UCLouvain

FACULTE DES SCIENCES

Faculté des sciences

Sex-specific mortality forecasting

Auteur : Christopher Viellevoye
Promoteur : Donatien Hainaut
Lecteur : Michel Denuit
Année académique 2020-2021

Préface.

Ce mémoire a pour objet l'étude du taux de mortalité en Belgique, en France et aux Pays-Bas à travers le modèle de Lee-Carter et de plusieurs de ses extensions. « L'Augmented Common Factor » à un niveau permet de faire une distinction entre homme et femme tandis que « L'Augmented Common Factor » à deux niveaux rajoute un terme de différenciation par rapport aux pays étudiés. Enfin, l'« Augmented Common Factor with cohort extension » permet l'analyse des potentiels effets de cohortes sur le taux de mortalité.

Une comparaison est faite entre les modèles de Lee-Carter et « L'Augmented Common Factor » à deux niveau via une calibration des paramètres suivie d'une analyse de leurs projections de l'espérance de vie vis-à-vis des données empiriques.

Enfin, une projection de l'espérance de vie jusqu'en 2060 pour chacuns des pays ainsi qu'un exemple d'application de pricing d'une rente viagère sont également donnés en fin de mémoire.

The purpose of this thesis is to study the mortality rate in Belgium, France and the Netherlands through the Lee-Carter model and several of its extensions. The "Augmented Common Factor" at one level makes it possible to make a distinction between men and women while "The Augmented Common Factor" at two levels adds a term of differentiation in relation to the countries studied. Finally, the "Augmented Common Factor with cohort extension" allows the analysis of the potential effects of cohorts on the mortality rate.

A comparison is made between the Lee-Carter models and the "Augmented Common Factor" at two levels via a calibration of the parameters followed by an analysis of their projections of life expectancy against empirical data.

Finally, a projection of life expectancy until 2060 for each country as well as an example of a pricing application for a life annuity are also given at the end of the thesis.

Remerciements.

Plusieurs personnes ont contribué, de près ou de loin, à la réalisation de ce mémoire et de mes études.

Tout d'abord, un grand merci à mon promoteur, le Professeur Donatien Hainaut pour ses conseils avisés et sa disponibilité.

Ensuite, je tenais à très sincèrement remercier l'ensemble du jury de m'avoir permis de me réinscrire au master en sciences actuarielles et de me permettre ainsi de terminer ce cursus.

Un merci particulier à Ludovic et Thomas, mes compagnons de route durant toutes ces années d'études, pour la relecture de ce mémoire. Votre amitié aura marqué de façon indélébile ce passage à l'UCL.

Merci au docteur Fairon pour la vérification des équations et pour son oeil averti, traquant les moindres erreurs d'indices.

Merci à Guillaume pour sa relecture et pour sa présence lors de mes nombreuses remises en question qui ont jalonné cette année.

Merci à Aline et Clémence pour leurs relectures et la chasse aux fautes d'orthographe, malgré le timing serré.

Merci également à la bande du cyclo : Rudy, Morgane, Alexstar, Mehdi et Diana. Les blocus sans vous n'auraient pas eu la même saveur et je vous dois en grande partie la réussite de mes études.

Merci à la Maison Des Sciences et au Lovaniensis Scientificus Ordo pour avoir été la bulle d'air me permettant de m'évader, quand la pression était trop forte.

Merci à Coralie et Julien pour les sorties en bateau, la pratique de la voile étant un excellent décompresseur.

Merci à Maxime et Marie, la « Team Cailloux », pour vos blagues qui parsemaient les longues soirées d'écriture.

Et enfin, un énorme merci à ma maman, Nathalie, pour son soutien sans faille et ses encouragements lors de mes (trop) nombreuses années d'études.

Table des matières

Chapitre 1

Introduction.

Dans le domaine de l'assurance, le coût futur d'un contrat, ce que l'on appelle sa sinistralité, est toujours incertain. Or, lorsqu'une compagnie d'assurance commercialise un nouveau produit, il lui revient de déterminer *a priori* quel est le montant de la prime que devra effectivement payer l'assuré. Si cette prime est insuffisante par rapport au nombre de sinistres rencontrés ainsi qu'à leurs coûts, la compagnie s'expose à un risque de défaut et par conséquent pourrait se retrouver dans une situation d'insolvabilité. Il est évident que ce type de situation conduirait à la disparition des indemnités prévues pour les assurés, ce qui serait évidemment catastrophique.

Couplée à la concurrence sévère entre assureurs, la juste estimation de la prime est donc d'une importance capitale afin qu'une compagnie puisse provisionner correctement les sinistres qu'elle va rencontrer tout en restant attractive pour les potentiels clients. Heureusement, l'actuaire dispose d'une palette d'outils mathématiques et statistiques qui lui permettent de fixer le montant de la prime en fonction du nombre de sinistres attendus et de la sévérité attendue de ces derniers.

Cette prime peut dépendre de nombreux paramètres en fonction du domaine étudié tel que l'âge du conducteur pour les assurances auto, la valeur de reconstruction à neuf d'un bâtiment pour une police incendie, la masse salariale en accident du travail ou encore la probabilité de survie dans de nombreuses assurances sur la vie. Cette probabilité dépend directement du taux de mortalité et c'est précisément ce taux de mortalité qui forme l'objet d'étude du présent mémoire.

Le taux de mortalité peut être défini comme étant le ratio du nombre de décès au sein d'une population. Comme expliqué plus haut, il est primordial de connaître avec une précision suffisante les taux de mortalité afin de pouvoir estimer correctement la prime d'assurance à demander aux clients.

Il existe dans la littérature différentes façons de modéliser mathématiquement les taux de mortalité. Ce travail s'arrête plus spécifiquement sur la méthode de Lee-Carter (appelée LC par la suite) ainsi que sur certaines de ses extensions.

Comme cela est montré par la suite, la méthode LC montre quelques lacunes car celle-ci fait un lissage sur le sexe des individus sans prendre en compte les différences génétiques entre les hommes et les femmes (les hommes décédant plus jeune que les femmes, voir [9]) et elle ne tient pas non plus compte d'une quelconque cohérence entre des populations de pays voisins.

Ces lacunes peuvent être corrigées successivement par, d'une part le modèle dit « Augmented Common Factor à un niveau » (appelé ACF-1 par la suite), qui prend en compte les

différences d'ordre génétique, et d'autre part le modèle dit « Augmented Common Factor à deux niveaux » (appelé ACF-2 par la suite) qui permet quant à lui de faire une modélisation sur plusieurs pays.

Enfin, un troisième modèle, dit « Augmented Common Factor with cohort extension » (appelé ACFC) prenant en compte le possible effet de cohorte peut également venir améliorer les prédictions du modèle LC

Les chapitres 2 et 3 se basent sur la méthodologie suivie dans l'article [19], ce dernier applique ces modèles sur la population du Royaume-Uni, ici l'analyse est faite sur l'axe Belgique - France - Pays-Bas.

Le chapitre 4 analyse les paramètres temporels des différents modèles afin de voir si ces derniers peuvent être considérés comme des séries temporelles suivant un modèle ARIMA. Cette analyse est faite via le test de Box-Jenkins et de la notion d'autocorrélation d'une part et du test de Jarque-Bera et de la normalité des séries temporelles d'autre part. Ensuite, ce chapitre définit la notion d'espérance de vie et compare l'extrapolation des séries temporelles analysées juste avant par rapport à l'espérance de vie réellement observée entre les modèles LC et ACF-2.

Le chapitre 5 fait une extrapolation des espérances de vie entre 2018 et 2060 pour les modèles LC et ACF-2, tout en analysant leurs différences.

Enfin, le chapitre 6 reprend un exemple de pricing d'assurance basé sur le taux de mortalité via un produit de type « rente viagère ». Il analyse ainsi l'impact des différentes projections de l'espérance de vie sur la tarification de ce produit.

Chapitre 2

Théorie.

2.1 Taux de mortalité et exposition au risque.

Comme vu lors de l'introduction, les assurances sur la vie se basent sur la probabilité de survie d'une personne. Cette dernière varie, de façon non exhaustive, en fonction de l'âge, du sexe, du pays, du niveau socio-économique, de l'état de santé et peut également varier d'une année à l'autre. Dans un premier temps et afin de ne pas complexifier l'exposé, l'accent est mis sur deux variables : l'âge et l'année considérée.

Comme expliqué dans [5], la probabilité de survie d'une personne est donnée par $p_x(t)$, la probabilité qu'une personne d'âge x survive jusqu'à l'âge $x+1$ durant l'année calendrier t.

De la même manière, la probabilité de décès $q_x(t)$ peut être définie comme la probabilité pour une personne d'âge x de mourir durant l'année calendrier t.

Puisqu'une personne d'âge x sera soit vivante, soit morte à la fin de l'année, le lien direct qui peut être fait entre la probabilité de survie et celle de décès est le suivant

$$p_x(t) = (1 - q_x(t)).$$

Étant donné le nombre de décès de personne d'âge x durant l'année t, $d_{x,t}$ et le nombre de personnes d'âge x en vie au début de l'année t, $l_{x,t}$, l'estimation de la probabilité de décès durant l'année t est donnée par

$$\hat{q}_x(t) = \frac{d_{x,t}}{l_{x,t}}.$$

Le taux de mortalité $\mu_{x,t}$, quant à lui, est défini comme étant le taux de décès instantané d'une personne d'âge x au temps t. Il est mathématiquement défini en fonction de la probabilité de survie comme suit

$$p_x(t) = exp(-\int_0^t \mu_{x,s} ds). \tag{2.1}$$

Théoriquement, $\mu_{x,s}$ devrait varier tout au long de l'année puisqu'il représente un taux instantané. Afin de pouvoir travailler plus facilement avec cette notion, la supposition d'un taux de mortalité constant par morceau est faite, c'est-à-dire qu'il est supposé que le taux de mortalité ne varie pas au cours d'une même année pour un âge donné. Pour $0 < \xi_1, \xi_2 < 1$,

$$\mu_{x+\xi_1,t+\xi_2} = \mu_{x,t}$$

avec x et t des entiers.
Ainsi, l'équation (2.1) devient, en version discrète

$$p_x(t) = exp(-\sum_{i=0}^{t} \mu_{x,i}).$$

Afin de pouvoir correctement prédire le taux de survie et donc de tarifer correctement les assurances sur la vie, il faut pouvoir estimer $\mu_{x,t}$. Pour ceci, il est nécessaire de faire appel à la notion d'exposition au risque (notée ETR « Exposure To Risk » par la suite). L'exposition au risque à l'âge x, dernier anniversaire durant l'année t, notée $ETR_{x,t}$, est le temps total vécu par les personnes d'âge x dont le dernier anniversaire tombe durant l'année calendrier t.

Il est par ailleurs intéressant de voir que l'estimateur non biaisé du taux de mortalité est donné par (la démonstration est faite dans [5], slide 102 et 103)

$$\widehat{\mu}_{x,t} = \frac{d_{x,t}}{ETR_x(t)}.$$

Les quantités $d_{x,t}$ et $ETR_{x,t}$ peuvent être trouvées dans les tables de données « The Human Mortality Database », disponible en [12]. Ce sont sur ces jeux de données que ce travail s'appuie afin d'estimer la valeur de $\mu_{x,t}$. Pour chaque pays analysé, les fichiers « Deaths » et « Exposure-to-risk » donnent, respectivement, les $d_{x,t}$ et les $ETR_{x,t}$.

2.2 Le modèle de Lee-Carter et l'approche Poisson.

Le modèle de Lee-Carter.

La quantité $d_{x,t}$, nombre de décès de personne d'âge x durant l'année t, peut-être intuitivement vue comme étant le produit de la population $N(x,t)$, nombre d'individus en vie âgée de x années à l'instant t, et d'un facteur multiplicatif $\alpha(x,t)$ qui dépend de l'âge x considéré et du temps t,

$$d_{x,t} = N(x,t)\alpha(x,t). \tag{2.2}$$

Dans le modèle développé par Lee et Carter [18], il a été décidé de modéliser ce paramètre $\alpha(x,t)$ comme étant une exponentielle dépendante de 3 facteurs a_x,B_x et K_t, tous réels,

$$\alpha(x,t) = e^{a_x} e^{B_x K_t} = e^{a_x + B_x K_t}.$$

Avec :
- a_x code l'influence de l'âge sur la mortalité ;
- K_t code l'évolution temporelle de la mortalité ;
- B_x code l'influence de l'âge sur le paramètre K_t. En effet, plus B_x sera grand, plus l'influence de K_t sera importante.

L'équation (2.2) peut se réécrire comme le quotient $\frac{d_{x,t}}{N(x,t)} = \alpha(x,t) = e^{a_x + K_t B_x}$. Ce ratio n'est autre que le taux de mortalité $\mu_{x,t}$ vu à la section 2.1. En appliquant la fonction logarithme népérien de chaque côté de l'équation, elle devient

$$ln(\mu_{x,t}) = a_x + B_x K_t. \tag{2.3}$$

Le modèle ainsi décrit ne rend pas compte de toutes les influences historiques liées à un âge spécifique. Pour remédier à cela, il faut introduire un paramètre d'erreur $\varepsilon_{x,t} \in \mathbb{R}$ qui va venir bruiter le modèle afin de mieux rendre compte de la réalité. Ce paramètre d'erreur est une variable aléatoire qui suit une loi gaussienne de moyenne 0 et de variance constante σ^2, $\varepsilon_{x,t} \sim N(0;\sigma^2)$. Le modèle complet de Lee-Carter devient alors

$$ln(\mu_{x,t}) = a_x + B_x K_t + \varepsilon_{x,t}.$$

Suite au trop grand nombre de degrés de liberté, il n'est pas possible d'identifier le modèle en l'état. Par exemple, multiplier les B_x par une constante $\alpha \neq 0$ et diviser les K_t par cette même constante donnerait un autre jeu de solutions. C'est pourquoi il est nécessaire de poser des contraintes sur les paramètres. Ces contraintes ont été choisies de la même manière que dans [23] et sont

$$\sum_x B_x = 1, \ \sum_t K_t = 0.$$

L'approche Poisson.

Bien que très populaire, le modèle historique de Lee-Carter pose plusieurs soucis. En effet, comme on peut le voir dans [5], le modèle de Lee-Carter suppose que les erreurs sont normalement distribuées et qu'elles sont homoscédastiques (c'est-à-dire que leurs variances sont identiques). Cette affirmation est fort contraignante et ne colle pas à la réalité.

Une solution, qui sera retenue ici, est de modéliser le nombre de décès $d_{x,t}$ par une loi de Poisson $D_{x,t}$ de paramètre $\lambda = E_{x,t}\mu_{x,t}$,

$$D_{x,t} \backsim Poisson(E_{x,t}\mu_{x,t}) \tag{2.4}$$

où

— $E_{x,t}$ est l'$ETR_{x,t}$ défini dans la section 2.1 ;

— $\mu_{x,t}$ est donné par l'équation (2.3).

Afin de pouvoir estimer les paramètres a_x, B_x et K_t, il faut maximiser la log-vraisemblance du modèle de Poisson. Sachant que la fonction de probabilité d'une loi de Poisson est donnée par $f(x,\lambda) = P_\lambda(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}$, la vraisemblance de (2.4) vaut

$$L(a_x, B_x, K_t) = \prod_{x,t} P_{E_{x,t}\mu_{x,t}}(D_{x,t} = d_{x,t}) = \prod_{x,t} e^{-E_{x,t}\mu_{x,t}} \frac{(E_{x,t}\mu_{x,t})^{d_{x,t}}}{d_{x,t}!}.$$

La log-vraisemblance $l(a_x, B_x, K_t)$ est donc donnée par

$$l(a_x, B_x, K_t) = ln(\prod_{x,t} e^{-E_{x,t}\mu_{x,t}} \frac{(E_{x,t}\mu_{x,t})^{d_{x,t}}}{d_{x,t}!}) = \sum_{x,t} ln(e^{-E_{x,t}\mu_{x,t}} \frac{(E_{x,t}\mu_{x,t})^{d_{x,t}}}{d_{x,t}!})$$

$$= \sum_{x,t}(-E_{x,t}\mu_{x,t} + ln((E_{x,t}\mu_{x,t})^{d_{x,t}}) - ln(d_{x,t}!)) = \sum_{x,t}(-E_{x,t}\mu_{x,t} + d_{x,t} ln(E_{x,t}\mu_{x,t})) + cst$$

$$= \sum_{x,t}(d_{x,t}ln(E_{x,t}) + d_{x,t}ln(\mu_{x,t}) - E_{x,t}\mu_{x,t}) + cst = \sum_{x,t}(d_{x,t}ln(exp(a_x + B_xK_t)) - E_{x,t}\mu_{x,t}) + cst.$$

D'où on conclut que

$$l(a_x, B_x, K_t) = \sum_{x,t}[(d_{x,t}(a_x + B_xK_t) - E_{x,t}exp(a_x + B_xK_t))] + cst. \tag{2.5}$$

2.3 Modèle ACF-1.

Afin d'améliorer le modèle de Log-Poisson, il a été proposé une première modification du modèle (voir [20]). Cette modification, appelée modèle ACF à 1 facteur, vise à prendre en compte les différences biologiques qu'il peut y avoir entre les hommes et les femmes. Comme vu dans l'introduction, les hommes meurent statistiquement plus jeunes que les femmes. Il est donc normal de vouloir modéliser cette différence. Pour ce faire, il est utile de redéfinir la manière de calculer le taux de mortalité en ajoutant deux nouveaux paramètres (i codant le sexe, masculin ou féminin) :

- $\kappa_{t,i}$ code l'évolution temporelle de la mortalité, en fonction du sexe ;
- $b_{x,i}$ code l'influence de l'âge sur le paramètre $\kappa_{t,i}$ en fonction du sexe. De la même manière que pour le rapport entre K_t et B_x, plus $b_{x,i}$ sera grand, plus l'influence de $\kappa_{t,i}$ sera importante.

L'équation (2.3) peut maintenant se réécrire

$$ln(\mu_{x,t,i}) = a_{x,i} + B_xK_t + b_{x,i}\kappa_{t,i}. \tag{2.6}$$

Suite au trop grand nombre de degrés de liberté, il n'est pas possible d'identifier le modèle en l'état. C'est pourquoi il est nécessaire de poser des contraintes sur les paramètres. De la même manière que dans la section précédente, ces contraintes sont les suivantes, pour i = 1,2

$$\sum_x B_x = 1,\ \sum_t K_t = 0$$

$$\sum_x b_{x,i} = 1,\ \sum_t \kappa_{t,i} = 0.$$

De la même manière que pour la relation (2.4), le nombre de décès $d_{x,t,i}$ peut être modélisé par une loi de Poisson de paramètre $E_{x,t,i}\mu_{x,t,i}$,

$$D_{x,t,i} \sim Poisson(E_{x,t,i}\mu_{x,t,i}).$$

Où $E_{x,t,i}$ est l'$ETR_{x,t}$ et $\mu_{x,t,i}$ est donné par (2.6).

Afin de pouvoir estimer les paramètres $a_{x,i}$, B_x, K_t, $b_{x,i}$ et $\kappa_{t,i}$ il faut maximiser la log-vraisemblance. En se basant sur le même développement que pour (2.5), cette dernière est donnée par

$$L(a, B, K, b, \kappa) = \sum_{x,t,i}[d_{x,t,i}(a_{x,i} + B_xK_t + b_{x,i}\kappa_{t,i}) - E_{x,t,i}exp(a_{x,i} + B_xK_t + b_{x,i}\kappa_{t,i})] + cst.$$

2.4 Modèle ACF-2.

Un autre défaut qui peut être reproché à l'approche Log-Poisson (mais également au modèle ACF-1), est le manque de continuité entre les mortalités des populations géographiquement voisines. En effet, des populations voisines ont, en général, en commun un même niveau de vie, un même niveau de sécurité sociale et plus globalement partagent plusieurs facteurs socio-économiques influençant la mortalité. Il est donc intuitivement logique de vouloir une certaine cohérence dans les probabilités de survie. Cette modification, appelée modèle ACF à 2 facteurs, vise à prendre en compte cette continuité tout en préservant les différences entre régions (voir [19]). Pour ce faire, il est utile de redéfinir la manière de calculer le taux de mortalité en ajoutant deux nouveaux paramètres (j codant le pays, Belgique, France ou Pays-Bas) :

— $\kappa_{t,i,j}$ code l'évolution temporelle de la mortalité, en fonction du sexe et du pays de résidence ;

— $b_{x,i,j}$ code l'influence de l'âge sur le paramètre $\kappa_{t,i,j}$ en fonction du sexe et du pays de résidence. De la même manière que pour le rapport entre K_t et B_x, plus $b_{x,i,j}$ sera grand, plus l'influence de $\kappa_{t,i,j}$ sera importante.

L'équation (2.6) peut maintenant être réécrite comme

$$ln(\mu_{x,t,i,j}) = a_{x,i,j} + B_x K_t + b_{x,i}\kappa_{t,i} + b_{x,i,j}\kappa_{t,i,j}. \tag{2.7}$$

Pour les mêmes raisons que plus haut, il est nécessaire de poser des contraintes sur les paramètres. Ces contraintes sont les suivantes, pour i = 1,2 et pour j = 1,2,3

$$\sum_x B_x = 1,\ \sum_t K_t = 0,\ \sum_x b_{x,i} = 1,\ \sum_t \kappa_{t,i} = 0,\ \sum_x b_{x,i,j} = 1,\ \sum_t \kappa_{t,i,j} = 0.$$

De la même manière que pour la relation (2.4), le nombre de décès $d_{x,t,i}$ peut être modélisé par une loi de Poisson de paramètre $E_{x,t,i,j}\mu_{x,t,i,j}$,

$$D_{x,t,i,j} \sim Poisson(E_{x,t,i,j}\mu_{x,t,i,j}).$$

Avec $E_{x,t,i,j}$ est l'$ETR_x(t)$ et $\mu_{x,t,i,j}$ est donné par (2.7).

Afin de pouvoir estimer les paramètres $a_{x,i,j}$, B_x, K_t, $b_{x,i}$, $\kappa_{t,i}$, $b_{x,i,j}$ et $\kappa_{t,i,j}$ il faut maximiser la log-vraisemblance en se basant sur le modèle (2.4). Cette dernière est donnée par

$$L(a, B, K, b, \kappa) =$$

$$\sum_{x,t,i,j} [d_{x,t,i,j}(a_{x,i,j} + B_x K_t + b_{x,i}\kappa_{t,i} + b_{x,i,j}\kappa_{t,i,j}) - E_{x,t,i,j} exp(a_{x,i,j} + B_x K_t + b_{x,i}\kappa_{t,i} + b_{x,i,j}\kappa_{t,i,j})] + cst.$$

2.5 Modèle ACFC.

Une dernière amélioration de notre modèle, proposée par [19] consiste à prendre en considération le possible effet de cohorte. Un effet de cohorte est un phénomène qui concerne une génération de personnes nées la même année et qui évoluent ensemble de manière différente

par rapport aux autres générations. Une génération présentant un effet de cohorte aurait donc une évolution de la mortalité significativement différente par rapport aux autres générations.

Cet effet est expliqué en détail dans [9]. Il est fort marqué au Royaume-Uni où les générations des années 1925-1945 ont connu une chute de leur mortalité. Plusieurs facteurs peuvent expliquer l'apparition de ces cohortes. Par exemple l'utilisation de la cigarette, beaucoup plus commune lors de la Grande guerre, a provoqué une augmentation de la mortalité des générations précédentes avant qu'une prise de conscience sur ces effets néfastes ne commence dans les années 1920, ce qui a diminué la mortalité de la génération suivante. Le régime alimentaire (dû aux privations d'après-guerre), la couverture sociale, le système de soins de santé ou encore la surmortalité des générations précédentes et suivantes (due à la première et seconde guerre mondiale) sont autant de facteurs propices à créer cet effet de cohorte.

Il est intéressant d'essayer de voir si ce phénomène de cohorte se retrouve de façon aussi marquée sur le continent.

Pour ce faire, il faut de nouveau rajouter un terme à l'équation (2.7), $g_{h,i}$ qui code l'effet de cohorte, en fonction du sexe i, pour chaque génération $h = t - x$

L'équation (2.7) devient donc

$$ln(\mu_{x,t,i,j}) = a_{x,i,j} + B_x K_t + b_{x,i}\kappa_{t,i} + b_{x,i,j}\kappa_{t,i,j} + g_{h,i}. \tag{2.8}$$

Comme pour le modèle ACF2, il faut poser les contraintes suivantes sur les paramètres du modèle ACFC afin de pouvoir l'identifier

$$\sum_x B_x = 1,\ \sum_t K_t = 0,\ \sum_x b_{x,i} = 1,\ \sum_t \kappa_{t,i} = 0$$

$$\sum_x b_{x,i,j} = 1,\ \sum_t \kappa_{t,i,j} = 0,\ \sum_h g_{h,i} = 0.$$

Le nombre de décès $d_{x,t,i,j}$ peut également être modélisé par une loi de Poisson de paramètre $E_{x,t,i,j}\mu_{x,t,i,j}$,

$$D_{x,t,i,j} \sim Poisson(E_{x,t,i,j}\mu_{x,t,i,j}).$$

Avec $E_{x,t,i,j}$ est l'$ETR_{x,t}$ et $\mu_{x,t,i,j}$ est donné par (2.8).

Afin de pouvoir estimer les paramètres a_x, B_x, K_t, $b_{x,i}$, $\kappa_{t,i}$, $b_{x,i,j}$, $\kappa_{t,i,j}$ et $g_{h,i}$ il faut maximiser la log-vraisemblance en se basant sur le modèle (2.4). Cette dernière est donnée par

$$L(a,B,K,b,\kappa,g) = \sum_{x,t,i,j,h=x-t} [d_{x,t,i,j}(a_{x,i,j} + B_x K_t + b_{x,i}\kappa_{t,i} + b_{x,i,j}\kappa_{t,i,j} + g_{h,i})$$

$$-E_{x,t,i,j}exp(a_{x,i,j} + B_x K_t + b_{x,i}\kappa_{t,i} + b_{x,i,j}\kappa_{t,i,j} + g_{h,i})] + cst.$$

Remarque sur les log-vraisemblances

Bien que leurs résolutions soient capitales dans l'identification des paramètres des modèles vus plus haut, aucune des log-vraisemblances ne peut être résolue de façon analytique. Leurs résolutions, forcément numérique, sera traitée en détail dans la section 3.2.

