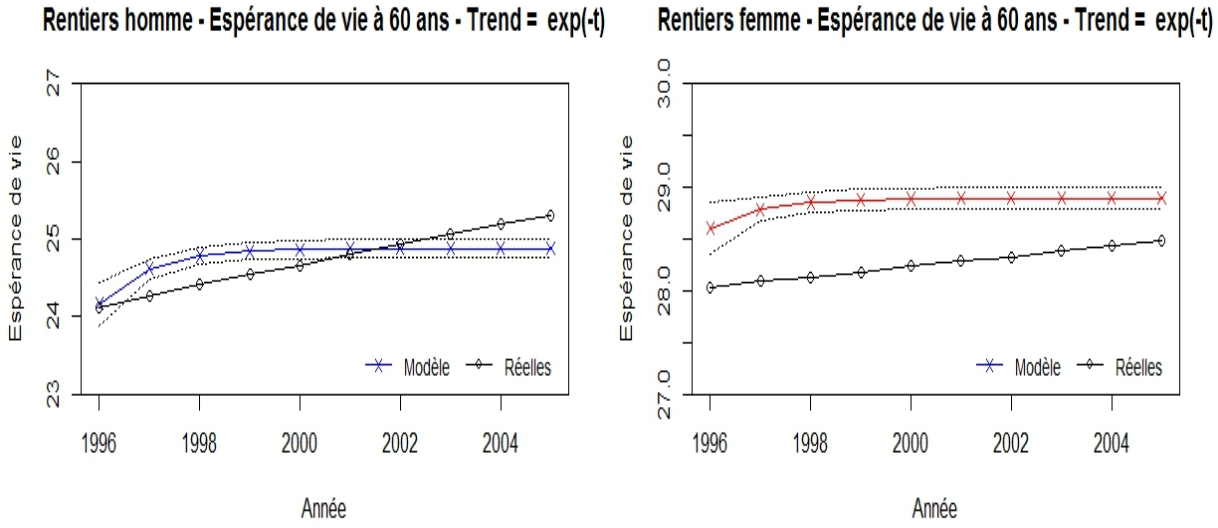

FIGURE 39 – Espérances de vies ajustées avec une tendance exponentielle inverse

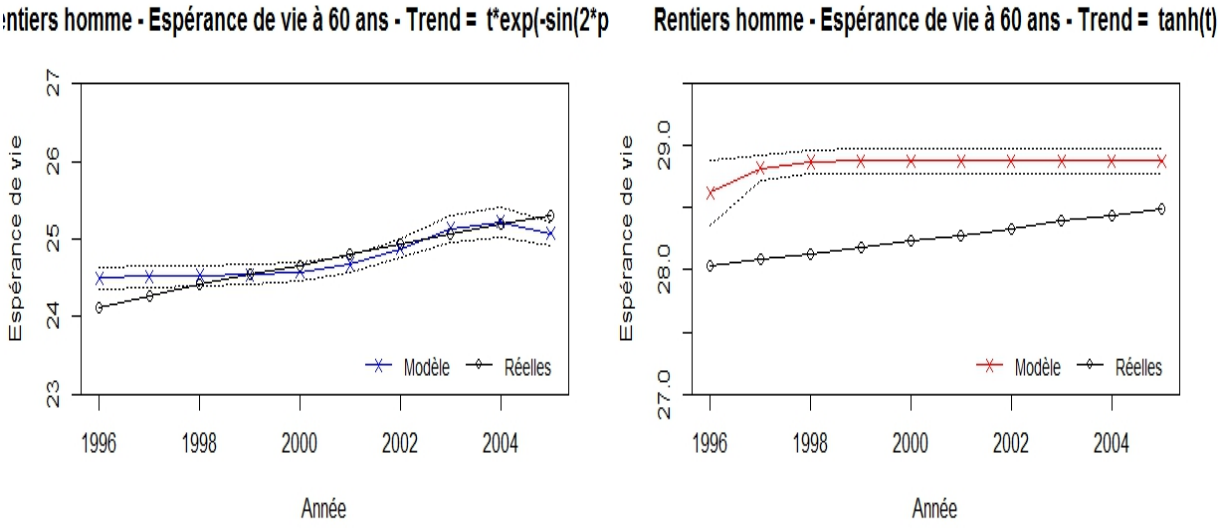

FIGURE 40 – Espérances de vies ajustées avec une tendance tangente hyperbolique

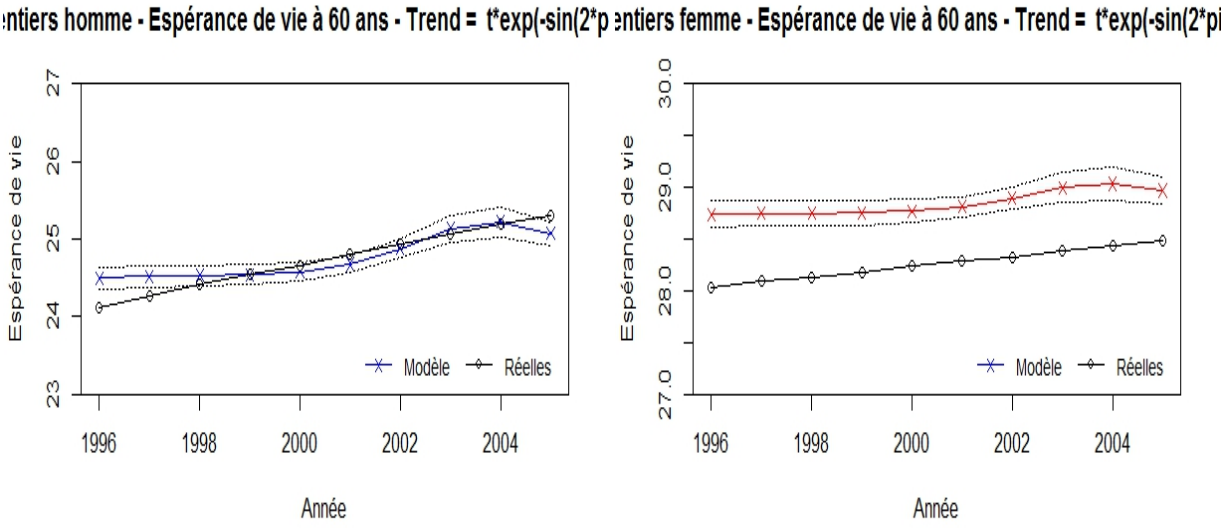

FIGURE 41 – Espérances de vies ajustées avec une tendance croissante sinusoïdale

	Espérances de vie ajustées			Espérances de vie réelles
	Quantile 99,5%	Central	Quantile 0,5%	
1996	24,48	24,30	24,11	24,11
1997	24,56	24,40	24,24	24,26
1998	24,64	24,50	24,37	24,41
1999	24,73	24,61	24,49	24,54
2000	24,82	24,71	24,60	24,66
2001	24,92	24,82	24,71	24,80
2002	25,04	24,92	24,80	24,93
2003	25,16	25,02	24,89	25,06
2004	25,28	25,12	24,97	25,19
2005	25,41	25,23	25,04	25,31
2006	25,53	25,33	25,12	25,46
2007	25,66	25,43	25,19	25,57
2008	25,79	25,53	25,27	25,72
2009	25,92	25,63	25,34	25,83
2010	26,05	25,73	25,41	25,95
