PSL UNIVERSITÉ PARIS

MIDO
MATHÉMATIQUES ET INFORMATIQUE
DE LA DÉCISION ET DES ORGANISATIONS

Dauphine
UNIVERSITÉ PARIS

Mémoire de Master 1 Mathématiques Appliquées de l'Université de Paris Dauphine

Titre : Modélisation prospective des causes de décès au moyen de copules

Par : Thomas Pain, Amélie de la Haye, Corentin Schmuck

Directeur de Mémoire : Quentin Guibert
Numéro de groupe : 5

Université Paris-Dauphine – MIDO – Bureau B536, Place du Maréchal de Lattre de Tassigny, 75016 PARIS

Table des matières

Introduction

L'objectif de ce mémoire est d'analyser différentes causes de mortalité en France, en déterminant des liens entre ces différentes causes. Ainsi, il est ensuite possible de déterminer des primes pour les engagements des futurs assurés envers une compagnie d'assurance. Nous pourrons alors mesurer les conséquences d'une éventuelle disparition ou forte diminution de la présence d'une pathologie sur les contrats d'assurance.

En premier lieu nous devrons établir les modèles et notations classiques inhérents à la modélisation de la mortalité. Nous avons choisi quelques causes pertinentes à analyser dans le cadre de notre étude. Nous présenterons une introduction à la théorie des copules, qui nous permettra de mettre des liens de probabilités entre nos différentes causes de mortalité afin de les étudier dans l'ensemble et non pas seulement de façon individuelle et indépendante. Le modèle de Lee-Carter sera décrit ici, il nous permettra de modéliser des projections futures sur les taux de mortalité, l'espérance de vie... Ce sera notre modèle de référence dans tout ce document.

Dans une deuxième partie, nous présentons les données que nous avons choisies d'utiliser. Des données épurées sur la mortalité en France permettant une manipulation simple et rapide avec R. Ces données seront soumises à quelques expériences et méthodes de statisques de base afin de pouvoir mieux les comprendre et mieux les prendre en main. Le principal problème de ces données est l'utilisation de tranches d'âges de 5 en 5 pour décrire les différentes données. Nous avons du mettre en place un système d'interpolation avec une loi de mortalité connue pour obtenir des données sur l'intégralité des âges entiers.

Une fois la question des données traitées, il faudra utiliser le modèle de Lee Carter. En première partie avec une copule d'indépendance, c'est à dire dans le cas où toutes nos causes sont indépendantes. Puis dans un second temps avec une structure de dépendance induite par des copules archimédiennes classiques. On simulera ensuite le cas où une de ces causes de mortalité disparaît, par exemple suite à un progrès de la médecine. On pourra retirer plus ou moins brutalement une cause de mortalité, pour y voir l'impact sur les fonctions de survie et sur l'espérance de vie de la population.

Enfin, nous avons utilisé nos modèles et prédictions pour calculer la prime correspondant à un produit de rentes viagières sur plusieurs années. Nous verrons alors les conséquences de la réduction ou de l'élimination d'une cause de décès sur la prime demandée à l'assuré.

1 Modélisations de la mortalité

1.1 Modélisation avec une seule cause de décès

On s'intéresse dans un premier temps aux modèles de durée utilisés en assurance et ici principalement en assurance vie pour mesurer la durée de vie humaine. Considérons un individu dont la durée de vie est modélisée par la variable aléatoire $T \in]0, +\infty]$. On note x son âge actuel. Pour simplifier les calculs, on utilisera parfois la notation $T_x = T - x$. Rappelons maintenant quelques définitions. **Dutang (2020-2021)**

Définition 1 *: La loi de probabilité de la variable aléatoire T peut être caractérisée des façons suivantes :*

- *La fonction de répartition $F_x(t) = P(T \leq t + x \,|\, T > x) = {}_tq_x$*
- *La fonction de survie $S_x(t) = P(T > t + x \,|\, T > x) = P(T_x > t \,|\, T_x > 0) = {}_tp_x$*
- *La densité de survie $f_x(t) = -S'_x(t) = F'_x(t)$*

Définition 2 (Taux de mortalité) *: On définit le taux instantané de mortalité (conditionné à l'âge x) ou taux de danger par :*

$$\mu_x(t) = \frac{f_x(t)}{S_x(t)} = \frac{-S'_x(t)}{S_x(t)} = -\frac{d}{dt}\, ln(S(t)).$$

Il en résulte que le taux de mortalité détermine la loi de T et on a la relation :

$$S(t) = exp\,(-\int_0^t h(s)\, ds)$$

Il est ensuite souvent nécessaire d'introduire des lois paramétriques pour estimer la valeur de T. Les plus communes pour modéliser la durée de vie humaine sont les lois de De Moivre, de Gompertz, de Makeham et de Weibull, nous allons briévement les présenter ici.

Loi de De Moivre (1729)

Soit ω fixé, on suppose $T \sim \mathscr{U}(0, \omega)$. Alors :

$$F_x(t) = {}_tq_x = \frac{P(0 < T_x \leq t)}{P(T_x > 0)} = \frac{\frac{t}{\omega}}{\frac{\omega - x}{\omega}} = \frac{t}{\omega - x} \text{ donc on a aussi } {}_tp_x = 1 - \frac{t}{\omega - x}$$

On remarquera que cette loi pose plusieurs problèmes pour la modélisation de la vie humaine. Tout d'abord il est nécessaire d'établir un âge maximum ω qu'aucun individu ne pourra dépasser. De plus, on sait très bien que la durée de vie n'est en général pas uniforme car la probabilité de décéder augmente fortement avec l'âge. On n'utilisera donc pas cette loi en pratique.

Loi de Gompertz (1825)

Soient $c \geq 1$ et $b > 0$. L'hypothèse de Gompertz est en fait $\mu_x = bc^x$. D'où :

$$\int_x^{x+t} \mu_u\, du = \int_x^{x+t} b \exp(u\, log(c))\, du = b \left[\frac{c^u}{log(c)}\right]_x^{x+t} = b\, \frac{c^{x+t} - c^x}{log(c)}$$

Ainsi, pour $x, t > 0$: ${}_tp_x = \exp\left(-b\, \frac{c^{x+t} - c^x}{log(c)}\right)$

Loi de Makeham (1867)

Soient $c \geq 1$, $b > 0$ et $a + b \geq 0$. L'hypothèse est proche de celle de Gompertz à une constante près, on suppose $\mu_x = bc^x + a$. Donc :

$$\int_x^{x+t} \mu_u\, du = \int_x^{x+t} a + b\,(\exp(u\, log(c)))\, du = at + b \left[\frac{c^u}{log(c)}\right]_x^{x+t} = at + b\, \frac{c^{x+t} - c^x}{log(c)}$$

Ainsi, pour $x, t > 0$: ${}_tp_x = \exp(-at)\exp\left(-b\,\frac{c^{x+t}-c^x}{log(c)}\right)$

Loi de Weibull (1939)

Soient $k > 0$ et $n > 0$. On suppose que $\mu_x = kx^n$:

$$\int_x^{x+t} \mu_u\,du = \int_x^{x+t} ku^n\,du = k\left[\frac{u^{n+1}}{n+1}\right]_x^{x+t} = k\,\frac{(x+t)^{n+1}-x^{n+1}}{n+1}$$

Ainsi ${}_tp_x = \exp\left(-k\,\frac{(x+t)^{n+1}-x^{n+1}}{n+1}\right)$

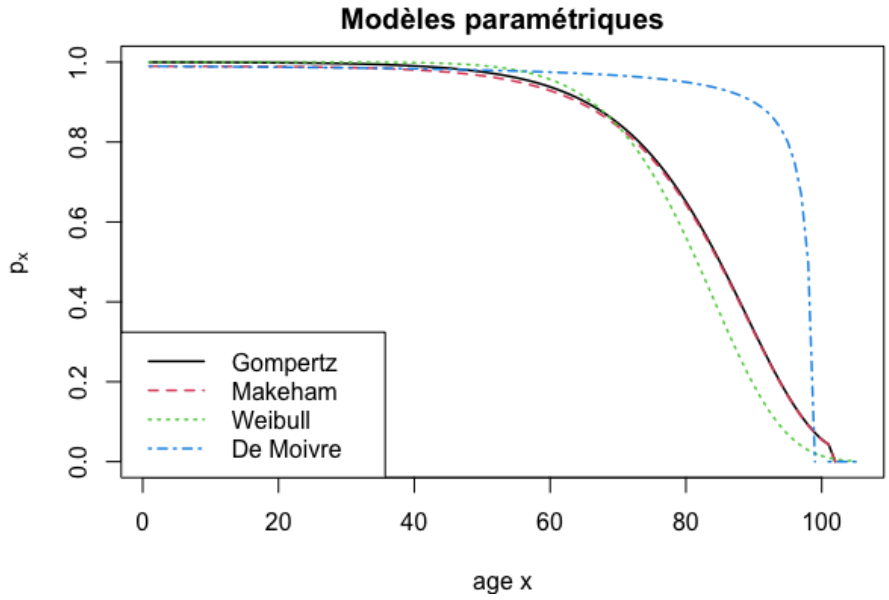


FIGURE 1.1 – Courbe $x \mapsto p_x$ pour des les lois paramétriques ; Gompertz $b = 0.002$, $c = 1.1$, Makeham $a = 0.01$, Weibull $k = 0.01448254, n = 10$, De Moivre $\omega = 100$.

Definitions (Taux de mortalité relatif à l'âge) (Age-specific Death Rate) :
Ce taux de mortalité correspond au nombre de morts pour un âge x donné, divisé par le total de la population ayant le même âge x, et cela dans une même région géographique donnée (par exemple la France). On appellera ce total de la population d'âge x, le taux d'exposition au risque (noté E_x), ou *exposure at risk* en anglais. Ce taux sera souvent multiplié par un coefficient de 100 000, la formule de son estimateur nous donne :

$$\frac{\text{Total Deaths in Specified Age Group}}{\text{Total Population in the Same Specified Age Group}} \quad \text{X} \quad 100{,}000$$

1.2 Modèle de Lee-Carter

Le modèle **Lee et Carter (1992)** est une méthode utilisée par de nombreux actuaires afin de prévoir la mortalité et d'estimer l'espérance de vie, il a été introduit par Ronald Lee et Lawrence Carter en 1992 dans "Modelling and forecasting the time series of US mortality".

Notations

On commence par définir les notations qui seront utilisées dans cette section. On note tout d'abord ${}_tp_x$ la probabilité de survivre t années pour un individu d'âge x, ${}_tq_x$ la probabilité de décès avant les t années.On note également μ_{xt} le taux de mortalité intantané pour un individu d'âge x à l'année t, l_x (resp d_x) le nombre de personnes qui vivent après x (resp qui meurent à l'âge x). On remarque notamment que $d_x = l_x - l_{x+1}$.

Définition du modèle

L'idée principale est de modéliser le logarithme népérien du taux de mortalité par :

$$ln(\mu_{xt}) = \alpha_x + \beta_x\kappa_t + \epsilon_{xt}$$

- ϵ_{xt} un terme d'erreur
- κ_t l'évolution générale de la mortalité
- β_x traduit la tendance de μ_{xt} en fonction de l'évolution de k_t. On a la relation suivante : $\beta_x = \frac{d\mu_{xt}}{dk_t}$ (en théorie β_x pourrait être négatif, ce qui signifierait que pour certains âges la mortalité chute, tandis qu'elle augmente pour d'autres).

On remarque que pour tout λ scalaire, si (α, β, κ) est solution du problème de minimisation alors $(\alpha - \lambda\beta, \beta, \kappa + \lambda)$ et $(\alpha, \lambda\beta, \kappa/\lambda)$ sont également solutions. Ainsi, dans le but de rendre le modèle identifiable, on fixe pour toute la suite $\sum_x \beta_x = 1$ et $\sum_t k_t = 0$. Basé sur ces contraintes on peut de fait interpréter α_x comme la valeur moyenne de $(\ln(\mu_{xt}))_t$.

On cherche donc pour la suite à déterminer $(\hat{\alpha}_x, \hat{\beta}_x, \hat{k}_t) = min(\sum_{x,t}(ln(\hat{\mu}_{xt}) - \alpha_x - \beta_x k_t))^2$, sous les contraintes imposées plus haut.

Estimation classique des paramètres

Une méthode classique de résolution est de réaliser une décomposition en valeurs singulières (SVD) de la matrice des logarithmes des taux, auxquels on soustrait la moyenne sur les années des taux pour un âge donné ; on obtient de cette façon une solution vérifiant le critère des moindres carrés. En suivant cette méthode on suppose que $\epsilon_{xt} \sim \mathcal{N}(0, \sigma^2)$,

Étape 1 : estimation de μ_{xt} et de α_x
On estime les taux μ_{xt} en posant $\hat{\mu}_{xt} = \frac{{}_td_x}{e_x}$ avec e_x le coefficient d'exposition à l'âge x et ${}_td_x$ le nombre de personnes qui décèdent entre x et $x+t$. On estime ensuite α_x en prenant la moyenne empirique des $\hat{\mu}_{xt}$

Étape 2 : estimation de β_x et de k_t
On note $\mathbf{1} = (1 \ldots 1)^T$, afin réaliser cette estimation, on réalise une décomposition en valeurs singulières de la matrice $Z = M - \alpha\mathbf{1}^T$. Avec $M = (ln(\hat{\mu}_{xt}))$ et $\alpha = (\alpha_x)$
Si on note u_1 (resp v_1) le vecteur propre de Z^TZ (resp ZZ^T) associé à la plus grande valeur propre λ_1, on a alors :

$$\hat{\beta} = \frac{v_1}{\sum_j v_{1j}} \quad , \quad \hat{k} = \sqrt{\lambda_1} \sum_j v_{1j} u_1$$

Le modèle de Lee Carter est énormément utilisé, néanmoins il présente des limites. La première est que la méthode repose sur la résolution d'un problème des moindres carrés ordinaires, or ceci nécéssite que les résidus soient homoscédastiques. En effet, pour un âgé fixé on devrait avoir une faible variation de la variance au cours du temps ce qui n'est pas forcément vrai. La seconde concerne le critère de sélection des paramètres. C'est pourquoi nous allons nous intéresser au modèle log Poisson qui n'est autre qu'une adaptation du modèle Lee Carter afin de pallier à ces problèmes. Le modèle de logPoisson est également le modèle présent dans le package StMoMo de R que nous utiliserons par la suite pour nos modélisations.

Modèle de LogPoisson

L'idée est simple, on note D_{xt} le nombre de décès à l'âge x l'année t, et on suppose $D_{xt} \sim \mathscr{P}(E_{xt}\mu_{xt})$, avec $\mu_{xt} = \exp(\alpha_x + \beta_x\kappa_t)$ Il est important de préciser que μ_{xt}, α_x, β_x et κ_t ont les mêmes interprétations que dans le modèle de Lee Carter. Le paramètre E_{xt} est le coefficient d'exposition au risque pour un individu d'âge x l'année t. On remarque que l'on est donc en présence d'un modèle linéaire généralisé ; le logarithme de la vraissemblance vaut :

$$ln(L(\alpha, \beta, \kappa)) = \sum_{x,t} (D_{xt}(\alpha_x + \beta_x\kappa_t) - E_{xt}exp(\alpha_x + \beta_x\kappa_t))$$

Dans notre cas, on ne peut pas résoudre analytiquement les équations de vraissemblance du modèle à cause du terme $\beta_x\kappa_t$, on peut néanmoins les résoudre numériquement.