Chapitre 3

Modélisation.

3.1 Analyse et traitement des données.

Les données utilisées proviennent des bases de données « The Human Mortality Database » [12]. Emanant des universités de Berkeley (USA) et de l'institut Max Planck pour la recherche démographique de Rostock (Allemagne), il s'agit d'un projet de mise à disposition gratuite de statistiques concernant les populations, pays par pays. Ce projet se base sur les chiffres de recensement officiels de chaque pays et fournit les paramètres nécessaires à cette étude, c'est-à-dire le nombre de décès $d_{x,t}$ et l'exposition au risque $ETR_{x,t}$. Chacune de ces valeurs étant disponible, par pays et aussi bien séparément pour les hommes que pour les femmes, que de façon agrégée. Au vu de leur méthode de travail et de la facilité à comparer différents pays (les données étant toutes présentées dans un même format), les bases de données HMD sont un très bon candidat sur lesquel se baser dans le cadre de ce mémoire.

Les données disponibles couvrent les années allant de 1850 à 2018 pour trois pays : la Belgique, les Pays-Bas et la France. Les données utilisées ici ont été extraites du site [12] en septembre 2020.

L'ensemble du code permettant la modélisation et l'identification des paramètres vus au chapitre 2 est quant à lui réalisé à l'aide du logiciel R et peut être trouvé en annexe.

Âges au-dessus de 100 ans.

Les grands âges peuvent poser des soucis. En effet, au vu de leur faible représentation, ils peuvent être source d'erreurs statistiques. De plus, lorsqu'il n'y a aucun représentant pour un couple x, t et que (voir 3.2.3 pour les notations) $E_{x,t} = ETR_{x,t} = 0$, alors $\hat{d} = 0 = f'$, et ce pour tous les modèles. Ceci donne un dénominateur nul dans la formule itérative et est générateur d'erreur du type « Not A Number » à cause de la division par 0.

C'est pourquoi, bien que les tables de données HMD aillent de 0 à 110 ans, la présente étude couvre 101 âges, de 0 à 100 ans.

Années étudiées.

Bien que les données disponibles sur le site HMD soient disponibles entre les années 1850 et 2018 pour les trois pays étudiés, la calibration des modèles est faite sur les données couvrant les années 1946 à 2013 inclues.

En effet les années avant 1946 ne sont pas pertinentes pour la modélisation future à cause des deux guerres mondiales. Durant les deux guerres, les pays étudiés ont connu un surcroît de mortalité qui n'est pas présent en temps de paix et qui peut fausser une évaluation future.

La modélisation s'arrête à l'année 2013 car les années 2014 à 2018 seront utilisées pour la validation du modèle dans le chapitre 4.

Nombre de cohortes.

Les cohortes x s'étalent sur 101 années et les années t sont au nombre de 68. Le nombre de cohortes $h = t - x$ est donné par $(t_{max} - x_{min}) - (t_{min} - x_{max}) + 1 = 168$.

3.2 Développement de l'algorithme.

Comme vu dans le chapitre 2, afin de pouvoir identifier les modèles qui y sont présentés, il est primordial de pouvoir maximiser les log-vraisemblances. Maximiser la log-vraisemblance revient à chercher les racines de sa dérivée, or cette solution ne saurait pas être analytique dans le cas présent, au vu du nombre important d'inconnues en jeu.

3.2.1 La méthode de Newton-Raphson.

Une approche calculatoire, choisie dans [5, 19, 23], consiste à utiliser la méthode de Newton-Raphson. Il s'agit d'une méthode itérative qui permet de trouver numériquement la racine d'une fonction en partant de l'approximation de cette fonction par son polynôme de Taylor.

Considérant une fonction $f(x)$, son polynôme de Taylor d'ordre un est donné par

$$T^1_{f(x)} = f(x_0) + f'(x_0)(x - x_0)$$

x étant le point (inconnu) où la fonction est évaluée et x_0 le point de départ de la méthode itérative. Idéalement x_0 n'est pas trop éloigné de la valeur x, ceci afin de minimiser le nombre d'itérations, et donc le temps de calcul. Une première approximation consiste à prendre $f(x) \backsimeq T^1_{f(x)}$.

Chercher la racine de cette fonction revient à solutionner l'équation

$$0 = f(x_0) + f'(x_0)(x - x_0).$$

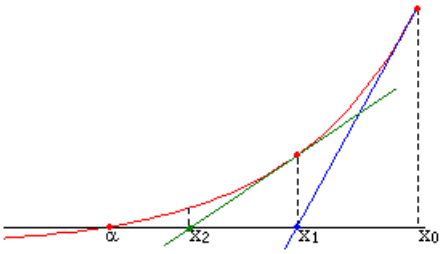
Isoler x donne

$$x = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

Le nombre x ainsi trouvé est la racine du polynôme de Taylor d'ordre 1 de la fonction $f(x)$ évaluée en x_0. Ce nombre x n'est bien évidemment pas la racine de la fonction $f(x)$ puisque le polynôme de Taylor est une approximation de la véritable fonction mais il se rapproche de la véritable racine α cherchée ([22]). Il est possible de tendre vers α en appliquant plusieurs fois de suite la méthode de Newton-Raphson et en travaillant de proche en proche

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}.$$

Si la racine α de la fonction f est isolée, alors la suite $(x_k)_{k\epsilon\mathbb{N}}$ va converger vers α.



(cette illustration vient de [11])

3.2.2 Application à la méthode Poisson.

Afin d'appliquer la méthode de Newton-Raphson au paramètre a_x, f_{a_x} est identifiée comme étant la dérivée partielle de l'équation 2.5 par rapport à a_x

$$f_{a_x} = \frac{\partial L(a,B,K)}{\partial a_x} = \sum_t \frac{\partial (D_{x,t}(a_x + B_x K_t) - E_{x,t} exp(a_x + B_x K_t))}{\partial a_x} = \sum_t D_{x,t} - E_{x,t} exp(a_x + B_x K_t).$$

Sa dérivée $f'(x)$ se calcule facilement

$$f'_{a_x} = \sum_t \frac{\partial (D_{x,t} - E_{x,t} exp(a_x + B_x K_t))}{\partial a_x} = -\sum_t E_{x,t} exp(a_x + B_x K_t).$$

Et le lien itératif est donné par

$$a_x^{k+1} = a_x^k - \frac{\sum_t (D_{x,t} - E_{x,t} exp(a_x + B_x K_t))}{-\sum_t E_{x,t} exp(a_x + B_x K_t)}.$$

Le même cheminement peut être mené en identifiant f_{B_x}et f_{K_t} à la dérivée partielle par rapport à $B_x$et par rapport à K_t

$$f_{B_x} = \sum_t \frac{\partial (D_{x,t}(a_x + B_x K_t) - E_{x,t} exp(a_x + B_x K_t))}{\partial B_x} = \sum_t (D_{x,t} - E_{x,t} exp(a_x + B_x K_t)) K_t.$$

$$f_{K_t} = \sum_x \frac{\partial (D_{x,t}(a_x + B_x K_t) - E_{x,t} exp(a_x + B_x K_t))}{\partial K_t} = \sum_x (D_{x,t} - E_{x,t} exp(a_x + B_x K_t)) B_x.$$

Les dérivées sont données respectivement par

$$f'_{B_x} = \sum_t \frac{\partial (D_{x,t} - E_{x,t} exp(a_x + B_x K_t)) K_t}{\partial B_x} = -\sum_t E_{x,t} exp(a_x + B_x K_t) (K_t)^2.$$

$$f'_{K_t} = \sum_x \frac{\partial (D_{x,t} - E_{x,t} exp(a_x + B_x K_t)) B_x}{\partial K_t} = -\sum_x E_{x,t} exp(a_x + B_x K_t) (B_x)^2.$$

Et le lien itératif pour ces deux paramètres est donné par

$$B_x^{k+1} = B_x^k - \frac{\sum_t (D_{x,t} - E_{x,t} exp(a_x + B_x K_t)) K_t}{-\sum_t E_{x,t} exp(a_x + B_x K_t)(K_t)^2}.$$

$$K_t^{k+1} = K_t^k - \frac{\sum_x (D_{x,t} - E_{x,t} exp(a_x + B_x K_t)) B_x}{-\sum_x E_{x,t} exp(a_x + B_x K_t)(B_x)^2}.$$

3.2.3 Extension aux autres modèles.

Afin de faciliter la lecture, les paramètres $\hat{d}$, D et E sont définis comme suit, respectivement pour chaque modèle :

	$\hat{d}$	D
Log-Poisson	$E_{x,t} exp(a_x + B_x K_t)$	$D_{x,t}$
ACF-1	$E_{x,t,i} exp(a_{x,i} + B_x K_t + b_{x,i}\kappa_{t,i})$	$D_{x,t,i}$
ACF-2	$E_{x,t,i,j} exp(a_{x,i,j} + B_x K_t + b_{x,i}\kappa_{t,i} + b_{x,i,j}\kappa_{t,i,j})$	$D_{x,t,i,j}$
ACFC	$E_{x,t,i,j} exp(a_{x,i,i} + B_x K_t + b_{x,i}\kappa_{t,i} + b_{x,i,j}\kappa_{t,i,j} + g_{h,i})$	$D_{x,t,i,j}$

Les log-vraisemblance peuvent ainsi se réécrire :

	log-vraisemblance
Log-Poisson	$\sum_{x,t} D(a_x + B_x K_t) - \hat{d} + cst$
ACF-1	$\sum_{x,t,i} D(a_{x,i} + B_x K_t + b_{x,i}\kappa_{t,i}) - \hat{d} + cst$
ACF-2	$\sum_{x,t,i,j} D(a_{x,i,j} + B_x K_t + b_{x,i}\kappa_{t,i} + b_{x,i,j}\kappa_{t,i,j}) - \hat{d} + cst$
ACFC	$\sum_{x,t,i,j,h=t-x} D(a_{x,i,j} + B_x K_t + b_{x,i}\kappa_{t,i} + b_{x,i,j}\kappa_{t,i,j} + g_{h,i}) - \hat{d} + cst$

En identifiant des fonctions f_α, dérivée partielle de la log-vraisemblance correspondante par rapport au paramètre α, pour chaque paramètre α de chaque modèle, il est aisé de calculer les fonctions f'_α ainsi que les fonctions itératives correspondantes. Toutes ces informations peuvent être trouvées dans les tableaux ci-dessous.

Log-Poisson.

α	f	f'	formule itérative
a_x	$\sum_t (D - \hat{d})$	$-\sum_t \hat{d}$	$a_x^{k+1} = a_x^k - \frac{\sum_t (D-\hat{d})}{-\sum_t \hat{d}}$
B_x	$\sum_t (D - \hat{d}) K_t$	$-\sum_t \hat{d} K_t^2$	$B_x^{k+1} = B_x^k - \frac{\sum_t (D-\hat{d}) K_t}{-\sum_t \hat{d} K_t^2}$
K_t	$\sum_x (D - \hat{d}) B_x$	$-\sum_x \hat{d} B_x^2$	$K_t^{k+1} = K_t^k - \frac{\sum_x (D-\hat{d}) B_x}{-\sum_x \hat{d} B_x^2}$

ACF-1.

α	f	f'	formule itérative
$a_{x,i}$	$\sum_{t,i}(D-\hat{d})$	$-\sum_{t,i}\hat{d}$	$a_{x,i}^{k+1} = a_{x,i}^{k} - \frac{\sum_{t}(D-\hat{d})}{-\sum_{t}\hat{d}}$
B_x	$\sum_{t,i}(D-\hat{d})K_t$	$-\sum_{t,i}\hat{d}K_t^2$	$B_x^{k+1} = B_x^{k} - \frac{\sum_{t,i}(D-\hat{d})K_t}{-\sum_{t,i}\hat{d}K_t^2}$
K_t	$\sum_{x,i}(D-\hat{d})B_x$	$-\sum_{x,i}\hat{d}B_x^2$	$K_t^{k+1} = K_t^{k} - \frac{\sum_{x,i}(D-\hat{d})B_x}{-\sum_{x,i}\hat{d}B_x^2}$
$b_{x,i}$	$\sum_{t}(D-\hat{d})\kappa_{t,i}$	$-\sum_{t}\hat{d}\kappa_{t,i}^2$	$b_{x,i}^{k+1} = b_{x,i}^{k} - \frac{\sum_{t}(D-\hat{d})\kappa_{t,i}}{-\sum_{t}\hat{d}\kappa_{t,i}^2}$
$\kappa_{t,i}$	$\sum_{x}(D-\hat{d})b_{x,i}$	$-\sum_{x}\hat{d}b_{x,i}^2$	$\kappa_{t,i}^{k+1} = \kappa_{t,i}^{k} - \frac{\sum_{x}(D-\hat{d})b_{x,i}}{-\sum_{x}\hat{d}b_{x,i}^2}$

ACF-2.

α	f	f'	formule itérative
$a_{x,i,j}$	$\sum_{t,i,j}(D-\hat{d})$	$-\sum_{t,i,j}\hat{d}$	$a_{x,i,j}^{k+1} = a_{x,i,j}^{k} - \frac{\sum_{t}(D-\hat{d})}{-\sum_{t}\hat{d}}$
B_x	$\sum_{t,i,j}(D-\hat{d})K_t$	$-\sum_{t,i,j}\hat{d}K_t^2$	$B_x^{k+1} = B_x^{k} - \frac{\sum_{t,i,j}(D-\hat{d})K_t}{-\sum_{t,i,j}\hat{d}K_t^2}$
K_t	$\sum_{x,i,j}(D-\hat{d})B_x$	$-\sum_{x,i,j}\hat{d}B_x^2$	$K_t^{k+1} = K_t^{k} - \frac{\sum_{x,i,j}(D-\hat{d})B_t}{-\sum_{x,i,j}\hat{d}B_t^2}$
$b_{x,i}$	$\sum_{t,j}(D-\hat{d})\kappa_{t,i}$	$-\sum_{t,j}\hat{d}\kappa_{t,i}^2$	$b_{x,i}^{k+1} = b_{x,i}^{k} - \frac{\sum_{t,j}(D-\hat{d})\kappa_{t,i}}{-\sum_{t,j}\hat{d}\kappa_{t,i}^2}$
$\kappa_{t,i}$	$\sum_{x,j}(D-\hat{d})b_{x,i}$	$-\sum_{x,j}\hat{d}b_{x,i}^2$	$\kappa_{t,i}^{k+1} = \kappa_{t,i}^{k} - \frac{\sum_{x,j}(D-\hat{d})b_{x,i}}{-\sum_{x,j}\hat{d}b_{x,i}^2}$
$b_{x,i,j}$	$\sum_{t}(D-\hat{d})\kappa_{t,i,j}$	$-\sum_{t}\hat{d}\kappa_{t,i,j}^2$	$b_{x,i,j}^{k+1} = b_{x,i,j}^{k} - \frac{\sum_{t}(D-\hat{d})\kappa_{t,i,j}}{-\sum_{t}\hat{d}\kappa_{t,i,j}^2}$
$\kappa_{t,i,j}$	$\sum_{x}(D-\hat{d})b_{x,i,j}$	$-\sum_{x}\hat{d}b_{x,i,j}^2$	$\kappa_{t,i,j}^{k+1} = \kappa_{t,i,j}^{k} - \frac{\sum_{x}(D-\hat{d})b_{x,i,j}}{-\sum_{x}\hat{d}b_{x,i,i}^2}$

ACFC.

α	f	f'	formule itérative
$a_{x,i,j}$	$\sum_{t,i,j}(D-\hat{d})$	$-\sum_{t,i,j}\hat{d}$	$a^{k+1}_{x,i,j} = a^{k}_{x,i,j} - \frac{\sum_{t,i,j}(D-\hat{d})}{-\sum_{t,i,j}\hat{d}}$
B_x	$\sum_{t,i,j}(D-\hat{d})K_t$	$-\sum_{t,i,j}\hat{d}K_t^2$	$B^{k+1}_{x} = B^{k}_{x} - \frac{\sum_{t,i,j}(D-\hat{d})K_t}{-\sum_{t,i,j}\hat{d}K_t^2}$
K_t	$\sum_{x,i,j}(D-\hat{d})B_x$	$-\sum_{x,i,j}\hat{d}B_x^2$	$K^{k+1}_{t} = K^{k}_{t} - \frac{\sum_{x,i,j}(D-\hat{d})B_t}{-\sum_{x,i,j}\hat{d}B_t^2}$
$b_{x,i}$	$\sum_{t,j}(D-\hat{d})\kappa_{t,i}$	$-\sum_{t,j}\hat{d}\kappa_{t,i}^2$	$b^{k+1}_{x,i} = b^{k}_{x,i} - \frac{\sum_{t,j}(D-\hat{d})\kappa_{t,i}}{-\sum_{t,j}\hat{d}\kappa_{t,i}^2}$
$\kappa_{t,i}$	$\sum_{x,j}(D-\hat{d})b_{x,i}$	$-\sum_{x,j}\hat{d}b_{x,i}^2$	$\kappa^{k+1}_{t,i} = \kappa^{k}_{t,i} - \frac{\sum_{x,j}(D-\hat{d})b_{t,i}}{-\sum_{x,j}\hat{d}b_{t,i}^2}$
$b_{x,i,j}$	$\sum_{t}(D-\hat{d})\kappa_{t,i,j}$	$-\sum_{t}\hat{d}\kappa_{t,i,j}^2$	$b^{k+1}_{x,i,j} = b^{k}_{x,i,j} - \frac{\sum_{t}(D-\hat{d})\kappa_{t,i,j}}{-\sum_{t}\hat{d}\kappa_{t,i,j}^2}$
$\kappa_{t,i,j}$	$\sum_{x}(D-\hat{d})b_{x,i,j}$	$-\sum_{x}\hat{d}b_{x,i,j}^2$	$\kappa^{k+1}_{t,i,j} = \kappa^{k}_{t,i,j} - \frac{\sum_{x}(D-\hat{d})b_{x,i,j}}{-\sum_{x}\hat{d}b_{x,i,i}^2}$
$g_{h,i}$	$\sum_{x,t,j,t-x=h} D-\hat{d}$	$-\sum_{x,t,j,t-x=h}\hat{d}$	$g^{k+1}_{h,i} = g^{k}_{h,i} - \frac{\sum_{x,t,j,t-x=h} D-\hat{d}}{-\sum_{x,t,j,t-x=h}\hat{d}}$
$g^*_{h,i}$	$\sum_{x,t,j,t-x=h}\omega_{x,t}(D-\hat{d})$	$-\sum_{x,t,j,t-x=h}\hat{d}$	$g^{k+1}_{h,i} = g^{k}_{h,i} - \frac{\sum_{x,t,j,t-x=h}\omega_{x,t}(D-\hat{d})}{-\sum_{x,t,j,t-x=h}\hat{d}}$

Les cohortes possédant peu d'observations sont non-significatives et risquent d'apporter des déviations non désirées. C'est pourquoi, à la place d'utiliser $g_{h,i}$, le modèle ACFC va utiliser $g^*_{h,i}$. Dans ce dernier, un paramètre $\omega_{x,t}$ a été rajouté comme facteur multiplicatif à chaque terme de la somme du numérateur afin de neutraliser les 5 premières et les 5 dernières cohortes. À cet effet, $\omega_{x,t} = 0$ pour $h \in \{1,2,3,4,5,164,165,166,167,168\}$ et $\omega_{x,t} = 1$ pour les autres valeurs de h.

3.2.4 Critère de convergence de la méthode de Newton-Raphson.

Afin de savoir quand la méthode de Newton-Raphson va converger, il faut faire appel à la notion de déviance.

La déviance est définie comme deux fois la différence entre la log-vraisemblance du modèle saturé (c'est-à-dire du modèle où chaque estimation est remplacée par son observation) et la log-vraisemblance du modèle classique ([19]). l étant la log-vraisemblance du modèle classique, et l_{sat} celle du modèle saturé :

$$l = \sum d \cdot ln(\hat{d}) - \hat{d}.$$

$$l_{sat} = \sum d \cdot ln(d) - d.$$

Et la déviance vaut

$$Dév = 2\sum(d \cdot ln(d) - d - d \cdot ln(\hat{d}) + \hat{d} = 2\sum(d \cdot ln(\frac{d}{\hat{d}}) - d + \hat{d}).$$

Ce qui donne, pour chaque modèle :

	Déviance
Log-Poisson	$2\sum_{x,t}(d_{x,t} \cdot ln(\frac{d_{x,t}}{\hat{d}_{x,t}}) - d_{x,t} + \hat{d}_{x,t})$
ACF-1	$2\sum_{x,t,i}(d_{x,t,i} \cdot ln(\frac{d_{x,t,i}}{\hat{d}_{x,t,i}}) - d_{x,t,i} + \hat{d}_{x,t,i})$
ACF-2	$2\sum_{x,t,i,j}(d_{x,t,i,j} \cdot ln(\frac{d_{x,t,i,j}}{\hat{d}_{x,t,i,j}}) - d_{x,t,i,j} + \hat{d}_{x,t,i,j})$
ACFC	$2\sum_{x,t,i,j,h=x-t}(d_{x,t,i,j} \cdot ln(\frac{d_{x,t,i,j}}{\hat{d}_{x,t,i,j}}) - d_{x,t,i,j} + \hat{d}_{x,t,i,j})$

L'algorithme calcule en continu la déviance et vérifie à chaque itération la différence entre la déviance de l'étape $k+1$ et celle de l'étape k et s'arrête dès que $|Dév^{k+1} - Dév^k| < 0,01$. Cette valeur de 0,01, bien qu'arbitraire est satisfaisante comme valeur de contrôle sur la diminution de la déviance. En effet, la déviance est d'au moins 6.000 (voir graphique en 3.3) et $\frac{0,01}{6000} \leq 10^{-5}$, ce qui est plus que suffisant.

3.2.5 Construction de l'algorithme.

L'algorithme est construit en 5 étapes, en se basant sur la même structure que [23].

Étape 0 Initialisation des paramètres, pour tous les $x, t, i, j, h = t - x$

$$\widehat{a}_{x,i,j} = \frac{1}{t}\sum_t ln(m_{x,t,i,j}) \quad \widehat{B}_x = \widehat{b}_{x,i} = \widehat{b}_{x,i,j} = 1/101$$
$$\widehat{K}_t = \widehat{\kappa}_{t,i} = \widehat{\kappa}_{t,i,j} = 0 \qquad g_{h,i} = 0$$

.

Étape 1 Modélisation de $\widehat{a}_{x,i,j} + \widehat{B}_x\widehat{K}_t$ à partir des formules d'itération de 3.2.3. S'ârreter à cette étape pour le modèle log-Poisson.

Étape 2 Modélisation de $\widehat{b}_{x,i}\widehat{\kappa}_{t,i}$ à partir des formules d'itération de 3.2.3. S'ârreter à cette étape pour le modèle ACF-1.

Étape 3 Modélisation de $g^*_{h,i}$ à partir des formules d'itération de 3.2.3. Ne faire cette étape que pour le modèle ACFC.

Étape 4 Modélisation de $\widehat{b}_{x,i,j}\widehat{\kappa}_{t,i,j}$ à partir des formules d'itération de 3.2.3. Cette étape donne le modèle ACFC si l'étape 3 a été faite ou le modèle ACF-2 si elle a été sautée.

Identification des paramètres. Afin de prendre en compte les contraintes d'idendifications, il y a un update des paramètres à chaque itération, pour chaque modèle. Cet update se fait de la manière suivante

$$\widetilde{K}_t = \left(\widehat{K}_t - \overline{K}\right) \sum_{x=x_1}^{x_{101}} \widehat{B}_x$$

$$\widetilde{B}_x = \frac{\widehat{B}_x}{\sum\limits_{x=x_1}^{x_{101}} \widehat{B}_x}$$

$$\widetilde{\alpha}_{x,i,j} = \widehat{\alpha}_{x,i,j} + \widehat{B}_x \overline{K}$$

Avec

— $\overline{K} = \sum\limits_{t=1}^{68} \widehat{K}_t$;

— $\widehat{K}_t$ qui représente $\widehat{K}_t$, $\widehat{\kappa}_{t,i}$ ou $\widehat{\kappa}_{t,i,j}$ en fonction de l'étape ;

— $\widehat{B}_x$ qui représente $\widehat{B}_x$, $\widehat{b}_{x,i}$ ou $\widehat{b}_{x,i,j}$ en fonction de l'étape ;

Cette update ne change pas la valeur des taux de mortalité. En effet, en prenant, par exemple

$$f(x,t) = \widehat{a}_{x,i,j} + \widehat{B}_x \widehat{K}_t \tag{3.1}$$

et en remplacant par les valeurs d'update associée

$$f(x,t) = \widetilde{\alpha}_{x,i,j} + \widetilde{B}_x \widetilde{K}_t$$

$$f(x,t) = \widehat{\alpha}_{x,i,j} + \widehat{B}_x \overline{K} + \frac{\widehat{B}_x}{\sum\limits_{x=x_1}^{x_{101}} \widehat{B}_x} \cdot \left(\widehat{K}_t - \overline{K}\right) \sum_{x=x_1}^{x_{101}} \widehat{B}_x$$

$$f(x,t) = \widehat{\alpha}_{x,i,j} + \widehat{B}_x \overline{K} + \widehat{B}_x \widehat{K}_t - \widehat{B}_x \overline{K}$$

$$f(x,t) = \widehat{a}_{x,i,j} + \widehat{B}_x \widehat{K}_t$$

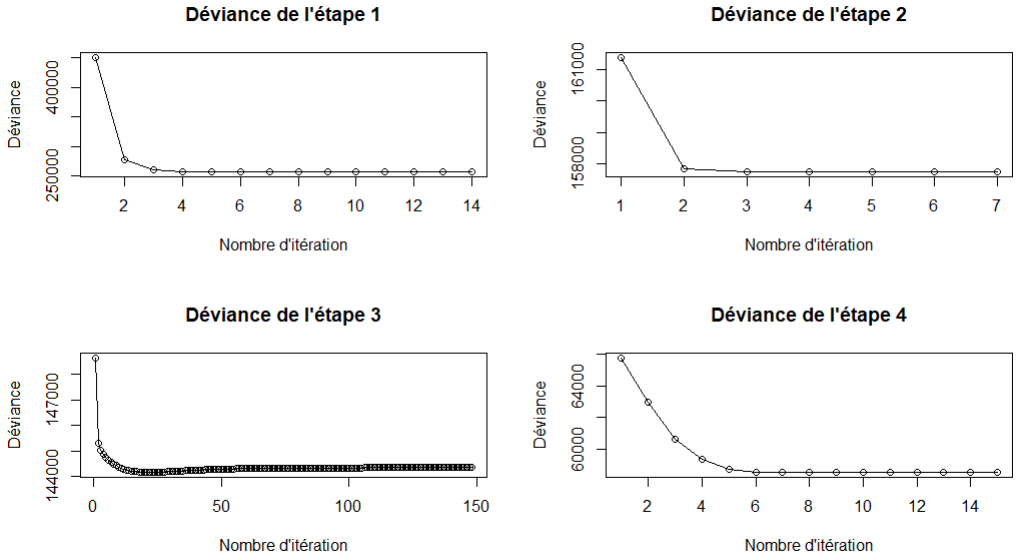
Ce qui donne bien l'équation (3.1) et ne modifie pas l'évaluation des taux de mortalité.