2011	26,18	25,83	25,48	26,09
2012	26,30	25,93	25,55	26,22
2013	26,43	26,03	25,62	26,33
2014	26,56	26,13	25,69	26,46
2015	26,68	26,23	25,76	26,59

TABLE 32 – Espérances de vie - *Trend* Linéaire (Hommes)

	Espérances de vie ajustées			Espérances de vie réelles
	Quantile 99,5%	Central	Quantile 0,5%	
1996	28,82	28,66	28,49	28,03
1997	28,84	28,70	28,56	28,09
1998	28,86	28,74	28,62	28,13
1999	28,89	28,78	28,68	28,18
2000	28,92	28,82	28,73	28,24
2001	28,96	28,87	28,77	28,28
2002	29,01	28,91	28,80	28,33
2003	29,07	28,95	28,83	28,39
2004	29,13	28,99	28,85	28,44
2005	29,19	29,03	28,87	28,49
2006	29,26	29,08	28,89	28,53
2007	29,32	29,12	28,91	28,61
2008	29,39	29,16	28,92	28,66
2009	29,45	29,20	28,94	28,75
2010	29,52	29,24	28,95	28,82
2011	29,59	29,28	28,97	28,89
2012	29,65	29,32	28,98	28,97
2013	29,72	29,36	29,00	29,06
2014	29,78	29,40	29,01	29,12
2015	29,85	29,44	29,03	29,21