1.3 Risques concurrents et causes de décès

Dans cette section, on essaye d'analyser la durée de vie selon différentes causes de mortalité. On note par T_j la variable aléatoire positive déterminant la durée de vie d'un individu avant de décéder de la cause $j = 1, ..., m$. On notera que ces variables aléatoires T_j peuvent être corrélées et ne sont pas forcément indépendantes. On suppose que $P(T_j < +\infty) = 1$. Soit $t_j \geq 0$ pour $j = 1, ..., m$, on définit la fonction de survie multivariée par :

$$S(t_1, ..., t_m) = P(T_1 > t_1, ..., T_m > t_m)$$

Définition 3 (Fonction de survie nette) *: Soit $j = 1, ..., m$, on appelle fonction de survie nette associée à la cause de mortalité j la probabilité suivante :*

$$S_j(t) = P(T_j > t)$$

Maintenant, imaginons que nous sommes parvenus à classifier la totalité des décès par m causes sous-jacentes. Alors la durée de vie T d'un individu est défini par $T = min(T_1, ..., T_m)$ et on peut identifier la cause de son décès par la variable aléatoire :

$$J = \sum_{j=1}^{m} j \, \mathbb{1}_{\{T=T_j\}}$$

Remarque : On peut déjà remarquer que les variables J et T sont observables dans la réalité tandis que les variables aléatoires T_j ne le sont pas. Il est à priori aussi évident que :

$$P(T_j = T_i) = 0 \quad \forall j \neq i$$

.

Définition 4 (Fonction de survie brute) *: Soit $j = 1, ..., m$, on définit la fonction de survie brute associée à la cause de mortalité j par la probabilité suivante :*

$$\tilde{S}_j(t) = P(T > t, J = j)$$

Par la suite, notre objectif sera de calculer $S_j(t)$, en supposant que la fonction $\tilde{S}_j(t)$ est connue, en effet cette dernière ne contient que des variables aléatoires observables. Il est donc nécessaire de trouver un "lien" entre ces deux fonctions. **Carriere (1995)**

1.4 Introduction à la théorie des copules

Définition 5 (Copule multivariée) *: Une copule multivariée est une application $C : [0,1]^d \mapsto [0,1]$ telle que*

- *$u_i \mapsto C(u)$ est croissante sur $[0,1]$ pour $i = 1, \ldots, d$.*
- *$\forall u \in [0,1]^d$, C est continue sur $[0,1]$.*
- *$\forall u \in [0,1]^d$, $C(u) = 0$ lorsque $u_i = 0$ pour $i = 1, \ldots, d$.*
- *$\forall u \in [0,1]^d$, $C(u) = u_i$, lorsque $u_j = 1$ pour tout $i \neq j$.*
- *C vérifie pour tout $0 \leq a_i \leq b_i \leq 1$ et $\forall u \in [0,1]^d$,*

$$\Delta_{a_1,b_1} \ldots \Delta_{a_d,b_d} C(u) \geq 0,$$
$$\text{où } \Delta_{a_i,b_i} C(u) = C(u_1, \ldots, u_{i-1}, b_i, u_{i+1}, \ldots, u_d) - C(u_1, \ldots, u_{i-1}, a_i, u_{i+1}, \ldots, u_d).$$

Proposition 1 *: Une fonction C définie sur $[0,1]^d$ est une fonction copule **si et seulement si** elle est la restriction à $[0,1]^d$ de la fonction de répartition d'un vecteur aléatoire $(U_1, \ldots, U_d)$ où les U_i ont la distribution uniforme sur $[0,1]$.*

Théorème 1 (Théorème de Sklar) *: Etant donné un vecteur aléatoire de fonction de distribution F et dont les composantes sont les fonctions de répartitions $(F_i)_{i=1,\ldots,d}$ pour $d \geq 2$.*

$$\text{Alors, il existe une copule } C \text{ telle que } \forall x \in \mathbb{R}^d,$$
$$F(x) = C(F_1(x_1), \ldots, F_d(x_d)).$$

De plus, si $F_1(x_1), \ldots, F_d(x_d)$ sont des fonctions continues, alors C est unique. Autrement, C est non-unique sur $[0,1]^d$, mais est unique sur le support de ses marges.

Démonstration : Les variables étant continues, on rappelle que l'on a bien $F_i(\mathbb{R}) = [0,1]$, $\forall u \in [0,1]^d$. Il existe donc bien une fonction copule d'après la défintion directe.
L'unicité résulte du fait que si pour deux fonctions copules C et D, nous avons
$\forall x \in \mathbb{R}^d, C(F_1(x_1), \ldots, F_d(x_d)) = D(F_1(x_1), \ldots, F_d(x_d))$.
Comme $F_i(\mathbb{R}) = [0,1]$, $\forall u_i \in [0,1]$, $\exists x_i \in \mathbb{R}$ tq $F_i(x_i) = u_i$.
Donc $\forall u_1, \ldots, u_d \in [0,1]$, $C(F_1(x_1), \ldots, F_d(x_d)) = D(F_1(x_1), \ldots, F_d(x_d))$, ce qui nous prouve l'unicité.

Définition 6 (Copule d'indépendance) *: On appelle la copule d'indépendance* Π *tel que*

$$\Pi(u_1, \ldots, u_d) = u_1 \ldots u_d.$$

Définition 7 (Copule de survie) *Nous allons ici travailler avec des fonctions de survie S plutôt qu'avec des fonctions de répartition et donc nous intéresser aux copules de survie afin d'établir une relation entre les fonctions de survie multivariées et les marginales.*
Soit $\boldsymbol{S}$ *la copule de survie*

$$\boldsymbol{S}(u_1, u_2) = u_1 + u_2 - C(u_1, u_2).$$

Remarque : Il existe deux familles de copules que l'on retrouve couramment en assurance et en finance : les copules elliptiques et les copules archimédiennes. Nous nous intéresserons ici aux familles de copules archimédiennes.

Famille Archimédienne

La famille archimédienne a été très étudiée par Christian Genest à partir du milieu des années 80 et est aujourd'hui très populaire. Ce succès s'explique par le fait que de nombreuses copules archimédiennes admettent une expression explicite. De plus, elles se généralisent très facilement à des dimensions supérieures à 2 en utilisant le notion de générateur. **Guibert (2020-2021)**

Définition 8 (Famille archimédienne quand d=1) *: Soit* $\phi : [0,1] \mapsto [0,\infty]$ *une fonction de classe* C^2*, strictement décroissante et convexe telle que* $\phi(1) = 0$*. Les copules de la famille archimédienne sont construites de la manière suivante*

$$C(u,v) = \begin{cases} \phi^{-1}(\phi(u) + \phi(v)) & si\ \phi(u) + \phi(v) < \phi(0) \\ 0 & sinon \end{cases}$$

La fonction ϕ *est appelé le générateur de la copule.*

Remarque :
- **quand d≥1** : Les copules archimédiennes sont contruites de la manière suivante :
$$C(u) = \phi^{-1}(\phi(u_1), \ldots, \phi(u_d)),$$
où $\phi : [0,1] \mapsto [0,\infty]$ une fonction continue telle que $\phi^{(i)}(t)(-1)^i \geq 0$ pour tout i.
- La copule archimédienne la plus simple est la copule indépendante $C^{\perp}$ pour $\phi(u) = ln(u)$.
- Les copules archimédiennes les plus connues sont celles de Clayton, Gumbel et Franck.

Proposition 2 *: Soit C une copule archimédienne de générateur* ϕ*. Alors elle vérifie :*
- *Symétrie :* $C(u,v) = C(v,u), \forall u, v \in [0,1]$.
- *Associativité :* $C(C(u,v),w) = C(u,C(v,w)), \forall u, v, w \in [0,1]$.

Le principal défaut des copules archimédiennes est la symétrie entre les variables. En effet, il n'y a, à priori, pas de symétrie entre les différentes variables dont on cherche à modéliser la dépendance.

Pour pallier ce défaut, il est possible d'introduire une structure hiérarchique, en liant, dans un premier temps, les causes semblables avant d'agréger les groupes ainsi constitués. Le regroupement des causes est subjectif. Il existe plusieurs façons de construire une hiérarchie.
On peut trouver une hiérarchie où toutes les copules sont liées appelées **les copules archimédiennes imbriquées** ou bien une hiérarchie où les copules ne sont pas toutes les liées mais sont

regroupées en deux groupes, deux vecteurs, qui sont donc **partiellement imbriqués.**
En dimension d= 5, nous pouvons voir un exemple de la différence entre ces deux structures avec la **figure 1.2**. Pour les copules partiellement imbriquées, le premier vecteur est composé des causes U_1, U_2, U_3 tandis que le deuxième regroupe U_4, U_5. **Nelsen (2006)**

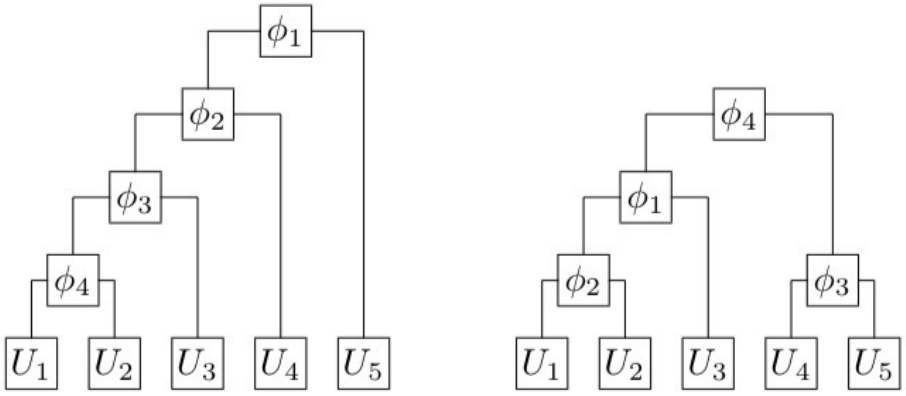


Figure 1.2 – A gauche les copules archimédiennes imbriquées et à droite, les copules archimédiennes partiellement imbriquées.

Symétrie

Rappel : Soit X une variable aléatoire et a un nombre réel, on dit que X est symétrique à a si les fonctions de distributions des variables aléatoires $X - a$ et $a - X$ sont les mêmes, c'est-à-dire si $\forall x \in \mathbb{R}, P(X - a \leq x) = P(a - X \leq x)$.
On dit qu'une copule est symétrique lorsque

$$P(U_1 < u_1 \mid U_2 < u_2) = P(U_2 < u_2 \mid U_1 < u_1)$$

Autrement dit, $C(u_1, u_2) = C(u_2, u_1)$.

1.5 Problème d'identifiabilité

Le but de cette section est de démontrer le lemme dit d'identifiabilité qui exprime la dérivée de la fonction de survie brute $\tilde{S}$ en fonction de la dérivée partielle d'une certaine fonction copule. Pour cela, on aura besoin de deux lemmes préliminaires. **Carriere (1995)**

Lemme 1 *: Si $S(t_1, \dots, t_m)$ est dérivable en chacune de ses coordonnées alors on a :*

$$\tilde{S}(t, j) = \int_t^{+\infty} -S_j(r, \dots, r)\ \mathrm{d}r$$

Où $S_j(r, \dots, r) = \partial t_j S(r, \dots, r)$.

Démonstration : Par définition de $\tilde{S}$ on a, $\tilde{S}(t, j) = P(T > t, J = j)$.
Or, $P(T > t, J = j) = P([\forall k \in [\![1\ ,\ m]\!], T_k > t \text{ et } T_j \leq T_k])$ J étant l'indice associé à la cause de décès, avoir $\forall k \in [\![1\ ,\ m]\!], T_k > t$ et $T_j \leq T_k$ est équivalent à avoir $T_j > t$ et $\forall k \neq j, T_k > T_j$.

En supposant l'absolue continuité de S, on peut affirmer qu'il existe une fonction f multivariée telle que :

$$S(t_1, \dots, t_m) = \int_{t_1}^{+\infty} \dots \int_{t_m}^{+\infty} f(t_1, \dots, t_m)\ \mathrm{d}t_1 \dots\ \mathrm{d}t_m$$

Avec ceci il vient :

$$\tilde{S}(t, j) = \int_t^{+\infty} (\int_{t_1}^{+\infty} \dots \int_{t_m}^{+\infty} f(t_1, \dots, t_m) \prod_{i \neq j} \mathrm{d}t_i)\ \mathrm{d}t_j$$

On reconnaît l'opposé de la dérivée partielle de S en t_j ; on obtient donc le résultat suivant :

$$\tilde{S}(t, j) = \int_t^{+\infty} -S_j(r, \dots, r)\ \mathrm{d}r$$

Lemme 2 *: On pose $C(u) = P(\bigcap_{j=1}^m S_j(T_j) \leq u_j)$ pour $u = (u_1, \dots, u_m)$ alors,*
C est l'unique copule telle que $S(t_1, \dots, t_m) = C(S_1(t_1), \dots, S_m(t_m))$.

Démonstration : Pour démontrer ce résultat on utilise le théoreme de Sklar énoncé plus haut. S_j étant continue, $S_j(T_j)$ suit une loi uniforme sur $[0, 1]$ et avec la proposition 1, C est une copule.
Démontrons à présent que $S(t_1, \dots, t_m) = C(S_1(t_1), \dots, S_m(t_m))$:

$$C(S_1(t_1), \dots, S_m(t_m)) = P(\bigcap_{j=1}^m S_j(T_j) \leq S_j(t_j)) = P(\bigcap_{j=1}^m T_j > t_j) = S(t_1, \dots, t_m)$$

En effet, on a que presque sûrement $\{S_j(T_j) \leq S_j(t_j)\}$=$\{T_j > t_j\}$.

On reprend dans cette section la fontion C définie précédement.

Théorème 2 *: Si C est différentiable en chacune de ses coordonnées, et S_j dérivable, alors :*

$$\frac{d\tilde{S}}{dt}(t,j) = C_j(S_1(t),\dots,S_m(t)) \times \frac{dS_j}{dt}(t)$$

Avec $C_j(u_1,\dots,u_m) = \partial_{u_j} C(u)$.

Démonstration : Avec le lemme 1, $\tilde{S}(t,j) = \displaystyle\int_t^{+\infty} -S_j(r,\dots,r)\,\mathrm{d}r$, donc en dérivant on obtient $\dfrac{d\tilde{S}}{dt} = S_j(t,\dots,t)$. Or, avec le lemme 2, $S(t_1,\dots,t_m) = C(S_1(t_1),\dots,S_m(t_m))$ donc on a :

$$S_j(t,\dots,t) = \partial_{t_j} S(t,\dots,t) = \partial_{u_j} C(S_1(t),\dots,S_m(t)) \times \tfrac{dS_j}{dt}(t)$$

Par la dérivation d'une composée de fonctions différentiables. D'où le résultat.

1.6 Application du thèorème d'identifiabilité aux copules archimédiennes

Dans cette section, on va appliquer les résultats du théorème d'identifiabilité à deux familles de copules archimidiennes qui sont les copules de Clayton et les copules de Franck. On va présenter une formule pratique du théorème pour cette famille de copules dépendant uniquement des taux de mortalité observables et des fonctions génératrices de nos copules.

Le théorème d'identiabilité est une méthode permettant de relier la fonction de survie brute et la fonction de survie nette, en résolvant une équation différentielle faisant intervenir le copule survie $\mathbf{S}$. Le théorème a été démontré plus haut. On rappelle l'équation (sous réserve d'avoir les bonnes hypothèses) :

$$\frac{d\tilde{S}}{dt}(t,j) = \mathbf{S}_j(S_1(t),\dots,S_m(t)) \times \frac{dS_j}{dt}(t)$$

Avec $\mathbf{S}_j(u) = \partial_{u_j}\mathbf{S}(u)$.
On ne peut néanmoins pas résoudre cette équation de manière exacte, on ne peut que l'approcher. Cependant, en présence de copules archimédiennes, nous avons une forme close qui nous permet, à partir des taux de mortalité estimés grâce aux données, d'obtenir une expression de la fonction de survie nette. **Li et Lu (2018)**.