3.3 Résultats et graphiques - Modèle ACFC.

La calibration a été réalisée avec les données couvrant les années 1946 à 2013 inclues. Les années 2014 à 2018 serviront pour la validation des modèles dans le chapitre 4.

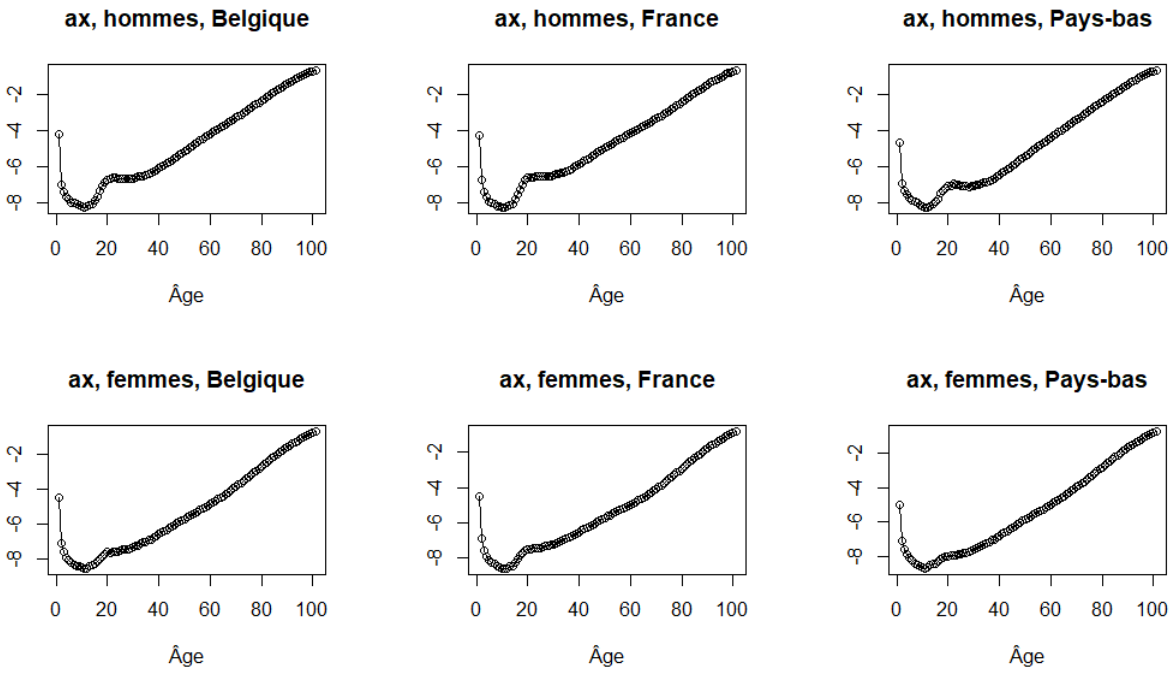
Convergence de la déviance.

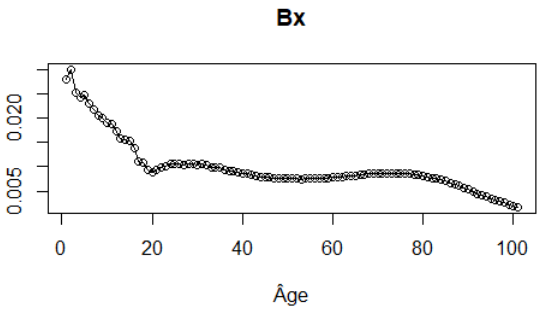
Voici les graphes des déviances à chaque étape de l'algorithme, en fonction du nombre d'itérations.

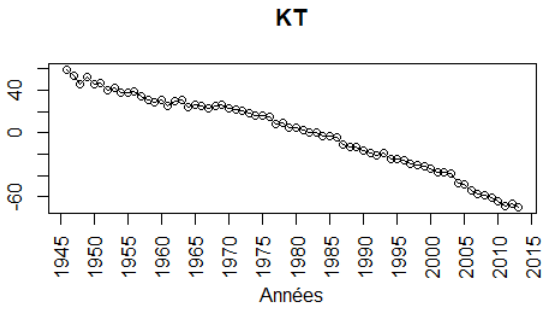


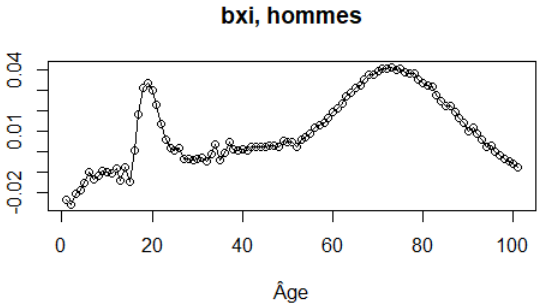
Evaluation des paramètres du modèle AFCF.

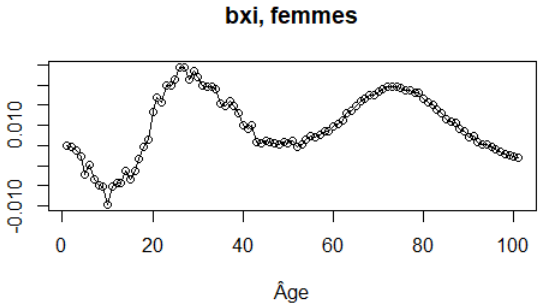
Ci-dessous, les graphiques des paramètres $a_{x,i,j}$, B_x, K_t, $b_{x,i}$, $\kappa_{t,i}$, $b_{x,i,j}$, $\kappa_{t,i,j}$ et $g_{h,i}$ du modèle ACFC.

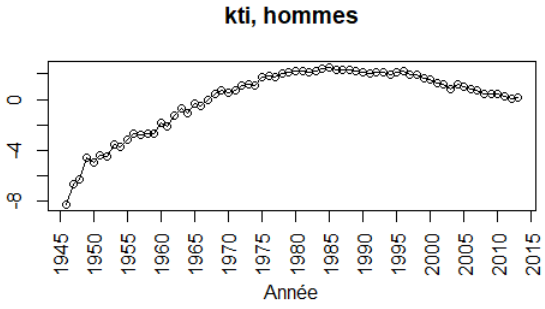


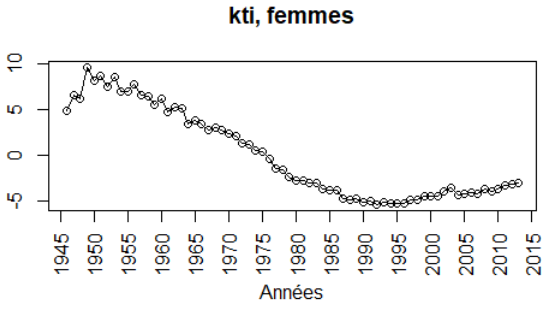

Bx
0.005
0.020
0
20
40
60
80
100
Âge

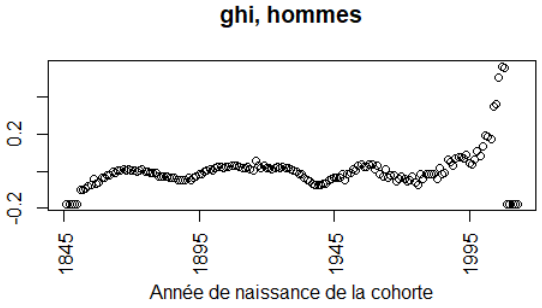

KT
-60
0
40
1945
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
Années

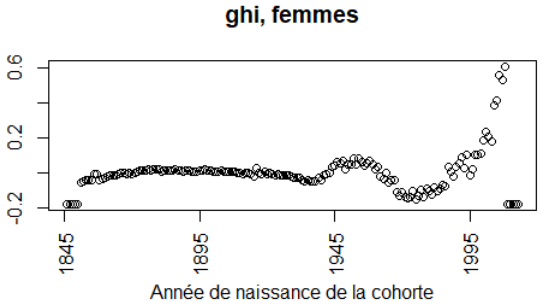

bxi, hommes
-0.02
0.01
0.04
0
20
40
60
80
100
Âge

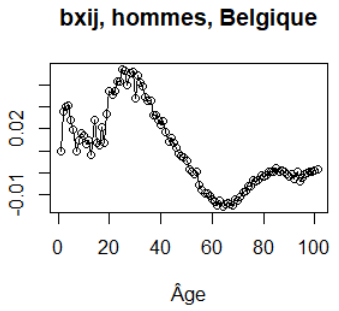

bxi, femmes
-0.010
0.010
0
20
40
60
80
100
Âge

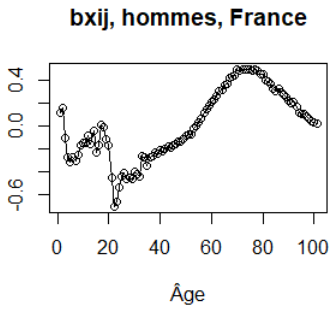

kti, hommes
-8
-4
0
1945
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
Année

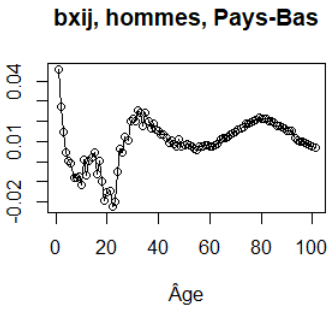

kti, femmes
-5
0
5
10
1945
1950
1955
1960
1965
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
Années

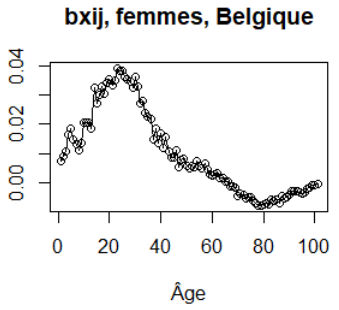

ghi, hommes
-0.2
0.2
1845
1895
1945
1995
Année de naissance de la cohorte

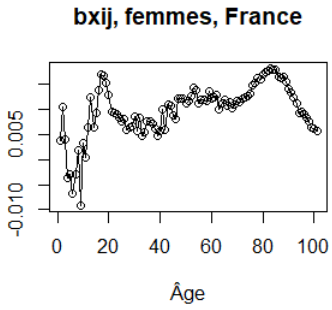

ghi, femmes
-0.2
0.2
0.6
1845
1895
1945
1995
Année de naissance de la cohorte

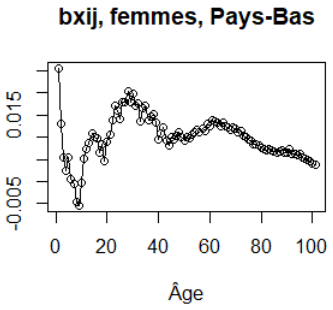

bxij, hommes, Belgique
-0.01
0.02
0
20
40
60
80
100
Âge

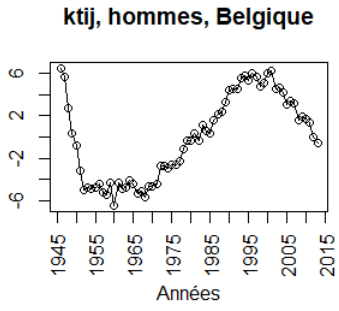

bxij, hommes, France
-0.6
0.0
0.4
0
20
40
60
80
100
Âge

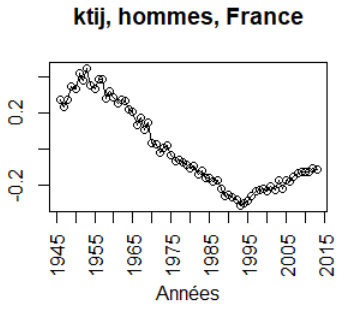

bxij, hommes, Pays-Bas
-0.02
0.01
0.04
0
20
40
60
80
100
Âge

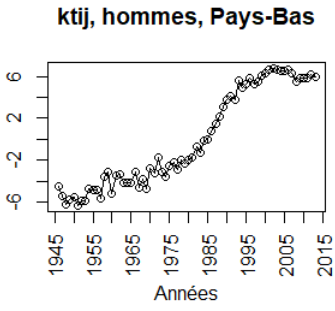

bxij, femmes, Belgique
0.00
0.02
0.04
0
20
40
60
80
100
Âge

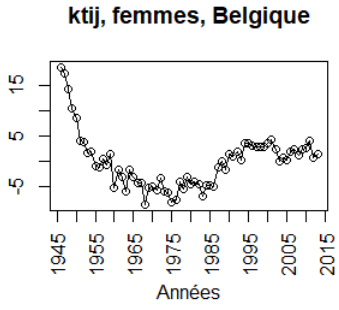

bxij, femmes, France
-0.010
0.005
0
20
40
60
80
100
Âge

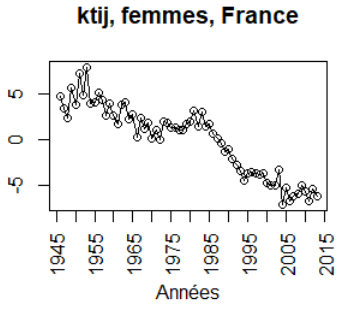

bxij, femmes, Pays-Bas
-0.005
0.015
0
20
40
60
80
100
Âge

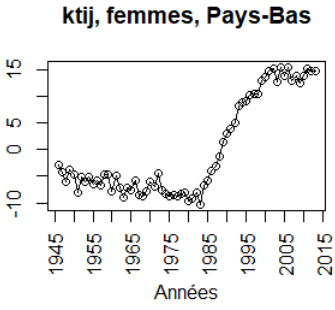

ktij, hommes, Belgique
-6
-2
2
6
1945
1955
1965
1975
1985
1995
2005
2015
Années


ktij, hommes, France
-0.2
0.2
1945
1955
1965
1975
1985
1995
2005
2015
Années


ktij, hommes, Pays-Bas
-6
-2
2
6
1945
1955
1965
1975
1985
1995
2005
2015
Années


ktij, femmes, Belgique
-5
5
15
1945
1955
1965
1975
1985
1995
2005
2015
Années


ktij, femmes, France
-5
0
5
1945
1955
1965
1975
1985
1995
2005
2015
Années


ktij, femmes, Pays-Bas
-10
0
5
15
1945
1955
1965
1975
1985
1995
2005
2015
Années

3.4 Interprétations des résultats.

Afin de pouvoir interpréter correctement les graphiques calibrés dans la section précédente, il est important de se rappeler le lien entre les probabilités de survie, les taux de mortalité $\mu_{x,t,i,j}$ et les différents paramètres modélisés : $a_{x,i,j}$, B_x, K_t, $b_{x,i}$, $\kappa_{t,i}$, $b_{x,i,j}$, $\kappa_{t,i,j}$ et $g_{h,i}$.

Pour rappel, les probabilités de survie $p_x(t)$ sont fonction de l'exponentielle des $-\mu_{x,t,i}$ qui eux-mêmes sont fonction positive de l'exponentielle des paramètres modélisés, le lien entre probabilités et paramètres est donc :

$$p_x(t) \propto -\mu_{x,t,i,j} \propto -\left(a_{x,i,j} + B_x K_t + b_{x,i}\kappa_{t,i} + b_{x,i,j}\kappa_{t,i,j} + g^*_{h,i}\right).$$

Plus les paramètres de modélisation sont élevés, plus la probabilité de survie diminue et donc plus le taux de mortalité augmente. L'interprétation des produits est plus compliquée car il suffit que le signe d'un des deux termes du produit change pour changer l'influence de ce produit sur $p_x(t)$. Néanmoins, si un des deux termes est fixé positivement, une augmentation du deuxième entraine une diminution de la probabilité de survie (et inversement en cas de diminution) tandis que s'il est fixé négativement, une augmentation du deuxième entraine une augmentation de la probabilité de survie (et inversement en cas de diminution).

Ces remarques étant faites, il est maintenant possible d'essayer d'interpréter les graphiques.

Interprétation générale.

- Les paramètres $a_{x,i,j}$, B_x et K_t sont lissés, ce qui ne correspond pas à la réalité (les probabilités réelles étant erratiques) mais ils donnent la tendance globale. Les autres paramètres sont là pour faire des ajustements en fonction du sexe et du pays et il est difficile de rattacher leurs évolutions à des facteurs sociétaux.
- Les paramètres $a_{x,i,j}$ croissent, ce qui montre une mortalité de plus en plus élevée avec l'avancement de l'âge.
- Dans la modélisation des $a_{x,i,j}$, il y a une forte mortalité infantile (mortalité des enfants de moins d'un an). Cette mortalité élevée dans la première année de vie décroît jusqu'à 11-12 ans, tranche d'âge représentant la population qui décède le moins.
- Toujours dans les graphiques des $a_{x,i,j}$, le « pic des 20 ans » est bien présent. Il s'agit d'une pente plus forte entre 15 et 20 ans qui indique une surmortalité à ces âges-là. Cette surmortalité est due à des accidents de la route dus à une mauvaise conduite et à des actes inconscients, deux comportements typiquement liés à cette tranche d'âge. À noter que ce phénomène est plus présent chez les hommes que chez les femmes, et que ceci est bien reflété dans les graphiques des $a_{x,i,j}$. (voir [2])
- Les paramètres K_t diminuent, ce qui indique une diminution de la mortalité avec les années qui passent, à âge constant.
- Les paramètres B_x ont tendance à diminuer, ce qui indique que plus une personne vieillit, moins l'influence du paramètre K_t sera importante. Intuitivement, les populations jeunes profitent plus de l'amélioration de la qualité de vie que les personnes plus âgées. Puisqu'il existe un âge maximum à la table des âges, l'influence positive de K_t sur la mortalité ne peut pas avoir la même répercussion sur tous les âges, au risque de dépasser l'âge maximum.

L'ensemble de ces constats donne une image cohérente de la répartition de la mortalité au sein d'une population et les modèles ont donc l'air de tourner correctement. Le chapitre suivant

permet de comparer les modèles à la réalité afin de savoir si, en plus d'être cohérents, ils rendent les bonnes valeurs passées et futures de la mortalité.

Cas particulier des $g^*_{h,i}$.

Maintenant que la modélisation des termes de cohortes $g^*_{h,i}$ est faite, la question posée en 2.5 peut être répondue. L'idée est de savoir si le trio Belgique-France-Pays-Bas présente un effet de cohorte similaire à celui observé au Royaume-Uni.

L'analyse des graphiques et des valeurs des paramètres $g^*_{h,1}$ et $g^*_{h,2}$ montre un premier effet cohorte pour les cohortes nées entre +- 1930 et +- 1947. Sur cette période, la valeur des paramètres est plus basse que les périodes qui précèdent et qui suivent. Cela veut dire que l'influence des paramètres fait augmenter la probabilité de survie et donc diminuer la mortalité pour les générations nées durant cette période. À noter que, pour les femmes, l'effet de cohorte de la période qui suit ce creux est plus élevé que celle de la période qui le précède, effet qui n'apparaît pas chez les hommes.

Il est très intéressant de remarquer que cette période correspond plus ou moins à la même période présentant un effet cohorte au Royaume-Uni, comme expliqué en [23]. Cette période, couvrant les cohortes nées entre 1925 et 1945, montrait également une diminution de la mortalité par rapport aux périodes précédentes et suivantes.

Une analyse socio-économique de cette période couplée à une analyse comparative entre le Royaume-Uni d'une part et l'axe Belgique-France-Pays-Bas d'autre part permettrait de cerner les causes et les similarités entre les deux situations. Cette question, dépassant le cadre de ce mémoire, ne sera pas traitée ici.

Un autre effet de cohorte, plus léger et concernant uniquement les hommes, est également présent pour les cohortes nées entre +- 1875 et +- 1895. La même question que celle posée pour le premier effet mis en évidence se pose également pour cette période, à savoir, quels sont les facteurs socio-économiques associés à cette diminution de la mortalité ? Une période similaire existe-t-elle aussi pour le Royaume-Uni ?

Une dernière particularité des paramètres $g^*_{h,i}$ est l'explosion de ceux-ci pour les cohortes de moins de 20 ans. Cet aspect sera traité dans la section 4.3.1.

Chapitre 4

Validation.

Afin de pouvoir valider les modèles présentés dans le chapitre 2 et de savoir lesquels correspondent le mieux aux données empiriques, il faut passer par la projection dans le temps des paramètres calibrés afin de voir si ces projections donnent des valeurs cohérentes par rapport aux données réelles. Cette vérification passe par la comparaison des espérances de vie, quantité qui permet de prendre en compte l'entièreté des paramètres descriptifs des modèles.

Ce chapitre aborde dans un premier temps la notion de séries temporelles (en particulier le modèle $ARIMA(p, d, q)$) et les outils utiles à leurs analyses ainsi que l'application de cette notion sur les variables dépendantes du temps des modèles LC et ACFC afin de pouvoir les projeter dans le futur.

Dans un second temps, ce chapitre compare les espérances de vie projetées et les espérances de vie empiriques via la notion d'écart quadratique moyen.

À noter que l'analyse comparative concerne uniquement les modèles LC et ACFC afin de pouvoir comparer le modèle plus classique (LC) et le modèle le plus compliqué parmi ceux développés (ACFC).

Pour la calibration des variables temporelles, les données utilisées sont celles comprises entre 1970 et 2013 et ceci afin de capturer au mieux la tendance actuelle de ces variables. En effet, l'utilisation de données trop anciennes comporte un risque de faire apparaitre des schémas mathématiques qui ne sont pas pertinents par rapport à la conjoncture actuelle, comme par exemple l'influence des différentes guerres mondiales ou de grosses épidémies. De plus, certaines données sont manquantes/non pertinentes quand l'on remonte trop loin dans les années (par exemple les années 1914-1918 sont manquantes en Belgique).

Pour la partie projection, les données utilisées sont celles des années 2014 à 2018.

4.1 Théorie sur les séries temporelles.

Une série temporelle est une suite de variables aléatoires qui représente l'évolution d'un paramètre au cours du temps. La particularité des séries temporelles par rapport aux séries classiques vient du fait que, bien que présentants un volet stochastique, les termes de la suite dépendent de la valeur des termes antérieurs. En ce sens, les paramètres K_t, $\kappa_{t,i}$, $\kappa_{t,i,j}$ et $g^*_{h,i}$ sont des séries temporelles puisque tous les quatre dépendent de la variable t (h étant donné par $t - x$) et pour lesquels est présupposé un lien historique entre les valeurs.

Afin de pouvoir faire des prédictions à partir des différents modèles calibrés, il est indispensable de pouvoir projeter ces séries dans le futur. Pour ce faire, il existe plusieurs manières

d'analyser mathématiquement les séries temporelles. L'analyse choisie ici est celle du modèle « Auto-Regressive – Integrated – Moving Average » développé par Box-Jenkis [4], dit modèle $ARIMA(p,d,q)$.

Partant d'une série temporelle donnée $(Y_t)_{t\in\mathbb{N}}$, le modèle $ARIMA(p,0,q)$ qui modélise cette série peut s'écrire comme

$$Y_t = \mu + \sum_{i=1}^{p} \varphi_i Y_{t-i} + \sum_{i=0}^{q} \theta_i \epsilon_{t-i}, \forall t \in \mathbb{N} \tag{4.1}$$

pour p,q des entiers et μ, φ_i et θ_i des réels à déterminer, avec$\theta_0 = 1$ et ϵ_t qui est un bruit blanc.

Étant une notion très souvent utilisée en modélisation stochastique, un bruit blanc est une suite de variables aléatoires indépendantes et identiquement distribuées de type gaussienne, de moyenne égale à 0 et de variance constante σ^2, $\epsilon_t \sim N(0,\sigma^2)$. Ce paramètre est très utile afin de pouvoir simuler les sauts aléatoires qui apparaissent dans des processus non déterministes.

4.1.1 Partie I(d) - Différenciation.

La modélisation d'un processus ARIMA suppose que la série soit stationnaire, c'est-à-dire que la moyenne et la variance de la série soient des paramètres constants au cours du temps. Or il est possible que la série analysée ne soit pas stationnaire. Afin de remédier à cela, une technique utile est celle de la différenciation. Pour ce faire, il faut remplacer la série temporelle d'origine Y_t par la série des différences adjacentes Z_t et travailler avec cette dernière, ceci dans l'objectif que Z_t ait une moyenne μ et une variance σ^2 constante.

Par exemple pour $d = 1$, il s'agit des différences successives

$$Z_t = Y_t - Y_{t-1} = \mu + \epsilon_t, \forall t \in \mathbb{N}.$$

Pour $d = 2$, il s'agit des différences de différences

$$Z_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} - Y_{t-2} = \mu + \epsilon_t, \forall t \in \mathbb{N}.$$

Et ainsi de suite pour les valeurs supérieures de d.