TABLE 33 – Espérances de vie - *Trend* Linéaire (Femmes)

Risque de tendance en Allemagne

Statistique de test

Test	Statistique	p-value
1995	581,719605	$4,80 \times 10^{-127}$
1996	530,61	$6{,}00\times10^{-116}$
1997	535,80	$4{,}48\times10^{-117}$
1998	596,61	$2{,}80\times10^{-130}$
1999	710,53	$5{,}12\times10^{-155}$
2000	873,37	$2{,}23\times10^{-190}$
2001	1079,61	$3{,}67\times10^{-235}$
2002	1322,82	$5{,}64\times10^{-288}$
2003	1596,16	0
2004	1892,74	0
2005	2206,02	0
2006	2530,01	0
2007	2859,44	0
2008	3189,82	0
2009	3517,42	0
2010	3839,28	0
2011	4153,06	0
2012	4457,05	0
2013	4750,02	0
2014	5031,17	0
2015	5300,03	0

TABLE 34 – Tests statistique - Modèle A (hommes)

Test	Statistique	p-value
1995	966,695171	$1{,}22\times10^{-210}$
1996	855,51805	$1{,}69\times10^{-186}$
1997	812,48535	$3{,}72\times10^{-177}$
1998	837,993366	$1{,}08\times10^{-182}$
1999	929,627387	$1{,}36\times10^{-202}$
2000	1082,46193	$8{,}84\times10^{-236}$
2001	1289,6366	$9{,}10\times10^{-281}$
2002	1543,05994	0
2003	1834,1036	0
2004	2154,19311	0
2005	2495,25079	0
2006	2849,98591	0
2007	3212,05089	0
2008	3576,09233	0
2009	3937,72763	0
2010	4293,47358	0
2011	4640,64864	0
2012	4977,26416	0
2013	5301,91549	0
2014	5613,6794	0
2015	5912,02164	0

TABLE 35 – Tests statistique - Modèle A (femmes)

Ajustement des espérances de vie

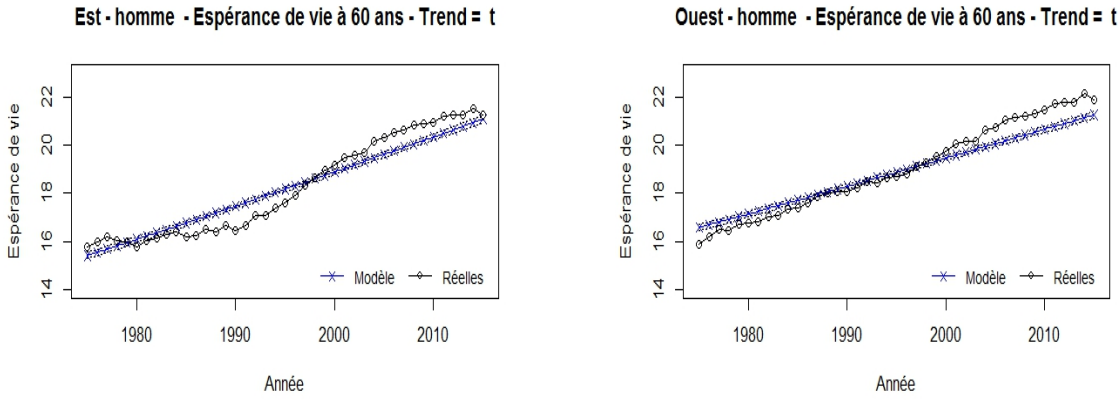

FIGURE 42 – Espérances de vies ajustées pour les hommes (Modèle A)

