DIPLÔME NATIONAL DE DOCTORAT

(Arrêté du 25 mai 2016)

Date de la soutenance : **12 juillet 2021**

Nom de famille et prénom de l'auteur : **Monsieur PIVETEAU Samuel**

Titre de la thèse : *Modélisation de la mortalité par cause de décès*

Résumé

Introduction

Cette thèse traite de la modélisation de la mortalité par cause de décès. Elle attaque le problème sous trois angles détaillés dans les chapitres 2, 3 et 4, lesquels seront résumés ci-après. Le chapitre 1 de cette thèse constitue l'introduction au sujet et propose une vue d'ensemble de la modélisation de la mortalité. Il débute naturellement par la présentation des enjeux sociaux de la modélisation de la mortalité. L'introduction se poursuit par l'histoire des modèles de mortalité générale et comment ces-derniers ont été raffinés pour inclure davantage d'information au cours des dernières décennies. Nous abordons ensuite la question de l'hétérogénéité de la mortalité sous deux aspects : la diversité des caractéristiques influençant la mortalité et les différentes causes de décès. Nous nous attardons principalement sur ce second point, qui constitue le cœur de cette thèse. Nous présentons les enjeux de la modélisation de mortalité par cause ainsi que les modèles les plus importants développés au cours des dernières années pour tenter de répondre à ces problématiques. Enfin nous présentons les données sur lesquelles nous avons été amenés à travailler, à savoir la mortalité aux Etats-Unis, et plus précisément la mortalité féminine. Nous détaillons, au moyen de statistiques descriptives les principales caractéristiques de la mortalité par cause féminine ainsi que son évolution au cours de la période récente.

Chapitre 2

Le deuxième chapitre traite de l'extrapolation de la mortalité par cause de décès aux âges avancés, sujet important en raison des incertitudes qui pèsent sur les données aux grands âges. L'objectif est d'extrapoler les forces de mortalité par cause aux âges avancés tout en maintenant une cohérence avec les méthodes usuelles d'extrapolation de la mortalité toutes causes aux grands âges. Pour ce faire, nous proposons une approche Top-down similaire à celle introduite dans Athanasopoulos et al. (2009) que nous adaptons à la mortalité par cause distribuée comme une variable de Poisson. L'approche Top-down est une méthode d'extrapolation des séries temporelles hiérarchiques dans laquelle la série principale est extrapolée puis fragmentée afin de retrouver les sous séries qui la composent. La méthode développée dans Athanasopoulos et al. (2009)

consiste à extrapoler la série principale et les contributions des séries qui la composent séparément avant de les recombiner pour obtenir l'intégralité des séries. Ainsi, nous sommes assurés que la somme des séries est égale à la projection de la série principale, chose qui peut s'avérer pertinente si le modèle d'extrapolation de la série principale est jugé important. Dans le cadre de la fermeture des tables de mortalité, certains modèles sont considérés comme des étalons dont il ne faut pas s'éloigner, justifiant ainsi le recours à cette approche Top-down pour l'extrapolation de la mortalité par cause aux grands âges. La vraisemblance du modèle de mortalité par cause est scindée en deux parties représentant respectivement la mortalité toutes causes et la contribution des causes de décès à celle-ci.