Théorème 3 *Soit m le nombre de causes de décès, $(\mu_1,\dots,\mu_m)$ les taux de mortalités relatifs à ces causes, et ψ la fonction génératrice archimédienne, alors on a :*

$$S_j(t) = \psi\left[-\int_0^t \frac{exp(-\int_0^s \sum_{i=1}^m \mu_i(u)du)}{\psi' \circ \psi^{-1}(exp(-\int_0^s \sum_{i=1}^m \mu_i(u)du))}\mu_j(u)ds\right]$$

Copule de Clayton

On définit le copule de Clayton en prenant comme fonction génératrice archimédienne, pour $\theta \geq 0$:

$$\psi : t \longmapsto 1/(1+t)^\theta$$

De ce fait, on a : $\quad \forall (u_i)_{i\in[\![1\ ;\ m]\!]} \in [0,1]^m \quad \mathbf{S}(u_1,\dots,u_m) = \sum_{i=1}^m (u_i^{-\frac{1}{\theta}} - m + 1)^\theta$

Démonstration : ψ étant inversible par construction, on a que $\psi^{-1}(u) = (1/u)^{\frac{1}{\theta}} - 1$,
donc $\sum_{i=1}^{m} \psi^{-1}(u_i) = \sum_{i=1}^{m} (u_i)^{\frac{-1}{\theta}} - m$.
Et en y appliquant la fonction ψ, on a alors $(\sum_{i=1}^{m} u_i^{-\frac{1}{\theta}} - m + 1)^{-\theta}$.

Or on a démontré précédemment que, en considérant pour $u = (u_i), C(u) = P(\bigcap_{j=1}^{m} S_j(T_j) \leq u_j)$
,
C est l'unique copule tel que $S(t_1, \ldots, t_m) = C(S_1(t_1), \ldots, S_m(t_m))$.
D'où le résultat.

En utilisant le fait que pour une fonction f bijective dérivable d'inverse dérivable $(f^{-1})' = \frac{1}{f' \circ f^{-1}}$
il vient :

$$\psi' \circ \psi^{-1} : u \longmapsto -\theta u^{\frac{1}{\theta}+1}$$

et donc :

$$S_j(t) = \psi \left(\int_0^t \theta \; exp(- \int_0^s \sum_{i=1}^{m} \mu_i(u)du) \mu_j(u)ds \right)$$

Copule de Franck

La copule de Franck est définie en considérant comme fonction génératrice archimédienne :

$$\psi : t \longmapsto (-1/\theta) \, log(1 + e^{-t}(e^{-\theta} - 1))$$

Ainsi on a comme fonction copule :

$$\forall (u_i)_{i \in [\![1 \; ; \; m]\!]} \in [0,1]^m, \quad \mathbf{S}(u_1, \ldots, u_m) = (-1/\theta) \; log \left(1 + \frac{(e^{-\theta u_1} - 1) \ldots (e^{-\theta u_m} - 1)}{(e^{-\theta} - 1)^{m-1}} \right)$$

Démonstration : Par construction, ψ est inversible et on a : $\forall u \in [0,1], \; \psi^{-1}(u) = -log \left(\frac{e^{\theta u} - 1}{e^{-\theta} - 1} \right)$,
donc en considérant $(u_i) \in [0,1]^m$, il vient : $\sum_{i=1}^{m} \psi^{-1}(u_i) = -log \left(\frac{\prod_{i=1}^{m} e^{-\theta u_i} - 1}{(e^{-\theta} - 1)^m} \right)$.

En appliquant à cette expression la fonction ψ on obtient : $(-1/\theta) \, log \left(1 + \frac{(e^{-\theta u_1} - 1) \ldots (e^{-\theta u_m} - 1)}{(e^{-\theta} - 1)^{m-1}} \right)$.

En réutilisant le même raisonnement que pour le copule de Clayton il vient l'égalité.
Puis en remplaçant dans le **théorème 3**, on obtient les formules suivantes :

$$\psi' \circ \psi^{-1} : u \longmapsto \frac{1}{\theta}(1 - e^{\theta u})$$

et donc :

$$S_j(t) = \psi \left(- \int_0^t \theta \; \frac{exp(- \int_0^s \sum_{i=1}^{m} \mu_i(u)du)}{1 - exp(- \int_0^s \sum_{i=1}^{m} \mu_i(u)du)} \mu_j(s))ds \right)$$

2 Présentation des données

2.1 Présentation

Dans un premier temps, il nous faut fixer quelles sont les données à utiliser. Nous avons ici deux possibilités, la première étant de prendre les données du site *NBER* : "Mortality Data – Vital Statistics NCSHS', Multiple Cause of Death Data, 1959-2017". Ces données sont "brutes" et contiennent un grand nombre d'informations et de variables, exposant les données de mortalités en France sur un grand nombre d'années et faisant le détail de chaque décès un par un. L'avantage de ces données est que les tables sont très bien documentées et qu'il est plutôt facile de s'y retrouver.

De plus, on peut y trouver plusieurs types de classification pour les décès. La plus courante, la classification ICD-10 de l'OMS est très précise, voir peut-être un peu trop précise si l'on souhaite étudier des causes de mortalité globales comme le cancer par exemple. On trouve également des classifications plus grossières codées sur respectivement 358 causes, 113 causes, 130 causes et 58 causes. Cela peut s'avérer utile pour la suite de notre étude. Cependant, ces données sont très volumineuses et donc difficile à traiter tant au niveau du nombre d'observations que du nombre de variables. De plus les tableaux sont séparés par année et il faudrait donc rassembler plusieurs de ces fichiers pour obtenir des données sur plusieurs années.

La deuxième option consiste à utiliser les données de *Causeofdeath.org* : "The Human Cause-of-death Database". Ces données ont déjà été traitées et sont bien moins volumineuses. Elles présentent uniquement le nombre de décès pour chaque cause de mortalité selon l'âge, le sexe et l'année. Les décès y sont répartis par tranches d'âges à la manière d'une table de mortalité. Il semble alors assez facile de faire une étude sur plusieurs années. Malheureusement, les données sont un peu moins bien documentées et sont seulement répartis sur 15 années. Dans un premier temps, ces données ont l'air d'être plus facilement exploitable, c'est pourquoi on commencera par faire les premiers tests sur celles-ci, avant de revenir éventuellement sur les données de *NBER* au besoin.

2.2 Exploitation des données

Nous allons tout d'abord retravailler quelque peu nos données et commencer par quelques tests simples. Celles-ci se présentent de cette façon (voir **figure 2.1**) :
On peut faire une analyse des différentes colonnes :

- "year" correspond à l'année des décès, c'est un entier compris entre 2000 et 2015
- "sex" correspond aux sexes des individus décédés, 1 correspond à un homme, 2 à une femme, et 3 prend en compte les deux sexes.
- "cause" correspond à la cause des décès, c'est un entier compris entre 0 et 16 (voir ci-dessous)
- "m0","m1","m5"... correspond aux nombre de décès pour différents intervalles d'âge respectivement : 0-1 ans,1-5ans, 5-10ans...

Par exemple, on peut ici prendre la cause de mortalité numéro 2, c'est à dire les néoplasmes ou autrement dit les cancers. On obtient à partir de ces données un tableau contenant le nombre de décès pour cette cause par tranche d'âge et par année. On peut également en tracer un histogramme sur la période.

country	year	sex	list	agf	cause	total	d0	d1	d5	d10	d15	d20	d25	d30	d35	d40	d45	d50	d55	d60	d65
FRATNP	2000	1	short	3	0	272043	1992	401	267	374	1342	1958	2314	2709	3943	6374	9782	13615	12962	17974	26844
FRATNP	2000	1	short	3	1	5757.26	78.92	27.77	2.14	8.33	7.45	12.11	55.23	137.38	295.78	294.87	237.85	261.56	235.23	315.94	469.18
FRATNP	2000	1	short	3	2	94230.24	19.99	67.76	69.42	56.26	123.45	133.25	208.52	319.04	663.84	1792.08	3876.37	6421.51	6547.46	9166.34	12918.05
FRATNP	2000	1	short	3	3	1040.84	16.84	22.22	7.48	8.33	6.39	8.81	9.02	11.35	10.01	16.32	31.07	37.06	29.54	37.54	78.89
FRATNP	2000	1	short	3	4	8296.02	46.3	13.33	8.54	10.42	8.51	14.32	27.05	22.71	57.82	99.02	172.5	280.62	296.41	432.72	805.5
FRATNP	2000	1	short	3	5	7429.72	0	3.33	0	0	5.32	30.83	50.72	110.13	200.15	348.19	505.71	452.17	369.2	361.82	450.5
FRATNP	2000	1	short	3	6	9149.77	65.24	33.32	19.22	28.13	64.92	56.16	64.25	99.91	128.99	175.18	240	293.33	251.05	374.33	687.16
FRATNP	2000	1	short	3	7	54789.37	44.2	8.89	7.48	10.42	29.8	45.15	71.01	120.35	321.36	739.9	1202.13	1878.59	1848.1	2815.28	4695.96
FRATNP	2000	1	short	3	8	16833.8	13.68	3.33	3.2	6.25	8.51	22.02	30.43	36.33	84.51	155.6	278.57	366.4	413.5	677.75	1247.69
FRATNP	2000	1	short	3	9	7511.96	0	1.11	0	0	4.26	6.61	9.02	21.57	38.92	87.05	148.93	225.56	270.04	428.55	678.86
FRATNP	2000	1	short	3	10	7662.32	11.58	6.66	0	4.17	5.32	4.4	7.89	14.76	12.23	44.61	78.21	113.31	130.8	191.86	353.96
FRATNP	2000	1	short	3	11	11649.23	12.63	6.66	4.27	6.25	12.77	18.72	16.91	26.11	40.03	65.29	127.5	256.27	283.75	514.05	1083.68
FRATNP	2000	1	short	3	12	13252.7	26.31	6.66	1.07	3.13	4.26	9.91	33.81	78.34	220.17	512.49	841.06	1141.55	969.41	1236.64	1550.79
FRATNP	2000	1	short	3	13	1935.77	2.1	3.33	2.14	1.04	3.19	3.3	1.13	5.68	14.46	13.06	18.21	37.06	45.36	63.6	121.45
FRATNP	2000	1	short	3	14	3850.55	9.47	0	0	1.04	0	3.3	7.89	7.95	10.01	16.32	37.5	56.12	72.78	136.59	236.67
FRATNP	2000	1	short	3	15	2098.37	1568.98	41.1	21.36	26.04	24.48	22.02	31.56	26.11	25.58	31.55	36.43	38.12	32.7	28.15	26.99
FRATNP	2000	1	short	3	16	26555.07	75.77	155.51	120.68	204.19	1033.37	1567.06	1689.57	1671.27	1819.16	1982.49	1949.97	1755.75	1166.66	1192.84	1438.68
FRATNP	2001	1	short	3	0	272274	1958	433	257	368	1429	1992	2221	2684	4027	6217	9903	14005	13750	17187	25482
FRATNP	2001	1	short	3	1	5448.88	35.73	21.17	8.75	5.26	15.18	15.61	30.44	113.94	291.6	291.03	244.1	278.95	245.59	281.91	392.09
FRATNP	2001	1	short	3	2	95306.89	16.82	74.11	65.62	75.7	105.17	144.99	184.9	326.87	640.84	1671.48	3845.98	6515.93	7049.14	8749.65	12475.08
FRATNP	2001	1	short	3	3	1041.37	10.51	10.59	4.37	5.26	6.51	6.69	2.25	4.6	21.47	19.77	26.04	38.33	33.87	54.5	74.04
FRATNP	2001	1	short	3	4	8554.16	39.94	24.35	12.03	16.82	13.01	21.19	29.31	27.62	65.55	118.61	197.45	316.21	308.05	430.72	769.59
FRATNP	2001	1	short	3	5	7735.74	0	0	1.09	0	8.67	34.58	48.48	110.49	219.26	316.29	482.78	564.29	424.49	376.23	402.52
FRATNP	2001	1	short	3	6	9792.91	72.52	45.52	14.22	32.59	67.22	78.07	66.52	85.17	136.76	197.68	240.85	316.21	302.76	411.86	665.31
FRATNP	2001	1	short	3	7	54219.29	29.43	6.35	6.56	15.77	27.11	29	82.3	138.11	335.68	712.74	1283.44	1798.27	1871.58	2612.63	4318.26
FRATNP	2001	1	short	3	8	16678.94	8.41	3.18	2.19	3.15	8.67	22.31	28.19	41.43	88.16	168.03	282.07	396.07	428.73	613.07	1220.08
FRATNP	2001	1	short	3	9	7400.51	3.15	0	0	1.05	3.25	11.15	5.64	31.08	32.78	70.29	164.91	251.27	262.53	363.65	688.25
FRATNP	2001	1	short	3	10	6578.18	15.76	3.18	1.09	3.15	0	3.35	6.76	16.11	28.26	51.62	81.37	110.73	109.03	186.54	296.16
FRATNP	2001	1	short	3	11	11403.52	5.25	10.59	4.37	5.26	9.76	15.61	21.42	28.77	38.43	71.38	135.61	231.04	295.35	505.13	930.18
FRATNP	2001	1	short	3	12	13732.81	24.17	3.18	3.28	6.31	8.67	6.69	22.55	77.11	212.48	548.01	875.52	1253.15	1028.95	1200.99	1519.37
FRATNP	2001	1	short	3	13	2105.08	0	1.06	0	3.15	4.34	1.12	4.51	5.75	5.65	10.98	22.78	40.46	53.99	53.45	103.24
FRATNP	2001	1	short	3	14	3822.84	3.15	0	0	0	1.08	2.23	6.76	4.6	11.3	15.38	29.29	51.11	68.81	117.37	213.78
FRATNP	2001	1	short	3	15	2125.39	1625.89	47.64	20.78	21.03	27.11	16.73	16.91	31.08	21.47	21.96	34.72	39.39	37.05	32.49	38.58

FIGURE 2.1 – Données brutes du site "causeofdeath.org"

No.	Title	Category codes according to ICD10
0	All causes	A00-Y98
1	Certain infectious diseases	A00-B99
2	Neoplasms	C00-D48
3	Diseases of the blood and blood-forming organs	D50-D89
4	Endocrine, nutritional and metabolic diseases	E00-E90
5	Mental and behavioral disorders	F00-F99
6	Diseases of the nervous system and the sense organs	G00-G44, G47-H95
7	Heart diseases	I00-I52
8	Cerebrovascular diseases	G45, I60-I69
9	Other and unspecified disorders of the circulatory system	I70-I99
10	Acute respiratory diseases	J00-J22, U04
11	Other respiratory diseases	J30-J98
12	Diseases of the digestive system	K00-K93
13	Diseases of the skin and subcutaneous tissue, musculoskeletal system and connective tissue	L00-M99
14	Diseases of the genitourinary system and complications of pregnancy, childbirth and puerperium	N00-O99
15	Certain conditions originating in the perinatal period and congenital malformations/anomalies	P00-Q99, R95
16	External causes	V01-Y98

FIGURE 2.2 – Les 16 causes répertoriées

Nous pouvons également construire une table de mortalité sur une année donnée. Avec l_x le nombre de personnes ayant survécu à l'âge x, d_x le nombre de personnes décédées dans la tranche d'âge, p_x la probabilité empirique de survie jusqu'à l'âge $x+5$ sachant que la cause de décès est le cancer, q_x la probabilité empirique de mourir avant l'âge $x+5$ sachant que la cause de décès est le cancer.

