UTILISATION DES ESTIMATEURS DE KAPLAN-MEIER PAR GÉNÉRATION ET DE HOEM POUR LA CONSTRUCTION DE TABLES DE MORTALITÉ PROSPECTIVES

Version 1.5 du 15/01/2018

Quentin Guibert[1] **Frédéric Planchet**[2]

ISFA - Laboratoire SAF [β]

Université de Lyon - Université Claude Bernard Lyon 1

RÉSUMÉ

Data quality is an overarching concern when it comes to building a mortality model or prospective mortality tables. This is even more significant when these procedures are based on a small population, as data may show major random fluctuations due to a lack of information for particular ages. Such situations arise frequently with the entry into force of Solvency II as insurers shall consider their own data sets, limited in size, to build best estimate tables. Since parametric methods are too rough to capture a realistic mortality pattern in two dimensions, the mortality profile is quite often adjusted using exogenous information, such as a table based on a national population. In light of this, the aim of this paper is to discuss the problem of choosing appropriate estimators for two-dimensional mortality rates or death rates in the presence of independent censoring. Indeed, practitioners currently use the Hoem estimator or the Kaplan-Meier estimator split by generation without questioning their relevance and reliability. We propose in this paper a comparative analysis of these estimators and try to give some criteria to choose one approach over another, and give some figures based on a real insurance portfolio and simulated data. Finally, we provided some non-parametric estimators for a direct estimation of death rates both with the cohort and the period approaches

[1] Quentin Guibert est post doctorant au sein du Laboratoire SAF et consultant chez PRIM'ACT.

[2] Frédéric Planchet est Professeur à l'ISFA et actuaire associé chez PRIM'ACT. Contact : frederic@planchet.net.

[β] Univ Lyon, Université Claude Bernard Lyon 1, Institut de Science Financière et d'Assurances (ISFA), Laboratoire SAF EA2429, F-69366, LYON, France.

1. Introduction

L'appropriation du modèle proposé par Lee et Carter, 1992 a conduit à partir du début des années 2000 à de très nombreux développements concernant la modélisation prospective de la mortalité (cf. Barrieu et *al*., 2012 ; Booth et Tickle, 2008 ; Cairns, Blake et Dowd, 2006 pour des exemples de revues). Les spécifications retenues s'appuient généralement sur des statistiques nationales tirées de recensements (observations du nombre de décès et d'expositions au risque) et utilisent des modèles liant, d'une part, des taux de mortalité ou des taux de hasard et, d'autre part, l'âge, le temps calendaire et l'année de naissance de la cohorte. Malgré cette forte émulation, Cairns et *al*., 2015 remarquent que la question de la qualité des données de base utilisée par ces modèles n'est pas triviale et montrent que les conventions retenues pour élaborer les taux de mortalité (ou les intensités de mortalité), variables de base pour l'immense majorité des modèles de mortalité des assureurs vie, peuvent présenter des biais dont les effets sont significatifs. Pour les données issues de recensements en Angleterre et au Pays de Galles, ces auteurs décrivent plusieurs anomalies venant affecter le calcul de l'exposition aux risques par âge et qui proviennent du mode de collecte, d'une mauvaise prise en compte des naissances (ou des nouvelles entrées dans la population) et des conventions de découpage de l'exposition par âge et par date de naissance entiers. Ces erreurs sont d'autant plus importantes qu'elles peuvent se propager au fil du temps, voir également Boumezoued, 2016 pour des propositions de corrections.

Les questions de qualité et de précision des quantités de base servant à estimer un modèle prospectif ou bien simplement à élaborer une table de mortalité de base sont cruciales. Elles apparaissent tout autant lorsqu'un assureur souhaite mettre à profit des données internes, construites sur la base d'observations ligne à ligne sur les assurés. Celles-ci sont usuellement observées en temps continu[3] sur une période de temps donnée et sont soumises à censure à droite et troncature à gauche. Contrairement aux données nationales, les échantillons considérés sont toutefois moins profonds et peuvent être soumis à des fluctuations d'échantillonnage plus importantes du fait de faibles volumes. Lorsque la taille de population assurée est trop modeste ou l'historique disponible trop peu profond, les données de l'assureur sont alors utilisées pour alimenter des approches par positionnement (p. ex. Ahcan et *al*., 2014 ; Tomas et Planchet, 2014). Dans ce contexte, l'estimation des quantités de base peut perdre en fiabilité et conduire, selon la méthode retenue, à des écarts dont l'effet est significatif lorsqu'il s'agit de projeter la mortalité.

Dans ce papier, nous analysons les estimateurs des taux de décès bruts[4] pouvant être retenus pour la construction et la validation de tables de mortalité prospectives, bâties à partir d'une population assurée. Plusieurs quantités peuvent en effet être retenues dans la

[3] Avec une discrétisation en pratique quotidienne.
[4] Au sens de probabilité conditionnelle de décès ou fonction de hasard en fonction du contexte.

pratique et il est utile d'examiner comment convenablement utiliser chacune lorsqu'une dimension âge et une dimension temps sont retenues.

Une littérature abondante en épidémiologie, en biostatistique et en démographie s'intéresse à l'inférence de modèles de survie prospectifs. Elle repose usuellement sur l'utilisation du diagramme de Lexis comme outil de base pour représenter la durée de vie d'un individu en fonction du temps calendaire. Le processus des décès dépend alors de la dynamique des naissances et l'inférence du modèle mise en œuvre peut être considérée de deux manières selon le mode d'échantillonnage retenu lors de la collecte des données (Lund, 2000). Lorsque le schéma d'observation donne une information sur les dates de naissance, la durée de vie d'un individu peut être estimée en fonction du temps calendaire et de son âge, en conditionnant par rapport à sa date de naissance (cf. Keiding, 1990 pour une présentation extensive des techniques d'inférence usuelles et Keiding, 1991 dans le cas de modèle d'invalidité). Ce mode d'inférence est celui habituellement retenu dans la littérature actuarielle et par les praticiens. Dans le cadre de données nationales par exemple, une information sur la date de naissance est généralement disponible, bien qu'elle puisse être imprécise (année de naissance) et ainsi nécessiter des corrections. Lorsque des données internes sont disponibles, cette information est plus fiable et les actuaires estiment la loi de survie en considérant l'année de naissance comme une covariable discrète, puis en appliquant des méthodes tirées de l'analyse de survie unidimensionnelle. Les estimateurs de Kaplan et Meier, 1958 ou de Hoem, 1971 sont alors classiquement utilisés dans la pratique pour l'évaluation de taux de décès bruts. Comme pour des données nationales toutefois, le choix de l'estimateur et des simplifications qui peuvent être retenues doit être examiné avec précaution, en particulier compte tenu des contraintes induites par la réglementation Solvabilité II concernant la fonction actuarielle. Il est également possible d'estimer la loi de survie en introduisant une hypothèse sur processus de naissance, mais comme le rappelle Lund, 2000, cette approche est plus fréquemment mise en œuvre dans un contexte où les dates de naissance ne sont pas disponibles (p. ex. données transversales), ce qui est rarement le cas des données collectées en assurance (cf. Guilloux, 2007) pour l'estimation de loi de survie avec ce type d'approche en présence d'un motif d'échantillonnage quelconque).