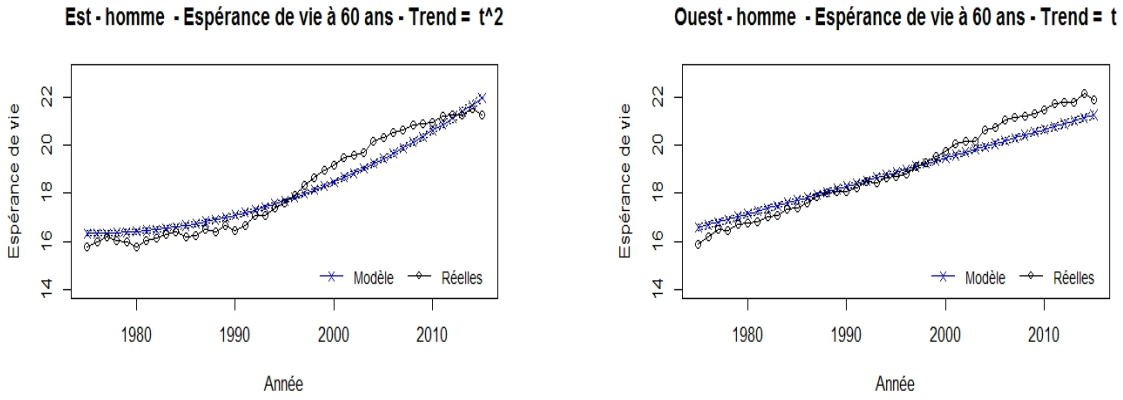

FIGURE 43 – Espérances de vies ajustées pour les hommes (Modèle B)

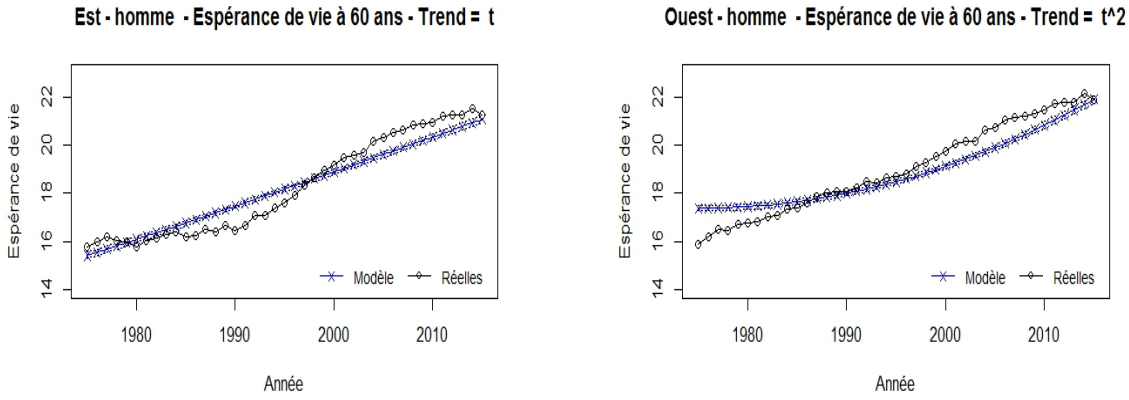

FIGURE 44 – Espérances de vies ajustées pour les hommes (Modèle C)

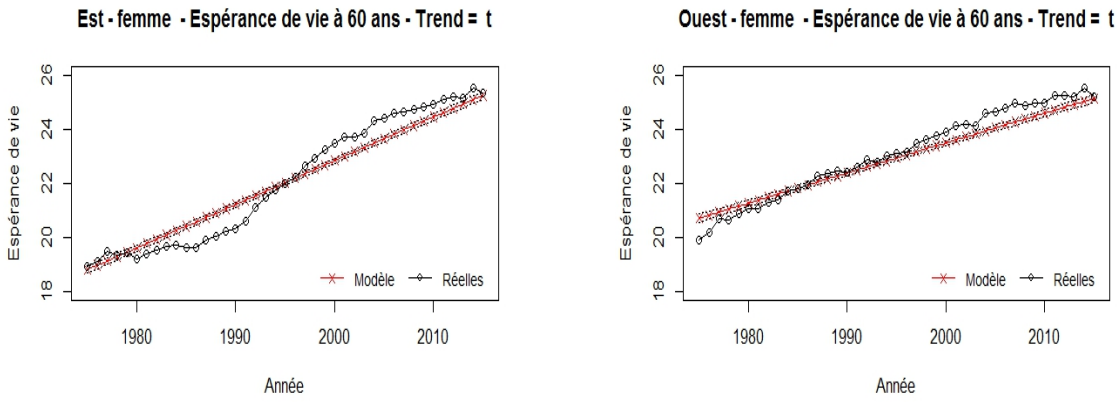

FIGURE 45 – Espérances de vies ajustées pour les femmes (Modèle A)