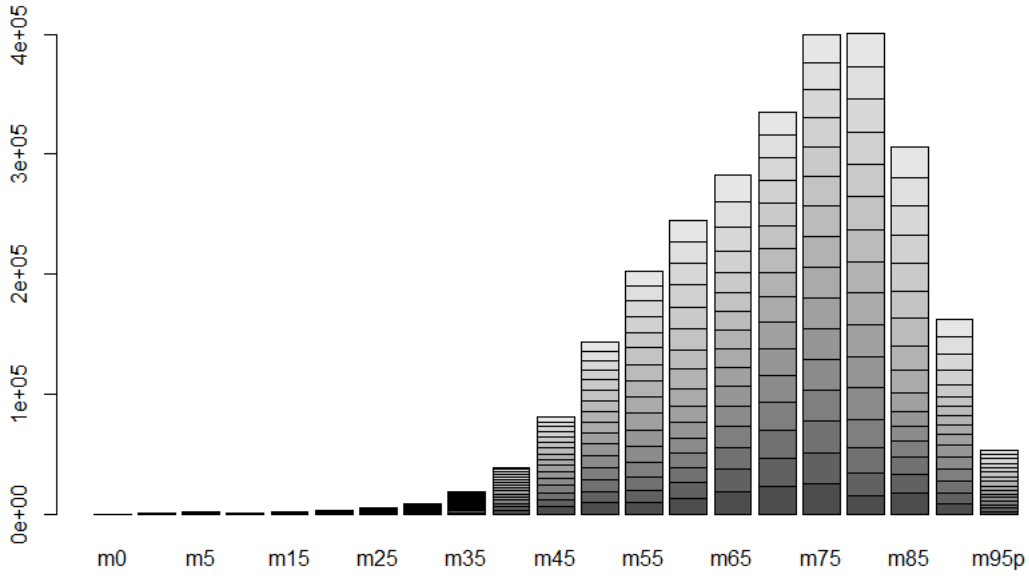


FIGURE 2.3 – Nombre de décès du cancer par tranche d'âge sur la période 2000-2015

	lx	dx	px	qx
m0	181879	19	0.9999	0.0001
m1	181860	92	0.9994	0.0006
m5	181768	123	0.9987	0.0013
m10	181645	103	0.9981	0.0019
m15	181542	136	0.9974	0.0026
m20	181406	199	0.9963	0.0037
m25	181207	304	0.9946	0.0054
m30	180903	586	0.9914	0.0086
m35	180317	1005	0.9859	0.0141
m40	179312	2038	0.9747	0.0253
m45	177274	4054	0.9524	0.0476
m50	173220	7510	0.9111	0.0889
m55	165710	12617	0.8417	0.1583
m60	153093	17536	0.7453	0.2547
m65	135557	22113	0.6237	0.3763
m70	113444	19244	0.5179	0.4821
m75	94200	22832	0.3924	0.6076
m80	71368	27386	0.2418	0.7582
m85	43982	24914	0.1048	0.8952
m90	19068	15118	0.0217	0.9783
m95p	3950	3950	0.0000	1.0000

FIGURE 2.4 – Table de mortalité de l'année 2015 pour les néoplasmes

La procédure est évident facilement reproductible pour chaque cause de mortalité (entre 0 et 16) et pour chaque année entre 2000 et 2015. Il suffit d'utiliser le même programme en changeant les valeurs initiales. Enfin, grâce aux données du HMD, nous avons pu récupérer un tableau des expositions aux risques pour les anneés 2000 à 2015, en voici un graphique démonstratif pour l'année 2015.

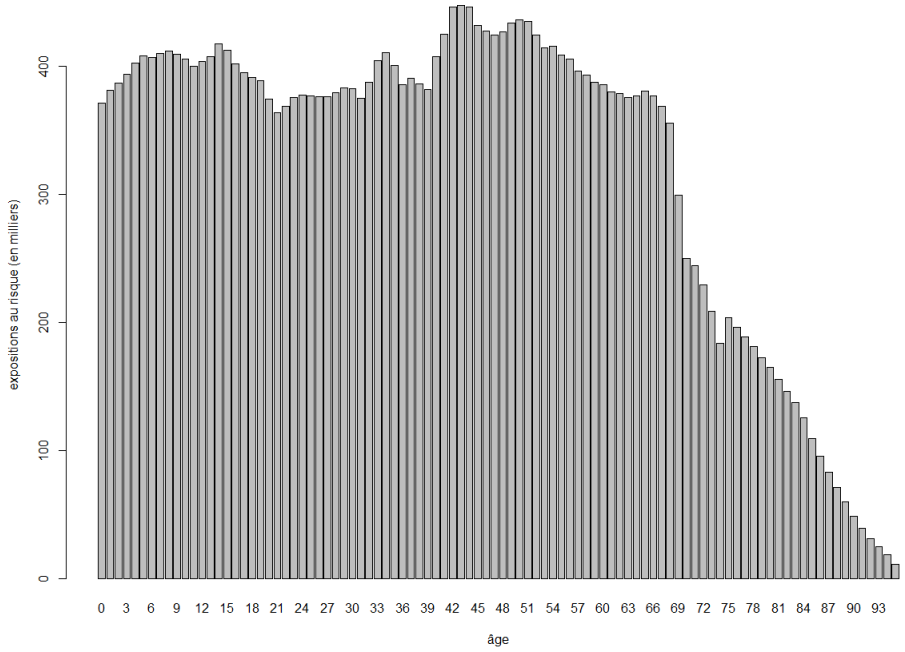


FIGURE 2.5 – Graphique des expositions aux risques par âge, pour l'année 2015.

2.3 Statistiques descriptives par cause

Nous avons considéré les 16 causes de décès qui étaient énumérées dans la short list et nous les avons regroupées, afin de nous faciliter la tâche, en 5 causes. Les 5 retenues sont les décès dûs : aux différents types de cancers ; à un dysfonctionnement de l'appareil circulatoire (par cela nous entendons : le cœur, les artères,les veines...) ; à un dysfonctionnement de l'appareil respiratoire ou du système digestif,à un événement externe (accident grave) ; la dernière cause regroupe les causes restantes telles que les maladies touchant le système nerveux ou encore celles de la peau.

Notre choix a été motivé par les proportions de morts concernant ces causes. En effet, les 4 plus grandes causes de mortalité en France selon wikipédia sont les 4 que nous avons énoncées plus haut, elles représentent les deux tiers des décès en France. Il ne reste donc plus qu'à déterminer les nombres de morts totaux pour chacune de ces causes.
Nous présentons ici les graphiques du nombre de morts pour chaque tranche d'âge, obtenus pour les individus de sexe masculin et cela pour l'année 2015, ainsi que des graphiques comparatifs des taux de mortalité pour les hommes et les femmes selon chaque cause.

Décès liés aux cancers

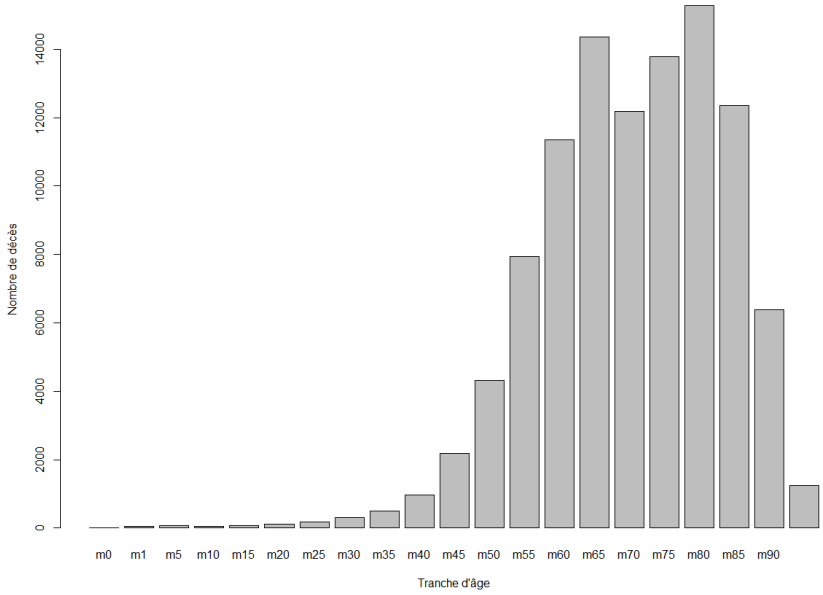


FIGURE 2.6 – Hommes, 2015. La moyenne de l'âge du décès est de 72,81 ans pour les hommes et de 75,46 ans pour les femmes. Le total de décès sur l'année 2015 s'élève à 181 878.

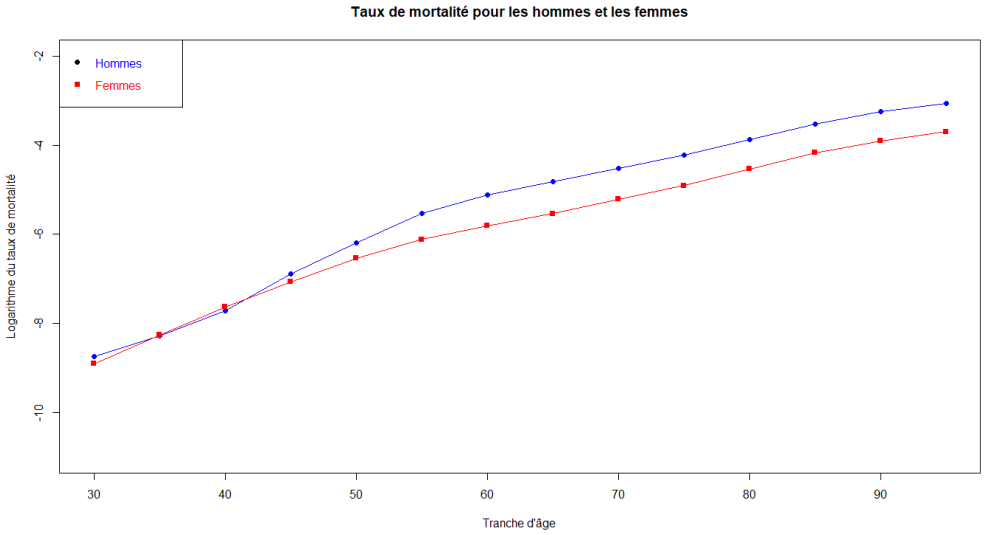


FIGURE 2.7 – Comparaison hommes et femmes, 2015.

Décès liés aux maladies de l'appareil circulatoire et accidents vasculaires cérébraux

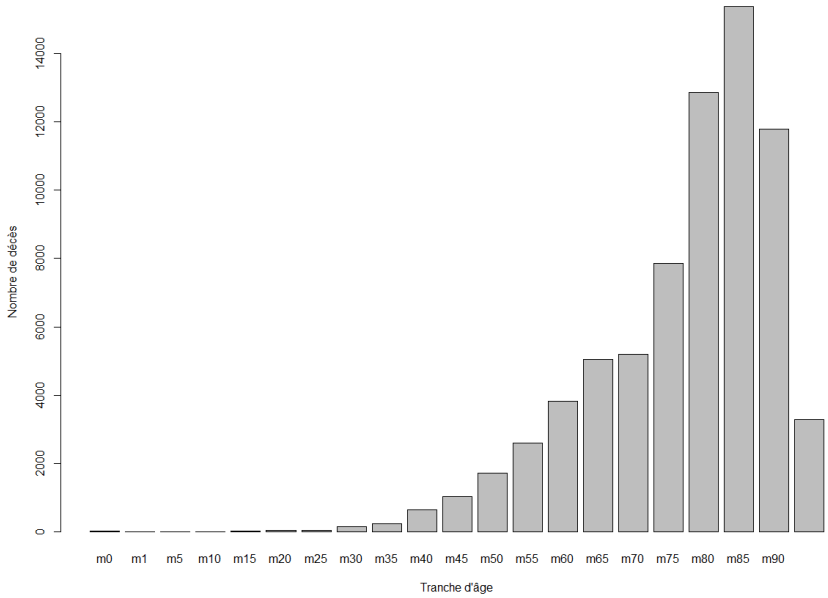


FIGURE 2.8 – Hommes, 2015. La moyenne de l'âge du décès est de 79,57 ans pour les hommes et de 86,91 ans pour les femmes. Le total de décès sur l'année 2015 s'élève à 157 386.

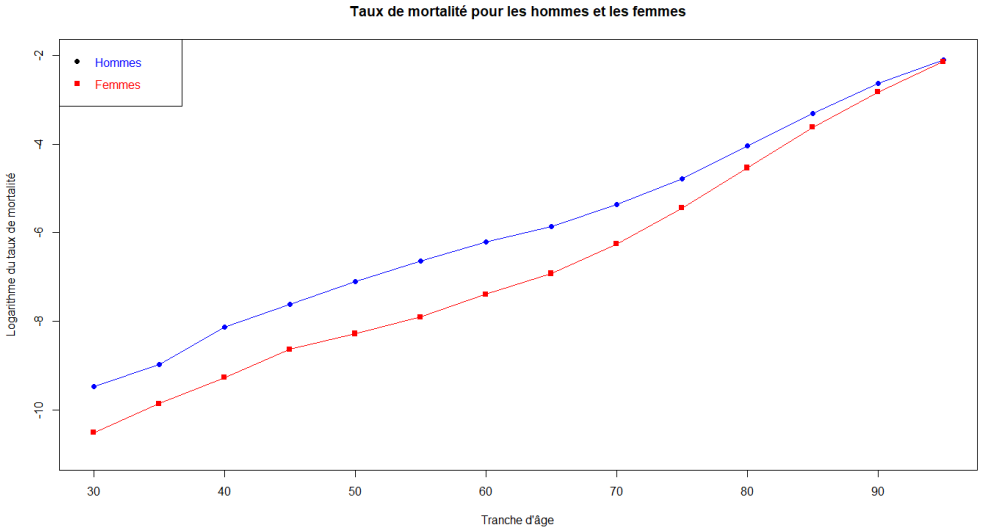


FIGURE 2.9 – Comparaison hommes et femmes, 2015.

Décès liés aux maladies de l'appareil respiratoire ou du système digestif

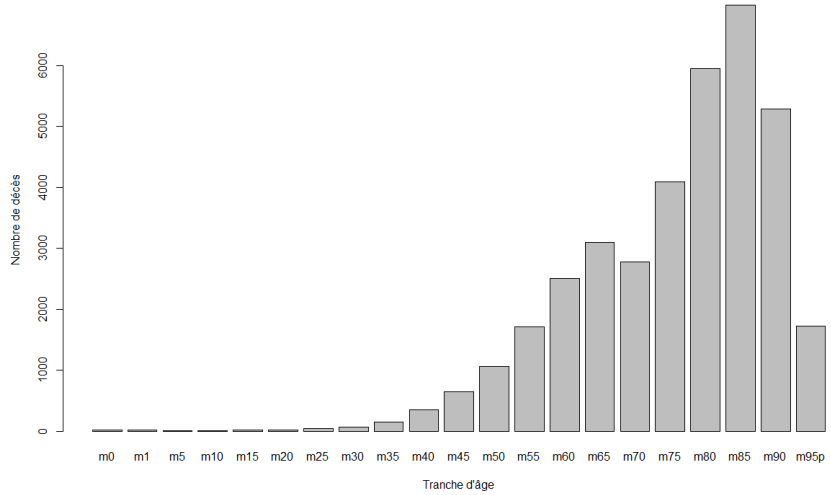


FIGURE 2.10 – Hommes, 2015. La moyenne de l'âge du décès est de 78,07 ans pour les hommes et de 84,89 ans pour les femmes. Le total de décès sur l'année 2015 s'élève à 72 417.