Le présent travail est organisé de la manière suivante. On rappelle dans la Section 2 la définition des probabilités d'intérêt pour les actuaires lorsque le cadre de travail utilisé est en temps discret. La Section 3 décrit ensuite le cadre d'inférence couramment rencontré dans la pratique et les différents estimateurs utilisés par les praticiens. La section 4 est consacrée à la comparaison de ces estimateurs numérique. La section 5 fournit enfin des estimateurs non-paramétriques généraux des taux de décès bruts en dimension deux, non utilisés des praticiens, et qui pourraient être présentés comme des alternatives intéressantes pour l'estimation et la validation de tables prospectives. La section 6 conclue ce papier.

Nous soulignons en outre que ce travail n'apporte pas d'innovations méthodologiques, en

dehors de la Section 5 où certains estimateurs présentés ne sont pas décrits dans la littérature à notre connaissance, concernant la construction de tables, mais permet d'examiner et de discuter, à partir d'une approche empirique, le rôle effectif des différents estimateurs utilisés par les praticiens pour l'élaboration de modèles prospectifs ou de tables de mortalité prospectives. Ces démarches étant désormais plus largement mise en œuvre avec la mise en application de Solvabilité II, ce papier comporte de ce point de vue une assistance à la profession d'actuaire dans le choix des quantités pertinentes à utiliser en fonction du contexte.

2. Probabilités de décès et diagramme de Lexis

La trajectoire de vie d'un individu est classiquement représentée dans un diagramme de Lexis où apparait en abscisse le temps calendaire et en ordonnée l'âge de l'individu. En reprenant les notations de Keiding, 1990, cette trajectoire est représentée dans ce système de coordonnées par le segment compris entre les points $(\sigma,0)$ et $(\sigma+X,X)$, où σ correspond à la date de naissance (aléatoire) de l'individu et X correspond à sa durée de vie (aléatoire). En supposant que les naissances surviennent selon un processus de Poisson indépendant de la durée de vie d'un individu, le processus ponctuel $(\sigma_i,X_i)_{i\in I}$, pour une population I d'individus i.i.d., admet une intensité $\mu(x,t)$ de décès (ou taux de décès) à l'âge exact x et à la date t (Brillinger, 1986 ; Lund, 2000), correspondant au nombre de décès observés sur le rectangle $[x,x+dx[\times[t,t+dt[$. Dans la suite, on note $Y=\sigma+X$.

Pour des raisons pratiques et notamment pour l'évaluation des primes et des provisions, les actuaires sont amenés à travailler en temps discret selon un découpage le plus souvent annuel en âge et en année. Dans cette configuration, le taux de décès est remplacé par une probabilité de décès annuelle $q(x,t)$ pour un âge x entier et une année t entière. Deux représentations du diagramme de Lexis sont alors couramment admises dans la pratique (Boumezoued, 2016), et aboutissent à une définition différente de ces probabilités de décès :

- selon une décomposition par période, i.e. en regroupant les individus atteignant l'âge x sur l'année t, soit

$$q_P(x,t)=P\left(X\le x+1,X+\sigma\le t+1\middle|X>x,X+\sigma>t\right),$$

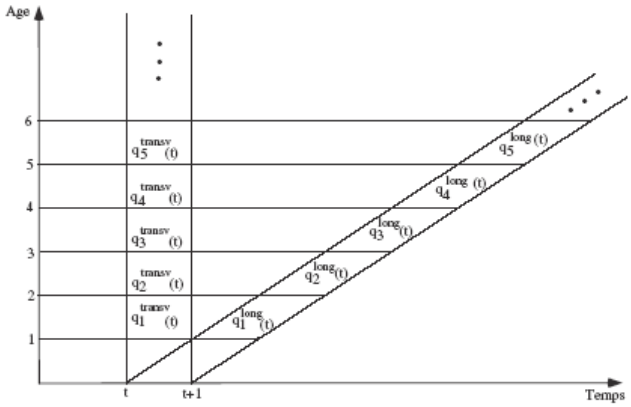
- selon une décomposition par cohorte ou générationnelle, i.e. en regroupant les individus nés l'année $t-x$, soit

$$q_C(x,t)=P\left(X\le x+1\middle|X>x,\sigma\in]t-x,t-x+1]\right).$$

La Fig. 1 : fournit une représentation de ces deux décompositions dans le diagramme de

Lexis.

Fig. 1 : Décomposition par période et par cohorte du diagramme de Lexis.



La décomposition par période conduit à mélanger des individus issus de deux cohortes d'individus nés les année $t-x-1$ et $t-x$, alors que la seconde décomposition concentre sur une seule génération et s'étale sur une période d'un an.

3. Estimation des taux de décès pour une table de mortalité prospective

Cette section décrit le cadre classique dans lequel les données sont observées, ainsi que les principales méthodes utilisées par les actuaires pour estimer les taux et probabilités de décès, ces derniers servant généralement de quantité de base pour la construction de tables prospectives dans la pratique.

a. Données observées en assurance

Les données issues de bases nationales (Human Mortality Database, 2016) permettent généralement d'estimer les taux de décès bruts sur la base d'une hypothèse de constance de la fonction d'intensité sur un carré ou un parallélogramme du diagramme de Lexis. L'estimateur considéré correspond au ratio *nombre de décès* sur *exposition au risque* pour lequel le calcul du dénominateur comprend un certain nombre d'approximations provenant de la non observation précise de la date de naissance des individus.