Le modèle mathématique à utiliser dépend donc du nombre de différenciations réalisées. Une fois le nombre adéquat de différenciations fait afin que Z_t soit stationnaire, cette série peut être modélisée à l'aide des termes $AR(p)$ et $MA(q)$ définis ci-dessous et ainsi suivre le modèle général $ARIMA(p,0,q)$ de l'équation (4.1), Z_t reprenant le rôle de Y_t.

4.1.2 Partie AR(p) - Auto-Regression.

La partie AutoRégressive du modèle $ARIMA(p,d,q)$, notée $ARIMA(p,0,0)$ ou $AR(p)$ est donnée par

$$Y_t = \sum_{i=1}^{p} \varphi_i Y_{t-i} + \epsilon_t$$

où ϵ_t est un bruit blanc.

Cette partie du modèle part du principe que chaque observation est composée d'une combinaison linéaire des observations précédentes et d'un choc aléatoire ϵ_t stochastique, φ_i étant un réel qui pondère le poids de chaque Y_{t-i} dans le Y_t calculé.

Plus la valeur de p est élevée, plus l'historique passé de la série temporelle a de l'importance dans les nouvelles valeurs de cette série.

4.1.3 Partie MA(q) - Moyenne Mobile.

La partie Moyenne-Mobile du modèle $ARIMA(p,d,q)$, notée $ARIMA(0,0,q)$ ou $MA(q)$ est donnée par

$$Y_t = \mu + \sum_{i=0}^{q} \theta_i \epsilon_{t-i}$$

où ϵ_t est un bruit blanc, les θ_i sont des réels et $\theta_0 = 1$.

Cette partie du modèle suppose que la série fluctue autour d'une valeur moyenne μ. La meilleure façon d'estimer Y_t est alors d'en faire une somme entre la moyenne μ et une fonction linéaire des résidus antérieurs ϵ_{t-i}. Ainsi, les différents Y_t sont dispersés autour de la valeur μ, d'où le nom de « Moyenne Mobile ».

De la même manière que φ_i dans le modèle $AR(p)$, le paramètre θ_i vient pondérer le poids des anciennes valeurs de ϵ_t dans la modélisation de Y_t. Plus la valeur de q est importante, plus l'historique des résidus antérieurs sera pris en compte dans la modélisation des nouvelles valeurs de la série temporelle.

4.1.4 Le modèle ARIMA (p,q,d).

Lors de l'analyse d'une série temporelle, il faut donc d'abord différencier la série le nombre adéquat de fois afin qu'elle soit stationnaire. Une fois cela fait, la série peut se voir comme la somme d'une partie d'Auto-Régression $AR(p)$ ainsi que d'une partie de Moyenne Mobile $MA(q)$, ce qui redonne l'équation générale (4.1) du modèle $ARIMA(p,0,q)$.

4.2 Outils d'analyse.

Afin de pouvoir identifier le modèle $ARIMA$ que suit une série temporelle et de savoir quels sont les paramètres p, d et q qui modélisent au mieux l'évolution des K_t, $\kappa_{t,i}$, $\kappa_{t,i,j}$ et $g^*_{h,i}$, il faut vérifier deux choses :

- la stationnarité de la série. Cela peut se faire en analysant l'allure générale de l'autocorrélation de la série via le test de Box-Jenkins. Pour ce faire, il est utile d'utiliser la notion de fonction d'autocorrélation (notée ACF pour *AutoCorrelation Function*) et celle de fonction d'autocorrélation partielle (notée PACF pour *Partial Autocorrelation Function*).
- la normalité des résidus, afin de pouvoir les simuler par un bruit blanc ϵ_t lors de la calibration du modèle. Pour rappel, le résidu dans une régression est ce qui n'est pas expliqué par le modèle et qui doit suivre, dans le cas présent, une loi Normale. Le test de Jarque-Bera sera utile afin de vérifier cette hypothèse.

4.2.1 Les fonctions ACF et PACF.

Fonction d'autocorrélation.

La fonction d'autocorrélation est une application de la notion de coefficient de corrélation entre deux variables X et Y à une série temporelle et à une version décalée dans le temps d'elle-même.

Pour rappel, en supposant deux variables aléatoires X et Y, le coefficient de corrélation est donné par la formule

$$r = \frac{\sum \left[(X_i - \overline{X})(Y_i - \overline{Y})\right]}{\sqrt{\sum (X_i - \overline{X})^2 (Y_i - \overline{Y})^2}}.$$

Ce coefficient, compris entre -1 et 1, code le lien de dépendance linéraire entre deux variables. Pour $r = 1$, les deux variables sont positivement corrélées, c'est-à-dire qu'une augmentation de la variable X entraîne une augmentation de la variable Y et qu'à contrario une diminution de la variable X entraîne une diminution de la variable Y.

À l'inverse, pour $r = -1$, les deux variables sont négativement corrélées, c'est-à-dire qu'une augmentation de la variable X entraîne une diminution de la variable Y et qu'une diminution de la variable X entraîne une augmentation de la variable Y.

Un coefficient de corrélation de 0 signifie que les deux variables ne sont pas linéairement dépendantes mais il pourrait très bien y avoir un autre lien de dépendance (en x^2 ou en $log(x)$, par exemple).

Appliquer cette formule à la série temporelle $(Y_t)_{t\in\mathbb{N}}$ et à sa version décalée $(Y_{t+k})_{t\in\mathbb{N}}$ donne pour l'ACF

$$ACF_k = \frac{\sum \left[(Y_i - \overline{Y})(Y_{i+k} - \overline{Y})\right]}{\sqrt{\sum (Y_i - \overline{Y})^2 (Y_{i+k} - \overline{Y})^2}}.$$

Où k est la valeur du lag, c'est-à-dire le décalage induit entre deux versions de la série temporelle $(Y_t)_{t\in\mathbb{N}}$.

Une valeur proche de 1/-1 de l'ACF_k entraîne donc une forte corrélation positive/négative entre la série temporelle $(Y_t)_{t\in\mathbb{N}}$ et sa version décalée $(Y_{t+k})_{t\in\mathbb{N}}$.

Fonction d'autocorrélation partielle.

La fonction d'autocorrélation partielle mesure également la corrélation entre une série $(Y_t)_{t\in\mathbb{N}}$ et sa version décalée $(Y_{t+k})_{t\in\mathbb{N}}$ mais cette fois en supprimant les effets des séries intermédiaires $(Y_{t+i})_{t\in\mathbb{N}}$ avec $1 \leq i < k$. Formellement, il s'agit du dernier coefficient de la projection linéaire de $(Y_{t+k})_{t\in\mathbb{N}}$ sur les valeurs de $(Y_{t+i})_{t\in\mathbb{N}}$ avec $0 \leq i < k$. Vu que cette projection linéaire doit être calculée pour chaque valeur de k, il n'existe pas de formule générale pour les $PACF_k$.

Par exemple, pour les 3 premières valeurs de k, les $PACF$ sont donnée par (voir [8])

$$p_1 = \rho_1$$

$$p_2 = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2}$$

$$p_3 = \frac{\rho_1^3 - \rho_1\rho_2(2-\rho_2) + \rho_3(1-\rho_1^2)}{1-\rho_2^2 - 2\rho_1^2(1-\rho_2)}$$

où les p_k représentent les $PACF_k$ et les ρ_k représentent les ACF_k.

4.2.2 Le test de Box-Jenkins.

Le test de Box-Jenkins consiste en une analyse des graphiques des séries temporelles ainsi que de leurs fonctions d'autocorrélations et de leurs fonctions d'autocorrélations partielles. Le but étant, grâce à cette analyse, de trouver les paramètres p, d et q qui calibrent le mieux le modèle ARIMA appliqué aux séries.

Deux étapes sont importantes dans ce processus (voir [15]) :

1. Vérifier la stationnarité, c'est-à-dire que la moyenne μ et la variance σ^2 soient constantes :
   - Analyser le graphique de la série afin de voir si la variance est constante. Ceci peut se voir visuellement par l'absence de saccades d'intensités fortement différentes dans ledit graphique. Si la variance n'est pas constante, il peut être utile d'appliquer une transformation (du type logarithme par exemple).
   - Analyser le graphique des ACF_k et regarder si les autocorrélations fortes disparaissent rapidement. Si ce n'est pas le cas, c'est qu'il y a une tendance dans la série qui crée les fortes autocorrélations, puisque cette tendance est présente dans chaque terme. (Une tendance linéaire par exemple, où les termes de la série oscillent autour d'une droite). La moyenne de la série n'est donc pas constante. À ce moment-là, il est utile de différencier la série afin d'obtenir une disparition rapide des autocorrélations fortes, typique d'une moyenne constante. Cette étape permet de trouver le paramètre d.
2. Analyser les graphiques des ACF_k et des $PACF_k$ des séries stationnaires afin de trouver les paramètres d'Auto-Régression p et de Moyenne Mobile q.
   - Si l'ACF présente des pics importants qui tendent vers zéro aux décalages initiaux et que la $PACF$ présente un seul pic important au premier décalage (et éventuellement au second), cela veut dire que les termes de la série sont fonction directement des termes qui les précèdent et donc qu'il s'agit d'un processus Auto-Régressif $AR(p)$.
   - A l'inverse, si l'ACF présente un seul pic important au premier décalage (et éventuellement au second) et que la $PACF$ présente des pics importants qui tendent vers zéro aux décalages initiaux, cela veut dire que les termes de la série ne sont pas fonction des termes qui les précèdent mais qu'ils fluctuent autour d'une moyenne et donc qu'il s'agit d'un processus à Moyenne Mobile $MA(q)$.
   - Si la série présente en même temps une ACF et une $PACF$ qui présentent des pics importants qui tendent vers zéro aux décalages initiaux, alors elle présente un volet Auto-Régressif $AR(p)$ et un volet à Moyenne Mobile $MA(q)$.

En couplant les analyses faites aux points 1 et 2, il est possible de créer le modèle $ARIMA(p,d,q)$ qui décrit le mieux la série temporelle.

4.2.3 Le test de Jarque-Bera.

Le test de Jarque-Bera est un test statistique servant à tester la normalité d'une distribution. Le test n'analyse pas directement si la distribution suit une loi normale mais teste les valeurs de Skewness et de Kurtosis.

Pour rappel, X étant une variable aléatoire :

— Le paramètre de Skewness $S = \mathbb{E}\left(\frac{X-\mu}{\sigma}\right)^3$ est le moment d'ordre 3 d'une distribution et caractérise son asymétrie. Si S est nul, comme pour la loi Normale, la distribution est symétrique. Elle sera soit décalée à droite de la médiane pour un $S < 0$ soit décalée vers la gauche de la médiane pour un $S > 0$.

— Le paramètre de Kurtosis $K = \mathbb{E}\left(\frac{X-\mu}{\sigma}\right)^4$ est le moment d'ordre 4 d'une distribution et caractérise l'aplatissement de celle-ci. Pour une loi Normale, le Kurtosis est de 3. S'il est plus élevé que 3, la distribution est plus haute et présente une queue plus épaisse que la loi Normale, indiquant des valeurs extrêmes plus fréquentes. S'il est plus bas que 3, la distribution est aplatie par rapport la loi Normale et les événements extrêmes sont plus rares.

Le test de Jarque-Bera va vérifier si les valeurs de Skewness et de Kurtosis sont, respectivement, différentes de 0 et de 3, (c'est-à-dire que $S \neq 0$ et $K \neq 3$).

Puisque les critères $S = 0$ et $K = 3$ sont des conditions nécessaires mais pas suffisantes pour qu'une distribution suive une loi normale, le test de Jarque-Bera peut montrer qu'une distribution ne suit pas une loi Normale mais ne peut pas montrer l'inverse.

Ceci pris en compte, le test de Jarque-Bera reste un très bon indicateur pour savoir si la distribution étudiée peut être considérée comme suivant une loi Normale.

La statistique associée au test de Jarque-Bera, JB est donnée par

$$JB = \frac{n-k}{6}\left(S^2 + \frac{(K-3)^2}{4}\right).$$

La statistique JB suit asymptotiquement une loi χ^2_2 et les deux hypothèses testées sont :

— H_0 : $JB = 0$ ou alternativement H_0 : $S = 0$ et $K = 3$.

— H_1 : $JB \neq 0$ ou alternativement H_1 : $S \neq 0$ et /ou $K \neq 3$.

La fonction R « JarqueBera.test », présente dans le package « tsoutliers » permet de calculer les trois statistiques suivantes, ainsi que leurs $p-$valeurs respectives :

— La statistique $JB = 0$.

— La statistique $S = 0$.

— la statistique $K = 3$.

4.3 Analyse des différentes séries temporelles.

Le test de Box-Jenkins permet de donner une estimation des paramètres p, d et q. Pour ce faire, il faut analyser visuellement les graphiques ACF_k et $PACF_k$. Ces graphiques présentent une ligne pointillée bleue. Si les valeurs des fonctions ACF et PACF sont au dessus de cette ligne, il s'agit de pics significatifs pour l'analyse présentée en 4.2.2.

Les sections suivantes vont analyser chaque variable temporelle des modèles LC et ACFC et tenter de trouver la meilleure calibration $ARIMA(p, d, q)$ possible pour chacune d'entre elles.

4.3.1 Analyse des $g^*_{h,i}$ du modèle ACFC.

En observant les graphiques de la section 3.3, un problème se pose lors de l'analyse des $g^*_{h,i}$. Il est clair que les valeurs de $g^*_{h,1}$ et $g^*_{h,2}$ explosent pour les 20 dernières cohortes.

Puisque chaque valeur de h représente une cohorte et que $h = t - x$, les dernières valeurs de h n'auront que peu de valeur de $\mu_{x,t}$ disponible pour la calibration. Par exemple la dernière cohorte avec $g^*_{h,i} \neq 0$, c'est-à-dire la cohorte $h = 163$, n'est composée que de 6 couples (x, t) qui ont comme valeurs $(0, 163)$, $(1, 164)$, $(2, 165)$, $(3, 166)$, $(4, 167)$ et $(5, 168)$. Plus le paramètre h est grand, moins il y a donc de $\mu_{x,t}$ disponible pour la modélisation, ce qui entraîne une sous-représentation des âges pris en compte dans la cohorte par rapport aux cohortes plus fournies.

Comme vu dans le chapitre 2, le taux de mortalité est fonction de l'exponentielle du terme de cohorte, $\mu_{x,t} \propto exp(g^*_{h,i})$. En se rappelant que la probabilité de survie est donnée par $p_x(t) = exp(-\sum_{i=0}^{t} \mu_{x,i})$, l'explosion positive des paramètres $g^*_{h,i}$ entraîne une augmentation significative du taux de mortalité et donc diminue la probabilité de survie des dernières cohortes. Une extrapolation à partir de ces données donnerait des termes de cohortes de plus en plus élevés et ainsi des probabilités de survie de plus en plus faibles. Les nouvelles générations auraient donc une espérance de vie à la naissance qui diminue, ce qui est en contradiction avec l'évolution positive constante de l'espérance de vie (voir [14] et [21]).

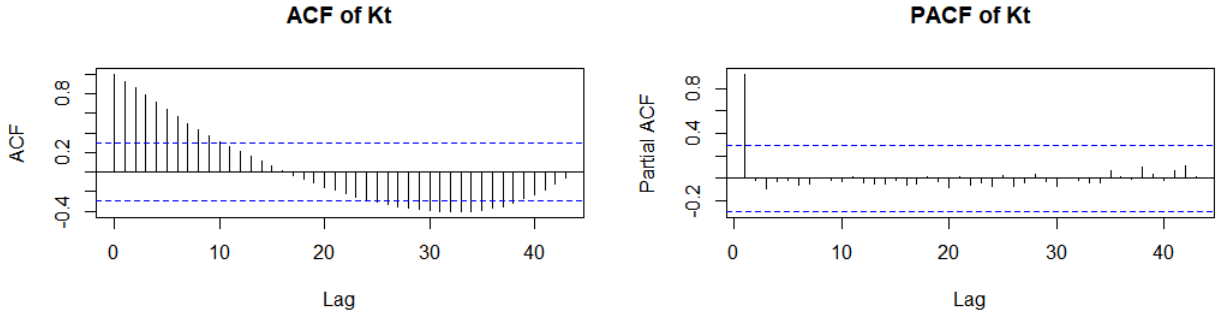
Dans [23], il a été décidé de mettre à zéro les 5 dernières et les 5 premières cohortes pour éviter cet effet de sous-représentation au sein d'une cohorte mais manifestement cela ne suffit pas pour pouvoir faire une prévision solide sur les valeurs de $g^*_{h,i}$.

Bien que l'analyse des $g^*_{h,i}$ soit intéressante d'un point de vue historique et au vu des remarques qui précèdent, la suite de ce travail se base sur l'utilisation du modèle ACF-2 au lieu du modèle ACFC afin que le pouvoir prédictif du modèle soit le plus efficace et le plus stable possible.

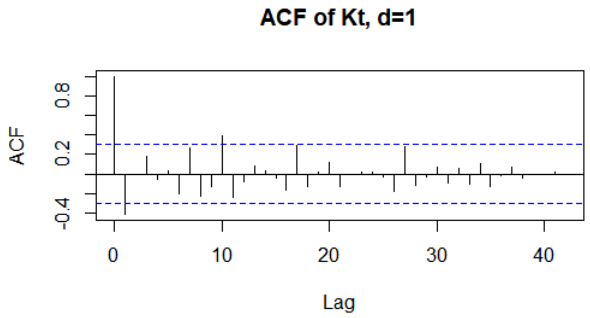
4.3.2 Analyse des K_t du modèle ACF-2.

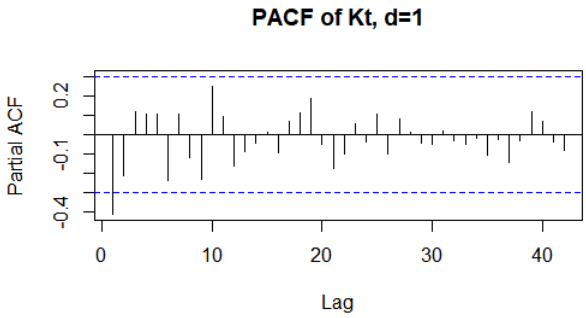
La première analyse porte sur le paramètre K_t qui est le paramètre général de dépendance temporelle du taux de mortalité, toutes régions et sexes confondus.

Les graphiques ACF et $PACF$ appliqués au paramètre K_t sont donnés ci-dessous. Pour rappel, le lag est la différence k entre la série $(K_t)_{t\in\mathbb{N}}$ et sa version décalée $(K_{t+k})_{t\in\mathbb{N}}$



Le graphique ACF montre que les autocorrélations fortes ne disparaissent pas rapidement, ce qui est le signe d'une tendance linéaire autour desquels les données fluctuent, le processus n'est donc pas stationnaire. Comme expliqué dans la section 4.2.2, une solution est de différencier une fois la série et de regarder à nouveau les graphiques ACF et $PACF$.





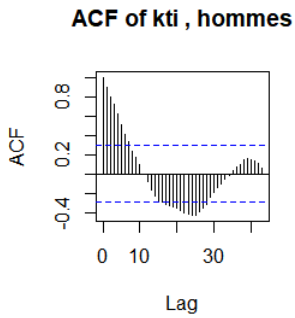
Les deux graphiques donnent cette fois-ci une allure stationnaire avec $d = 1$. Le graphique ACF montre une dépendance de type $AR(p)$ avec $p = 1$ puisqu'il y a un pic d'autocorrélation au premier lag. Le graphique $PACF$ montre qu'il n'y a pas de dépendance de type $MA(q)$ puisqu'il n'y a pas de pic après le premier lag et donc pas de retour à la moyenne.

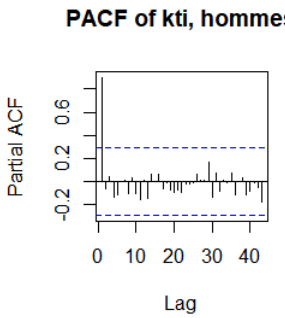
La série des K_t va donc être simulée selon un modèle $ARIMA(1, 1, 0)$.

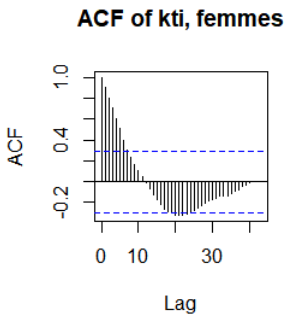
4.3.3 Analyse des $\kappa_{t,i}$ du modèle ACF-2.

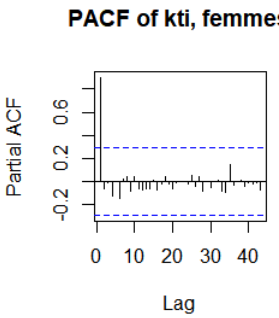
La seconde analyse porte sur les paramètres $\kappa_{t,i}$ qui sont les paramètres de dépendance temporelle du taux de mortalité par sexe, toutes régions confondues.

Les graphiques ACF et $PACF$ appliqués au paramètre $\kappa_{t,i}$ sont donnés ci-dessous. De la même façon qu'en 4.3.2, le lag est la différence k entre la série $(\kappa_{t,i})_{t\in\mathbb{N}}$ et sa version décalée $(\kappa_{t+k,i})_{t\in\mathbb{N}}$.

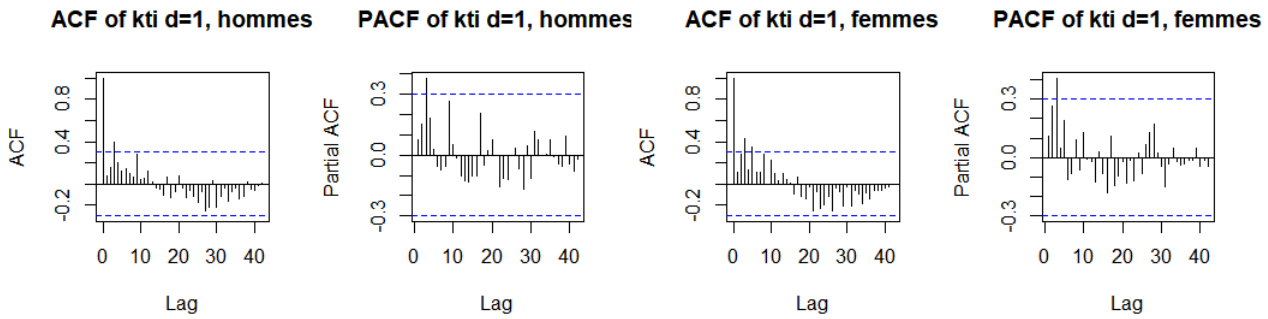








Le constat concernant la stationnarité est le même que pour les K_t et il faut donc différencier une fois les séries $(\kappa_{t,i})_{t\in\mathbb{N}}$, ce qui donne les graphiques ACF et $PACF$ suivant :

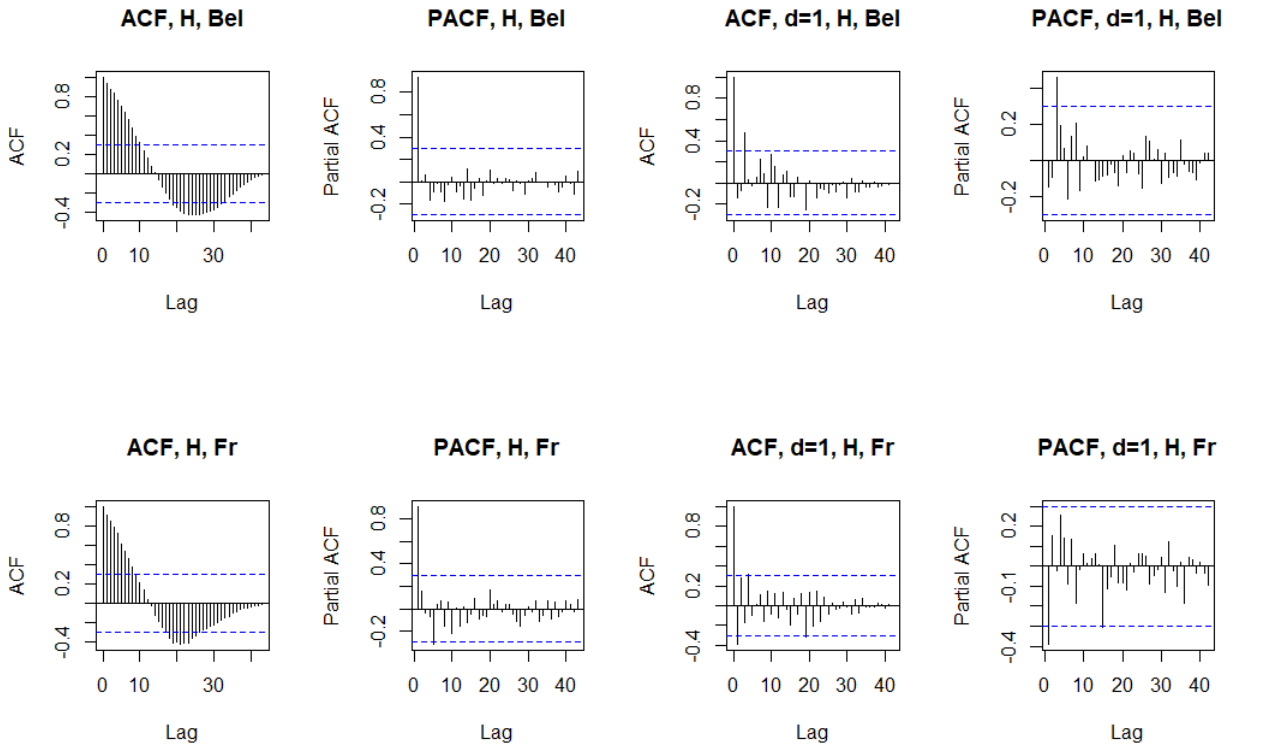


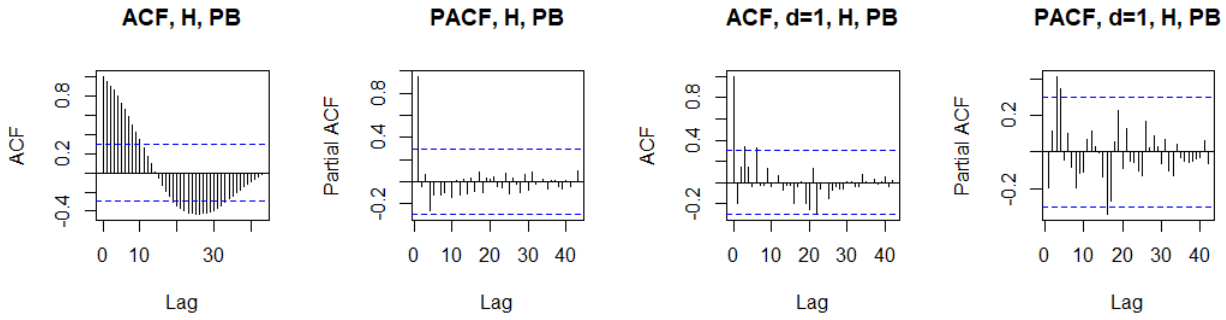
Les deux graphiques ont une allure stationnaire et donc $d = 1$. Le graphique ACF montre clairement une dépendance de type $AR(p)$ avec $p = 1$ et le graphique $PACF$ montre qu'il n'y a pas de dépendance de type $MA(q)$. Les séries des $\kappa_{t,i}$ vont donc être simulées selon un modèle $ARIMA(1, 1, 0)$.