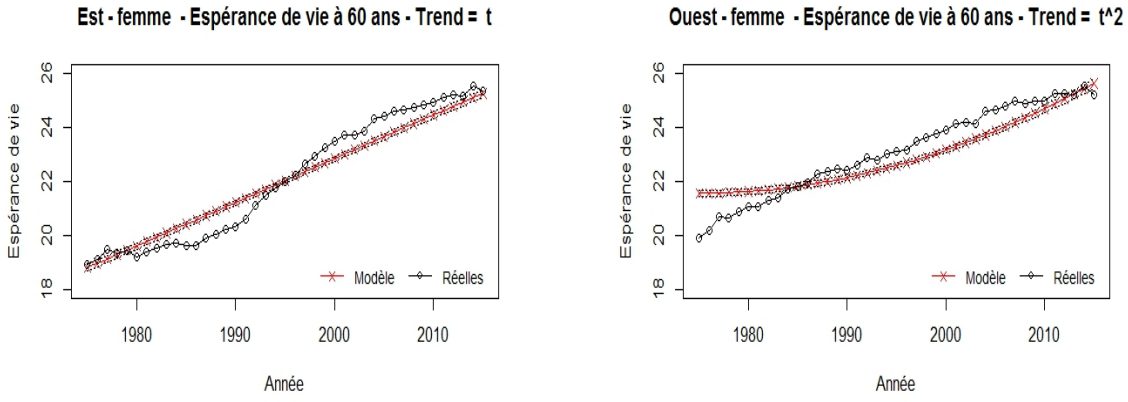

FIGURE 46 – Espérances de vies ajustées pour les femmes (Modèle B)

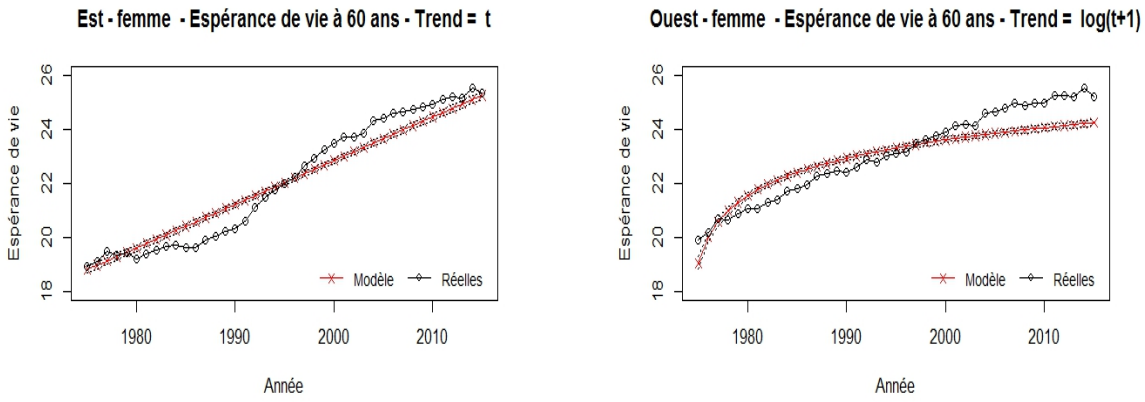

FIGURE 47 – Espérances de vies ajustées pour les femmes (Modèle C)