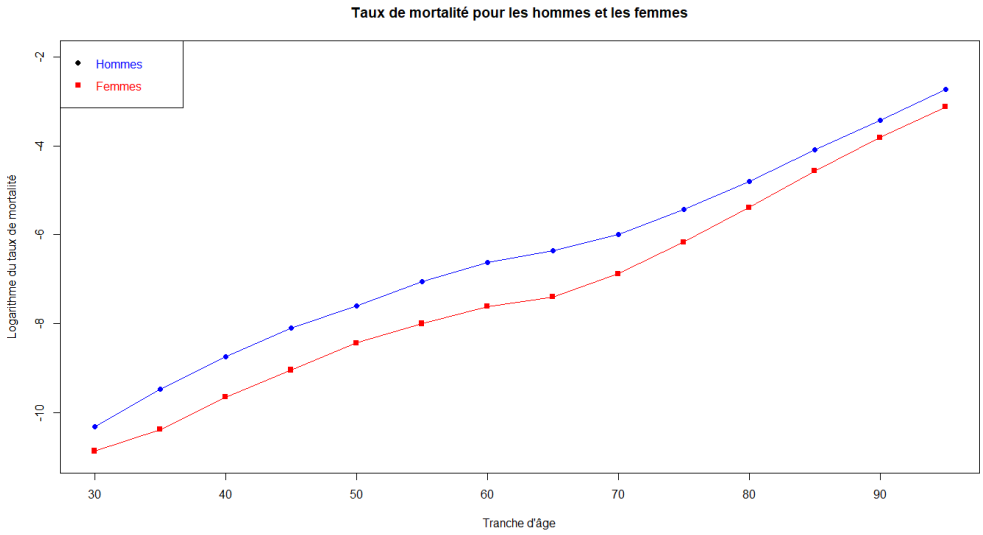


FIGURE 2.11 – Comparaison hommes et femmes, 2015.

Décès liés à des causes externes (accidents...)

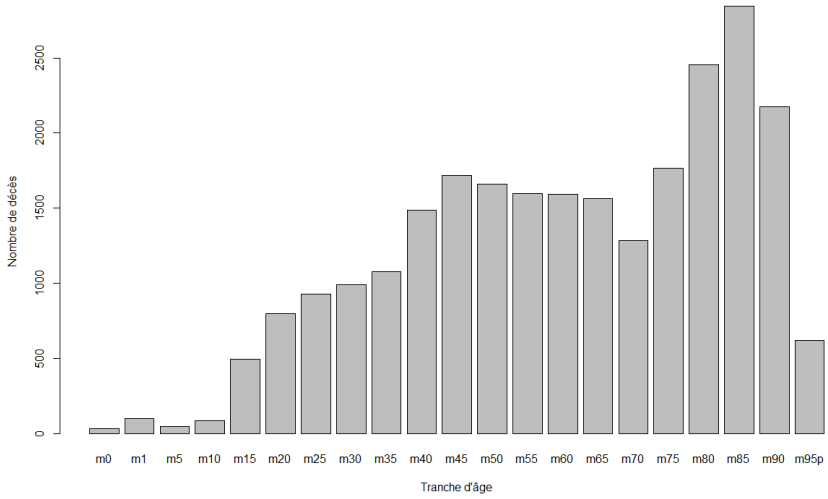


FIGURE 2.12 – Hommes, 2015. La moyenne de l'âge du décès est de 63,14 ans pour les hommes et de 78,15 ans pour les femmes. Le total de décès sur l'année 2015 s'élève à 42 834.

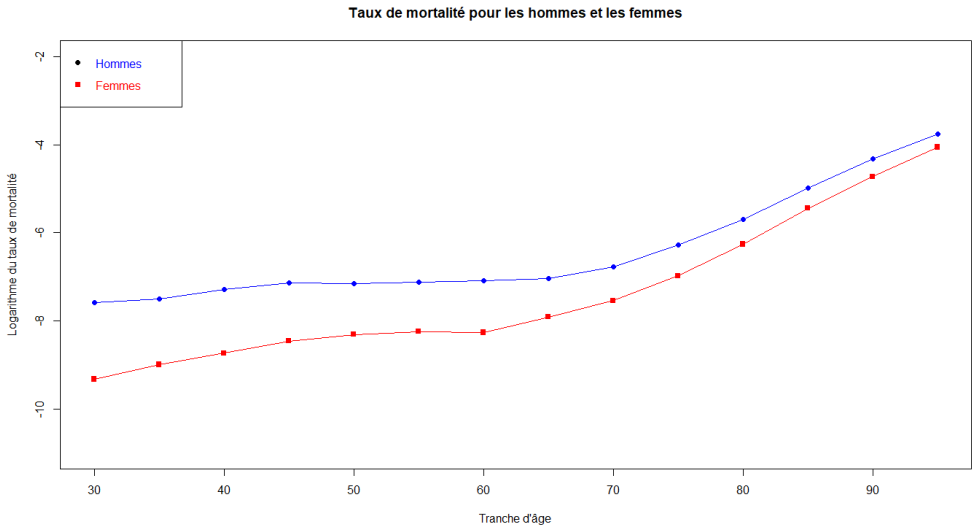


FIGURE 2.13 – Comparaison hommes et femmes, 2015.

Décès liés à d'autres causes

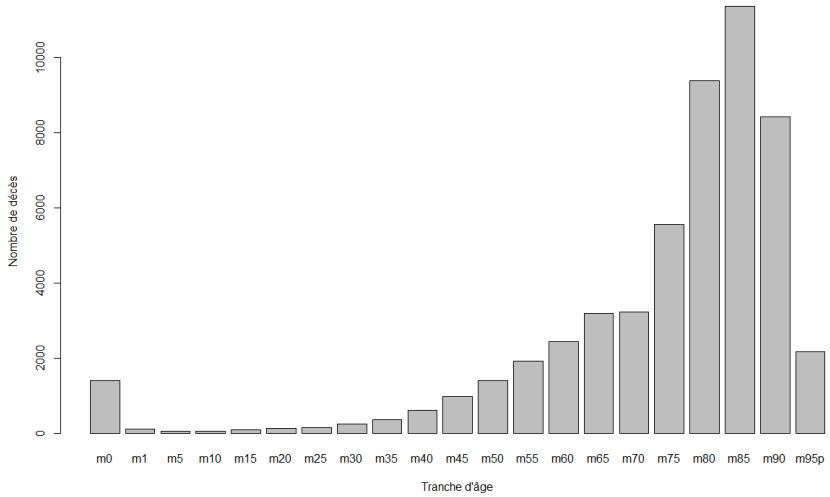


FIGURE 2.14 – Hommes, 2015. La moyenne de l'âge du décès est de 76,61 ans pour les hommes et de 84,31 ans pour les femmes. Le total de décès sur l'année 2015 s'élève à 127 253.

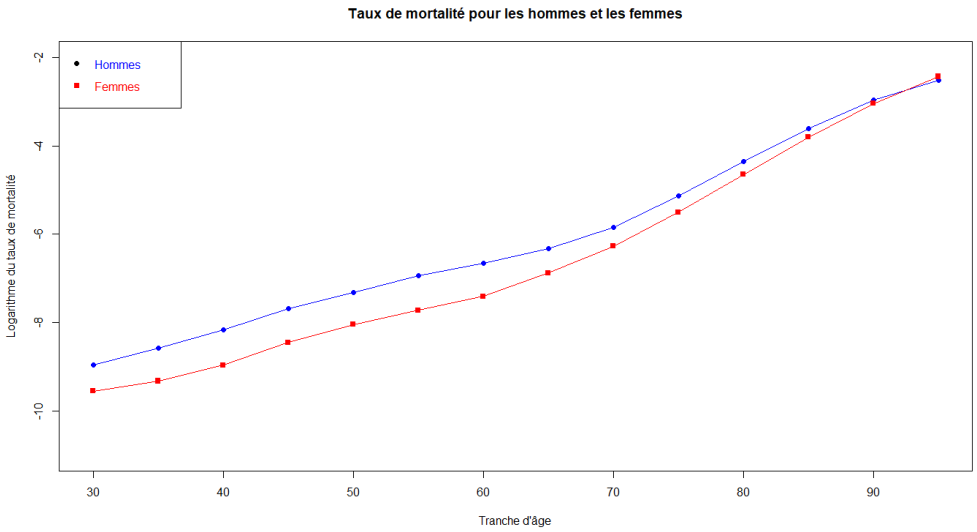


FIGURE 2.15 – Comparaison hommes et femmes, 2015.

Tableau récapitulatif

Causes de décès	Age moyen de décès (hommes)	Age moyen de décès (femmes)	Total de décès	Pourcentage (%)
Néoplasmes	72,81	75,46	181878	31
Maladies de l'appareil circulatoire et AVC	79,57	86,91	157386	27
Maladies de l'appareil respiratoire ou digestif	78,07	84,89	72417	12
Causes externes	63,14	78,15	42834	7
Autres causes	76,61	84,31	127253	22
Toutes causes confondus	75	82	581768	100

On s'aperçoit ici que les décès liés au cancer touchent en moyenne des gens beaucoup plus jeunes (8-10 ans plus jeunes) que les décès liés à des dysfonctionnements ou maladies des organes vitaux. Nos deux premières causes sont d'ailleurs de loin les 2 causes principales de décès en France en 2015, avec pratiquement 60% des décès dû à celles-ci. On peut aussi remarquer que les femmes vivent en moyenne 7 ans de plus que les hommes, cela se traduit sur les graphiques des taux de mortalité où l'on voit une progression "parallèle" des 2 courbes mais avec un delta plus ou moins conséquent.

Ce delta a l'air d'être moins important pour les néoplasmes **(figure 2.7)**, notamment pour les cancers précoces (moins de 50 ans), c'est à dire que les hommes et les femmes peuvent mourir du cancer à des âges assez similaires. D'autre part, les décès par accident et causes externes ont un delta très élevé avec pas moins de 15 ans d'écart en moyenne pour l'âge du décès **(figure 2.13)**.

En effet, les hommes ont une plus grosse tendance au risque, par exemple sur le choix de leurs métiers qui sont souvent plus dangereux que celui des femmes. Militaires, pompiers, gendarmes sont des professions où les taux du nombre d'hommes par rapport aux femmes sont encore très élevés (environ 85% contre 15%). Cependant cette analyse est à nuancer car les raisons de cet écart sont certainement multiples, et mériterait une recherche plus approfondie.

2.4 Interpolation par âge entier

Comme nous n'avons les valeurs des nombres de décès que pour des intervalles de 5 ans, nous avons choisi de faire une interpolation afin d'avoir les taux de mortalité pour chaque année de décès. Nous avons choisi de faire cette interpolation avec la loi de **Gompertz** (**1.1**). Pour une année donnée, nous avons pris deux taux de mortalité dont les âges sont divisible par 5 afin de résoudre un système pour trouver les paramètres b et c. Cela nous conduit au système suivant : $\forall x_1, x_2 \in \mathbb{R}$ divisible par 5 et tel que $x_1 < x_2$,

$$\begin{cases} \mu_{x_1} = bc^{x_1} \\ \mu_{x_2} = bc^{x_2} \end{cases}$$

Autrement dit,

$$\begin{cases} b = (\mu_{x_1}^{x_2} / \mu_{x_2}^{x_1})^{1/(x_2 - x_1)} \\ c = (\mu_{x_2} / \mu_{x_1})^{1/(x_2 - x_1)} \end{cases}$$

Nous avons ensuite calculé les taux de mortalité entre ces deux points avec ces paramètres là grâce à la formule de la loi de Gompertz. Et nous avons appliqué ce processus aux autres années pour obtenir les taux de mortalité pour chaque année et pour chaque âge.
Il est par la suite facile d'obtenir les nombres de décès interpolés grâce aux tableaux des expositions aux risques et à la formule liant ces 3 paramètres.
Voici un aperçu de toutes ces valeurs :

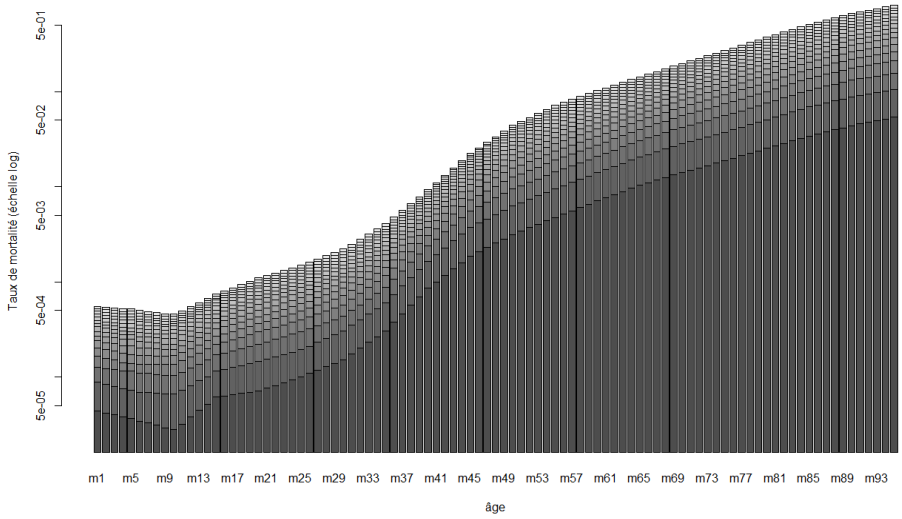


FIGURE 2.16 – Les taux de mortalité du cancer pour les hommes après interpolation (échelle log).

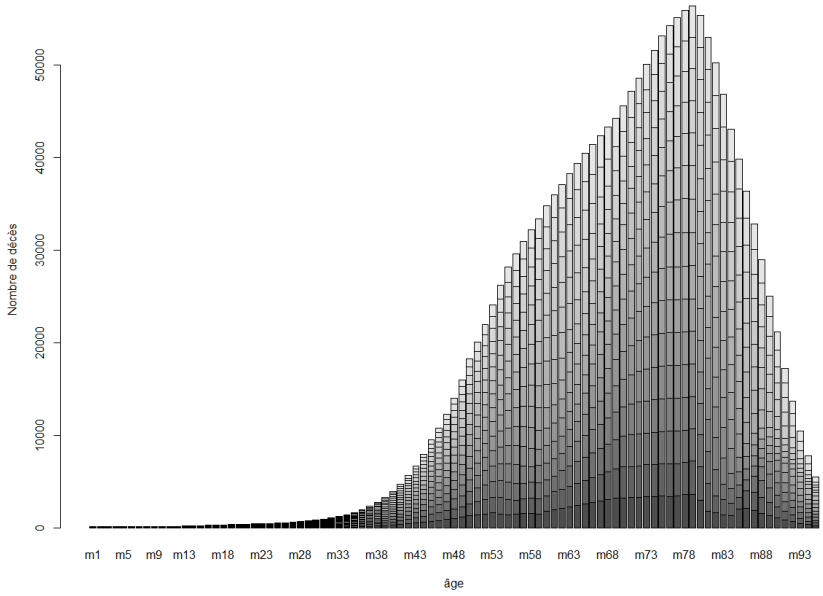


FIGURE 2.17 – Le nombre de décès dû au cancer pour les hommes après interpolation.

3 Modélisation de la mortalité par cause avec des copules

3.1 Modèle de Lee-Carter avec une copule indépendante

Nous allons faire dans un premier temps l'hypothèse simple que chaque cause de mortalité est indépendante l'une de l'autre. Dans ce cas, les taux nets de mortalité pour chaque cause sont égaux aux taux bruts, comme cela est expliqué dans **Li et Lu (2018)**. Nous allons donc utiliser le modèle de Lee-Carter séparemment pour chaque cause. Pour se faire, nous avons utilisé le package StMoMo de R qui permet d'utiliser différents modèles de mortalité classiques et ainsi de faire des projections sur l'avenir. **P. Millosovich et Kaishev (2018)**

Le modèle de Lee-Carter est un modèle très proche des modèles linéaires généralisés classiques. On initialise donc dans R un lien "logit" pour ce modèle. Ensuite il faut créer un objet de type "StMoMoData" à l'aide des données que nous possédons. Nous avions besoin d'un tableau contenant les taux de mortalité par âge et par année, celui-ci est obtenu après nos interpolations. Mais également d'un tableau contenant les expositions aux risques par âge et année, des données trouvables facilement sur le site du HMD.