Nous faisons l'hypothèse que la force de mortalité toutes causes est une fonction de l'âge et dépend d'un ensemble de paramètres que nous cherchons à estimer, tandis que la contribution des différentes causes de décès à la force de mortalité totale, dépend d'un autre ensemble de paramètres. Sous certaines hypothèses, la fonction de log-vraisemblance se scinde en deux parties : la première est relative à la force de mortalité toutes causes confondues et la seconde aux contributions des causes à la mortalité totale. Ainsi l'extrapolation de la mortalité par cause aux âges avancés est obtenue en deux étapes. La première consiste à extrapoler la mortalité toutes causes aux âges avancés en maximisant la première partie de la log-vraisemblance. Cela peut être accompli en utilisant les techniques standards de fermeture de table de mortalité telles que proposées dans Delwarde et Denuit (2005). Nous avons sélectionné trois approches : le modèle de Gompertz-Makeham, le modèle de Kannisto (voir Kannisto et al. (1994)) et le modèle Coale-Kisker (1990) revisité sous forme paramétrique ainsi que détaillé dans Flici (2016). Chacun de ces modèles repose sur des hypothèses de mortalité variées, ce qui conduit à des extrapolations de mortalité sensiblement différentes. La deuxième partie de l'algorithme consiste à extrapoler la contribution des causes de décès à la mortalité générale en maximisant la seconde partie de la log-vraisemblance. Dans le cadre qui est le nôtre, une telle extrapolation revient à considérer les décès par cause comme des réalisations de variables multinomiales, dont la somme est égale au nombre de décès toutes causes confondues. Afin d'extrapoler de la façon la plus souple possible, nous proposons d'utiliser un modèle P-splines adapté au modèle multinomial. La méthode des P-splines est adaptée au cadre GLM conformément à Eilers et Marx (1996). Nous montrons également dans ce chapitre que le modèle GLM multinomial peut être étendu au moyen d'autres fonctions de lien issues de la théorie CoDA (voir Aitchison (1986)). En recombinant les extrapolations de la force de mortalité aux grands âges et des contributions des causes à la mortalité générale, nous obtenons des extrapolations des mortalités par cause cohérentes avec le choix de la méthode de fermeture de table de mortalité. Une telle exigence peut s'avérer nécessaire lorsque la fermeture de la table de mortalité est fondée sur des hypothèses spécifiques que l'on souhaite conserver. Nous proposons une analyse des résultats obtenus sur la population féminine américaine en 2017. Nous constatons que les différents GLMs ont peu d'influence sur le résultat final, contrairement au choix de la méthode de fermeture. Celle-ci impose une cadence de décès plus ou moins rapide à la population, modifiant substantiellement l'importance de certaines causes, sans pour autant bouleverser la hiérarchie des principales causes de décès. Une telle méthode a été développée afin de fournir aux experts médicaux des bases complètes sur lesquelles appliquer leurs jugements d'experts. Elle peut servir de support à une projection temporelle par cause de décès et par âge telle que proposée dans le chapitre 4.

Chapitre 3

Dans le chapitre 3 de cette thèse, nous proposons un algorithme permettant de constituer des groupes tels que l'ajustement par un modèle de Lee-Carter sur chaque groupe soit optimal. Le cadre d'étude est le suivant : nous avons à disposition des séries temporelles de forces de mortalité calculées pour un ensemble fini de caractéristiques, par exemple l'âge, la cause de décès ou le pays, et que nous souhaitons ajuster par un modèle de Lee-Carter.

En raison de la dynamique propre à certains groupes de caractéristiques, il peut être utile de diviser la base de données en plusieurs groupes, sur chacun desquels un modèle de Lee-Carter est calibré afin d'obtenir un meilleur ajustement et une plus juste appréciation de la dynamique de l'ensemble des séries. Des méthodes

de classification de séries temporelles existent déjà dans la littérature statistique, ainsi Paparrizos et Gravano (2015) propose une approche de type K-centroids pour les séries temporelles, permettant de grouper les séries temporelles présentant des formes proches au sein de K groupes différents. Hélas, ce type d'approche n'est pas spécifiquement adapté aux modèles de mortalité stochastiques et ne fonctionnent que sur des séries standardisées. Pour ces raisons, la question demeure ouverte de savoir comment constituer les groupes optimisant l'erreur d'ajustement sur des modèles de Lee-Carter. Mathématiquement parlant, le problème revient à minimiser la fonction de coût complexe, problème qui n'admet pas de solution sous forme de formule fermée.

Nous proposons un algorithme dérivé des K-centroids (voir Macqueen (1967)) et adapté au modèle de Lee-Carter que nous nommons les K-LC. A partir d'un algorithme d'apparence complexe, nous montrons que, moyennant une modification des contraintes usuelles sur les paramètres du modèle, la méthode revient à un algorithme des K-centroids avec une fonction de distance particulière. Nous proposons également deux variantes des K-LC adaptées aux données erratiques : les K-LC contraints, inspirés de Leisch et Grün (2020) et les K-LC pénalisés.

Deux applications sont proposées afin d'illustrer l'algorithme. La première porte sur la division par sexe dans les prédictions de mortalité. Il est usuel (voir Lee (2000)) de diviser la population selon le sexe avant d'appliquer le modèle de Lee-Carter sur chacune des sous-populations séparément. Nous montrons sur la population américaine entre les années 1979 et 2012 que cette division n'est pas la plus pertinente et qu'un découpage générationnel est plus intéressant pour cette période. La seconde application porte sur le découpage de groupes au sein de la mortalité féminine américaine par âges et causes de décès. Nous montrons que 6 groupes différents mêlant plusieurs causes peuvent être constitués afin d'obtenir un meilleur ajustement. Nous effectuons ensuite des prévisions de mortalité à l'aide d'une marche aléatoire et constatons que les résultats obtenus après découpage par K-LC sont plus proches des réalisations observées sur les années allant de 2013 à 2017 que les résultats obtenus par classification usuelle (sexes séparés et causes séparées). Les deux variantes des K-LC sont également appliquées et permettent d'améliorer les résultats.