4.3.4 Analyse des $\kappa_{t,i,j}$ du modèle ACF-2.

La troisième analyse porte sur les paramètres $\kappa_{t,i,j}$ qui sont les paramètres de dépendance temporelle du taux de mortalité par sexe et par région.

Ci-dessous, les résultats pour les hommes dans les trois pays avec à gauche, les ACF et $PACF$ de la série concernée et à droite, ceux de la série différenciée une fois. Il est clair que les $\kappa_{t,i,j}$ suivent également un modèle $ARIMA(1, 1, 0)$, de la même manière que les K_t et les $\kappa_{t,i}$. Afin de faciliter la lecture, les résultats concernants les femmes sont en annexes.





4.3.5 Analyse des K_t du modèle LC.

En faisant tourner le modèle de Lee-Carter sur chaque sous-ensemble (en ventilant par sexe et par pays), les paramètres K_t ainsi calculés sont assez similaires aux K_t généraux du modèle ACF-2.

Leurs traitements via le test de Box-Jenkis donnent également des résultats similaires et ils peuvent donc être simulés via un modèle $ARIMA(1,1,0)$.

Afin de ne pas alourdir la lecture de ce travail, les graphiques et tableaux afférents sont repris en annexe.

4.3.6 Analyse des résidus via le test de Jarque-Bera.

Afin de pouvoir correctement simuler les valeurs futures des séries chronologiques, il est important d'analyser les résidus. Comme discuté en 4.2.3, le test de Jarque-Bera permet de savoir si une variable aléatoire ne suit pas une loi normale.

Ci-dessous, le tableau avec les résultats des p-valeurs et des statistiques résultantes du test de Jarque-Bera appliqué aux variables K_t, $\kappa_{t,i,}$ et $\kappa_{t,i,j}$ du modèle ACF-2.

	$JB=0$		$S=0$		$K=3$	
	p-valeur	Stat	p-valeur	Stat	p-valeur	Stat
K_t	0.03996	6.4399	0.03558	0.78509	0.155	4.0625
$\kappa_{t,1}$	0.02982	7.0254	0.03589	0.78378	0.1053	4.2099
$\kappa_{t,2}$	0.1858	3.3661	0.07297	0.66976	0.6973	2.7094
$\kappa_{t,1,1}$	0.4459	1.6153	0.235	0.44365	0.651	2.662
$\kappa_{t,1,2}$	0.6164	0.96781	0.7159	0.13597	0.3607	2.3172
$\kappa_{t,1,3}$	0.6858	0.75425	0.4159	0.30387	0.761	2.7728
$\kappa_{t,2,1}$	0.7202	0.65634	0.5929	0.1997	0.5427	2.5452
$\kappa_{t,2,2}$	0.007809	9.7048	0.07731	0.65988	0.01029	4.917
$\kappa_{t,2,3}$	0.5926	1.0464	0.3755	0.33106	0.6095	2.6183

Au seuil de significativité 5%, il faut rejeter la normalité des séries K_t, $\kappa_{t,1}$ et $\kappa_{t,2,2}$.

Il est intéressant de remarquer que la non-normalité des séries K_t, $\kappa_{t,1}$ vient surtout du paramètre de Skewness qui est significativement différent de 0 alors que le paramètre de Kurtosis n'est pas significativement différent de 3.

Pour le paramètre $\kappa_{t,2,2}$, c'est l'inverse et c'est le paramètre de kurtosis qui est significativement différent de 3.

Les différentes statistiques des autres séries ne permettent pas de rejeter l'hypothèse nul du test de Jarque-Bera. Bien que cela ne signifie pas que ces séries suivent des lois Normales, leurs résidus n'ont pas un comportement significativement différent d'une loi Normale en ce qui concerne leurs Skewness et leurs Kurtosis. La loi Normale est donc une bonne approximation pour simuler le comportement de ces séries.

Concernant les séries K_t du modèle LC, le tableau équivalent est disponible en annexe. Il montre que la série $K_{t,1,2}$ et la série $K_{t,2,2}$ (qui sont les paramètres hommes et femmes pour la France) ne suivent pas une loi Normal. Néanmoins, ils sont quand même simulés par une modèle ARIMA afin d'avoir une unité de traitement, tout en sachant qu'il s'agit d'une approximation.

4.4 Projection des différentes séries temporelles.

Suite aux analyses faites dans la section 4.3, le choix de modélisation s'est porté sur le modèle $ARIMA(1,1,0)$ pour simuler les valeurs futures des différentes séries temporelles. En effet, il y a été montré qu'il s'agissait du modèle qui présente le meilleur rapport entre une modélisation correcte du comportement des séries et la complexité dudit modèle.

À noter toutefois que, comme vu dans la section 4.3.6, les résidus des séries K_t, $\kappa_{t,1}$ et $\kappa_{t,2,2}$ du modèle ACF-2 et ceux des séries $K_{t,1,2}$ et $K_{t,2,2}$ du modèle LC ne peuvent être parfaitement modélisés par une loi Normale et que donc le modèle $ARIMA(1,1,0)$ est une approximation pour ces séries.

4.4.1 Modélisation pratique du modèle ARIMA (1,1,0).

Le modèle $ARIMA(1,1,0)$ est donné par

$$Y_t = \mu + \alpha Y_{t-1} + \epsilon_t, \forall t \in \mathbb{N}.$$

Afin de pouvoir projeter les séries temporelles, il faut trouver les estimateurs $\widehat{\mu}$ et $\widehat{\alpha}$ qui correspondent le mieux aux données empiriques. Pour ce faire, la méthode des moindres carrés sera utilisée. Celle-ci consiste à minimiser la somme (T étant la valeur de la dernière année d'observation)

$$S(\mu,\alpha) = \sum_{t=1}^{T} \left(Y_t - \mu - \alpha Y_{t-1}\right)^2. \tag{4.2}$$

Ceci étant un simple problème d'optimisation, il suffit de calculer les dérivées partielles et de trouver les valeurs $\widehat{\mu}$ et $\widehat{\alpha}$ pour lesquelles elles s'annulent afin de trouver le minimum de la fonction $S(\mu,\alpha)$.

Dérivée partielle par rapport à μ.

$$\frac{\partial S(\mu,\alpha)}{\partial \mu} = -2\sum_{t=1}^{T}\left(Y_t - \mu - \alpha Y_{t-1}\right).$$

En posant cette dérivée partielle égale à 0

$$\sum_{t=1}^{T}(Y_t-\mu-\alpha Y_{t-1})=0.$$

$$-T\mu+\sum_{t=1}^{T}(Y_t-\alpha Y_{t-1})=0.$$

$$\mu=\frac{1}{T}\sum_{t=1}^{T}(Y_t-\alpha Y_{t-1}) \qquad \mu=\overline{Y}_t-\alpha\overline{Y}_{t-1}.$$

L'estimateur $\widehat{\mu}$ est donc donné par $\widehat{\mu}=\overline{Y}_t-\alpha\overline{Y}_{t-1}$.

Dérivée partielle par rapport à α.

Partant de (4.2) et en remplaçant μ par son estimateur

$$S(\widehat{\mu},\alpha)=\sum_{t=1}^{T}\left(Y_t-\widehat{\mu}-\alpha Y_{t-1}\right)^2$$

$$=\sum_{t=1}^{T}\left(\left(Y_t-\overline{Y}_t\right)-\alpha\left(Y_{t-1}-\overline{Y}_{t-1}\right)\right)^2.$$

$$\frac{\partial S(\widehat{\mu},\alpha)}{\partial\alpha}=-2\sum_{t=1}^{T}\left[\left(Y_t-\overline{Y}_t\right)-\alpha\left(Y_{t-1}-\overline{Y}_{t-1}\right)\right]\left(Y_{t-1}-\overline{Y}_{t-1}\right).$$

En posant cette dérivée partielle égale à 0 et en distribuant le terme $\left(Y_{t-1}-\overline{Y}_{t-1}\right)$

$$\sum_{t=1}^{T}\left(Y_t-\overline{Y}_t\right)\left(Y_{t-1}-\overline{Y}_{t-1}\right)=\alpha\sum_{t=1}^{T}\left(Y_{t-1}-\overline{Y}_{t-1}\right)^2$$

$$\alpha=\frac{\sum_{t=1}^{T}\left(Y_t-\overline{Y}_t\right)\left(Y_{t-1}-\overline{Y}_{t-1}\right)}{\sum_{t=1}^{T}\left(Y_{t-1}-\overline{Y}_{t-1}\right)^2}\frac{T-1}{T-1}$$

$$\alpha=\frac{cov\left(Y_t,Y_{t-1}\right)}{Var\left(Y_{t-1}\right)}$$

L'estimateur $\widehat{\alpha}$ est donc donné par $\widehat{\alpha}=\frac{cov(Y_t,Y_{t-1})}{Var(Y_{t-1})}$.

Extrapolation des Y_t.

Afin de calculer la suite de la série, les paramètres Y_t pour $t>T$ sont calculés de la manière suivante

$$Y_t=\widehat{\mu}+\widehat{\alpha}Y_{t-1}+\epsilon_t.$$

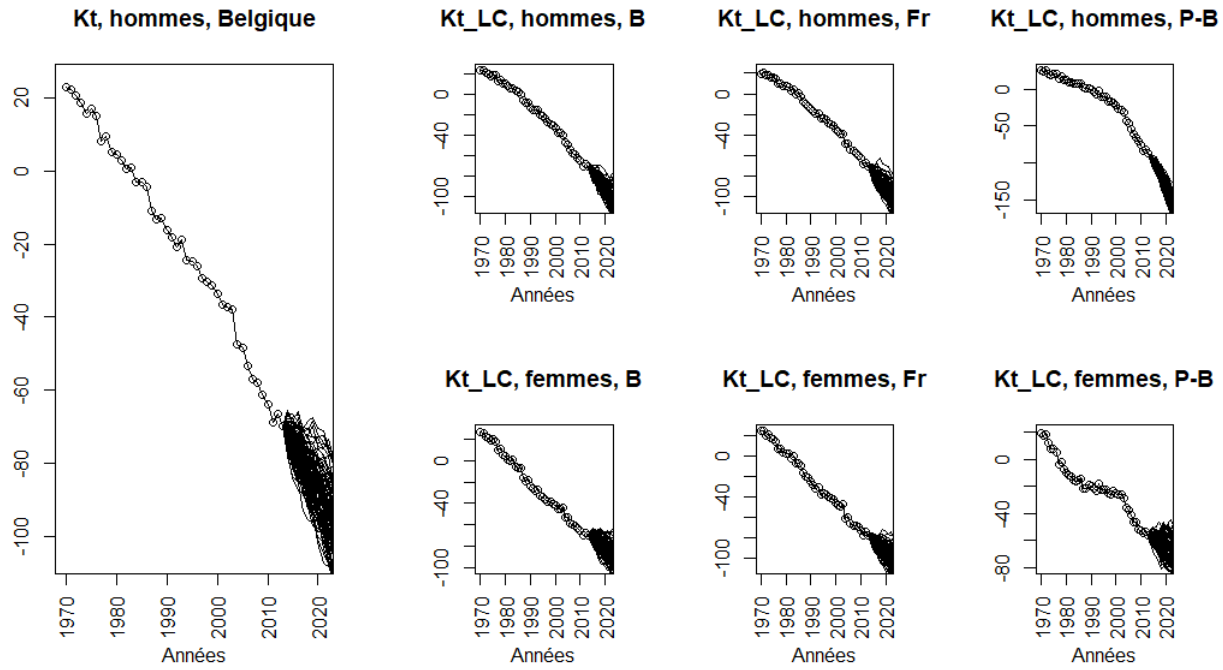
Avec $\epsilon_t\sim N(0,\sigma^2)$ où σ^2 est la variance empirique obtenue à partir des résidus de la série $(Y_t)_{1<t<T}$

4.4.2 Application aux variables K_t, $\kappa_{t,i}$ et $\kappa_{t,i,j}$.

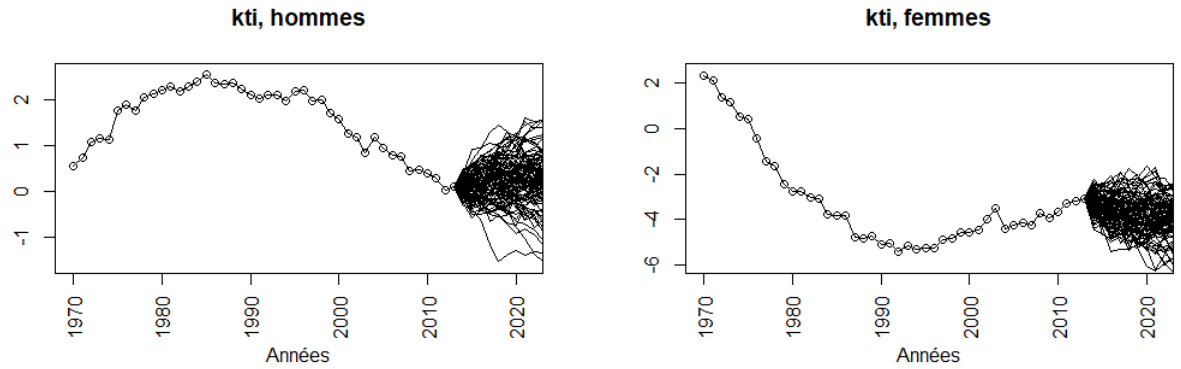
Afin d'illustrer visuellement l'extrapolation des séries temporelles, $N = 100$ scénarios ont été simulés sur une période $t = 10$ années à partir de 2013. Les graphiques associés sont donnés ci-dessous.

Les scénarios ont été simulés d'une part pour les K_t, $\kappa_{t,i}$ et $\kappa_{t,i,j}$ du modèle ACF-2 et d'autre part pour les K_t ventilés par sexe et par pays dans le cadre du modèle de Lee-Carter. Ces derniers sont notés Kt_LC dans les titres des graphiques.

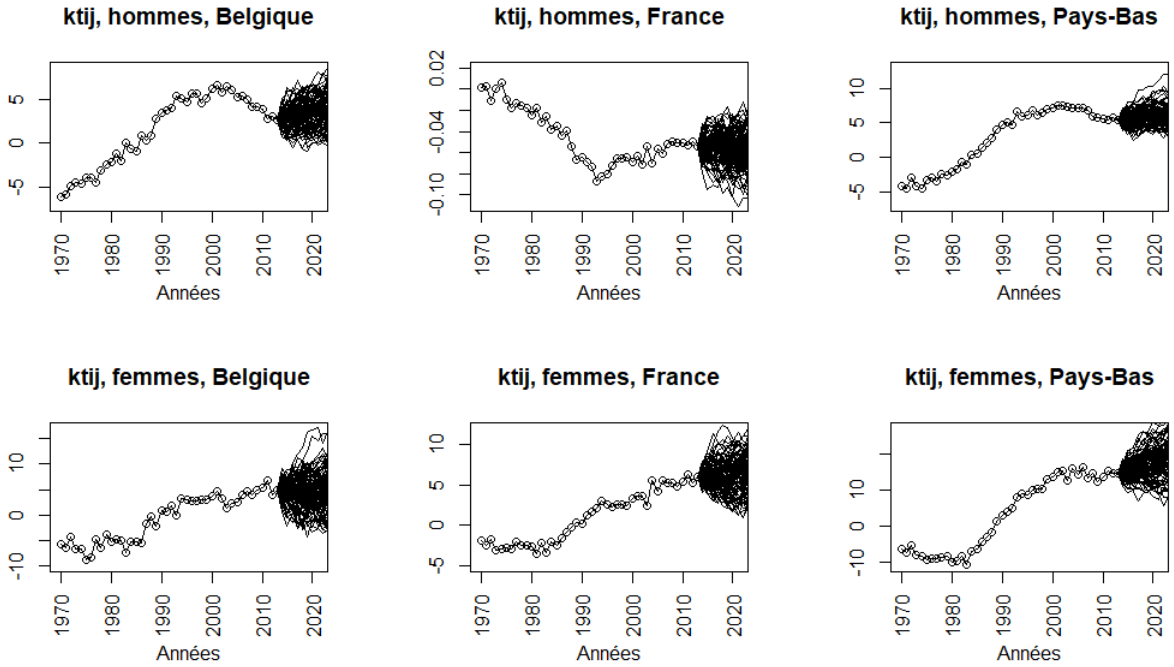
K_t des modèles ACF-2 et LC.



Application aux $\kappa_{t,i}$ du modèle ACF-2.



Application aux $\kappa_{t,i,j}$ du modèle ACF-2.



Analyse.

Ces graphiques permettent de visualiser la volatilité des prédictions des différents paramètres temporels. Les paramètres K_t, que cela soit ceux du modèle ACF-2 ou du modèle LC, sont beaucoup moins volatiles que les paramètres $\kappa_{t,i}$ et $\kappa_{t,i,j}$. Il s'agit d'un résultat attendu puisque les paramètres K_t codent la tendance générale de l'évolution temporelle de la mortalité alors que les autres paramètres viennent corriger cette tendance générale afin de coller au mieux à la réalité, d'où une volatilité plus élevée.

4.5 Comparaison entres les modèles et la réalité.

Afin de pouvoir comparer les modèles LC et ACF-2, il va falloir comparer leurs prédictions de l'espérance de vie par rapport à l'espérance de vie réellement observée.

C'est pourquoi les modèles ont été calibrés à partir des données des années 1946 à 2014 et puis extrapolés pour les années 2015 à 2018 selon la méthode décrite en 4.2.

Les extrapolations ont été réalisées avec N=10.000 scénarios.

4.5.1 Calcul de l'espérance de vie - méthode transversale.

Comme expliqué en 2.1, la probabilité de survie $p_x(t)$ est donnée par $p_x(t) = exp(-\mu_{x,t})$, la probabilité de survie à un an d'une personne d'âge x, durant l'année t. Une façon de calculer l'espérance de vie est de la calculer de manière « transversale », c'est-à-dire en gelant l'année t et en prenant en compte les différentes valeurs de $p_x(t)$ pour un t fixé. C'est la méthode

la plus simple pour connaître l'espérance de vie, puisqu'elle ne repose que sur les données de l'année étudiée. Une autre méthode, dite « diagonale » et qui suit l'évolution de la variable t nécessiterait d'avoir des données sur 101 années différentes (101 puisque la table s'étend sur 101 âge x différents). Cette seconde méthode sera utilisée dans le cadre du pricing du chapitre 6.

Avec la méthode transversale, connaissant l'année de référence t, $p_x(t)$ peut se réécrire p_x.

Afin de calculer l'espérance de vie, il est nécessaire d'avoir la probabilité de survie, non pas à un an, mais à plusieurs années. T_x étant une variable aléatoire codant le temps restant à vivre pour un individu d'âge x, ${}_tp_x = P(T_x > t)$ est définie comme étant la probabilité pour cette personne d'âge x d'être en vie après t années.

D'après [7], slides 10 et 11,

$$ {}_tp_x = p_x \cdot p_{x+1} \cdot p_{x+2} \cdot \ldots \cdot p_{x+t-1}. $$

Intuitivement, cela peut être compris comme le fait que la probabilité de survivre durant t années est la probabilité de survivre durant la première année, puis durant la deuxième année et ainsi de suite jusqu'à survivre durant l'année $t-1$. Ces probabilités étant indépendantes (le fait que j'ai survécu à une année n'influence pas le fait de survivre à la suivante), il faut en faire le produit pour avoir ${}_tp_x$.

L'espérance de vie, quant à elle, est donnée par l'espérance mathématique de ${}_tp_x$. Ce qui donne (voir également [7], slides 19)

$$ \mathring{e}_x = \int_0^{w-x} {}_up_x du $$

w étant la fin de la table de mortalité, $w = 101$ dans le cas présent.

Puisque les taux de mortalité ont été supposés constants, l'espérance de vie d'un individu d'âge x peut se réécrire en version discrète comme

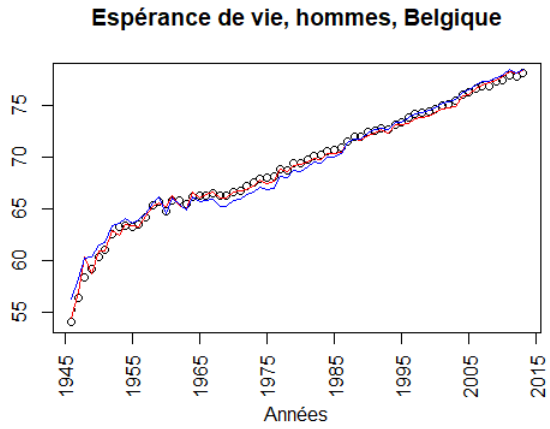
$$ \mathring{e}_x = \sum_{t=1}^{101-x} {}_tp_x. $$

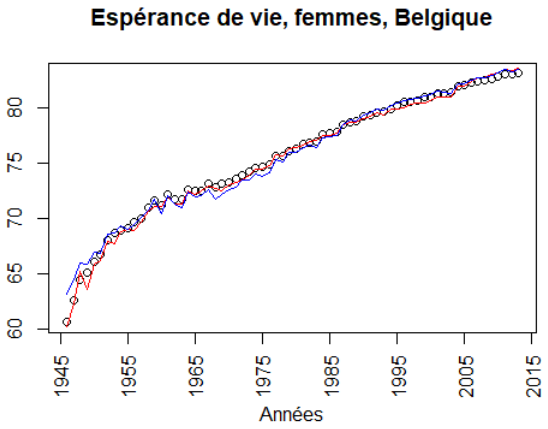
Afin de pouvoir comparer les modèles, les espérances de vie seront calculées à la naissance, c'est-à-dire avec $x = 0$.

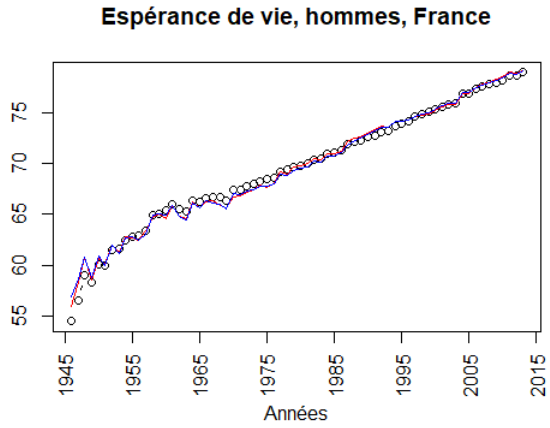
4.5.2 Espérance de vie calibrée, entre 1946 et 2013.

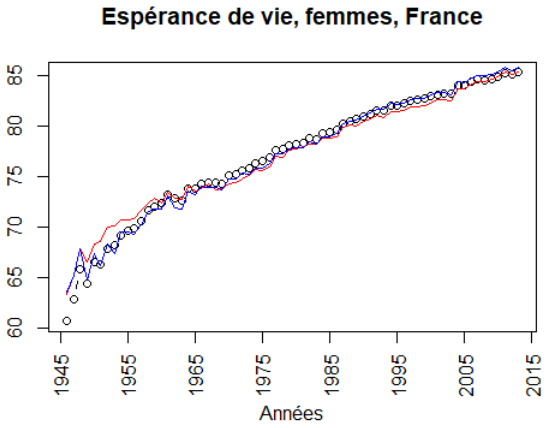
Les graphiques ci-dessous montrent la superposition des espérances de vie calculées par, respectivement, les données empiriques, le modèle ACF-2 et le modèle LC. Ces espérances de vie ont été calibrées à partir des années 1946 à 2013 inclus.

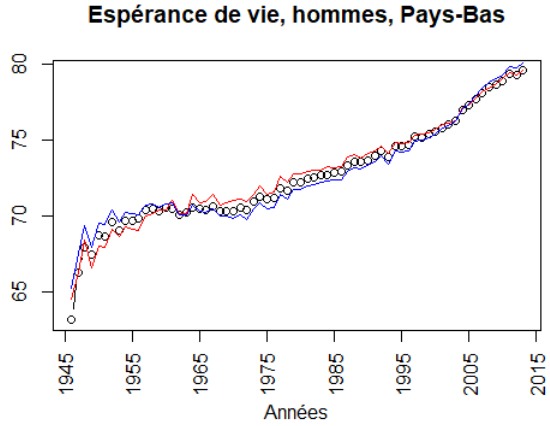
Dans cette section et dans celle qui suit, la légende pour tous les graphiques est la suivante : noir pour l'espérance de vie empirique, rouge pour le modèle ACF-2 et bleu pour le modèle LC.