Une fois cela fait, il suffit d'utiliser les fonctions adéquates du package afin d'obtenir différents résultats sur les estimateurs, les résidus, ou des projections des taux de mortalité par âge. Nous présentons ici les graphiques concernant les décès des hommes en France dû au cancer et leurs projections sur plusieurs années.

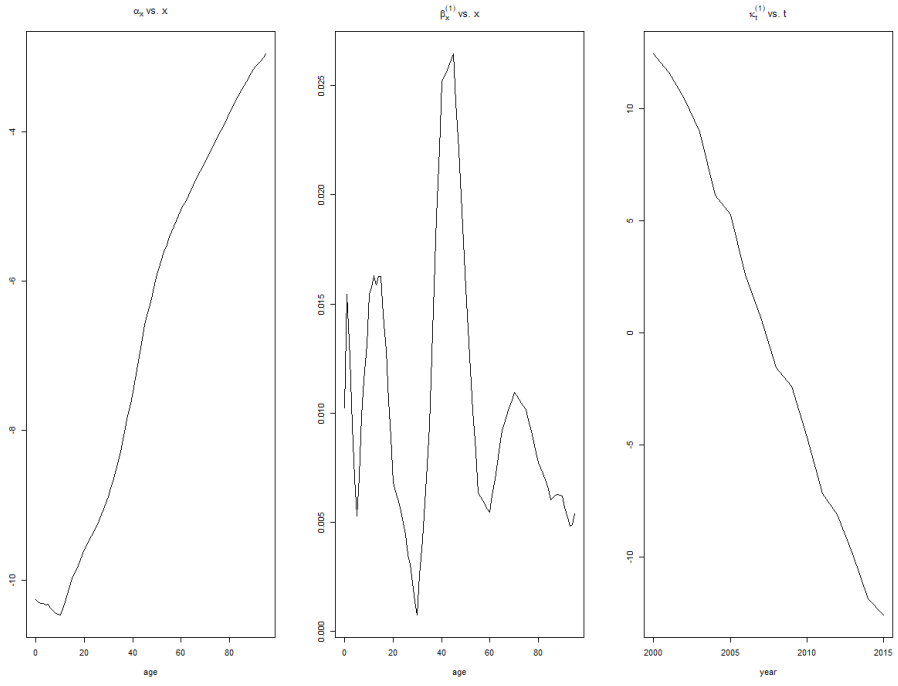


FIGURE 3.1 – Les estimateurs de Lee-Carter en fonction de x (pour alpha et bêta) ou de t (pour kappa)

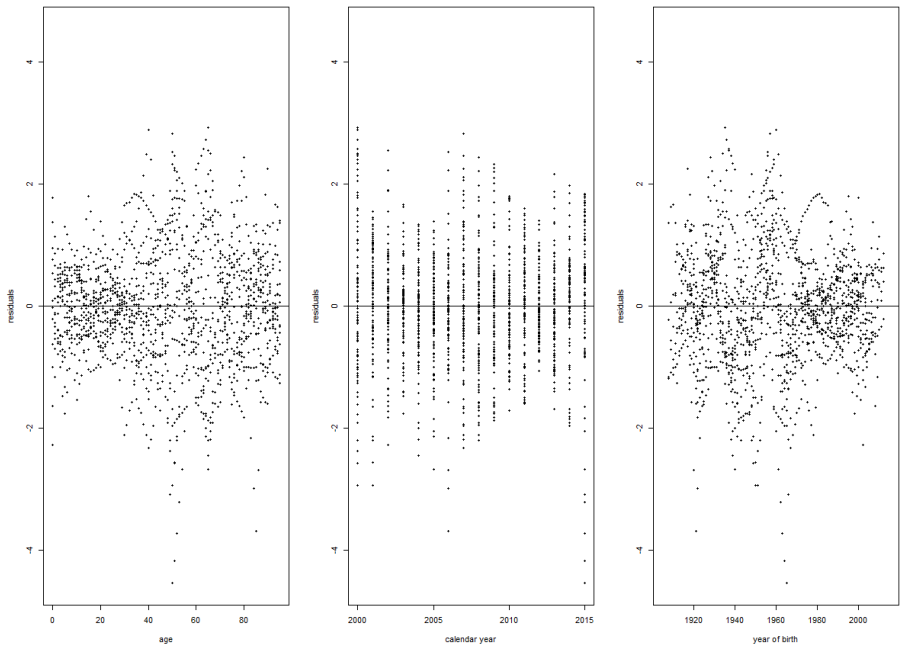


FIGURE 3.2 – Les résidus en fonction de l'âge, l'année calendaire et l'année de naissance.

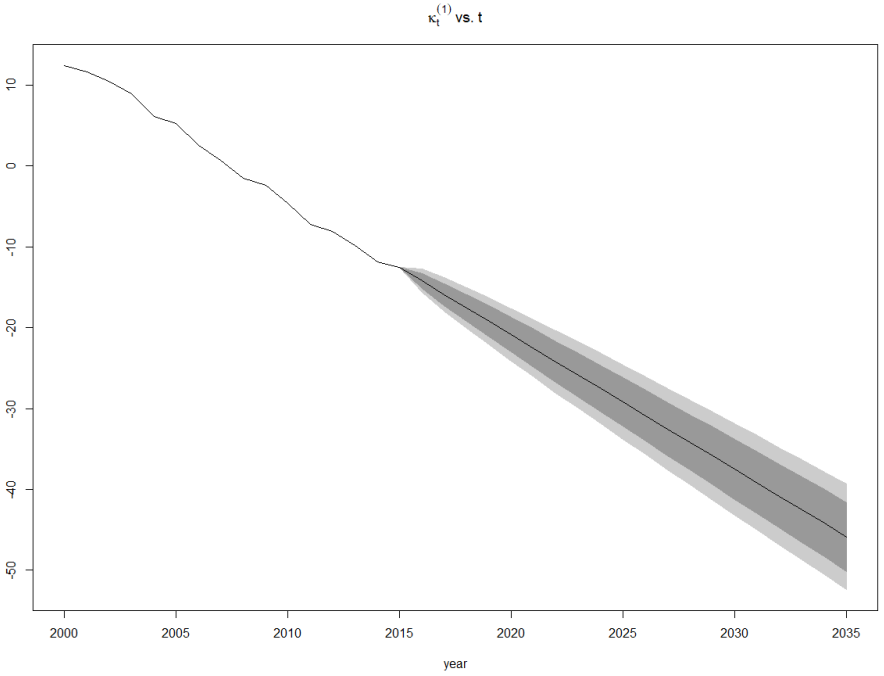


FIGURE 3.3 – Une projection probable de l'estimateur kappa dans plusieurs années.

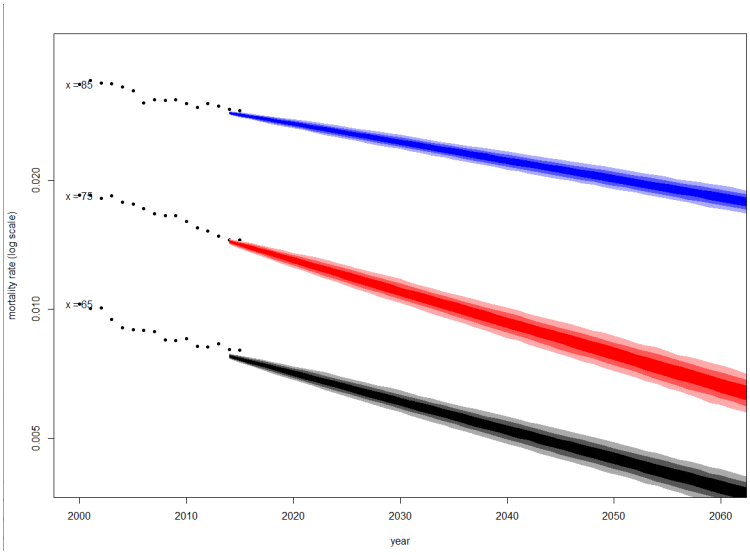


FIGURE 3.4 – Une projection du taux de mortalité pour les cancers selon le modèle de Lee-Carter. Les projections concernent les hommes âgés de 65ans, 75ans et 85ans.

Ici on peut déjà remarquer que les résidus sont assez stables **(figure 3.1)** et n'ont pas l'air d'avoir de tendances particulières, ce qui est bon signe pour la validation du modèle. Concernant les taux de mortalité, on remarque qu'ils vont subir une diminution progressive au fur à mesure des années selon un certain intervalle de confiance **(figure 3.4)**. Cela paraît cohérent puisque la médecine est de plus en plus efficace dans le traitement des différents cancers. Comparons maintenant les taux de mortalité projetés des 5 grandes causes pour une cohorte donnée. Nous allons aussi regarder les coefficients kappa projetés sur plusieurs années.

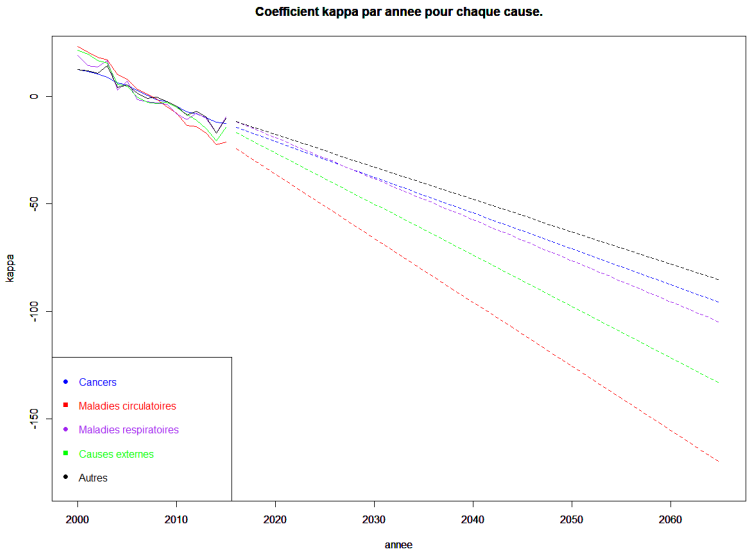


FIGURE 3.5 – Une projection du coefficient kappa pour les 5 grandes causes, par année calendaire.

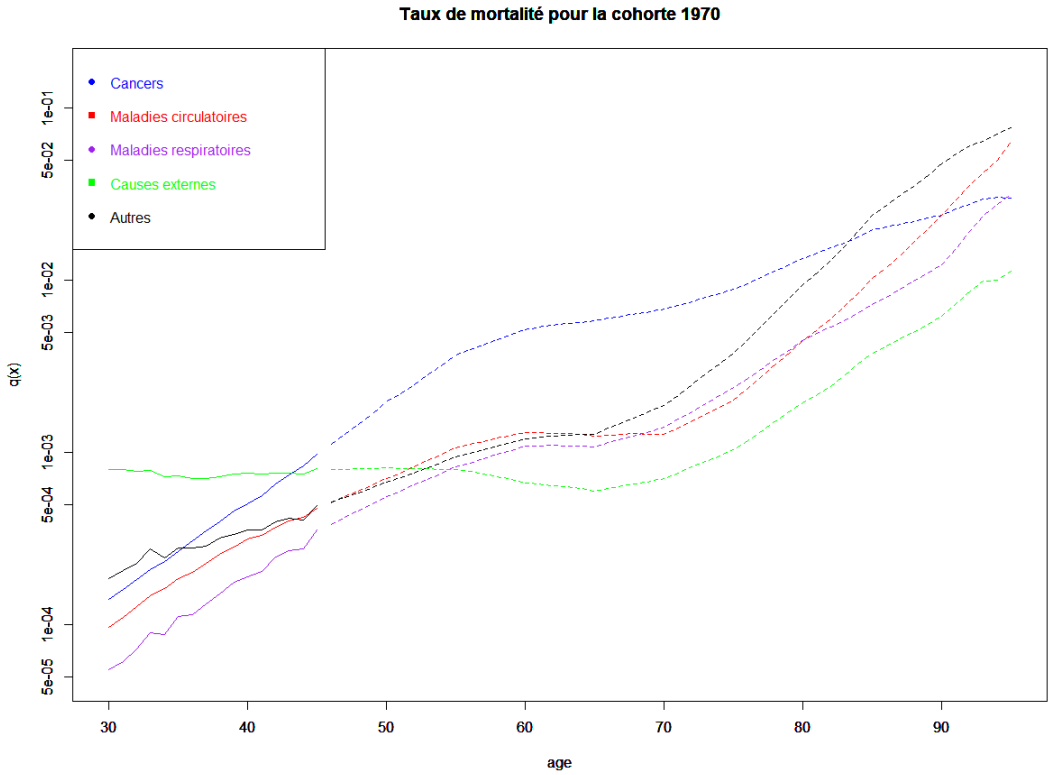


FIGURE 3.6 – Une projection des taux de mortalité pour les 5 causes pour la cohorte 1970.

On remarque sur le premier graphique **3.5** que la mortalité dû aux maladies du système circulatoire est en forte diminution comparée aux autres causes comme les cancers par exemple. La mortalité par accident elle aussi est en forte baisse. De manière générale le coefficent kappa est à la baisse quelque soit la cause ce qui parait cohérent.

Sur le second graphique **3.6**, on peut observer les taux de mortalité projetés pour les personnes nées en 1970, le graphique est assez proche des observations obtenues dans la section "statistiques descriptives". On peut noter que le comportement général de la mortalité par accident est relativement différent des autres courbes et touche en moyenne des personnes plus jeunes. Les cancers ont un taux très supérieurs aux autres causes pour la tranche d'âge 50-75 ans, il sera intéressant par la suite de voir l'impact de l'élimination de cette cause sur l'espérance de vie générale. **CARRIERE (1994)**

3.2 Modélisations avec des copules archimédiennes

Dans cette section, nous allons considérer le modèle en utilisant une copule de Franck et une copule de Clayton. Chacune de ces copules est caractérisée par un seule paramètre θ. Pour la copule de Franck, lorsque θ tend vers 0, les taux nets deviennent identiques aux taux bruts et la copule de Franck se réduit au cas de la copule d'indépendance. C'est le même constat lorsque θ tend vers l'infini dans le cas de la copule de Clayton. Inversement, si θ est grand (rsp.petit) dans la copule de Franck (rsp. Clayton), alors on a un grand facteur de dépendance entre les différentes causes.

Tout d'abord nous allons évaluer l'impact de l'introduction d'une copule archimédienne sur les taux de mortalité. On compare les taux nets que l'on a au préalable calculés avec la formule close **(theorème 3)**, avec les taux de mortalité bruts. Nous prenons des θ bien choisis afin que la dépendance soit forte pour pouvoir observer ou non un changement significatif sur les taux. On voit que pour chacune des copules, l'intensité nette de chaque cause est légèrement plus importante que l'intensité brute. Pour les âges élevés, on observe un rapprochement significatif des taux entre les différentes causes. Cela pourrait s'expliquer par le fait que les personnes agées sont généralement plus fragiles et ont plus tendance à multiplier les maladies et donc à décéder de causes multiples.