Étude de sensibilité

Sensibilité à la taille de données

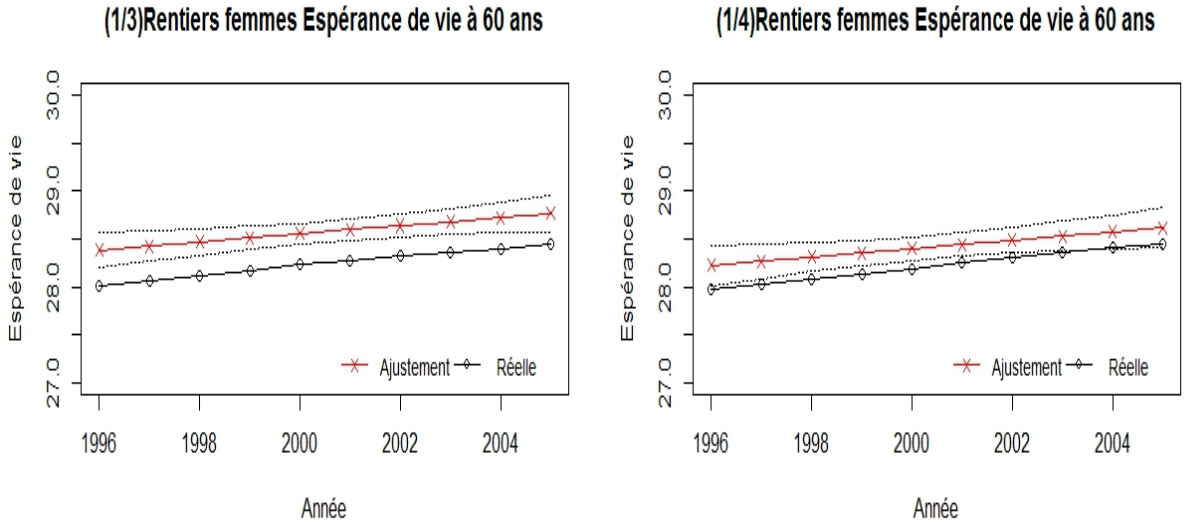

FIGURE 48 – Espérances de vies ajustées (Sensibilité aux données)

Sensibilité aux tranches d'âges de calibration

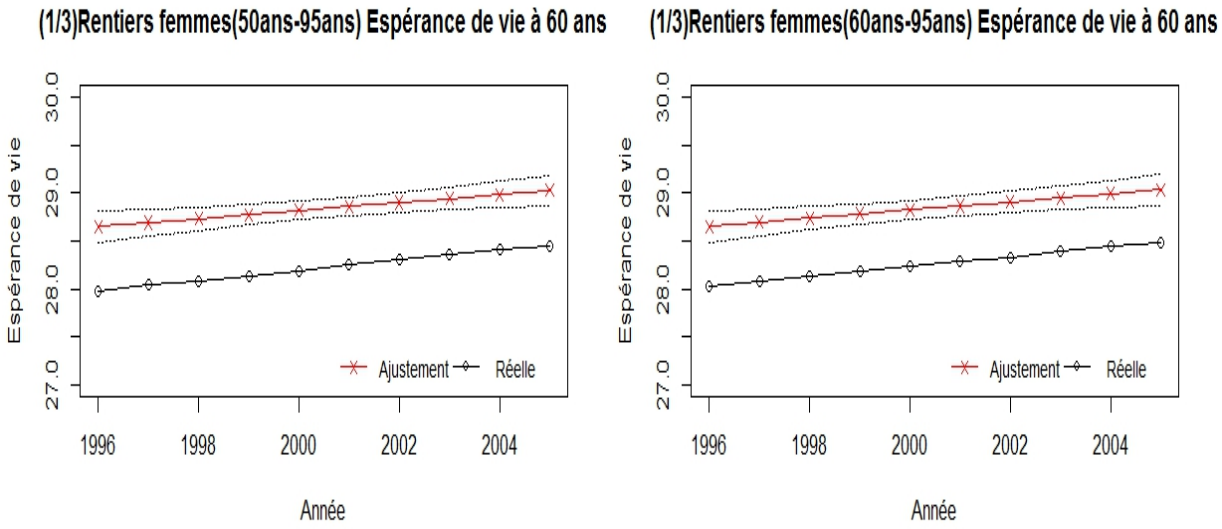

FIGURE 49 – Espérances de vies ajustées (Sensibilité aux tranches d'âges)

Bibliographie

Altman Douglas G., Andersen Per Kragh. A Note on the Uncertainty of a Survival Probability Estimated from Cox's Regression Model // Biometrika. 1986. 73, 3. 722–724.

Andersen P. K., Gill R. D. Cox's Regression Model for Counting Processes : A Large Sample Study // The Annals of Statistics. 1982. 10, 4. 1100–1120.

Anderson J. A., Senthilselvan A. A Two-Step Regression Model for Hazard Functions // Journal of the Royal Statistical Society. Series C (Applied Statistics). 1982. 31, 1. 44–51.

Augustin T. An exact corrected log-likelihood function for Cox's proportional hazards model under measurement error and some extensions // Sonderforschungsbereich 386. 2002. Discussion Paper 487.

Barbieri Magali. La mortalité départementale en France // I.N.E.D | Population. 2013. 68, 3. 433–479.