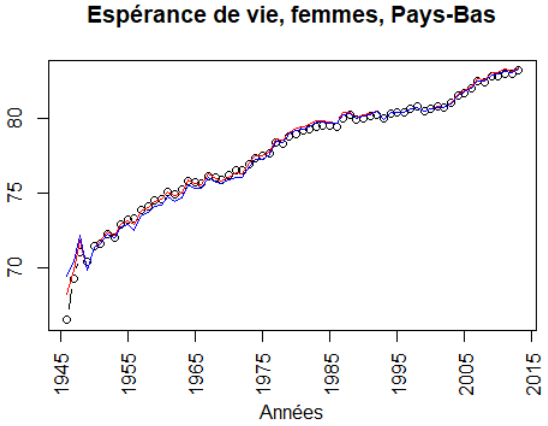












Visuellement, le modèle ACF-2 a l'air de mieux coller aux données empiriques, sauf pour les femmes en France. Afin de confirmer ou d'infirmer cette analyse, une technique consiste à

calculer l'écart quadratique moyen. Ce nombre est calculé comme étant la moyenne du carré des différences entre d'une part, l'espérance de vie calculée à partir des modèles LC et ACF-2 respectivement, et d'autre part l'espérance de vie empirique.

Si $\mathring{e}_t$ est l'espérance de vie empirique et $\widetilde{e_t}$ l'espérance de vie trouvée par modélisation (LC ou ACF-2), l'écart quadratique moyen, noté EQ est donné par

$$EQ = \frac{1}{N} \sum_{i=1}^{N} (\mathring{e}_t - \widetilde{e_t})^2$$

où N est le nombre d'années prises en compte (68 pour la partie calibration, 5 pour la partie projection).

Le modèle qui minimise EQ est donc le modèle qui modélise au mieux l'espérance de vie empirique puisque c'est celui qui donnera les valeurs projetées de l'espérance de vie les plus proches de ce qui est observé dans la réalité.

Écart quadratique entre les modèles et les données empiriques.

Le tableau ci-dessous donne les valeurs de l'EQ pour chacun des deux modèles, LC et ACF-2, ventilé par sexe et par pays. Le modèle qui calibre le mieux (c'est-à-dire où l'EQ est le plus bas) est en vert, l'autre est en rouge.

	LC	ACF-2
Hommes, Belgique	0.4925664	0.1673707
Hommes, France	0.3405787	0.2670826
Hommes, Pays-Bas	0.2848098	0.21621894
Femmes, Belgique	0.3620244	0.1423806
Femmes, France	0.4314610	0.8952173
Femmes, Pays-Bas	0.2311295	0.09282435
Total	*2.14257*	1.781094

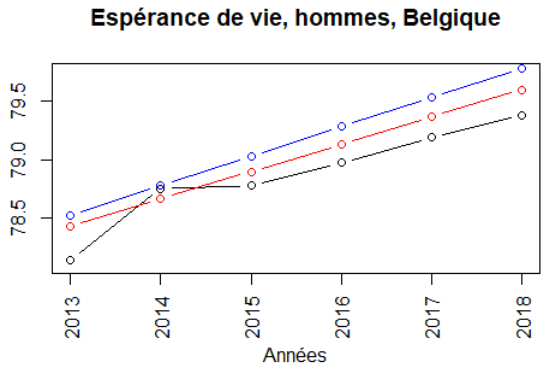
L'analyse des écarts quadratiques moyens corrobore donc l'analyse visuelle qui avait été faite préalablement, c'est-à-dire que le modèle ACF-2 est celui qui modélise le mieux les valeurs passées de l'espérance de vie.

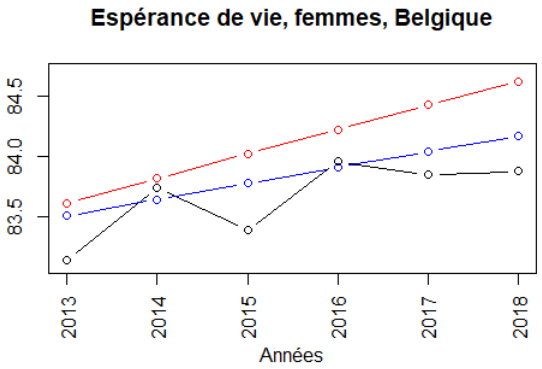
À noter que pour le cas des femmes en France, c'est le modèle classique de Lee-Carter qui est le meilleur.

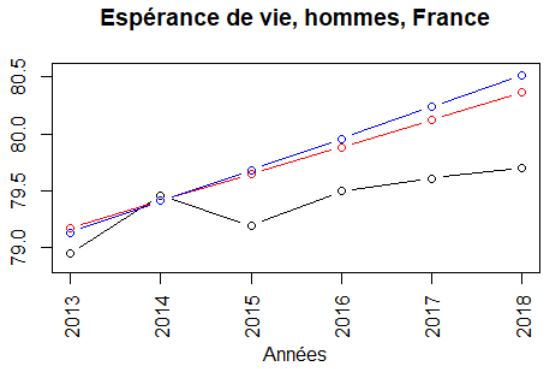
4.5.3 Espérance de vie projetée, entre 2014 et 2018.

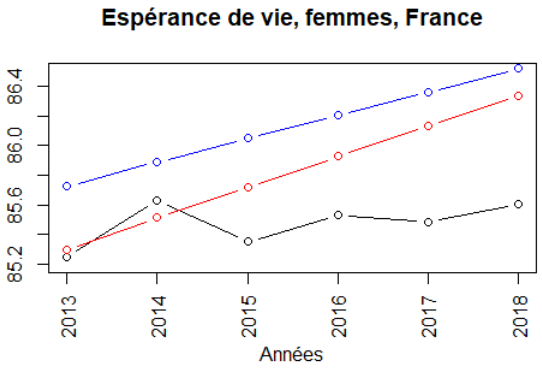
Espérance de vie calculée avec les valeurs médianes des séries temporelles.

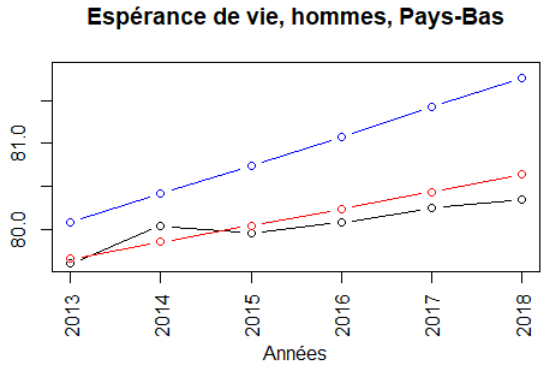
Les graphiques ci-dessous montrent la superposition des espérances de vie calculées par, respectivement, les données empiriques, le modèle ACF-2 et le modèle LC. Ces espérances de vie ont été calculées à partir des projections des séries temporelles K_t, $\kappa_{t,i}$, $\kappa_{t,i,j}$ pour les années 2014 à 2018. L'année 2013 est visuellement présente afin de faciliter la lecture mais ne fait pas partie de la projection.

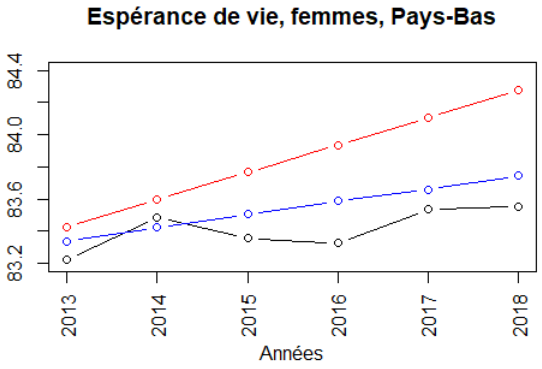












Écart quadratique moyen entre les projections et les données empiriques.

Visuellement, le modèle ACF-2 a l'air de mieux coller aux données empiriques, sauf pour les femmes en Belgique et au Pays-Bas. De la même manière que pour le point précédent, il est important de calculer l'EQ afin de confirmer ou d'infirmer ce constat.

Ci-dessous le tableau des EQ, ventilé par sexe et par pays (le code couleur étant le même qu'au point précédent).

EQ	LC	ACF-2
Hommes, Belgique	0.08854674	0.02548446
Hommes, France	0.3037305	0.2123095
Hommes, Pays-Bas	1.02894186	0.03846807
Femmes, Belgique	0.05639950	0.27231069
Femmes, France	0.5205504	0.2490462
Femmes, Pays-Bas	0.02907017	0.28029727
Total	*2.027239*	*1.077916*

Dans ce cas-ci également, l'analyse des écarts quadratiques moyens corrobore l'analyse visuelle.

Le modèle ACF-2 est donc celui qui modélise le mieux de façon générale les valeurs futures de l'espérance de vie, sauf dans le cas des femmes en Belgique et au Pays-Bas.

Il est toutefois important de remarquer que quel que soit le modèle utilisé, ils ont tendances à surévaluer l'espérance de vie future par rapport à la réalité.

Ceci pourrait s'expliquer par une stagnation récente de l'espérance de vie, qui fait que la règle des « 3 mois d'espérance de vie gagnés par an » ne peut durer éternellement (voir [21]) et que nous pourrions atteindre un plafond d'espérance de vie. Cette stagnation, datant de 2015 et donc très récente, n'est pas prise en compte par les modèles développés. En effet, ces derniers ne font que capturer les tendances passées et ne peuvent pas prévoir des impacts inhabituels sur l'espérance de vie.

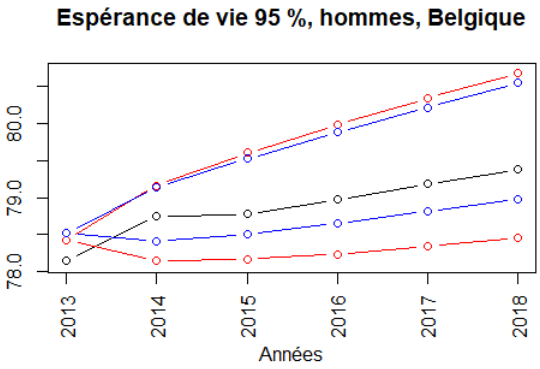
D'après [13], plusieurs causes pourraient expliquer cet arrêt de l'augmentation linéaire de l'espérance de vie :

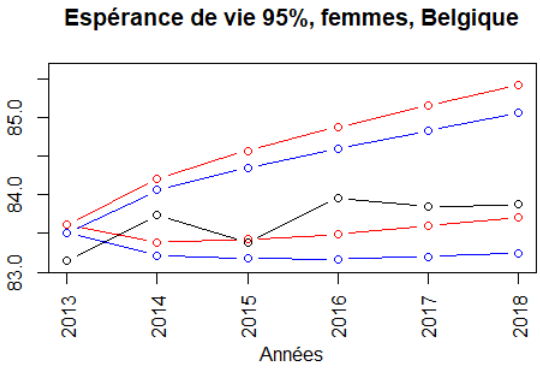
— Plusieurs épidémies de grippe ont touché les populations les plus âgées, surtout l'épidémie de 2015 qui a été la plus meurtrière ([14] pour la Belgique).
— Les campagnes de prévention des maladies cardio-vasculaires atteignent leurs limites et ne permettent plus de diminuer le nombre de décès dus à cette cause.
— Le tabagisme qui augmente, particulièrement chez les femmes, augmente le nombre de décès dans certaines tranches de la population ([3] pour la France).
— Les événements climatiques extrêmes : canicules, hiver plus rudes, inondations, ... sont de plus en plus fréquents et augmentent donc la mortalité ([9]).

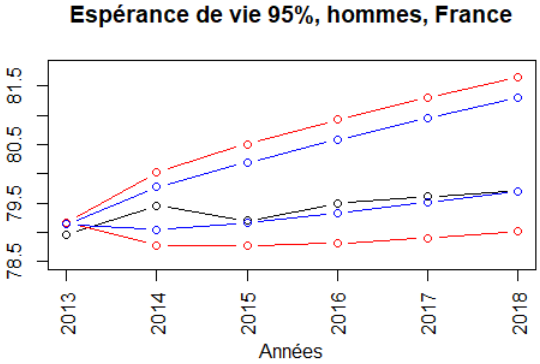
Ces événements entraînent une surmortalité et donc une augmentation des paramètres $\mu_{x,t}$ associés. Une augmentation des $\mu_{x,t}$ entraine une diminution des probabilités de survie p_x puisque $p_x(t) = exp(-\mu_{x,t})$ et donc une diminution de l'espérance de vie calculée. C'est pourquoi les événements cités ci-dessus ont un impact direct sur le calcul de l'espérance de vie.

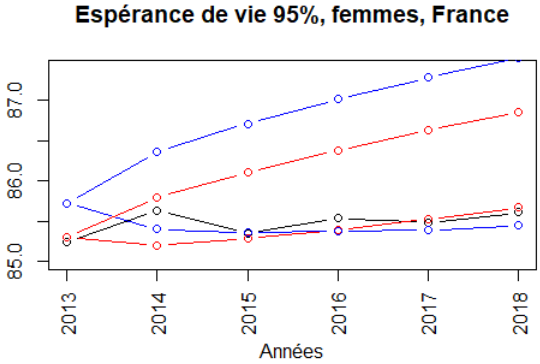
Espérance de vie calculée avec les quantiles 2,5% et 97,5 % des séries temporelles.

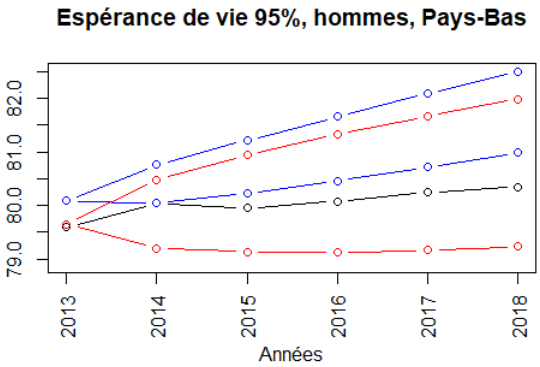
Les graphes suivants comparent l'espérance de vie empirique à un intervalle de confiance 95% des valeurs possibles pour les modèles ACF-2 et LC. Ces fourchettes ont été calculées en utilisant les quantiles 0,025 et 0,975 des scénarios simulés.

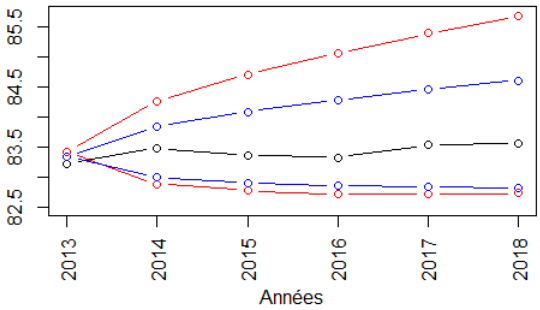












Deux remarques sont à faire à partir de ces différents graphiques :

— la volatilité du modèle ACF-2 est plus importante que celle du modèle LC

— l'espérance de vie réelle après 5 ans est comprise dans la fourchette de valeurs obtenues avec le modèle ACF-2, sauf pour les femmes en France. Pour les hommes en France et aux Pays-Bas, le modèle LC a du mal à donner une fourchette cohérente avec la réalité.

4.5.4 Résumé de la comparaison entre les modèles ACF-2 et LC.

Voici un tableau récapitulatif entre les modèles ACF-2 et LC.

— la colonne « Calibration » est la capacité du modèle à répliquer les données de 1946 à 2013.
— la colonne « Projection » est la capacité du modèle à prédire les valeurs 2014 à 2018.
— la colonne « 95% » est la capacité du modèle à prédire un intervalle de confiance à 95% qui contient l'année 2018.

Pour les colonnes « Calibration » et « Projection », le modèle qui explique le mieux les données est indiqué par un V, l'autre par un X. Pour la colonne « 95% », le cas où l'intervalle de confiance à 95% pour l'année 2018 du modèle contient l'observation réelle de l'espérance de vie est noté par un V et par un X si ce n'est pas le cas.

	Calibration		Projection		95%	
	LC	ACF-2	LC	ACF-2	LC	ACF-2
Hommes, Belgique	X	V	X	V	V	V
Hommes, France	X	V	X	V	X	V
Hommes, Pays-Bas	X	V	X	V	X	V
Femmes, Belgique	X	V	V	X	V	V
Femmes, France	V	X	X	V	V	X
Femmes, Pays-Bas	X	V	V	X	V	V

En combinant les différentes analyses, c'est le modèle ACF-2 qui donne l'impression de mieux correspondre à la réalité tout en donnant les prédictions les plus cohérentes. Attention toutefois que ce dernier est plus volatil que le modèle LC, se baser sur le modèle ACF-2 peut donc amener à des sur/sous-estimations importantes de l'évolution de l'espérance de vie.

Maintenant que ce constat sur la solidité des modèles est posé, il reste deux facettes intéressantes à comparer

— Les prédictions à long terme de l'espérance de vie.
— Analyser la différence d'impact en matière de pricing.

Le premier point sera analysé dans le chapitre 5 tandis que le second sera traité dans le chapitre 6.

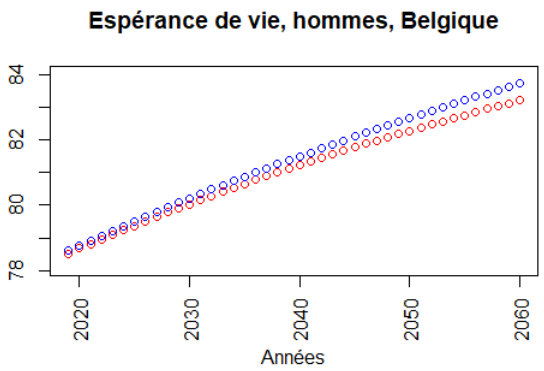
Chapitre 5

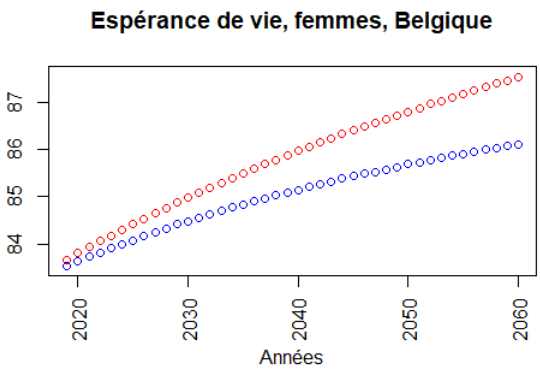
Prédiction 2020-2060.

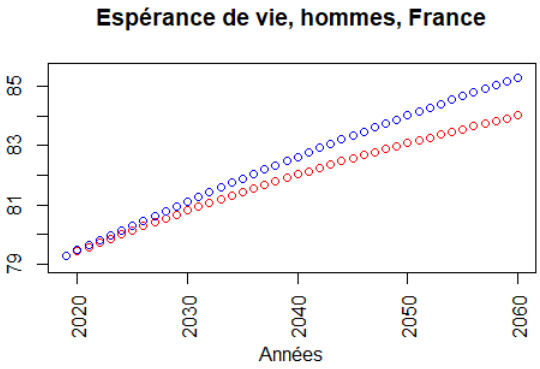
Le présent chapitre va permettre de comparer les projections de l'espérance de vie entre les modèles LC et ACF-2 pour les années 2019 à 2060. Les espérances de vie sont calculées à la naissance, de façon transversale, comme expliqué dans la section 4.5.1. Toutes les simulations présentées sont faites à partir de la calibration des modèles sur les données 1946-2018, comme développé dans le chapitre 3. Ensuite les paramètres temporels ont été extrapolés pour les années 2019 à 2060 et les espérances de vie calculées selon les méthodes présentées dans le chapitre 4. Pour ce faire, $N = 10.000$ scénarios ont été simulés.

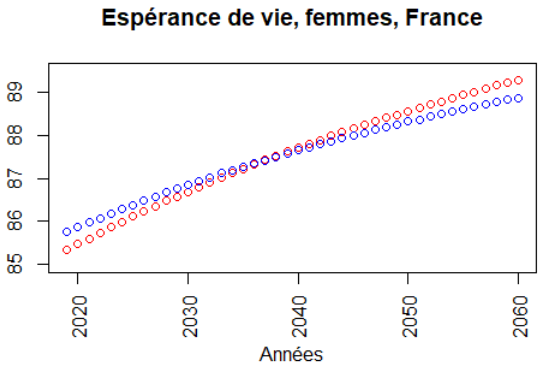
5.1 Valeurs médianes.

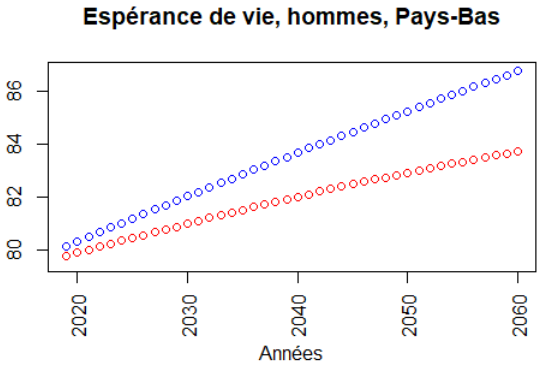
Ci-dessous, les graphiques des médianes des projections pour les modèles LC et ACF-2. Comme précédemment, la couleur bleue représente le modèle LC et la couleur rouge représente le modèle ACF-2.

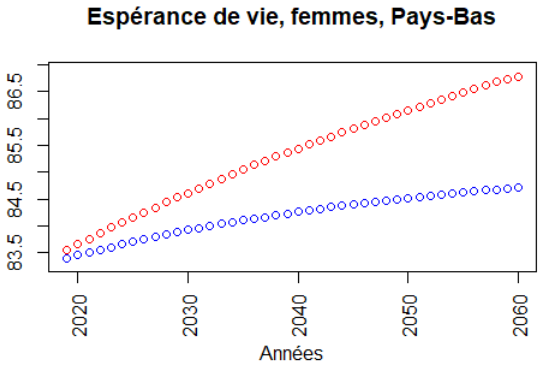












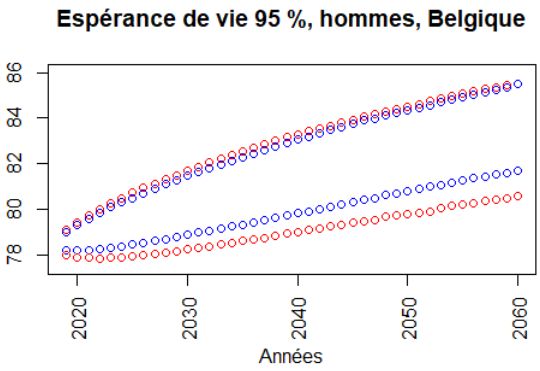
Afin de pouvoir se donner une idée des âges atteints par chaque modèle, voici un tableau comparatif entre les modèles LC et ACF-2 pour les années 2020, 2040 et 2060. Les données de ce tableau sont arrondies au centième près.

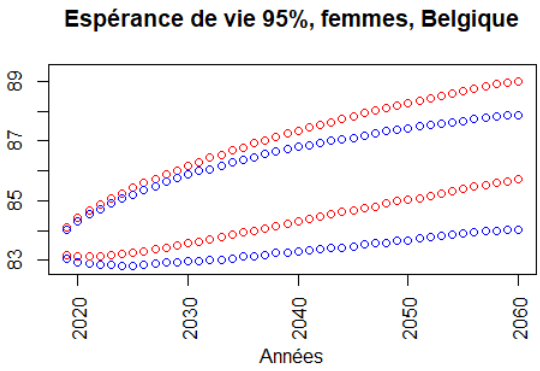
	2020		2040		2060	
	LC	ACF-2	LC	ACF-2	LC	ACF-2
Hommes, Belgique	78.76	78.68	81.50	81.22	83.71	83.20
Hommes, France	79.47	79.44	82.62	82.03	85.29	84.01
Hommes, Pays-Bas	80.33	79.90	83.70	82.02	86.76	83.75
Femmes, Belgique	83.64	83.80	85.15	85.97	86.11	87.54
Femmes, France	85.87	85.48	87.66	87.72	88.89	89.31
Femmes, Pays-Bas	83.45	83.66	84.26	85.45	84.71	86.79

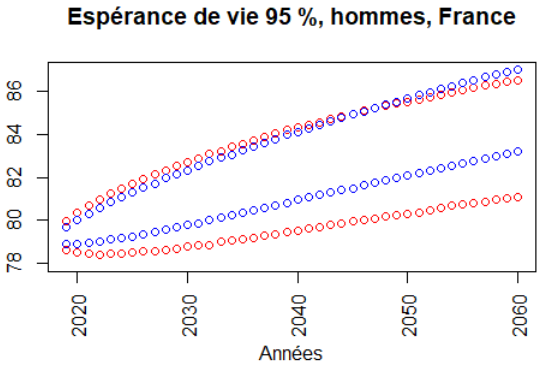
Le modèle ACF-2 prévoit donc en général une plus faible évolution médiane de l'espérance de vie que le modèle LC sauf dans le cas des femmes en Belgique et aux Pays-Bas, deux catégories où l'écart de prédiction est vraiment important (plus de deux ans aux Pays-Bas !). Ce constat est à mettre en parallèle avec les conclusions de la section 4.5.4 qui donnaientt le modèle LC comme meilleur modèle pour ces deux catégories.

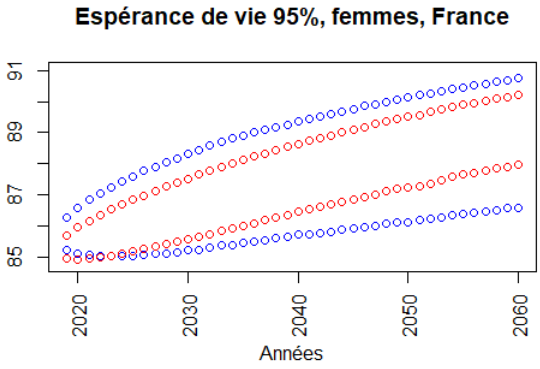
5.2 Intervalles de confiance 95%.