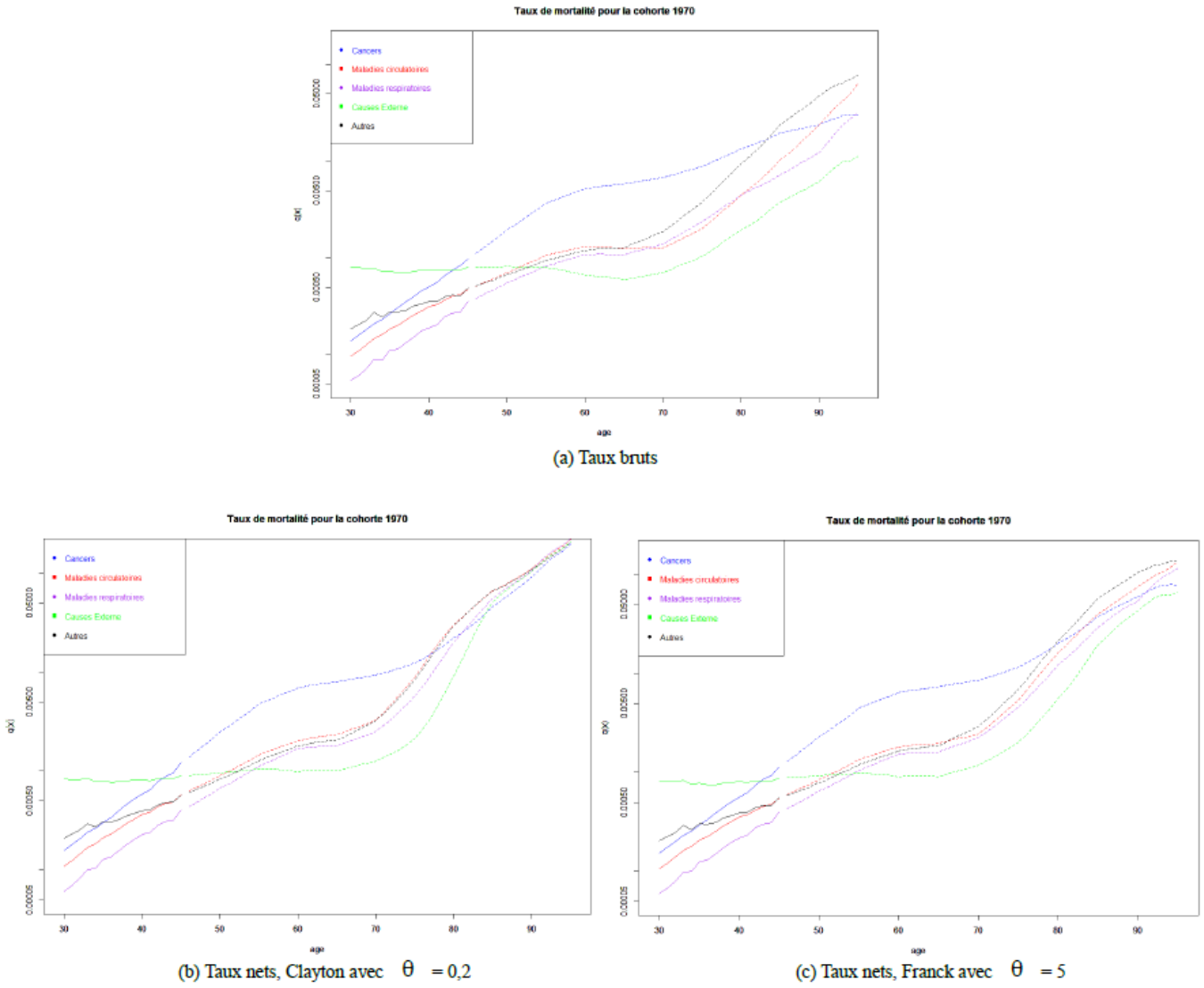
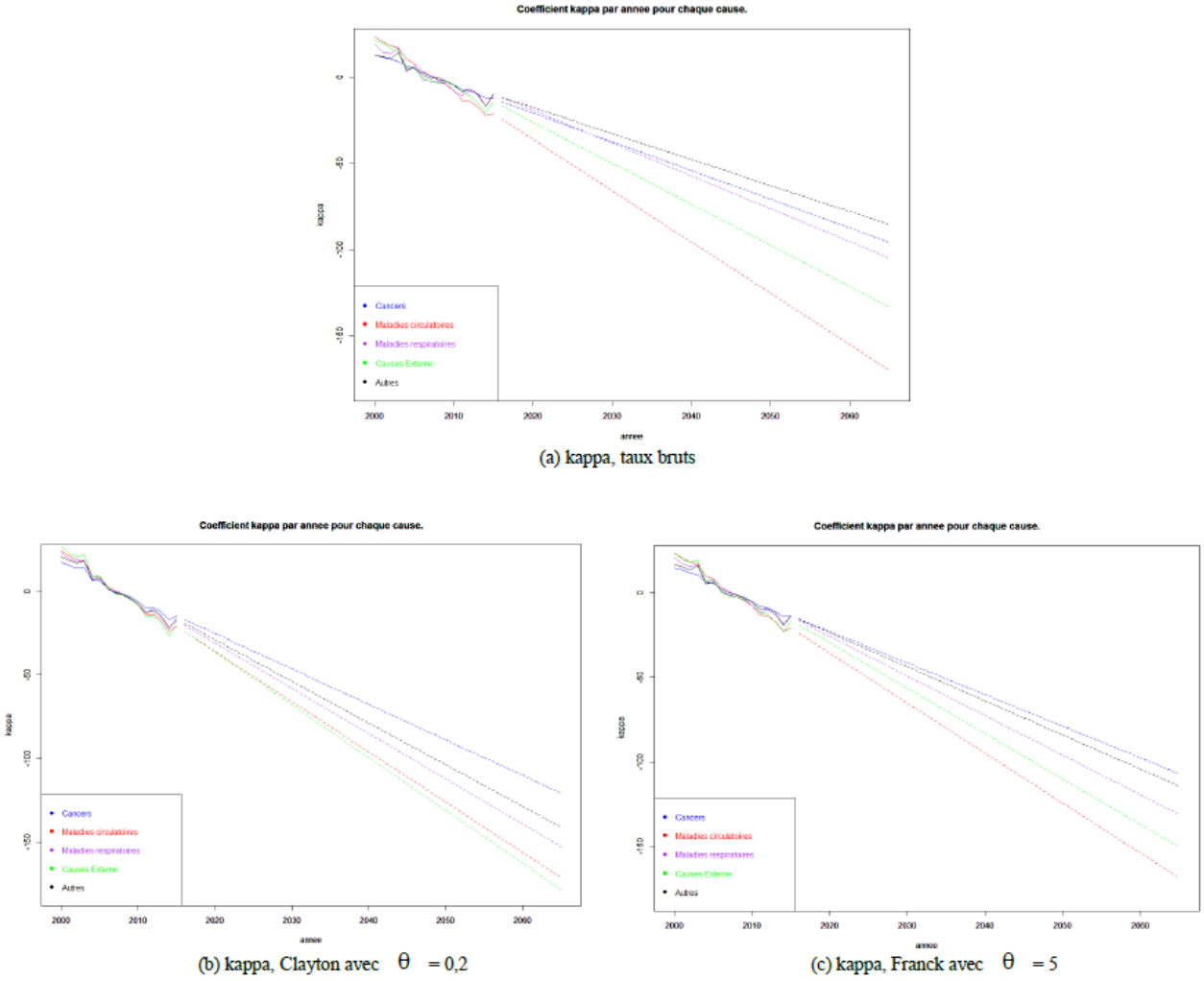


FIGURE 3.7 – Comparaison des taux de mortalité nets et bruts et de leur projection, pour la cohorte 1970.



(a) kappa, taux bruts

(b) kappa, Clayton avec $\theta = 0{,}2$

(c) kappa, Franck avec $\theta = 5$

FIGURE 3.8 – Comparaison des coefficients kappa et de leur évolution future pour les taux nets et bruts.

Sur la **figure 3.8** nous pouvons remarquer un phénomène déjà observé dans l'article de Li et Lu. En effet, introduire une structure de dépendance réduit fortement l'écart entre le κ des maladies cardiovasculaires et le κ des autres causes et notamment celle du cancer. D'après l'article, les maladies cardiovasculaires et les cancers partagent des facteurs de risque communs, comme par exemple le tabagisme et l'obésité. Il explique également que les efforts faits pour réduire la mortalité dû aux cancers sont en partis couverts par la forte diminution de la mortalité dû aux maladies du système circulatoire. Effectivement, les personnes sauvés des maladies cardiaques seront plus souvent sujettes à décéder du cancer par la suite. **LI ET LU (2018)**

3.3 Réduction et élimination d'une cause de mortalité

Dans un premier temps nous avons besoin de nous intéresser plus amplement aux fonctions de survie et à l'espérance de vie. Tout d'abord, une définition équivalente de la copule archimédienne **(définition 8)**, dans notre cadre est la suivante :

$$S(t) = \mathbb{P}[T_1 > t, \dots, T_5 > t] = \psi(\psi^{-1}(S_1(t)) + \cdots + \psi^{-1}(S_5(t))) \textbf{ Li et Lu (2018)}$$

avec ψ la fonction génératrice de la copule archimédienne, S_i la fonction de survie nette de la cause de mortalité i, et T_i le temps de survie avant de décéder de la cause $i \in [\![0, 5]\!]$. Retrouver la fonction de survie dans le cas de la copule indépendante est encore plus simple car les T_i sont indépendants et donc on a directement : $S(t) = \mathbb{P}[T_1 > t, \dots, T_5 > t] = S_1(t) \times \cdots \times S_5(t)$.

Soit $T > 0$ la durée de vie d'un individu, on rappelle que $T = min(T_1, \dots, T_5)$ **(section 1.3)**. On peut maintenant facilement montrer que $\mathbb{E}(T) = \sum_{t>0} S(t)$, avec t des âges entiers.

Notre objectif maintenant est de calculer l'impact de la suppression ou de la réduction d'une cause de décès sur la fonction de survie et l'espérance de vie. Cela nous permettra de calculer des produits de rentes dans la dernière partie de ce mémoire. Pour se faire, nous avons utilisé la méthode décrite dans **V. Kaishev (2005)**, section 6. Soit j la cause que l'on souhaite réduire, on se sert de l'identité suivante :

$S_j(k) = S_j(1)/S_j(0) \times S_j(2)/S_j(1) \times \cdots \times S_j(k)/S_j(k-1) = (1-q_0) \times (1-q_1) \times \cdots \times (1-q_{k-1})$
en utilisant les notations actuarielles (on omet l'indice j pour ne pas surcharger l'écriture).

On introduit ensuite une fonction linéaire par morceaux $L_x(a, b, c, d)$ qui va nous permettre de retirer la cause de manière brutale ou bien de manière plus lente, $\forall x = 0, 1, 2, \dots, 120$:

$$\begin{aligned} L_x(a, b, c, d) &= a, && \text{si } x \in [0, c] \\ &= a + \frac{b-a}{d-c} \times (l - c), && \text{si } x \in [c, d] \\ &= b, && \text{si } x \in [d, 120] \end{aligned}$$

où a, b, c, d sont des paramètres telles que $0 \leq a, b \leq 1$ et $0 < c < d < 120$.

Nous allons utiliser cette fonction pour modifier la fonction de survie nette. On note $S_j^*(t)$ cette nouvelle fonction modifiée et on pose $q_x^* = (1 - L_x)\, q_x$, elle vaudra alors :

$$S_j^*(t) = (1 - q_0^*) \times (1 - q_1^*) \times \cdots \times (1 - q_{k-1}^*)$$

Nous allons maintenant présenter les résultats qui découle de cette stratégie. Nous commençons par le cas où $a = b = 1$, dans cette situation $L_x = 1$, et donc on enlève brutalement la cause de mortalité j. Encore une fois pour notre exemple, nous allons prendre des θ qui induisent une structure de dépendance forte, ce ne sont pas nécessairement les plus réalistes mais ils permettent d'identifier plus clairement les différences entre le modèle avec et sans les copules.

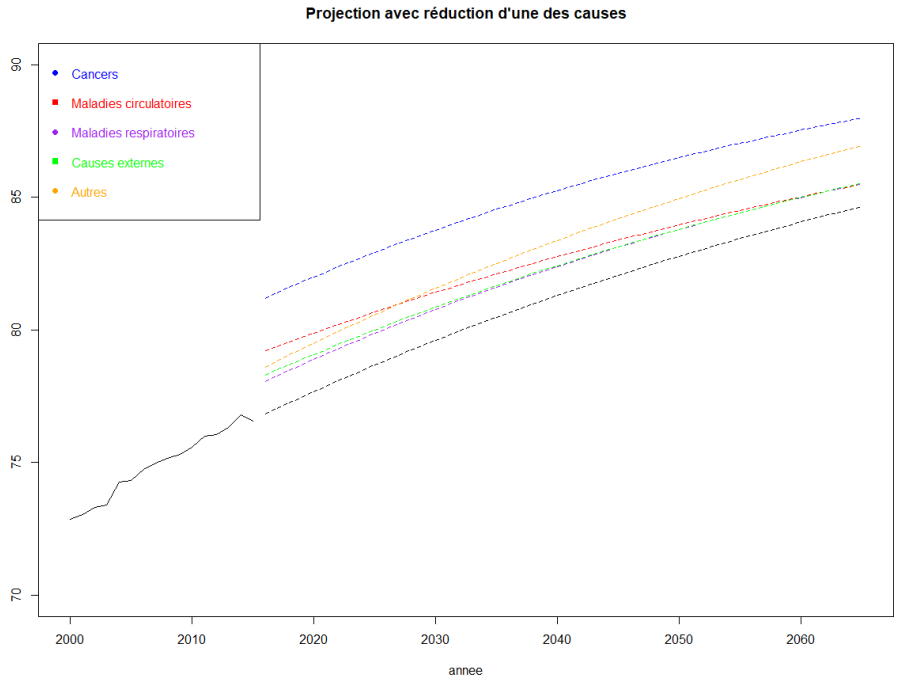


FIGURE 3.9 – Evolution de l'espérance de vie dans le cas d'une élimination brutale d'une cause spécifique, pour la copule indépendante.

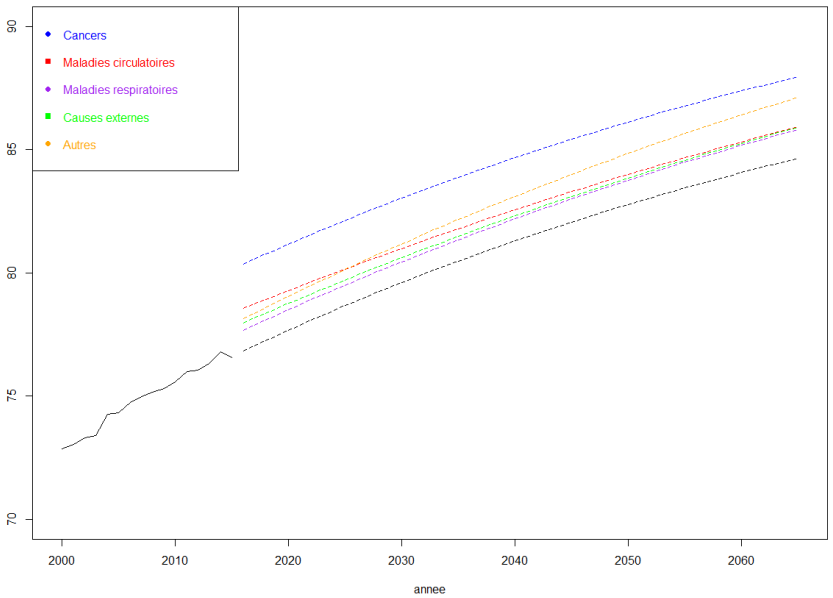


FIGURE 3.10 – Evolution de l'espérance de vie dans le cas d'une élimination brutale d'une cause spécifique, pour la copule de Clayton avec $\theta = 1$.