Booth Heather, Tickle Leonie. Mortality modelling and forecasting : a review of methods // Australian Demographic Social Research Institute. 2008. 3.

Borucka Jadwiga. Methods for handling tied events in the cox proportional hazard model // Studia oeconomica posnaniensia. 2014. 2, 263.

Brooks-Wilson Angela R. Genetics of healthy aging and longevity // Springerlink.com. 2013.

Cairns Andrew J.G., Blake David, Dowd Kevin, Coughlan Guy D., Epstein David, Ong Alen, Balevich Igor. A quantitative comparison of stochastic mortality models using data from England & Wales and the United States // The Pensions Institute Cass Business School City University. 2007. 58. 236–244.

Carnes Bruce A., Olshansky S. Jay, Grahn Douglas. Biological evidence for limits to the duration of life // 4. 2003. 31–45.

Chauvel Cecile. Empirical processes for inference in the non-proportional hazards model. 2014.

Rising life expectancy in England has slowed since recession. // . 2017. New Scientist Live.

Cox D. R. Regression Models and Life-Tables // Journal of the Royal Statistical Society. 1972. 34, 2. 187–220.

Currie Iain. Smoothing and Forecasting Mortality Rates with P-splines // Talk given at the Institute of Actuaries. June 2006.

Doblhamme Gabriele, Vaupel James W. Lifespan depends on month of birth // Proceedings of the National Academy of Sciences of the United States of America. 2001. 98, 5. 2934–2939.

Duroux Roxane, O'Quigley John. Inference for changepoint survival models. 2016.

Efron Bradley. The efficiency of Cox's likelihood function for censored data // National Science Foundation grant MPS74-21416. 1975. 15.

Européenne Commission. RÈGLEMENT DÉLÉGUÉ (UE) 2016/467 DE LA COMMISSION du 30 septembre 2015 // Journal officiel de l'Union européenne. 2016.

GREY Aubrey D.N.J. DE, AMES Bruce N., Andersen Julie K., BARTKE Andrzej, CAMPISI Judith, HEWARD Christopher B., Mccarter Roger J. M., STOCKh Gregory. Time to Talk SENS : Critiquing the Immutability of Human Aging // New York Academy of Sciences. 2002. 959. 452–462.

Gibbons Jean D., Pratt John W. P-values : Interpretation and Methodology // The American Statisticia. 1975. 29, 1.

Goris Jérémy. Influence de l'environnement social sur la longévité des rentiers // Mémoire d'actuariat, Institut des Actuaires. 2016.

Grambsch Patricia M., Therneau Therry M. Proportional Hazards Tests and Diagnostics Based on Weighted Residuals // Biometrika. 1994. 81, 3. 515–526.

Guérette Marie-Josée, Caradec Vincent, Bergman Howard, Joanette Yves, Blanchard François. La vieillesse, c'est le déclin, on n'y peut rien // Santé, Société et Solidarité. 2006. 5, 1. 23–31.

Hartigan J. A., Wong M. A. A K-Means Clustering Algorithm // Journal of the Royal Statistical Society. 1979. 28, 1. 100–108.

Horny Guillaume. Modèles de hasards proportionnels et hétérogénéité non observée // Bulletin Français d'actuariat. 2008. 8, 15. 4–31.

Hosmer David W., Royston Patrick. Using Aalen's linear hazards model to investigate time-varying effects in the proportional hazards regression model // The Stata Journa. 2002. 2, 4. 331–350.

INSEE . 21 000 centenaires en 2016 en France, 270 000 en 2070 ? // Insee Première. 2016. 1620.

Joe Ward Jr. Hierarchical grouping to optimize an objective function // Journal of the American Statistical Association. 1963. 58. 236–244.

Kalbfleisch J. D., Prentice R. L. Marginal Likelihoods Based on Cox's Regression and Life Model // Biometrika. 1973. 60, 2. 267–278.

Kamega Aymric, Planchet Frédéric. Hétérogénéité : mesure du risque d'estimation dans le cas d'une modélisation intégrant des facteurs observables // Bulletin Français d'actuariat. 2011. 11, 21. 99–129.

Kaplan E. L., Meier Paul. Non parametric Estimation from Incomplete Observations // Journal of the American Statistical Association. 1958. 53, 282. 457–481.