En revanche, les données issues d'un portefeuille d'assurance permettent un suivi individuel de la population assurée. Les dates de naissance, d'entrée dans le portefeuille et de décès peuvent être connues avec précision. On constate de manière empirique que ces informations sont soumises à censure à droite et à troncature à gauche Généralement, elles couvrent un historique nettement plus court et correspondent à des volumétries significativement inférieures aux données nationales.

Dans cette section nous explicitons deux cadres d'estimations permettant d'aboutir aux estimateurs classiquement retenus par les praticiens. On se place pour cela dans le cadre d'un modèle d'échantillonnage en présence de censure aléatoire droite non informative et

indépendante, voir par exemple Planchet et Thérond, 2011. Afin d'alléger les notations, les quantités qui suivent doivent être comprises comme étant relatives aux observations tronquées, bien que la présence de troncature ne soit pas explicitement mentionnée. En notant C la variable de censure à droite de fonction de répartition G, nous disposons d'un échantillon de taille $n \geq 1$ où sont observés (σ_i, T_i, D_i), $i = 1, \ldots, n$, tel que

$$T_i = X_i \wedge C_i \text{ et } D_i = \begin{cases} 1 \text{ si } X_i \leq C_i \\ 0 \text{ si } X_i > C_i \end{cases}.$$

On note $Z_i = T_i + \sigma_i$. Remarquons que la variable σ n'est pas soumise à censure et est nécessairement observée pour l'ensemble des individus.

Remarque 1. *La référence explicite à la troncature gauche (indépendante) implique d'introduire $(E_1, \ldots, E_n)$ les instants d'entrée dans l'échantillon de chaque individu. Dans ce cas, il s'agit de remplacer la loi de X_i par celle de $X_i \mid X_i > E_i$, ce qui conduit en pratique à remplacer l'échantillon de données observé par un ensemble de données indépendantes, mais non identiquement distribuées, du fait du conditionnement.*

b. Estimateurs de Hoem

Une hypothèse couramment employée que ce soit sur des données longitudinales ou des données nationales consiste à supposer que les intensités de transition sont constantes sur un carré (décomposition par période) ou sur un parallélogramme (approche cohorte) du diagramme de Lexis. L'estimateur du maximum de vraisemblance des taux de décès correspond s'obtient alors comme (Hoem, 1969 ; Boumezoued, 2016)

$$\mu_P(x,t) = \frac{N_P(x,t)}{E_P(x,t)} \text{ et } \mu_C(x,t) = \frac{N_C(x,t)}{E_C(x,t)},$$

où $N_P(x,t)$ et $E_P(x,t)$ (resp. $N_C(x,t)$ et $E_C(x,t)$) correspond au nombre de décès et à l'exposition au risque attachés au carré (x,t) (resp. au parallélogramme (x,t)). Ces estimateurs sont usuellement appelés estimateurs de Hoem. Cette approche paramétrique, reposant sur l'hypothèse de constance par morceaux des taux décès, est issue d'une longue tradition en sciences actuarielles. Les taux estimés, qualifiés de « taux bruts », font généralement l'objet d'une régularisation dans un second temps, par exemple en utilisant des modèles linéaires généralisés, cf. par exemple Delwarde et Denuit, 2005.

Pour calculer ces quantités, introduisons les processus de comptage :

- $R_i(a,s) = I(T_i \geq a, \sigma_i + T_i \geq s)$, l'indicatrice de présence sous risque, comptabilisant les individus vivants et non censurés au point (a,s) ;

- $N_i(a,s) = I\left(T_i \le a, \sigma_i + T_i \le s, D_i = 1\right)$, le processus ponctuel des décès non censurées au point (a,s) ;

et les processus agrégés :

- $r(a,s) = \sum_{i=1}^{n} R_i(a,s)$, comptabilisant l'effectif sous risque ;
- $n(a,s) = \sum_{i=1}^{n} N_i(a,s)$, comptabilisant le nombre d'évènements survenus non censurés.

Avec ces notations, on obtient

$$N_P(x,t) = \int_x^{x+1}\int_t^{t+1} n(a,s)\,dads \text{ et } E_P(x,t) = \int_x^{x+1}\int_t^{t+1} r(a,s)\,dads,$$

et

$$N_C(x,t) = \int_x^{x+1}\int_{t+a-x}^{t+a-x+1} n(a,s)\,dads \text{ et } E_C(x,t) = \int_x^{x+1}\int_{t+a-x}^{t+a-x+1} r(a,s)\,dads.$$

On remarque que l'exposition au risque avec l'approche par période s'écrit également

$$E_P(x,t) = E_U(x,t) + E_D(x,t),$$

où $E_U(x,t) = \int_x^{x+1}\int_t^{t+a-x} r(a,s)\,dads$ correspond au temps passé dans $[x,x+1[\times[t,t+1[$ par les individus nés en $t-x-1$ avant leur anniversaire de l'année t et $E_D(x,t) = \int_x^{x+1}\int_{t+a-x}^{t+1} r(a,s)\,dads$ correspond au temps passé dans $[x,x+1[\times[t,t+1[$ par les individus nés en $t-x$ après leur anniversaire de l'année t (cf. Planchet et Thérond, 2011, p. 145). De la même manière, le nombre de décès s'écrit

$$N_P(x,t) = N_U(x,t) + N_D(x,t).$$

Il convient de noter que les estimateurs en vision par période ou générationnelle ne sont pas liés par une relation simple, et en particulier $\mu_P(x,t) \neq \mu_C(x,t)$.

Puisque les intensités de transition sont supposées constantes par morceaux, les estimateurs naturellement associés de la probabilité conditionnelle de décès sont alors

$$q_P(x,t) = 1 - \exp\left(-\mu_P(x,t)\right) \text{ et } q_C(x,t) = 1 - \exp\left(-\mu_C(x,t)\right).$$

Certains auteurs (par exemple Cairns, Blake et Dowd, 2008) proposent de considérer une

approximation du type $q(x,t) \approx \mu(x,t) \Big/ \left(1+0,5\times\mu(x,t)\right)$. En pratique, les actuaires utilisent également une version approximée des estimateurs des taux de décès, parfois qualifiée par abus de langage d'estimateurs de Hoem, i.e. $q(x,t) \approx \mu(x,t)$. Comme $1-e^{-x} \leq x$ lorsque $x \geq 0$, cet estimateur simplifié surestime la probabilité de décès estimée par les relations faisant intervenir l'exponentielle. Les praticiens modifient souvent le calcul de l'exposition $E(x,t)$ pour diminuer cet écart. Dans la suite, nous retiendrons cette convention et nous réaliserons également cet abus de langage.