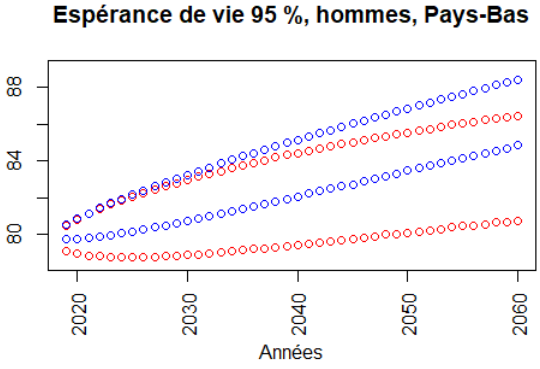
Les mêmes graphiques ont été réalisés pour l'intervalle de confiance 95% et sont présentés ci-dessous. La branche haute représente le quantile 0.975 et la branche basse le quantile 0.0275, le code couleur des modèles reste inchangé.

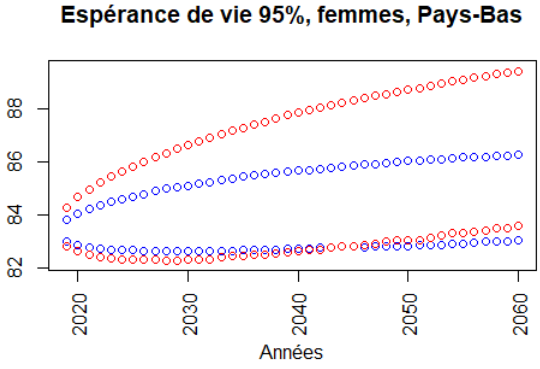












Afin de pouvoir comparer les médianes et les intervalles de confiance entre les deux modèles, voici les tableaux des valeurs maximales et minimales prises par l'espérance de vie pour chaque

modèle, aux années 2020, 2040 et 2060. Les colonnes B sont les estimations basses (percentile 0.5%) de l'espérance de vie et les colonnes H en sont les estimations hautes (percentile 99.5%).

	LC						ACF-2					
	2020		2040		2060		2020		2040		2060	
	B	H	B	H	B	H	B	H	B	H	B	H
H, Be	78.17	79.33	79.82	83.06	81.70	85.48	77.88	79.44	79.00	83.28	81.70	85.51
H, Fr	78.90	80.03	80.97	84.13	83.25	87.06	78.49	80.36	79.55	84.33	83.25	86.52
H, PB	79.76	80.87	82.06	85.14	84.83	88.40	78.95	80.82	79.45	84.42	84.83	86.41
F, Be	82.93	84.31	83.28	86.80	84.03	87.89	83.12	84.42	84.31	87.35	84.03	89.01
F, Fr	85.13	86.57	85.72	89.36	86.59	90.74	84.94	85.95	86.45	88.63	86.59	90.20
F, PB	82.84	84.04	82.70	85.66	83.02	86.23	82.60	84.65	82.62	87.83	83.02	89.40

Il est intéressant de remarquer que

— Les intervalles de confiance se recoupent en partie et ne sont donc pas mutuellement exclusifs.

— La médiane d'un des modèles tombe toujours dans l'intervalle de confiance de l'autre modèle, à l'exception de la médiane ACF-2 dans l'intervalle LC pour 2040 et de la médiane LC dans l'intervalle ACF-2 pour 2060 pour les hommes aux Pays-Bas, en rouge dans le tableau ci-dessus. Hormis ces deux cas de figure, cela veut dire que prendre la médiane de n'importe lequel des modèles afin de simuler l'espérance de vie serait cohérent avec l'intervalle de confiance 95% de l'autre modèle.

Évidemment, l'ensemble de ces valeurs suppose une évolution de l'espérance de vie en continuité avec l'évolution passée. Comme pour toute projection, il faut prendre celles faites ici pour ce qu'elles sont, c'est-à-dire uniquement des estimations et non une prévision certaine de l'espérance de vie future. Les modèles, quels qu'ils soient, ne font qu'extrapoler un comportement passé et ne peuvent pas prédire une modification de ce comportement.

Par exemple, il suffit qu'il survienne une découverte majeure dans le domaine de la médecine afin que l'espérance de vie augmente de façon importante. A contrario, des causes inhabituelles de décès peuvent diminuer l'espérance de vie, comme c'est le cas actuellement depuis 2015 (voir section 4.5.3).

En cas de pricing long terme à partir de ces modèles, il faut donc être le plus prudent possible afin de ne pas mettre la compagnie d'assurance en danger, comme nous allons le voir dans le prochain chapitre.

Chapitre 6

Pricing.

Afin d'éviter au maximum le risque de faillite, un assureur doit présenter un ratio de solvabilité au-dessus de 100% afin d'être financièrement solvable. Ce ratio est défini comme le quotient entre les actifs de la compagnie et ses passifs. Or, dans une compagnie d'assurance, les actifs sont fonction, en partie, des primes payées par le client et le passif est fonction des provisions réalisées, provisions qui elles-mêmes dépendent de l'espérance de perte du produit sous-jacent. Il faut donc que la tarification donne une prime plus élevée que l'espérance de perte (aussi appelé « Value at Risk ») afin de garder un ratio au-dessus de 100%. Ceci est bien sûr une simplification du risk management d'une compagnie. Dans la réalité, la loi oblige un ratio de solvabilité plus élevé que les 100% et le jeu de vases communicants entre actif et passif dépend de bien plus de paramètres.

Les produits de type vie/décès pouvant se baser, en partie, sur la probabilité de survie, ce dernier chapitre a pour objectif d'en donner un exemple de tarification. Ce dernier se fera sur un produit fictif de type « rente viagère ».

Une dernière remarque concernant le pricing concernant le principe antidiscrimination en Belgique. Bien qu'il soit utile de comparer les différences de pricing possible entre les hommes et les femmes afin de provisioner correctement les primes, il est interdit de faire une différence de tarification lors de la vente de contrat d'assurance sur la vie, comme le stipule l'Arrêté Royal du 14 novembre 2003 ([16]). L'assureur devra donc s'arranger pour présenter une prime unique quel que soit le sexe tout en couvrant correctement ses besoins en capitaux.

6.1 Outils de calcul.

6.1.1 Calcul de l'espérance de vie - méthode diagonale.

La méthode de calcul utilisée pour l'espérance de vie dans la section 4.5.1, bien que facile à calculer, présente le défaut de ne pas suivre un individu tout au long de sa vie mais de composer avec les $p_x(t)$ d'une année donnée. Cela peut poser problème : par exemple, pour une personne de 35 ans en 2015, $p_{40}(2015)$ sera utilisé dans le calcul de son espérance de vie alors que l'individu de 35 ans en 2015 n'aura 40 ans qu'en 2020 et que donc le paramètre $p_{40}(2020)$ devrait être utilisé à la place puisque rien ne garantit d'avoir $p_{40}(2015) = p_{40}(2020)$. La méthode diagonale diffère de la méthode transversale par le fait que l'espérance de vie de l'individu de 35 ans en 2015 est calculée avec les paramètres $p_{35}(2015)$,$p_{36}(2016)$,$p_{37}(2017)$ et ainsi de suite, au lieu d'utiliser uniquement les paramètres de l'année 2015.

Pour des données historiques, la méthode diagonale fonctionne mais pour un individu naissant aujourd'hui, il faudrait simuler les $p_x(t)$ sur les 100 prochaines années afin de connaître son espérance de vie. Or les chapitres précédents ont montré que plus les extrapolations des séries temporelles sont lointaines, plus l'erreur sur l'espérance de vie réelle sera importante et plus l'intervalle de confiance augmente. Néanmoins, lorsque l'extrapolation des $p_x(t)$ est limitée dans le temps, les valeurs trouvées sont assez fiables et peuvent être utilisées à des fins de pricing, même si ces valeurs seront biaisées par rapport à la réalité.

Formellement, les deux méthodes varient quant à la manière de calculer le paramètre ${}_tp_x$. Pour la méthode transversale, il sera le produit des termes en bleu dans le tableau ci-dessous. Pour la méthode diagonale, il sera le produit des termes en rouge.

	t	$t+1$	$t+2$	$t+3$	$t+4$	$t+5$
$x+5$	$p_{x+5}(t)$					$p_{x+5}(t+5)$
$x+4$	$p_{x+4}(t)$				$p_{x+4}(t+4)$	
$x+3$	$p_{x+3}(t)$			$p_{x+3}(t+3)$		
$x+2$	$p_{x+2}(t)$		$p_{x+2}(t+2)$			
$x+1$	$p_{x+1}(t)$	$p_{x+1}(t+1)$				
x	$p_x(t)$					

Hormis cette différence, l'espérance de vie de la méthode diagonale se calcule de la même manière que pour la méthode transversale, c'est-à-dire

$$\mathring{e}_x = \sum_{t=1}^{101-x} {}_tp_x.$$

6.1.2 Rente viagère.

Une rente viagère est un produit d'assurance sur la vie qui consiste au paiement d'un certain montant, défini à l'avance, à intervalles réguliers et tant que l'assuré est en vie. La prime est censée être plus petite que la somme des paiements futurs actualisés car la mutualisation des risques permet de payer les sommes dues avec les primes des personnes décédées avant la fin de la période couverte par le contrat. Il faut donc faire appel aux probabilités de survie et à leurs extrapolations afin de pouvoir tarifer correctement la rente viagère.

Définition du terme de rente.

Afin de pouvoir avoir un euro à distribuer l'an prochain, il faut investir une certaine somme aujourd'hui, cette somme dépendant du taux d'intérêt qu'il est possible d'obtenir sur le marché obligataire. Pour représenter ce phénomène dans les calculs actuariels de rente viagère, il est intéressant de définir le terme de rente $a_{\overline{k}|}$. En supposant un taux d'intérêt i constant au cours du temps, la somme à investir aujourd'hui pour avoir un euro dans un an vaut $\frac{1}{1+i}$, celle pour avoir un euro dans deux ans vaut $\left(\frac{1}{1+i}\right)^2$ et ainsi de suite pour les années suivantes.

Ainsi, en supposant un paiement d'un euro par an démarrant l'an prochain durant k années, la valeur actuelle de l'ensemble des paiements futurs, appelé terme de rente $a_{\overline{k}|}$ est donné par

$$a_{\overline{k}|} = v + v^2 + v^3 + ... + v^k = \frac{1-v^k}{i}.$$

Avec $v = \frac{1}{1+i}$.

Exemple de rente viagère.

Pour l'exemple de ce chapitre, la rente viagère considérée est une rente à terme échu limitée à 20 ans, pour un individu de 35 ans qui donne droit à un paiement de 1 € par année à partir de l'année qui suit la signature du contrat. Une telle rente est donnée par la formule suivante (Voir [6])

$$ {}_{n}a_x = \sum_{k=1}^{n-1} a_{\overline{k}|}\, {}_{k}p_x q_{x+k} + a_{\overline{n}|}\, {}_{n}p_x. \tag{6.1} $$

avec

— ${}_{k}p_x$, la probabilité pour une personne d'âge x d'être en vie après k années, ${}_{t}p_x$ étant calculé comme pour la méthode diagonale de l'espérance de vie (voir la section 6.1.1).

— q_{x+k}, la probabilité de décéder durant l'année $x+k$.

— $a_{\overline{k}|}$ est le terme de rente.

Intuitivement, chaque terme de la somme représente l'histoire de la proportion individus $({}_{k}p_x)$ qui se voit verser un euro par an actualisé $(a_{\overline{k}|})$ jusqu'à l'année qui précède leurs décès (q_{x+k}). Le dernier terme $a_{\overline{n}|}\, {}_{n}p_x$ représente tous les individus qui décèdent après l'année n.

6.1.3 SCR et Value-at-Risk.

Afin de pouvoir maintenir un ratio de solvabilité au dessus de 100%, il est important de définir les actifs et passifs qui permettent le calcul de ce ratio. Dans ce cadre, il faut définir les notions de SCR et de Value-at-Risk

SCR.

Le « Solvency Capital Requirement (SCR) » ou « Capital de Solvabilité Requis » est défini au paragraphe 3 de l'article 101 de la directive européenne Directive 2009/138/CE comme suit ([10]) :

« ... Le capital de solvabilité requis correspond à la valeur en risque (Value-at-Risk) des fonds propres de base de l'entreprise d'assurance ou de réassurance, avec un niveau de confiance de 99,5 % à l'horizon d'un an. »

Il représente l'argent que doit avoir en réserve une compagnie d'assurance afin de pouvoir encaisser les chocs majeurs. Ces chocs peuvent être de plusieurs natures et sont répartis entre minimums 6 types de risques définis au paragraphe 4 de l'article précité.

Les faibles taux d'intérêt actuels sont exemple de choc du type « risque de marché » : les garanties de type « accident du travail » étant techniquement déficitaires (le taux de prime n'est pas assez élevé pour couvrir les obligations de l'assureur), il est habituel de couvrir une partie de ce risque avec le marché obligataire. À cause des taux faibles, les compagnies d'assurance rentrent moins d'argent que prévu et commencent à se retrouver en défaut d'actifs couvrant les sinistres, ce qui se traduit par une hausse des taux de primes techniques dans les branches concernées.

En restreignant le calcul du SCR au « risque de souscription en vie » et en ne considérant que la prime technique de l'exemple de rente viagère imaginé ici, le SCR peut être considéré comme étant le montant d'actifs nécessaires pour couvrir le risque lié au produit.

Value-at-Risk.

Les normes européennes, dites *Solvency II*, obligent les assureurs à pouvoir faire face au risque de faillite annuelle 199 sur 200. C'est-à-dire calculer les provisions liées aux produits commercialisés de telle manière à ce que la compagnie d'assurance ne puisse faire faillite qu'une fois tous les 200 ans en moyenne, c'est-à-dire dans 0.5% des cas.

Afin de calculer ces provisions, il est nécessaire de faire intervenir la notion de Value-at-Risk, ou VaR. Une définition classique de la VaR, au niveau α, représente la perte maximale liée à une distribution, dans α% des cas. Elle est mathématiquement définie comme

$$VaR_\alpha(X) = inf\left\{x : P(X > x) \leq 1 - \alpha\right\}.$$

Ceci peut se comprendre comme le fait que l'on veut que la perte liée à la distribution X ne puisse dépasser la valeur $VaR_\alpha(X)$ que dans $(1-\alpha)$% des cas.

Dans le cas de l'application de la Value-at-Risk à la rente viagère, la perte maximum représente la valeur maximum, au niveau 99.5%, que peut prendre ${}_na_x$ puisque c'est ce montant-là que la compagnie d'assurance devra débourser.

Pour rappel

$$p_x(t) \propto -\left(B_xK_t + b_{x,i}\kappa_{t,i} + b_{x,i,j}\kappa_{t,i,j}\right) \qquad et \qquad q_x(t) = (1 - p_x(t)),$$

le montant maximun pour la rente sera donc trouvé en prenant le percentille 0.5% des paramètres temporels sous-jacents $K_t, \kappa_{t,i}, \kappa_{t,i,j}$ puisque ces derniers vont maximiser les valeurs ${}_kp_x$ tout en minimisant les valeurs q_{x+k}, ce qui donnera la $VaR_{99.5}$ de la rente viagère.

Ratio de solvabilité.

En identifiant le SCR à l'actif nécessaire pour couvrir les provisions liées à la rente viagère et en identifant la $VaR_{99.5}$ au passif créé par ces provisions, la compagnie d'assurance ne sera solvable que si

$$\frac{actifs}{passifs} \approx \frac{SCR}{VaR_{99.5}} = 100\%.$$

C'est-à-dire que $SCR = VaR_{99.5}$.

6.2 Application aux données.

Grâce aux outils développés dans la section précédente, il est maintenant possible de calculer le SCR nécessaire pour couvrir les obligations liées au produit fictif de rente viagère. Celui-ci s'obtient en calculant sa Value-at-Risk 99.5% sur base de l'équation (6.1), dans lequel les ${}_kp_x$ sont calculés, de façon diagonale, comme étant le percentile 99.5% de leurs extrapolations selon les modèles LC et ACF-2. Il est également intéressant de comparer la $VaR_{99.5}$ à la VaR_{50}, Value at Risk de la médiane, afin de voir les différences de primes que cela impliquerait.

En supposant avoir un individu de 35 ans et un contrat, signé en 2018, de rente viagère d'une durée de 20 ans qui donne droit à un euro par an à partir de la fin de la première année, les tableaux ci-dessous font la comparaison entre les $VaR_{99.5}$ et VaR_{50} des modèles LC et ACF-2, tout en donnant l'augmentation de primes relatives Δ entre les deux.

Pour un taux d'intérêt de $i = 0.01$:

	LC			ACF-2		
	VaR_{50}	$VaR_{99.5}$	Δ	VaR_{50}	$VaR_{99.5}$	Δ
Hommes, Belgique	17.78771	17.82192	0.192 %	17.76650	17.81424	0.269 %
Hommes, France	17.74196	17.77790	0.203 %	17.72254	17.76165	0.221 %
Hommes, Pays-Bas	17.89871	17.92549	0.150 %	17.85559	17.88483	0.164 %
Femmes, Belgique	17.88650	17.91193	0.142 %	17.88455	17.92033	0.200 %
Femmes, France	17.89598	17.92179	0.144 %	17.88104	17.90870	0.155 %
Femmes, Pays-Bas	17.89565	17.91679	0.118 %	17.89656	17.93110	0.193 %

Pour un taux d'intérêt de $i = 0.0125$:

	LC			ACF-2		
	VaR_{50}	$VaR_{99.5}$	Δ	VaR_{50}	$VaR_{99.5}$	Δ
Hommes, Belgique	14.23017	14.25754	0.192 %	14.21320	14.25139	0.269 %
Hommes, France	14.19357	14.22232	0.203 %	14.17803	14.20932	0.221 %
Hommes, Pays-Bas	14.31897	14.34040	0.150 %	14.28447	14.30787	0.164 %
Femmes, Belgique	14.30920	14.32955	0.142 %	14.30764	14.33626	0.200 %
Femmes, France	14.31678	14.33743	0.144 %	14.30483	14.32696	0.155 %
Femmes, Pays-Bas	14.31652	14.33343	0.118 %	14.31725	14.34488	0.193 %

En se rappelant que $VaR_{99.5} = SCR$, plusieurs constats sont à faire :

— Comme attendu, le SCR est plus important pour le modèle qui prévoit une espérance de vie plus élevé. Puisque les individus vivent plus longtemps, il faut prévoir plus d'argent afin de couvrir les obligations du contrat. Le modèle LC prévoit donc un SCR plus élevé pour les hommes dans les trois pays et pour les femmes en France, alors que le modèle ACF-2 prévoit un SCR plus élevé pour les femmes en Belgique et aux Pays-Bas. En se référant à la section 4.5.3, les modèles qui calculent le SCR le plus élevé sont justement ceux qui prédisaient le moins bien l'espérance de vie entre 2014 et 2018, puisqu'ils avaient tendance à la surévaluer. En couplant ceci au fait que l'augmentation de l'espérance de vie diminue, il paraît judicieux de partir du modèle au SCR le plus bas comme référence de pricing.

— La différence de pricing entre le modèle LC et le modèle ACF-2 est assez faible, de l'ordre de quelques dixièmes de pourcent.

— L'augmentation relative de la VaR entre la médiane et le percentile 99.5 est plus important pour le modèle ACF-2, quel que soit le sexe ou le pays. Cela peut s'expliquer par le fait que le modèle ACF-2 dépend de 3 paramètres temporels extrapolés (K_t, $\kappa_{t,i}$et$b_{x,i,j}$) contre un seul pour le modèle LC (K_t), ce qui implique une volatilité plus élevée pour le premier modèle.

— La différence entre les simulations faites avec un taux d'intérêt de 1% ou un taux d'intérêt de 1.25% sont de l'ordre de $\pm$25%. En comparaison, l'augmentation de prime relative Δ entre les VaR_{50} et les $VaR_{99.5}$ est de l'ordre de $\pm$0.2%. Il y a donc un facteur 100 entre les augmentations relatives de primes, ce qui est énorme. Le risque le plus important pour l'assureur, dans le cas de la rente viagère présentée ici, n'est donc pas tellement le risque d'augmentation de l'espérance de vie mais plutôt le risque de variation des taux d'intérêt.

Chapitre 7

Conclusion.

Bien qu'il n'y ait pas de réponse sans équivoque à la question de savoir si un des modèles présentés dans ce mémoire est meilleur qu'un autre, plusieurs constats ont été posés tout au long des chapitres et sont récapitulés ci-dessous.

Le modèle classique de Lee-Carter pour la modélisation des taux de mortalité peut être complexifié en rajoutant des paramètres dépendant du sexe (modèle ACF-1), du pays (ACF-2) et de la cohorte considérée (ACFC).

L'analyse du paramètre de cohorte montre un effet de sous-mortalité pour les cohortes nées entre 1925 et 1945 dans l'axe Belgique/France/Pays-Bas, cet effet est similaire à l'effet de sous-mortalité présent au Royaume-Uni pour les cohortes nées entre 1930 et 1947, les causes pouvant être communes. Un deuxième effet de cohorte, plus léger, est également présent pour les hommes entre 1875 et 1895.

L'analyse rétrospective des paramètres de cohorte permet une compréhension des évolutions passées de la mortalité par cohorte. Par contre, lorsqu'il s'agit d'en faire une extrapolation, ces paramètres explosent à cause du petit nombre de représentants des dernières cohortes, le modèle ACFC n'est donc pas un bon candidat pour la modélisation des taux futurs.

Les résidus des paramètres temporels des différents modèles suivent presque tous une loi Normale ou peuvent y être associés. Afin d'extrapoler les espérances de vie, ces paramètres sont modélisés par le modèle $ARIMA(1, 10)$.

Pour ce qui est de la comparaison entre les modèles LC et ACF-2, l'espérance de vie de 1946 à 2013 est mieux captée par le modèle AFC, sauf dans le cas des femmes en France.

Dans le cas de la projection de ces espérances pour les années 2014 à 2018, le modèle ACF-2 donne une meilleure prédiction dans 4 cas sur 6, le modèle LC étant meilleur pour le cas des femmes en Belgique et aux Pays-Bas.

Les prédictions long terme de l'espérance de vie pour les années 2019 à 2060 montrent que le modèle ACF-2 donne une espérance de vie plus faible sur le long terme dans 4 cas sur 6 et que le modèle LC donne une espérance plus faible pour les deux autres cas, la répartition étant la même que pour les années 2014 à 2018.

Ces constats, couplés au fait que l'augmentation de l'espérance de vie stagne depuis 2015 permettent de conclure que l'utilisation du modèle ACF-2 pour les hommes (tous pays confondus) et pour les femmes en France et l'utilisation du modèle LC pour les femmes en Belgique et aux Pays-Bas peut être considérée comme une façon cohérente de prédire l'espérance de vie future.

Enfin, l'analyse des différences de pricing sur une rente viagère à terme échu montre que les différences entre les deux modèles sur la tarification sont de l'ordre de moins d'un pourcent. De même, les différences entre les estimations médianes et hautes au sein d'un modèle sont très faibles, de l'ordre de $0,2\%$. Le risque de souscription du SCR n'est donc pas le risque le plus important pour ce produit, le risque de marché via l'influence des taux d'intérêt étant beaucoup plus fort.

Bibliographie

[1] Bera A, Jarque C (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals : Economics Letters, 1980, vol. 6, issue 3, 255-259

[2] Boureau C (2014). Prise en compte de l'heterogeneite des individus dans le modèle de Lee-Carter. Gestion des risques [q-fin.RM]. dumas-01073547, 15

[3] Bourdillon F (2018). Les pathologies liées au tabac chez les femmes : une situation préoccupante. Bull Epidémiol Hebd.(35-36) :682.

[4] Box G, Jenkins G (1970). Time Series Analysis : Forecasting and Control. San Francisco : Holden-Day.

[5] Brouhns N, Denuit M, Vermount JK (2002). A poisson log-bilinear regression approach to the construction of projected lifetables. Insurance : Mathematics and Economics, 31(3) :373-393.

[6] Dickson D, Hardy M, Waters H (2009). Actuarial Mathematics for life Contingent Risks. International Series on Actuarial Science, 2nd Ed. :113

[7] Hainaut D, LACTU 2030 : Life Insurance, Université catholique de Louvain : https ://sites.google.com/view/donatienhainaut/home

[8] Hurlin C, Econométrie Appliquée, Université d'Orléans : https ://www.univ-orleans.fr/deg/masters/ESA/CH/CoursSeriesTemp_Chap3.pdf

[9] https ://climat.be/changements-climatiques/consequences/sante

[10] https ://eur-lex.europa.eu/LexUriServ/LexUriServ.do ?uri=OJ :L :2009 :335 :0001 :0155 :fr :PDF

[11] https ://fr.wikipedia.org/wiki/M%C3%A9thode_de_Newton

[12] https ://www.mortality.org/

[13] https ://observatoire-des-seniors.com/lesperance-de-vie-des-francais-stagne-depuis-2015/

[14] https ://statbel.fgov.be/fr/nouvelles/lesperance-de-vie-atteint-84-ans-pour-les-femmes-et-79-ans-pour-les-hommes-en-2016

[15] https ://support.minitab.com/fr-fr/minitab/18/help-and-how-to/modeling-statistics/time-series/supporting-topics/time-series-models/fit-an-arima-model/

[16] https ://www.etaamb.be/fr/arrete-royal-du-14-novembre-2003_n2003023014.html

[17] Kalben B (2000) Why men Die Younger. N am Acteur J 4(4) :83-111

[18] Lee R, Carter L (1992) Modeling and forecasting US mortality. J Am Stat Assoc 87(14) :659-675

[19] Li J (2013) A Poisson common factor model for projecting mortality and life expectancy jointly for females and males. Popul stud 67(1) :111-126

[20] Li N, Lee R (2005) Coherent mortality forecasts for a group of populations : an extension of the Lee-Carter method. demography 42(3) :575-594

[21] Meslé F, Vallin J (2010). Espérance de vie : peut-on gagner trois mois par an indéfiniment ? Population & Société 473

[22] Rudin W (1976). Principles of mathematical analysis. Third Edition. International Series in Pure and Applied Mathematics. McGraw-Hill Book Co., New York-Auckland-Düsseldorf : x+342

[23] Sex-specific mortality forecasting for UK countries : a coherent approach. Ree Yongquing Chen, Pietro Millossovich. European Actuarial Journal (2017) 7 :405-434

[24] Willets R(2004) The cohort effect : insights and explanations. Br Actuar J 10(4) :833-877

Annexe.