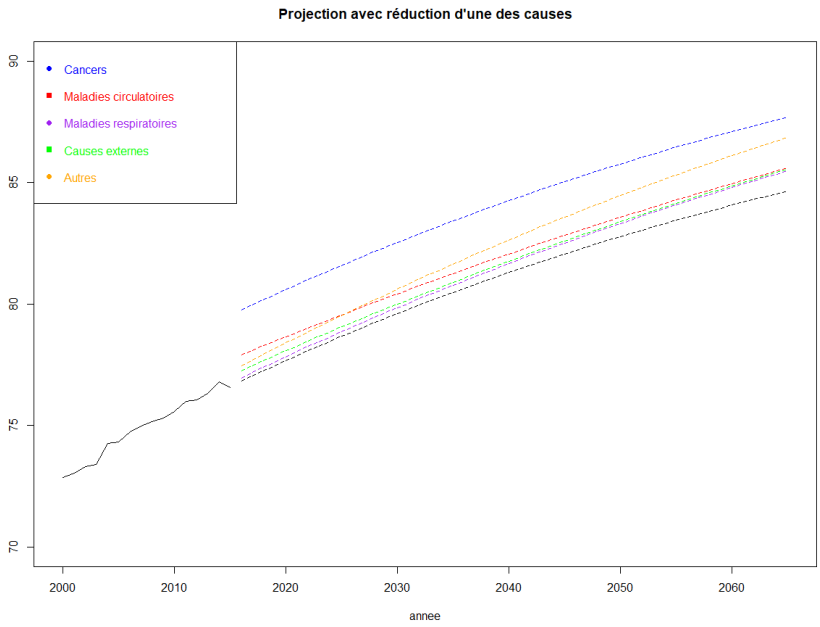
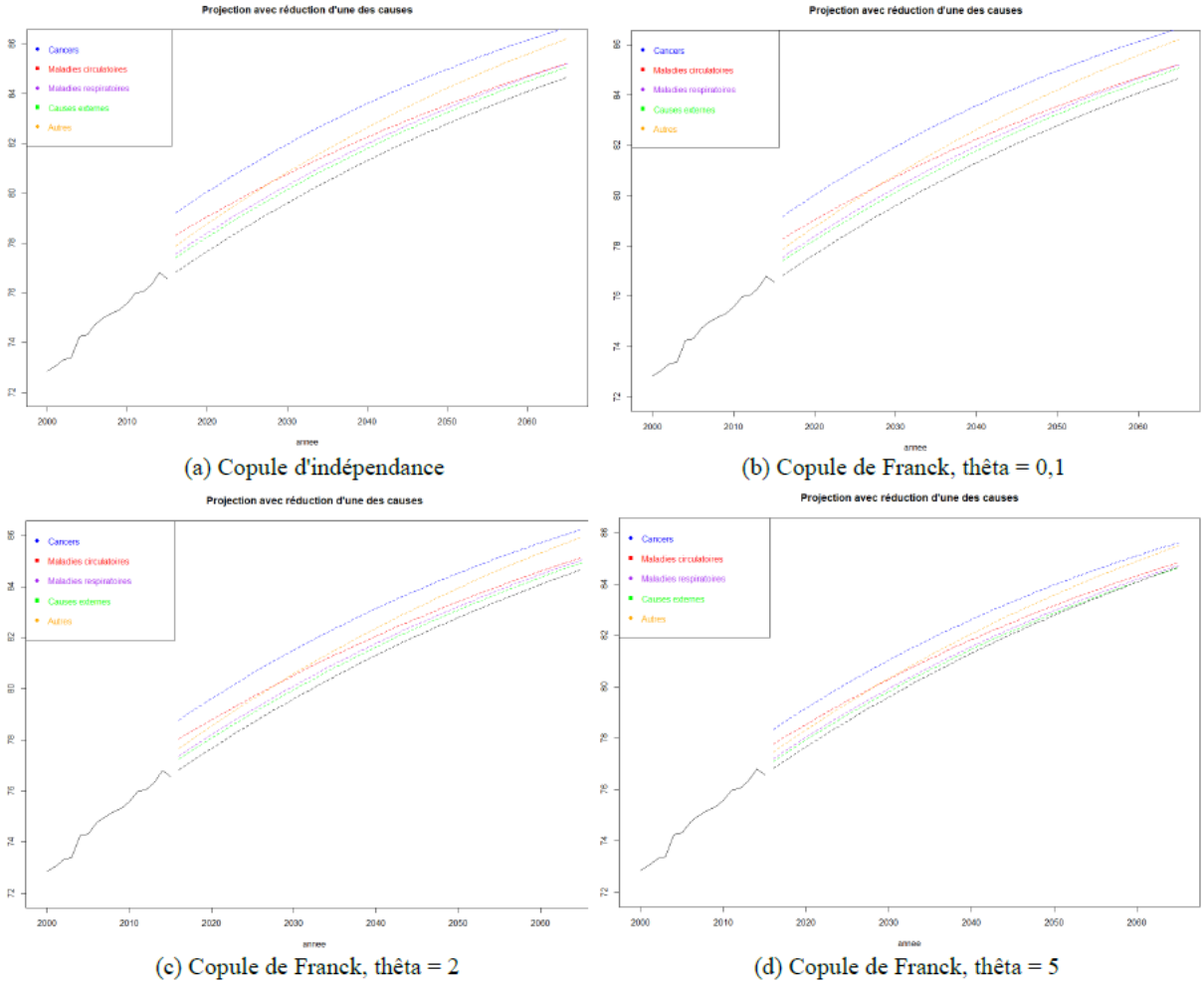


FIGURE 3.11 – Evolution de l'espérance de vie dans le cas d'une élimination brutale d'une cause spécifique, pour la copule de Franck avec $\theta = 5$.

Dans ces graphiques, nous avons à chaque fois fais cinq projections de l'espérance de vie en enlevant la cause de décès correspondant à la couleur de la courbe, la courbe noire représentant le scénario classique sans enlever de cause. On observe premièrement un affaissement des différentes projections dans le cadre des copules. Retier l'une ou l'autre cause semble moins bénéfique que pour la copule d'indépendance. Cependant on peut remarquer que dans tout les scénarios, une disparition soudaine des cancers aurait un impact spectaculaire sur l'espérance de vie général de la population ; gain immédiat de quasiment cinq ans d'espérance de vie.

Dans la **figure 3.12**, on a utilisé la technique de réduction vu au début de la section. En prenant pour la fonction L_x les paramètres : $a = 0.3, b = 0.8, c = 55, d = 75$. C'est donc une réduction partielle d'une cause de décès ; la diminution sera faible dans les âges 0 à 55, forte de 55 à 75 ans, et très forte après 75 ans. On s'apperçoit que les scénarios sont meilleurs pour le cas d'indépendance, et que les scénarios sont de moins en moins favorables lorsque θ devient grand pour la copule de Franck. En d'autres termes, plus la structure de dépendance est forte dans la copule, moins la réduction partielle (ou complète) d'une cause n'aura d'impact sur l'espérance de vie. Nous avons aussi pu le voir dans le cas de la copule de Clayton, même si nous n'illustrons que celle de Franck ici.



(a) Copule d'indépendance

(b) Copule de Franck, thêta = 0,1

(c) Copule de Franck, thêta = 2

(d) Copule de Franck, thêta = 5

FIGURE 3.12 – Comparaison de la diminution d'une cause de mortalité, dans le cas de l'indépendance et avec une copule de Franck de paramètre θ

Dans la dernière partie du mémoire, nous allons utiliser les résultats trouvés dans les 3 sections de ce chapitre pour établir des primes dans différents scénarios futurs : indépendance, copules, élimination du cancer... En effet, le calcul des fonctions de survie avec la formule **(3.3)** vu précédemment, nous sera très utile pour calculer les valeurs actuelles probables (VAP) de l'assureur et de l'assuré.

4 Application des modèles au domaine de l'assurance

Définition 9 (Rente viagère) *Dans* **Dutang** ***(2020-2021)****, une rente viagère à termes anticipés, notée $\ddot{z}_x$, est définie comme une série annuelle de flux d'un montant fixé à l'avance jusqu'au décès de l'individu à commencer de maintenant.*
Pour rappel, x désigne l'âge de l'individu, T_x la durée de vie de l'individu, r le taux d'intérêt, et $v = 1/(1+r)$ le facteur d'actualisation.
Si le montant versé périodiquement vaut 1€, alors on a :

$$\ddot{z}_x = \sum_{k=0}^{+\infty} v^k \mathbb{1}_{T_x > k} = \sum_{k=0}^{\lfloor T_x \rfloor} v^k.$$

$\ddot{z}_x$ est la valeur actuelle.
La valeur actuelle probable (VAP) correspondante est notée :

$$\ddot{a}_x = \mathbb{E}(\ddot{z}_x \mid T_x > 0) = \sum_{k=0}^{+\infty} v^k \mathbb{E}(\mathbb{1}_{T_x > k} \mid T_x > 0) = \sum_{k=0}^{+\infty} v^k {}_kp_x$$

Construction de notre produit de rente :

Hypothèses : Nous allons faire l'hypothèse que le montant versé periodiquement vaut 1, pour plus de simplicité. On considère ici que l'individu (assuré) paiera alors une prime unique, notée π. Nous allons supposer que l'assuré est agé de x années en 2016.

Engagements : Par le principe d'équité actuarielle, VAP assuré = VAP assureur.
Ici, la VAP de l'assuré vaut π et la VAP de l'assureur vaut

$$\sum_{k=0}^{+\infty} v^k {}_kp_x$$

Impact de l'élimination d'une cause de décès : Nous avons choisi de réduire partiellement la cause liée aux cancers pour se rapprocher d'une situation réaliste. Nous avons donc juste recalculer les ${}_kp_x$ pour calculer notre prime pour le modèle avec réduction partielle des cancers en changeant les valeurs de x, de θ pour les copules de Franck et Clayton, et Les résultats sont présentés dans la **figure 4.1** avec le modèle inchangé et le modèle avec réduction de cause. Nous pouvons remarquer qu'en prenant un taux d'intérêt r égal à 0, notre taux d'actualisation v vaut alors 1 et notre prime π correspond alors à l'espérance de vie de l'âge résiduel. Nos π trouvés semblent très cohérents et réalistes pour une espérance de vie.

Nous avons remarqué que l'impact du taux d'actualisation v (et donc de r) sur la prime est plus fort lorsque l'assuré est plus jeune. De plus, on voit que la prime est légèrement modifiée lorsque l'on utilise une copule, mais la différence reste peu importante à priori même avec un facteur multiplicatif de 10 ou 100. Par contre, on observe une réelle différence lorsque l'on réduit la mortalité liée aux cancers. Dans ce scénario, la prime augmente dans le cas d'indépendance de 7 à 10 % selon l'âge de départ de l'assuré, ce qui peut s'avérer considérable pour une compagnie d'assurance gérant un grand nombre de contrats. Cependant, cette inflation de la prime est significativement réduite (aux alentours de 2 à 5 %) lorsque l'on utilise une copule avec un haut facteur de dépendance.

	Année de naissance **Age en 2016** Taux d'intérêt	1941 **75** r = 0%	r = 1%	1951 **65** r = 0%	r = 1%	1961 **55** r = 0%	r = 1%
MODELE INCHANGE	**Copule indépendante**	13,18	12,12	21,58	19,03	30,62	25,84
	Copule Clayton						
	θ = 0,5	13,01	11,97	21,49	18,97	30,69	25,90
	θ = 2	13,15	12,10	21,66	19,09	30,86	26,00
	θ = 5	13,18	12,12	21,63	19,07	30,75	25,93
	Copule Franck						
	θ = 0,5	13,14	12,09	21,54	19,00	30,59	25,81
	θ = 2	13,04	12,00	21,41	18,90	30,45	25,72
	θ = 5	12,91	11,89	21,20	18,74	30,18	25,53
MODELE AVEC REDUCTION PARTIELLE DES CANCERS	**Copule indépendante**	14,51	13,28	23,66	20,72	32,94	27,57
	Copule Clayton						
	θ = 0,5	13,29	12,22	22,53	19,85	32,06	26,97
	θ = 2	14,04	12,88	23,33	20,47	32,82	27,48
	θ = 5	14,31	13,10	23,53	20,62	32,91	27,55
	Copule Franck						
	θ = 0,5	14,37	13,16	23,51	20,61	32,79	27.47167
	θ = 2	13,98	12,82	23,06	20,26	32,35	27,16
	θ = 5	13,45	12,36	22,36	19,70	31,60	26,63

FIGURE 4.1 – Comparaison des primes π en fonction des différents modèles avec les différentes valeurs de θ

Conclusion

Nous avons pu observer que les taux de mortalité pour les cancers avec ou sans utilisation des copules sont supérieurs aux autres causes chez les 50-70 ans. On le voit encore mieux lorsqu'on réduit brutalement cette cause : l'espérance de vie augmente alors considérablement, entre deux à cinq années de plus selon le coefficient de réduction appliqué.

Nous avons constaté que notre modèle de Lee-Carter avec les copules archimédienne réagissait en accord avec le papier de **Li et Lu (2018)** : la structure de dépendance met en lumière le fait que les maladies cardiovasculaires et les cancers partagent des facteurs de risques communs. De manière générale, les résultats de projection trouvés avec cette base de données de la mortalité en France, restent très semblables aux observations réalisées dans les différents papiers de recherche les plus récents sur le sujet. On remarquera au final que l'impact majeur pour les sociétés d'assurances vie serait une élimination complète ou partielle des cancers. On considère ce scénario comme le scénario extrême sur lequel on pourrait se baser pour les primes.

La différence majeure avec les recherches actuelles par rapport à nos analyses sont les chiffres de l'espérance de vie. En effet, ceux-ci semblent assez bas par rapport aux chiffres trouvés par les instituts de statistiques. Il semble exister un biais de l'espérance de vie dans nos observations, qui pourrait être dû au manque de données sur les plus de 95 ans, ou sur la méthode d'interpolation, qui nous a donné des approximations des taux de mortalité entre chaque tranche d'âge.

Le deuxième point que nous aimerions soulever est la méthode de réduction des causes. Même si celle-ci semble nous donner une bonne idée de la projection future dans un scénario sans cancer par exemple, c'est une méthode un peu brutale qui réduit d'une année à l'autre la cause concernée. Dans la réalité, cette réduction serait progressive au fur et à mesure des années, mais nous n'avons trouvé aucun autre document où une méthode plus douce est mis en place.

Nous pourrions pour plus de curiosité appliquer le modèle de Lee-Carter à d'autre copules archimédiennes ou gaussiennes mais surtout aux copules hierarchiques dont nous avons briévement parlé dans notre la section 1.4. Ces copules ont été utilisé dans le document de Li et Lu, ils y donnent également des formules closes du théorème d'identifiabilité, permettant de les utiliser facilement dans notre cadre.

Références

Carriere, J. (1994). Dependent decrement theory. *Transactions of the Society of Actuaries* 46, p. 45-74.

— (1995). Removing Cancer when it is Correlated with other Causes of Death, p. 45-74.

Dutang, C. (2020-2021). Cours d'Actuariat 1.

Guibert, Q. (2020-2021). Methodes de Simulation en assurance.

Lee, R. et Carter, L. (1992). Modeling and forecasting U.S. Mortality. *Journal of the American Statistical Association* 7.419, p. 659-671.

Li, H. et Lu, Y. (2018). Modeling and Forecasting U. S. Mortality. *Scandinavian Actuarial Journal*.

Nelsen, R. B. (2006). An Introduction to Copulas. Springer Series in Statistics, p. 32-34.

P. Millosovich, A. V. et Kaishev, V. (2018). StMoMo : An R Package for Stochastic Mortality Modelling. *Journal of Statistical Software*.

V. Kaishev V. Haberman, S. D. (2005). Modelling the joint distribution of competing risks survival times using copula functions.