Remarque 2. *L'approximation utilisée pour le calcul des probabilités de décès est comparable à l'estimateur introduit par Harrington et Flemming pour la fonction de survie (univariée), i.e.* $\hat{S}_{HF}(t) = \mathbf{exp}\left(-\hat{H}_{NA}(t)\right)$, avec $\hat{H}_{NA}(t)$ l'estimateur de Nelson-Aalen de la fonction d'intensité cumulée, voir Planchet et Thérond, 2011.

c. Estimateurs de Kaplan-Meier par cohorte

Dans la pratique courante, les praticiens estiment également les taux de décès par cohorte à partir de l'estimateur de Kaplan-Meier avec une stratification par cohorte. Pour obtenir cette décomposition, nous considérons la sous-population appartenant à l'ensemble $S_C^{x,t} = \left\{i : T_i > x, t-x < \sigma_i \leq t-x+1\right\}$ et estimons l'estimateur de la fonction de répartition Kaplan-Meier de X sur cette sous-population, qui est consistant pour $x+1 \leq \tau_T$.

L'estimateur du taux de décès s'obtient par $\hat{q}_C(x,t) = 1-\hat{S}_{g,x}(1)$ où $\hat{S}_{g,x}(1)$ est l'estimateur de Kaplan-Meier de la fonction de survie conditionnelle (à l'âge *x* pour la génération $g = t-x$) au point 1.

Remarque 3. *L'estimateur de Kaplan-Meier comporte un biais de sous-estimation de la probabilité conditionnelle de sortie. On a* $S(t) \leq E\left(\hat{S}_{KM}(t)\right)$ en particulier*, ce qui donne* $q(x) \geq E\left(\hat{q}_{KM}(x)\right)$*, en notant que* $q(x) = 1-S_x(1)$ *avec* S_x *la fonction de survie conditionnelle à l'âge x.*

Remarque 4. *L'estimateur de Harrington-Flemming de la fonction de survie est supérieur à l'estimateur de Kaplan-Meier, soit* $\hat{S}_{KM}(t) \leq \hat{S}_{HF}(t)$*.On a donc* $\hat{q}_{HF}(x) \leq \hat{q}_{KM}(x)$*. Par ailleurs, du fait du biais de l'estimateur de Nelson-Aalen et de l'inégalité de Jensen, l'estimateur de Harrington-Flemming est biaisé positivement,* $S(t) \leq E\left(\hat{S}_{HF}(t)\right)$*,ce qui donne* $q(x) \geq E\left(\hat{q}_{HF}(x)\right)$*. Au global, on en déduit donc que* $q(x) \geq E\left(\hat{q}_{KM}(x)\right) \geq E\left(\hat{q}_{HF}(x)\right)$*. Ce biais de sous-estimation a pour conséquence l'introduction mécanique d'une certaine prudence dans la construction de tables des mortalité prospectives destinées à être utilisées*

pour évaluer des engagements de régimes de rentes[5].

Remarque 5. *Dans la pratique courante, aucun estimateur des taux de décès par période n'est proposé.*

4. Comparaison des estimateurs des taux de décès

Nous avons vu dans la section précédente que trois estimateurs principaux peuvent être utilisés dans la pratique pour évaluer les taux de décès bruts dans une logique de construction d'une table de mortalité prospective : ceux dits de Hoem par période et par cohorte, vu comme des approximations des taux de décès par période et par cohorte, et celui par cohorte obtenus au moyen de l'estimateur de Kaplan-Meier. Dans cette section, on se propose de comparer ces différentes estimateurs pour mettre en exergue leurs principales différences et leurs conditions d'utilisation. Cette analyse permet ainsi de clarifier le rôle de ces estimateurs, qui n'est pas toujours bien appréhendé par les praticiens.

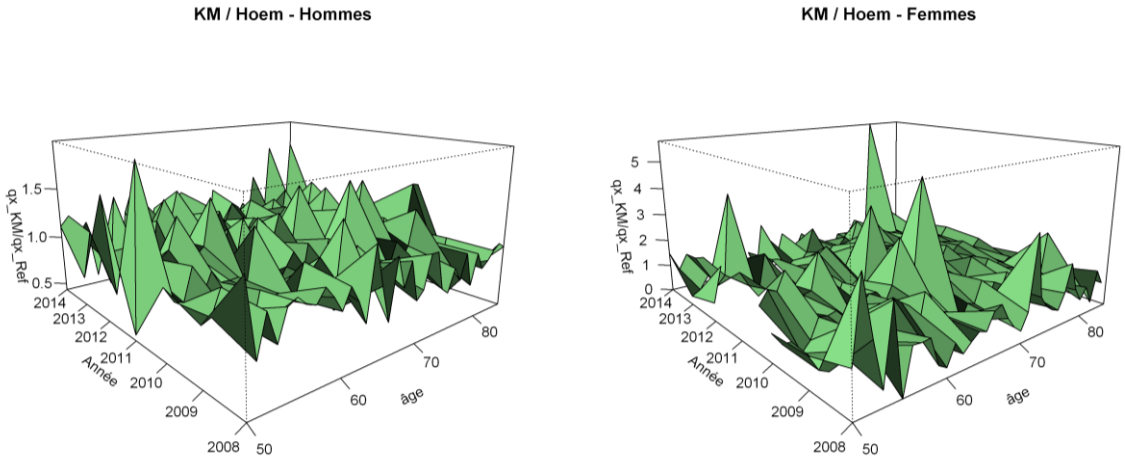
a. Comparaison empirique des logiques par cohorte et par période sur données d'assurance

Bien qu'il s'agisse de quantité proche en apparence, les logiques par période et par cohorte aboutissent à des estimations des taux bruts qui divergent de manière importante, en particulier pour de petites populations. Elles répondent cependant à un cadre de modélisation différent, bien qu'il soit délicat d'identifier une pratique claire en ce qui concerne l'utilisation de tables prospectives pour le calcul de tarif ou en matière de provisionnement. En effet, si l'approche par période est généralement reprise pour la construction de modèle de mortalité stochastique, les deux approches subsistent, et sont parfois utilisées sans précaution pour élaborer des tables prospectives. Notons en particulier, que les méthodes de construction par positionnement (Planchet et Tomas, 2014), très fréquentes pour de petits des portefeuilles d'assurance de petites tailles, requièrent l'utilisation d'une table de référence, dont il est important de savoir si elle a été élaborée selon une logique par période ou bien une logique par cohorte, afin de choisir l'estimateur approprié des taux de décès brut.