A1. Code et explications.

L'entièreté des codes utilisés dans le cadre de ce mémoire a été réalisée à l'aide du logiciel R. Ils ont été annexés sous forme de fichier .R.

Les codes « ACFC », « ACF-2 » et « LC » permettent d'avoir les paramètres de calibration de leurs modèles respectifs. Ils se basent sur les fichiers .txt suivant : « deaths_Belgique », « deaths_France », « deaths_ndls », « exposures_Belgique », exposures_France », exposures_ndls » qui sont également annexés au présent mémoire.

Calibration.

Les graphiques utilisés dans le chapitre 3 sont issus du code « ACFC ».

Le code « ACF-2 », utlisé pour les années 1946 à 2013 permet d'obtenir les fichiers .txt :

— « base68 » qui reprend les paramètres $d_{x,t}$ et $l_{x,t}$ empiriques pour les années 1946 à 2013.

— « kt_ACF2 » qui reprend les paramètres calibrés K_t, $\kappa_{t,i}$ et $\kappa_{t,i,j}$ du modèle ACF-2 pour les années 1946 à 2013.

— « data_ACF2 » qui reprend les paramètres calibrés $a_{x,i,j}$, B_x, K_t, $b_{x,i}$, $\kappa_{t,i}$, $b_{x,i,j}$ et $\kappa_{t,i,j}$ du modèleACF-2 pour les années 1946 à 2013.

Le code « ACF-2 », utilisé pour les années 1946 à 2018 permet d'obtenir les fichiers :

— « base73 » qui reprend les paramètres $d_{x,t}$ et $l_{x,t}$ empiriques pour les années 1946 à 2018.

— « kt_ACF2_2018 » qui reprend les paramètres calibrés K_t, $\kappa_{t,i}$ et $\kappa_{t,i,j}$ du modèle ACF-2pour les années 1946 à 2018.

— « data_ACF2_2018 » qui reprend les paramètres calibrés $a_{x,i,j}$, B_x, K_t, $b_{x,i}$, $\kappa_{t,i}$, $b_{x,i,j}$ et $\kappa_{t,i,j}$ du modèle ACF-2 pour les années 1946 à 2018.

Le code « LC », utilisé pour les années 1946 à 2013 permet d'obtenir les fichiers :

— « kt_LC » qui reprend les paramètres calibrés K_t du modèle LC pour les années 1946 à 2013.

— « data_LC » qui reprend les paramètres calibrés $a_{x,i,j}$, B_x, et K_t du modèle LC pour les années 1946 à 2013.

Le code « LC », utilisé pour les années 1946 à 2018 permet d'obtenir les fichiers :

— « kt_LC_2018 » qui reprend les paramètres calibrés K_t du modèle LC pour les années 1946 à 2018.

— « data_LC_2018 » qui reprend les paramètres calibrés $a_{x,i,j}$, B_x, et K_t du modèle LC pour les années 1946 à 2018.

Extrapolation.

Le code « extrapolation des Kt » utilise :

— le fichier « kt_ACF2 » afin de simuler 10.000 scénarios sur les années 2014 à 2018. Il en resort le fichier « ktS » qui reprend les paramètres K_t, $\kappa_{t,i}$ et $\kappa_{t,i,j}$ du modèle ACF-2 pour les années 2014 à 2018.

— le fichier « kt_ACF2_2018 » afin de simuler 10.000 scénarios sur les années 2019 à 2060. Il en resort le fichier « ktS_2018 » qui reprend les paramètres K_t, $\kappa_{t,i}$ et $\kappa_{t,i,j}$ du modèle ACF-2 pour les années 2019 à 2060.

Le code « extrapolation des Kt_LC » utilise :

— le fichier « kt_LC » afin de simuler 10.000 scénarios sur les années 2014 à 2018. Il en resort le fichier « ktS_LC » qui reprend les paramètres K_t, $\kappa_{t,i}$ et $\kappa_{t,i,j}$ du modèle LC pour les années 2014 à 2018.

— le fichier « kt_LC_2018 » afin de simuler 10.000 scénarios sur les années 2019 à 2060. Il en resort le fichier « ktS__LC_2018 » qui reprend les paramètres K_t, $\kappa_{t,i}$ et $\kappa_{t,i,j}$ du modèle LC pour les années 2019 à 2060.

Ces deux codes permettent également d'obtenir

— les graphiques des fonctions ACF et PACF de la section 4.3.

— de calculer les résultats du test de Jarque-Bera de la section 4.3.6.

— les graphiques des extrapolations des K_t, $\kappa_{t,i}$ $\kappa_{t,i,j}$ du modèle ACF-2 et K_t du modèle LC de la section 4.4.

Calcul des espérances de vie.

Le code « espérance » utilise les fichiers « data_ACF2 », « data_LC », « base73 », « ktS » et « ktS_LC » afin de calculer :

— Les espérances de vie empiriques entre 1946 et 2013.

— Les espérances de vie basée sur le modèle ACF-2 entre 1946 et 2013.

— Les espérances de vie basée sur le modèle LC entre 1946 et 2013.

— Les estimations hautes de l'espérance de vie des modèles ACF-2 et LC entre 2014 et 2018.

— Les estimations basses de l'espérance de vie des modèles ACF-2 et LC entre 2014 et 2018.

Ce code fournit également, pour le chapitre 4 :

— les graphiques d'espérance de vie, médiane et 95%

— les calculs d'écart quadratique, médiane et 95%

Le code « espérance_2060 » utilise les fichiers « data_ACF2_2018 », « data_LC_2018 », « base73 », « ktS_2018 » et « ktS_LC_2018 » afin de calculer :

— les estimations médianes de l'espérance de vie des modèles ACF-2 et LC entre 2019 et 2060.

— les estimations hautes de l'espérance de vie des modèles ACF-2 et LC entre 2019 et 2060.

— les estimations basses de l'espérance de vie des modèles ACF-2 et LC entre 2019 et 2060.

Ce code fournit également, pour le chapitre 5 :

— les graphiques d'espérance de vie, médiane et 95%

— les donnés numériques des tableaux.

Cacul du pricing

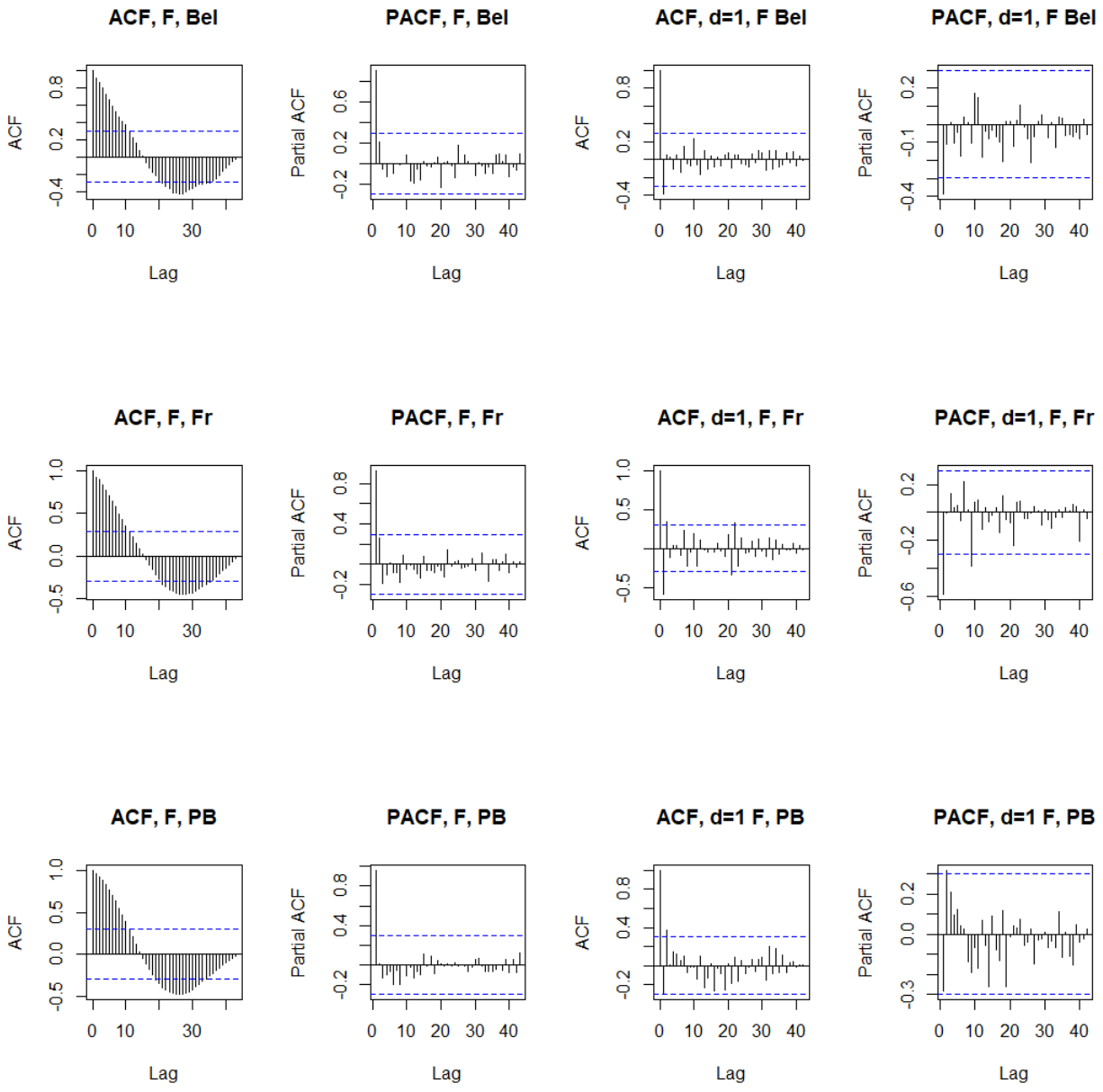
Le code « rente viagère » utilise les fichiers « data_ACF2_2018 », « data_LC_2018 », « ktS_2018 » et « ktS_LC_2018 » afin de calculer :

— les valeurs $VaR_{99.5}$ et VaR_{50} pour les modèles ACF-2 et LC.

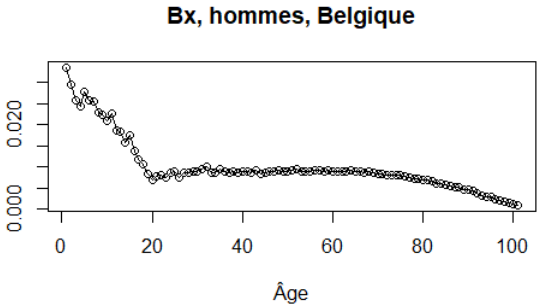
— les écarts entre relatives entre ces valeurs.

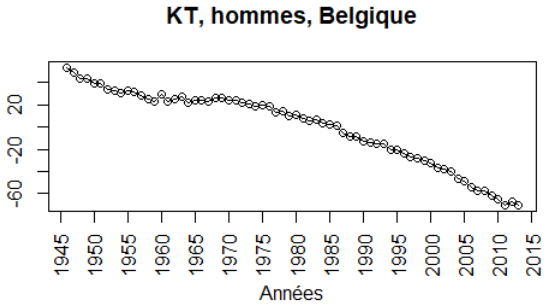
A2. Différents graphiques.

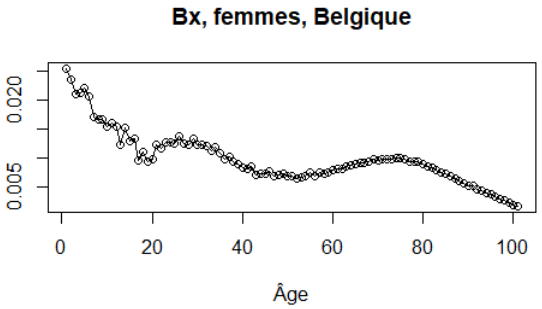
ACF et PACF pour les ktij, femmes, ACF-2.

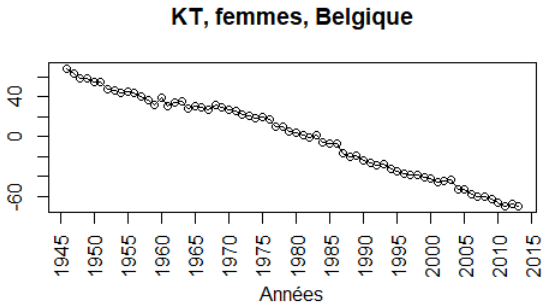


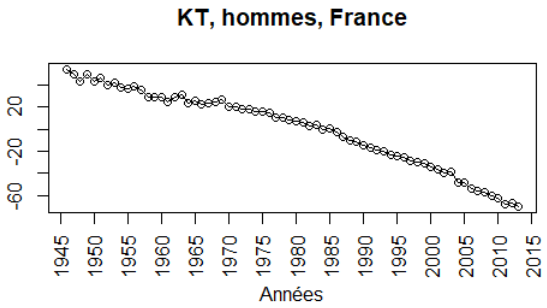
Graphiques des Bx et Kt de la méthode Lee-Carter entre 1946 et 2014, ventilé par pays et par sexe.

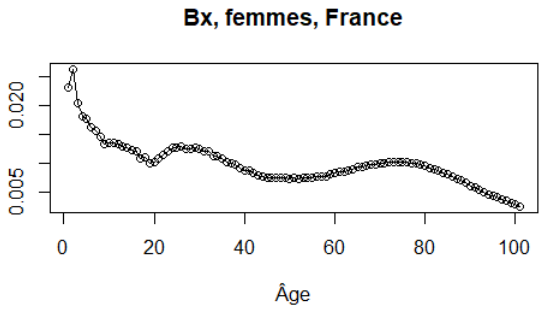


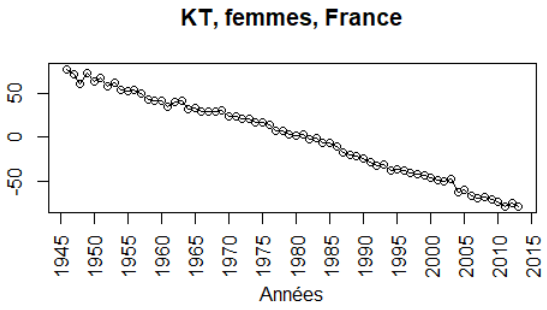


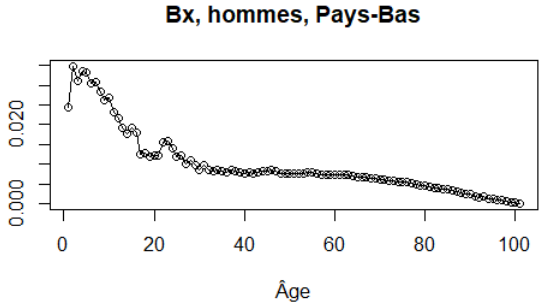


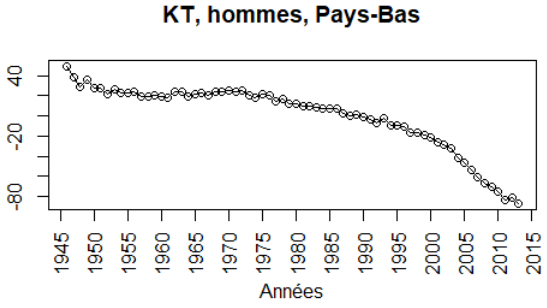


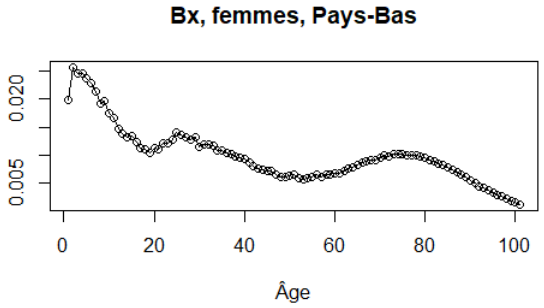


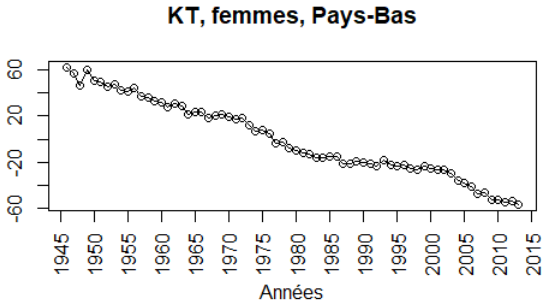




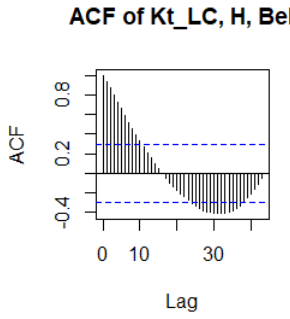


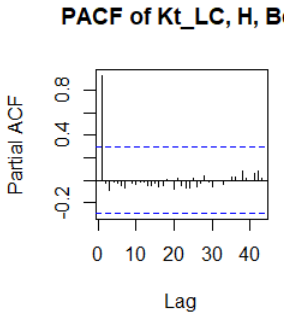


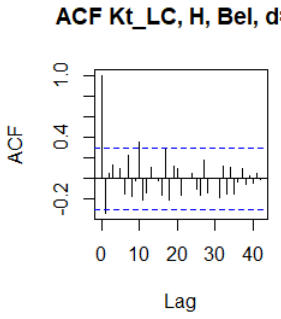


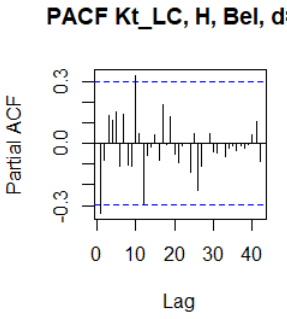


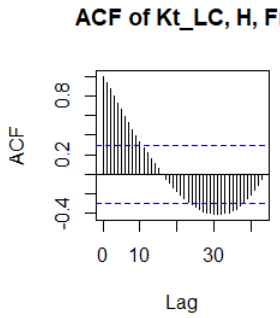
ACF et PACF pour les Kt_LC.

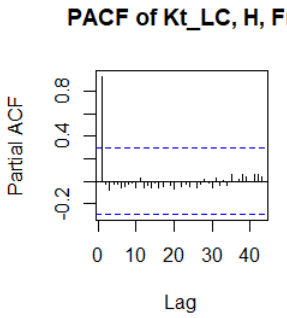


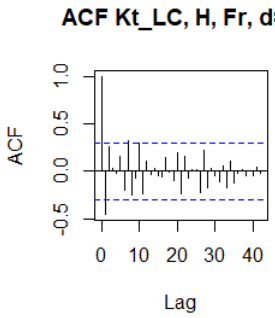


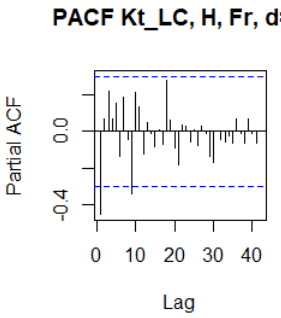


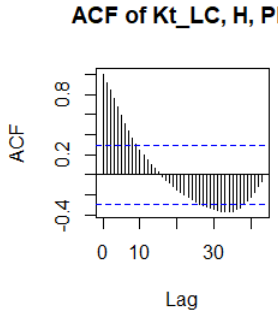


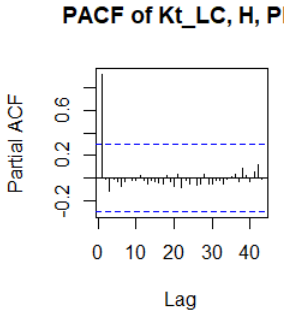

ACF of Kt_LC, H, Fr
ACF
Lag

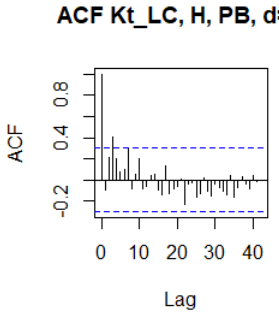

PACF of Kt_LC, H, Fr
Partial ACF
Lag

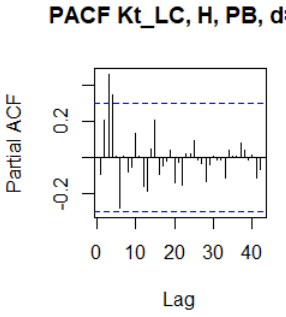

ACF Kt_LC, H, Fr, d=1
ACF
Lag

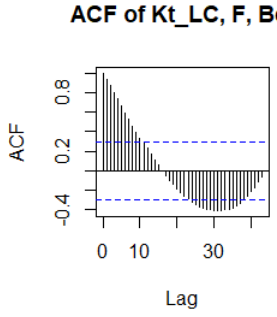

PACF Kt_LC, H, Fr, d=1
Partial ACF
Lag

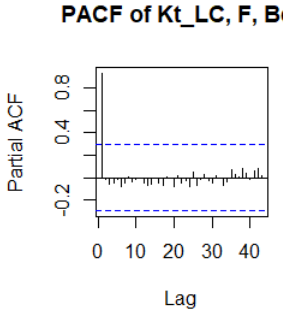

ACF of Kt_LC, H, PB
ACF
Lag

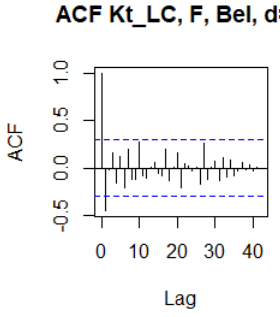

PACF of Kt_LC, H, PB
Partial ACF
Lag

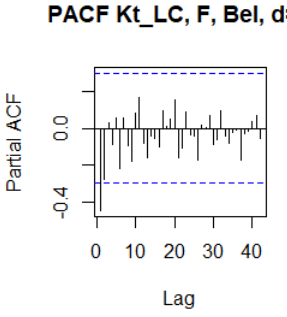

ACF Kt_LC, H, PB, d=1
ACF
Lag

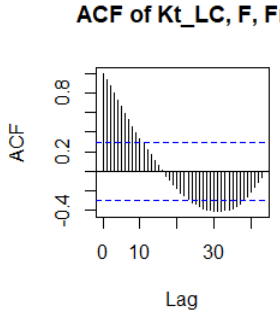

PACF Kt_LC, H, PB, d=1
Partial ACF
Lag

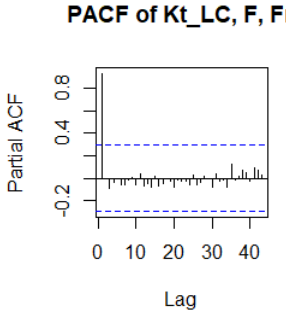

ACF of Kt_LC, F, Bel
ACF
Lag

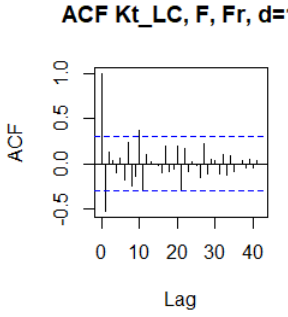

PACF of Kt_LC, F, Bel
Partial ACF
Lag

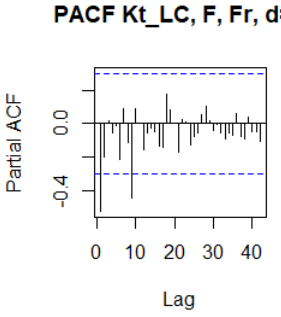

ACF Kt_LC, F, Bel, d=1
ACF
Lag

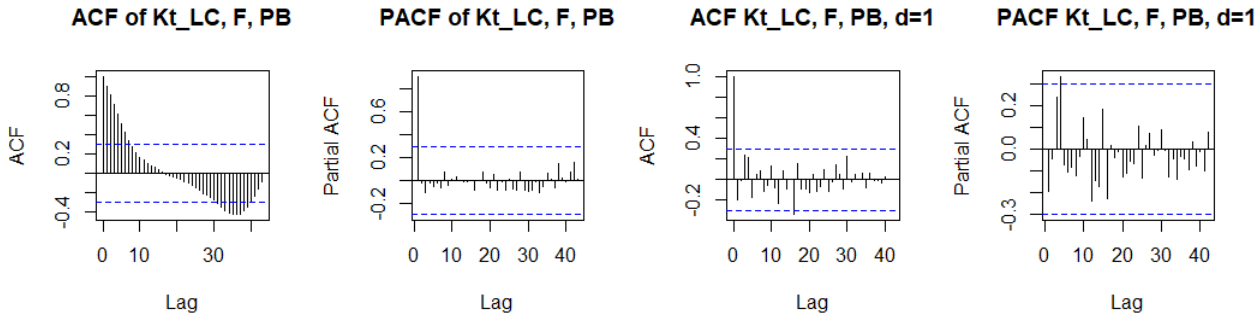

PACF Kt_LC, F, Bel, d=1
Partial ACF
Lag


ACF of Kt_LC, F, Fr
ACF
Lag


PACF of Kt_LC, F, Fr
Partial ACF
Lag


ACF Kt_LC, F, Fr, d=1
ACF
Lag


PACF Kt_LC, F, Fr, d=1
Partial ACF
Lag



A3. Tableau de l'analyse Jarque-Bera pour les Kt_LC.

	$JB = 0$		$S = 0$		$K = 3$	
Kt_LC	p-valeur	Stat	p-valeur	Stat	p-valeur	Stat
$K_{t,1,1}$	0.8741	0.26905	0.6729	0.15772	0.7632	2.7749
$K_{t,1,2}$	1.364e-05	22.405	0.001807	1.1656	0.0003719	5.6591
$K_{t,1,3}$	0.8704	0.27756	0.8045	0.092483	0.6419	2.6526
$K_{t,2,1}$	0.1559	3.7173	0.0554	0.7156	0.8277	3.1626
$K_{t,2,2}$	1.004e-07	32.228	0.0009429	1.2353	3.945e-06	6.4472
$K_{t,2,3}$	0.8066	0.42989	0.5551	0.22047	0.7752	3.2133

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Faculté des sciences
Place des sciences, 2 bte L6.06.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/sc**