Kontis Vasilis, Bennett James E, Mathers Colin D, Li Guangquan, Foreman Kyle, Ezzati Majid. Future life expectancy in 35 industrialised countries : projections with a Bayesian model ensemble // thelancet. 2017. 389.

Lee Ronald D., Carter Lawrence R. Modeling and Forecasting U. S. Mortality // Journal of the American Statistical Association. 1992. 87, 419. 659–671.

Lin D. Y., Ying Zhiliang. Semiparametric Analysis of General Additive-Multiplicative Hasard Models for Counting Processes // The Annals of Statistics. 1995. 23, 5. 1712–1734.

Lopez Olivier. Réduction de dimension en présence de données censurées. 2007.

Medeiros Francisco M. C., Silva-Junior Antonio H. M. da, Valença Dione M., Ferrari Silvia L. P. Testing Inference in Accelerated Failure Time Models // International Journal of Statistics and Probability. 2014. 3, 2.

Monde Le. Pour la première fois depuis 1993, l'espérance de vie a reculé aux Etats-Unis // Le Monde. 2016.

Musset Anne-Sophie. Risque de longévité : la référence à la tendance de la population nationale est-elle justifiée ? // Mémoire d'actuariat, Institut des Actuaires. 2016.

Oeppen Jim, W.Vaupel James. Broken Limits to Life Expectancy // Science's Compass. 2002. 296. 1129–1131.

Olshansky S. Jay, Carnes Bruce A. The Future of Human Longevity // P. Uhlenberg (ed.), International Handbook of Population Aging. 2009.

O'Hare Colin. Essays in modelling mortality rates. 12 2012.

PLSA . Longevity trends : Does one size fit all ? // Pensions and Lifetime Savings Association. 2017.

Pechholdová Markéta, Grigoriev Pavel, Meslé France, Vallin Jacques. Espérance de vie : les deux Allemagne ont-elles convergé depuis la réunification ? // Population Sociétés, bulletin mensuel d'information de l'Institut national d'études démographiques. 2017. 544.

Pitacco Ermanno, Denuit Michel, Haberman Steven, Olivieri Annamaria. Modelling Longevity Dynamics for Pensions and Annuity Business. 2009.

Planchet Frédéric. Tables de mortalité d'expérience pour les portefeuilles de rentiers // Institut des Actuaires. 2006.

Planchet Frédéric. Statistique des modèles non paramétriques. 9 2016a.

Planchet Frédéric. Statistique des modèles paramétriques et semi-paramétriques. 12 2016b.

Pouchard Alexandre. Six chiffres-clés pour comprendre le suicide en France // Le Monde. 2016.

Renshaw A.E., Haberman S. A cohort-based extension to the Lee–Carter model for mortality reduction factors // Insurance : Mathematics and Economics. 2006. 38. 556–570.

Riley Kathleen. We're Entering a New Age, One Where the Life Expectancy Is 90+ // futurism. 2017.

Robin Jean-Pierre. Notre esperance de vie augmente de 5 heures et demi chaque jour // Le Figaro. economie. 2013.

Rougerie Catherine. Le projet de loi des retraites est la réforme clé de la fin du mandat de Nicolas Sarkozy // France Télévisions. 2010.

Schemper Michael, Stare Janez. Explained Variation in Survival Analysis // Statistics in Medicine. 1996. 15. 1999–2012.

Schoenfeld David. Partial residuals for the proportional hazards regression model // Biometrika. 1982. 69. 239–241.

Therneau Therneau Terry M., Grambsch Patricia M. Modeling Survival Data : Extending the Cox Model. 2000.

Tsiatis Anastasios A. A large sample study of Cox's regression model // The Annals of Statistics. 1981. 9, 1. 93–108.

Vaupel James W. La Longévité : Passé, présent et putur // Revue d'économie financière. 2016. 122. 41–56.

Viallon Vivian. Processus empiriques, estimation non paramétrique et données censurées. 2006.

Wasserman Larry. All of Statistics : A Concise Course in Statistical Inference. 2004.

Wei L. J. The Accelerated Failure Time Model : A useful Alternative to the Cox Regression Model in Survival Analysis // Statistics in medecine. 1992. 11. 1871–1879.

West Mike. Modelling time-varying hasards and covariate effects. 1991.

How Long Can We Live ? // . 2001.

tdg . Hommes et femmes pas égaux face à la maladie // Tribune de Genève,. 2016.