Voyons au travers de données réelles, observées sur la période 2008-2014 où sont distingués les hommes et les femmes, l'impact de la référence de taux bruts. La Fig. 2 : présente le rapport entre les estimations de Hoem par période et celles issues de Kaplan-Meier et permet d'illustrer l'effet d'une utilisation inappropriée de l'estimateur de Kaplan-Meier pour l'élaboration de table par période. Le ratio des deux quantités, quoique globalement centré sur 1, visiblement est très erratique.

[5] On peut noter toutefois que les modèles d'ajustement utilisés dans le contexte de la construction de tables pour des portefeuilles d'assurance, typiquement le modèle de Brass, induisent un biais (de sous-estimation pour les probabilités inférieures à ½) sur la valeur ajustée des probabilités conditionnelles de décès (cf. Planchet et Thérond, 2011).

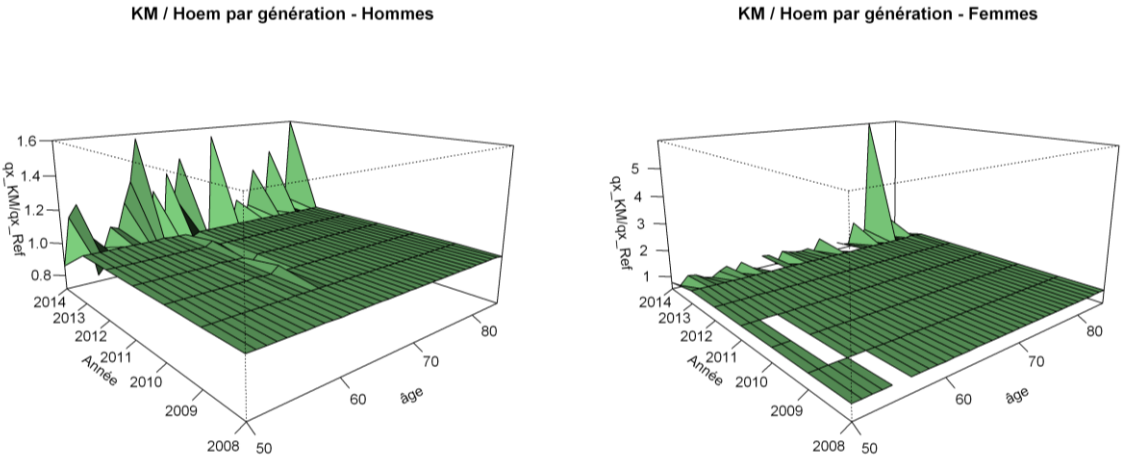
Fig. 2 : Rapport des estimateurs de Hoem « du moment » et de Kaplan-Meier pour une population masculine et une population féminine sur la période 2008-2014.



Deux sources écarts sont susceptibles d'expliquer ces variations : les différentes approximations introduites pour transformer l'estimateur de Hoem en un estimateur des taux de décès, et la règle de découpage, par période pour l'approche de Hoem et par cohorte pour l'approche s'appuyant sur l'estimateur de Kaplan-Meier. L'analyse de ces deux effets potentiels est complexifiée par le fait que l'estimateur de Hoem, calculé à chaque âge *x* et pour l'année *t*, fait intervenir les membres des deux générations $t-x$ et $t-x-1$, alors que le second type d'estimateur est calculé par génération, puis réordonné en utilisant $g=t-x$.

C'est cependant très clairement la règle de découpage qui est la source majeure de ces écarts. L'utilisation de l'estimateur de Hoem par cohorte permet en effet de réduire très sensiblement l'écart avec l'estimateur de Kaplan-Meier stratifié par cohorte comme l'illustre la figure suivante.

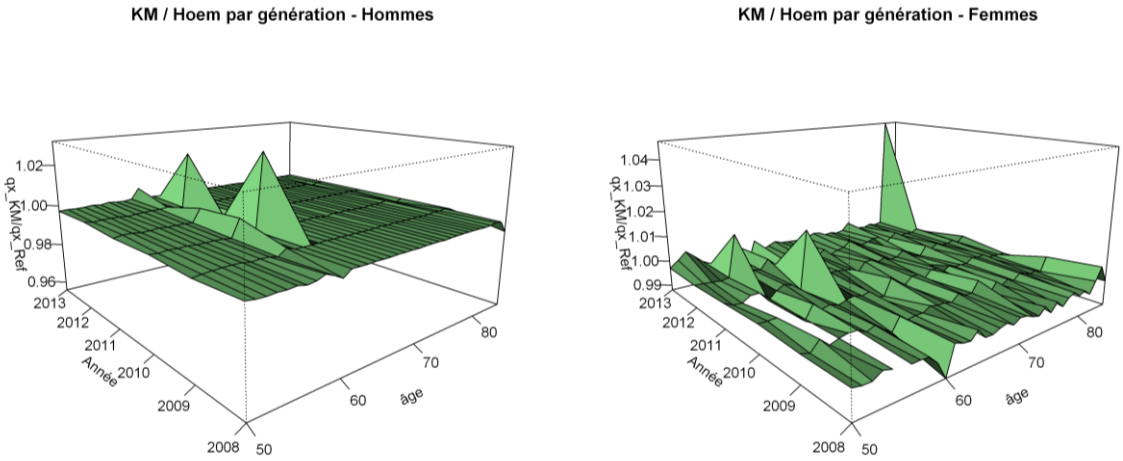
Fig. 3 : Rapport des estimateurs de Hoem par cohorte et de Kaplan-Meier pour une population masculine et une population féminine sur la période 2008-2014.



En supprimant la dernière année d'observation, pour laquelle les écarts sont plus

significatifs, on observe la très bonne cohérence entre les deux estimateurs, comme il est possible de le voir sur la Fig. 4 :.