Chapitre 4

Le chapitre 4 traite des prévisions des taux bruts mortalité par cause. Nous avons proposé un modèle permettant de traiter de trois problèmes fondamentaux dans la projection de la mortalité par cause, à savoir les changements de tendances, le problème de la dépendance temporelles des causes de décès et la présence d'un biais dans les prévisions de mortalité.

Nous montrons que le modèle défini s'écrit comme un modèle Espace-Etat non gaussien. Le problème des biais est résolu par l'usage d'un modèle Espace-Etat pour lequel les variables d'espace ne sont pas les logarithmes des forces de mortalité mais les décès par cause (voir Fung et al. (2017) pour un modèle Espace-Etat appliqué aux logarithmes des forces de mortalité). Afin de rester dans un cadre connu, nous employons pour chacune des causes le modèle de Lee-Carter-Poisson tel que présenté dans Brouhns et al. (2002). Les paramètres temporels du modèle de Lee-Carter sont considérés comme des variables d'état auxquelles nous attribuons une structure de dépendance multivariée afin d'intégrer la dépendance temporelle entre les causes. Enfin, le problème des changements de tendances des paramètres temporels du modèle est pris en considération par l'ajout d'une dynamique au paramètre de pas.

La calibration du modèle reprend la méthode fréquentiste proposée par Schön et al. (2011), consistant à estimer les paramètres du modèle au moyen d'un algorithme Espérance-Maximisation. Nous adaptons cette méthode au modèle présenté, et montrons que pour un certain nombre de paramètres, le recours à des méthodes d'optimisation numérique n'est pas nécessaire. Une application à la population féminine américaine entre 1979 et 2012 est ensuite proposée. Nous détaillons les structures de dépendance obtenues et nous en mesurons l'impact sur la dépendance entre causes de décès au moyen de simulations. Nous effectuons ensuite des prévisions sur les années 2012 à 2017, que nous comparons avec celles d'un modèle de Lee-Carter standard appliqué sur chacune des causes séparément. En comparant les résultats issus des

modèles aux réalisations de décès observées, nous constatons une nette supériorité du modèle Espace-Etat, que nous estimons en grande partie due à sa capacité à prendre en considération les changements de tendances.

Références

J. Aitchison. The statistical analysis of compositional data. Journal of the Royal Statistical Society. Series B (Methodological), 44(2) :139–177, 1982.

G. Athanasopoulos, R. Ahmed, and R. Hyndman. Hierarchical forecasts for australian domestic tourism. International Journal of Forecasting, 25(1) :146–166, 2009.

N. Brouhns, M. Denuit, and J. Vermunt. A poisson log-bilinear regression approach to the construction of projected life tables. Insurance : Mathematics and Economics, 31 :373–393, 2002.

A.J. Coale and Ellen Kisker. Defects in data on old-age mortality in the United States : New procedures for calculating mortality schedules and life tables at the highest ages. Asian & Pacific Population Forum, 4 :1–31, 1990.

A. Delwarde and M. Denuit. Construction de tables de mortalité périodiques et prospectives. ECONOMICA, 2005.

P. Eilers and B. Marx. Flexible smoothing with b-splines and penalties. Statistical Science, 11(2) :89–102, 1996.

F. Flici. Closing-out the Algerian life tables : for more accuracy and adequacy at old-ages. 2016.

M.C. Fung, G. Peters, and P. Shevchenko. A unified approach to mortality modelling using state-space framework : characterisation, identification, estimation and forecasting. Annals of Actuarial Science, 11(2) :343–389, 2017.

B. Gompertz. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. Philosophical Transactions of the Royal Society of London, 115 :513–583, 1825.

V. Kannisto, J. Lauritsen, A.R. Thatcher, and J. Vaupel. Reductions in mortality at advanced ages : Several decades of evidence from 27 countries. Population and Development Review, 20(4) :793–810, 1994.

R. Lee. The lee-carter method for forecasting mortality, with various extensions and applications. North American Actuarial Journal, 4 :80–91, 2000.

F. Leisch and B. Grün. Extending standard cluster algorithms to allow for group constraints. 2020.

J. MacQueen. Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1 : Statistics, pages 281–297. University of California Press, 1967.

W. Makeham. On the law of mortality and the construction of annuity tables. The Assurance Magazine, and Journal of the Institute of Actuaries, 8(6) :301–

310, 1860.

John P. and Luis G. k-shape : Efficient and accurate clustering of time series.
In SIGMOD Conference, 2015.

T. Schön, A. Wills, and B. Ninness. System identification of nonlinear state-space
models. Automatica, 47 :39–49, 2011.