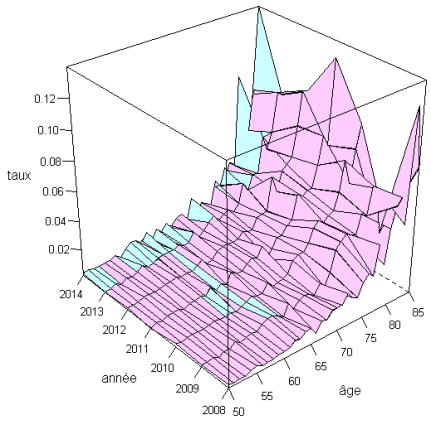
Fig. 4 : **Rapport des estimateurs de Hoem par cohorte et de Kaplan-Meier pour une population masculine et une population féminine sur la période 2008-2013.**



L'écart plus important pour la dernière année de la période d'observation $t_M = 2014$ est la conséquence d'un effet de bord qui conduit à réarranger les estimations par année du moment à l'aide de $g = t - x$, et pour laquelle un seul des deux termes de la somme est utilisé pour $t = t_M$.

Par ailleurs, les écarts entre l'approche de Hoem par cohorte et celle s'appuyant sur l'estimateur de Kaplan-Meier présentent un biais pour ces données dont l'allure est représentée par la Fig. 5 :.

Fig. 5 : **Représentation des estimateurs de taux de décès de la population masculine de Hoem par cohorte (en bleu) et obenus par Kaplan-Meier (rose) sur la période 2008-2014.**

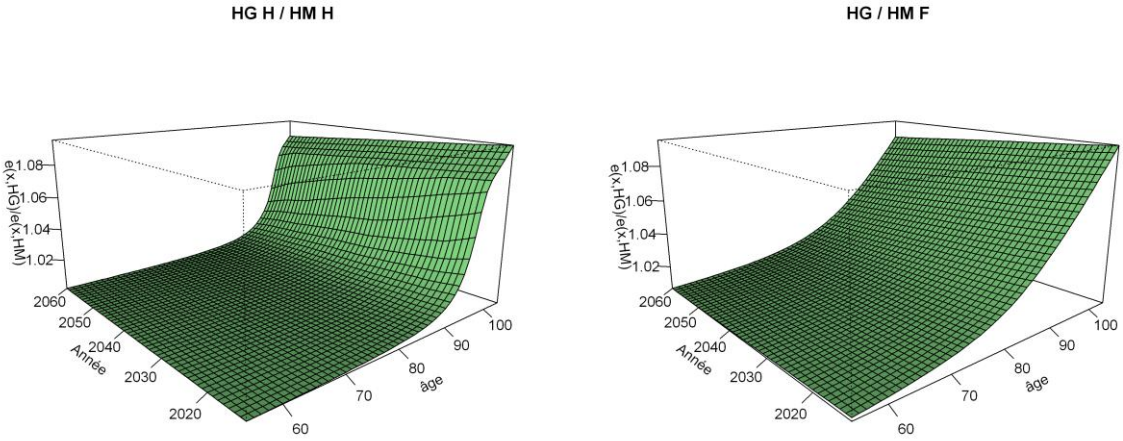


Il apparaît que l'estimateur obtenu par Kaplan-Meier est plus grand que l'estimateur de Hoem par génération sur les carrés pour lesquelles les deux estimateurs sont proches. À l'inverse, le signe de l'écart est opposé lorsque l'écart est plus important en valeur absolue.

Les sources d'écarts résiduelles entre les deux approches proviennent des approximations réalisées au travers de l'approche décrite dans la section 3.b. Ces approximations n'étant pas spécifiques au fait de travailler en dimension deux, nous examinons dans la section suivante les biais qui subsistent en dimension un.

Pour clôturer cette analyse, on construit un ajustement sur la base de chacun des deux estimateurs, « Hoem du moment » et « Hoem par génération » en utilisant l'approche de Planchet et Tomas, 2014 avec comme références[6] les tables périodiques IA 2013, lissées avec un modèle de vraisemblance locale. On compare à la Fig. 6 : les espérances de vie prospectives issues de ces deux ajustements.

Fig. 6 : <u>Rapport des espérances de vie prospectives résiduelles par période et par cohorte, calculées après ajustements sur les taux décès bruts selon le genre.</u>



On observe que l'écart entre les deux ajustements est faible dans les zones de faible volatilité des estimateurs sous-jacents, c'est-à-dire pour des tranches situées avant 80 ans où les observations sont abondantes et augmente rapidement lorsque celle-ci s'accroît. Pour les âges élevés, les écarts atteignent des niveaux très significatifs dans la perspective de l'utilisation des tables pour provisionner des rentes. De plus, le biais associé à l'utilisation indue ici des estimateurs par cohorte induit une surestimation, dans cet exemple, des espérances de vie résiduelles, ce qui peut s'avérer pénalisant pour l'assureur selon le cadre d'utilisation.

b. Analyse d'écarts dans un cadre unidimensionnel : une approche sur données simulées

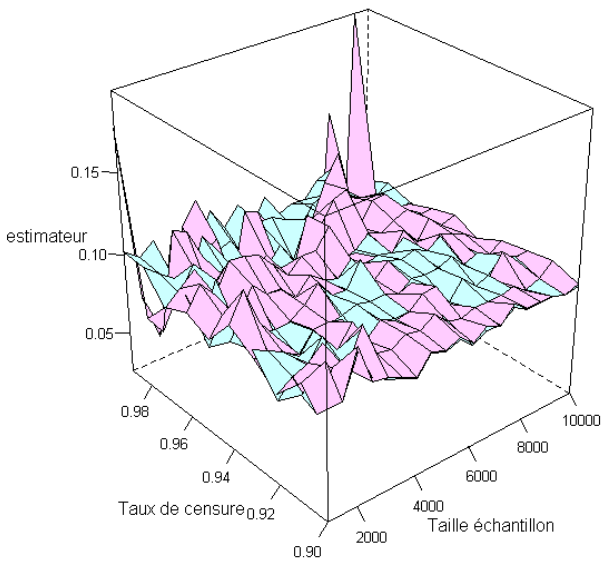
Les écarts entre les approches de Hoem et de Kaplan-Meier ne sont pas spécifiques à une analyse des taux de décès en dimension deux. Les écarts peuvent donc être examinés sans perte de généralité en dimension un. Le choix de l'estimateur utilisé pour les taux bruts répond à une logique de type biais-variance. En effet, l'estimateur de Hoem, paramétrique, comprend un biais qui peut être plus important car il dépend du choix du pas de

[6] Voir http://www.ressources-actuarielles.net/gtmortalite.

discrétisation utilisé (annuel en pratique pour la mortalité) et du respect sur chaque carré du diagramme Lexis de l'hypothèse de constance des taux de hasard. En revanche, lorsque cette dernière est satisfaite, on peut s'attendre à ce que sa variance soit plus faible, puisqu'il s'agit d'un estimateur paramétrique.

Afin de poursuivre cette comparaison des estimateurs, on utilise une approche numérique en dimension un. On se place sans perte de généralité dans le cas de l'intervalle $[x, x+1[= [0,1[$, et on suppose que X (resp. C) suit une loi exponentielle de paramètre μ_X (resp. μ_C). Le taux de censure (sur $[0,+\infty[$) dans cette expérience est noté $\beta = P(X > C) = \frac{\mu_C}{\mu_X + \mu_C}$. Pour l'illustration numérique, on prend $\mu_X = 1$ et $\mu_C = \frac{\beta}{1-\beta}$ et on fait varier le taux de censure entre 0 et 1. Dans cette expérience, l'hypothèse de constance par morceaux du taux de hasard est naturellement satisfaite, ce qui conduite à contrôler fortement le biais produit par l'approche de Hoem et à le limiter aux seules approximations décrites *supra* conduisant à l'assimiler à un taux de décès.On observe ensuite le comportement des estimateurs obtenus par les approches de Hoem et de Kaplan-Meier pour différentes valeurs de μ_X en fonction du taux de censure et de la taille de l'échantillon. La figure suivante se focalise sur la région où les écarts les plus importants apparaissent, à savoir pour des taux de censure supérieurs à 90 %.

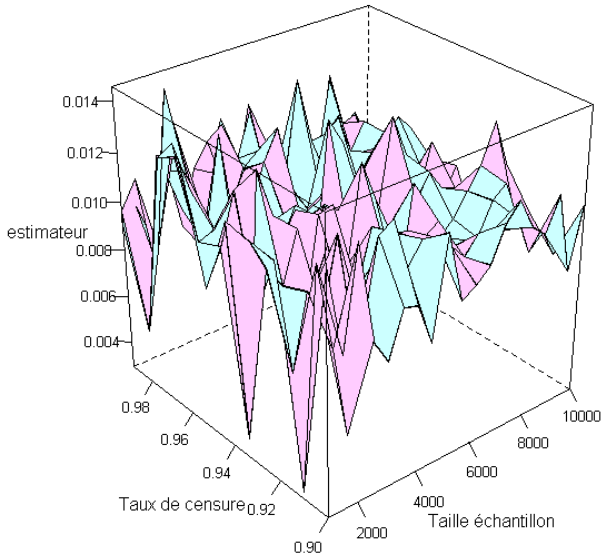
Fig. 7 : **Comparaison des estimateurs de Hoem (en bleu) et Kaplan Meier (en rose) pour μ_X = 1.**



On remarque que les deux estimateurs sont assez proches l'un de l'autre. L'estimateur de Kaplan-Meier est plus volatile, ce qui se voit plus particulièrement dans les zones de très forte censure. L'utilisation d'une probabilité de sortie plus faible conduit qualitativement à la même conclusion, avec une volatilité globale plus importante, comme l'illustre la figure

suivante.

Fig. 8 : **Comparaison des estimateurs de Hoem (en bleu) et Kaplan Meier (en rose) pour μ_X = 0,01.**



Pour conclure, il ressort de ces illustrations numériques que l'estimateur des taux de décès obtenu *via* Kaplan-Meier est plus volatile que l'estimateur de Hoem, ce dernier étant donc en ce sens plus robuste, à condition que l'hypothèse de constance par morceaux soit satisfaite, même pour des niveaux de censure élevés.

5. Extension du cadre d'estimation non-paramétrique par période et par cohorte

On se propose dans cette section de présenter une méthodologie d'estimation non-paramétriques des taux de décès selon les logiques période et par cohorte, permettant de disposer des estimateurs présentés ci-dessus dans un cadre unifié.

Pour cela, nous réécrivons les probabilités de décès comme l'espérance de transformations simples de (σ, X), en suivant l'idée développée par Meira-Machado, de Uña-Álvarez et Cadarso-Suárez, 2006. On a avec $x \le t$,

$$q_P(x,t) = \frac{E\left[I\left(x < X \le x+1, t < X+\sigma \le t+1\right)\right]}{E\left[I\left(X > x, X+\sigma > t\right)\right]},$$

et

$$q_C(x,t) = \frac{E\left[I\left(x < X \le x+1, t-x < \sigma \le t-x+1\right)\right]}{E\left[I\left(X > x, t-x < \sigma \le t-x+1\right)\right]}.$$

avec $I(\cdot)$, la fonction indicatrice. À partir de cette construction et en exploitant des résultats obtenus sur les intégrales Kaplan-Meier, on obtient naturellement des estimateurs non-

paramétriques de ces probabilités

$$q_P\left(x,t\right)=\frac{\sum_{i=1}^{n}W_{in}I\left(x<T_{i:n}\leq x+1,t<T_{i:n}+\sigma_{[i:n]}\leq t+1\right)}{\sum_{i=1}^{n}W_{in}I\left(T_{i:n}>x,T_{i:n}+\sigma_{[i:n]}>t\right)},$$

et

$$q_C\left(x,t\right)=\frac{\sum_{i=1}^{n}W_{in}I\left(x<T_{i:n}\leq x+1,t-x<\sigma_{[i:n]}\leq t-x+1\right)}{\sum_{i=1}^{n}W_{in}I\left(T_{i:n}>x,t-x<\sigma_{[i:n]}\leq t-x+1\right)},$$

où $T_{1:n}\leq T_{2:n}\leq\ldots\leq T_{n:n}$ correspondent aux valeurs ordonnées de T et $\left(D_{[i:n]},\sigma_{[i:n]}\right)$ correspond aux valeurs concomitantes des variables D et σ. Le poids associé à la i-ème observation s'écrit

$$W_{in}=\frac{D_{[i:n]}}{n-i+1}\prod_{j=1}^{i-1}\left(\frac{n-j}{n-j+1}\right)^{D_{[j:n]}},$$

et correspond au poids de Kaplan Meier relatif à la variable de durée X. Dans ces équations, la variable σ, non soumise à censure, joue le même rôle qu'une covariable, ce qu'il explique qu'elle n'intervienne pas dans l'expression des poids W_{in}. Notons que ces derniers valent $W_{in}=n^{-1}$ en l'absence de censure. Ce type de raisonnement est également exploité par Guibert et Planchet, 2016 pour la construction d'estimateurs non-paramétriques de probabilités de transition pour des modèles multi-états non-markoviens.

Ces estimateurs satisfont aux conditions de Stute, 1993, 1996 et présentent par conséquent des propriétés de convergence et de normalité asymptotique dans des conditions que nous discutons. On note τ_A, la borne supérieure du support de la fonction de répartition associée à une variable aléatoire A. La convergence des estimateurs est acquise lorsque $\tau_X\leq\tau_C$. Cependant, ils sont systématiques biaisés si ce n'est pas le cas. En suivant le raisonnement de Uña-Álvarez et Meira-Machado, 2015, deux alternatives simples pour corriger ces estimateurs seraient envisageables. Toutefois, afin de faire correspondre nos estimateurs avec ceux utilisés par les praticiens, nous choisissons de ne considérer que l'une d'elles, obtenue en sélectionnant des sous-population d'individus.

Introduisons l'ensemble $S_P^{x,t}=\left\{i:T_i>x,T_i+\sigma_i>t\right\}$, de cardinal $n_P^{x,t}$. L'estimateur des probabilités de décès selon une décomposition par période peut être redéfinie par

$$q_P\left(x,t\right)=\sum_{i=1}^{n_P^{x,t}}W_{in_P^{x,t}}I\left(T_{i:n_P^{x,t}}\le x+1,T_{i:n_P^{x,t}}+\sigma_{\left[i:n_P^{x,t}\right]}\le t+1\right),$$

avec $W_{in_P^{x,t}}$ i-ème poids Kaplan-Meier de la fonction de survie de X, estimé sur la sous-population $S_P^{x,t}$. Cet estimateur est convergent pour $x+1\le\tau_T$ et $t+1\le\tau_Z-\tau_\sigma$, et correspond à un estimateur de Kaplan-Meier multivarié.

S'agissant de la décomposition par cohorte, nous considérons l'ensemble $S_C^{x,t}=\left\{i:T_i>x,t-x<\sigma_i\le t-x+1\right\}$, de cardinal $n_C^{x,t}$, et l'estimateur

$$\hat{q}_C\left(x,t\right)=\sum_{i=1}^{n_C^{x,t}}W_{in_P^{x,t}}I\left(T_{i:n_C^{x,t}}\le x+1\right),$$

avec $W_{in_c^{x,t}}$ i-ème poids Kaplan-Meier de la fonction de survie de X, estimé sur la sous-population $S_C^{x,t}$. Cette quantité correspond exactement à l'estimateur de la fonction de répartition Kaplan-Meier de X sur cette sous-population, qui est consistant pour $x+1\le\tau_T$. Cet estimateur est le plus souvent calculé $\hat{q}_C\left(x,t\right)=1-\hat{S}_{g,x}\left(1\right)$ où $\hat{S}_{g,x}\left(1\right)$ est l'estimateur de Kaplan-Meier de la fonction de survie conditionnelle (à l'âge *x* pour la génération $g=t-x$) au point 1.

Remarque 6. *L'estimateur $\hat{q}_C\left(x,t\right)$ correspond à celui usuellement utilisé par les praticiens. La construction rencontrée en pratique consiste à ne filtrer que la population appartenant à la cohorte née en $t-x$. L'estimateur obtenu peut varier alors légèrement en fonction du choix effectué pour la règle d'arrondi des valeurs non entières. On utilise ici systématiquement la partie entière.*

Remarque 7. *Le calcul de l'estimateur $\hat{q}_C\left(x,t\right)$ est plus accessible aux praticiens, car $q_P\left(x,t\right)$* requiert préalablement de calculer des poids Kaplan-Meier alors que $\hat{q}_C\left(x,t\right)$ *est obtenu directement à l'aide des logiciels statistiques usuels en utilisant la relation* $\hat{q}_C\left(x,t\right)=1-\hat{S}_{g,x}\left(1\right)$, sans calcul explicite des poids Kaplan-Meier.

6. Conclusion et discussion

La présente étude se focalise sur le choix de l'estimateur des taux de décès brut à retenir pour la construction de tables de mortalité prospectives ou pour l'estimation de modèles de mortalité stochastiques. Lorsque des taux périodiques sont modélisés, l'étude insiste sur le fait d'utiliser l'estimateur de Hoem par période, plutôt que celui par cohorte ou l'estimateur de Kaplan-Meier des taux de décès, sous peine d'introduire un biais important. À l'inverse, elle montre que dans les études par cohorte, le choix d'un estimateur de type Hoem ou

Kaplan-Meier a peu d'impact.

Ce papier propose enfin plusieurs variantes bi-dimensionnelle de l'estimateur de Kaplan-Meier par cohorte, et en particulier introduit une version par période, qui a notre connaissance, n'est pas utilisé dans la littérature actuarielle ou en biostatistique. Ainsi, une piste de réflexion serait de développer des approches de construction de tables s'appuyant sur ces estimateurs.

7. Références

Ahcan A., Medved D., Olivieri A. et Pitacco E. (2014), « Forecasting mortality for small populations by mixing mortality data », Insurance: Mathematics and Economics, vol. 54, pp. 12-27.

Barrieu P., Bensusan H., El Karoui N., Hillairet C., Loisel S., Ravanelli C. et Salhi Y. (2012), « Understanding, modelling and managing longevity risk: key issues and main challenges », Scandinavian Actuarial Journal, vol. 2012, n°3, pp. 203-231.

Booth H. et Tickle L. (2008), « Mortality Modelling and Forecasting: a Review of Methods », Annals of Actuarial Science, vol. 3, n°1-2, pp. 3–43.

Boumezoued A. (2016), « Improving HMD mortality estimates with HFD fertility data »,.

Brillinger D.R. (1986), « A Biometrics Invited Paper with Discussion: The Natural Variability of Vital Rates and Associated Statistics », Biometrics, vol. 42, n°4, pp. 693-734.

Cairns A.J.G., Blake D. et Dowd K. (2006), « A Two-Factor Model for Stochastic Mortality with Parameter Uncertainty: Theory and Calibration », Journal of Risk and Insurance, vol. 73, n°4, pp. 687-718.

Cairns A.J.G., Blake D. et Dowd K. (2008), « Modelling and management of mortality risk: a review », Scandinavian Actuarial Journal, vol. 2008, n°2-3, pp. 79-113.

Cairns A.J.G., Blake D., Dowd K. et Kessler A. (2015), « Phantoms never die: Living with unreliable mortality data », Discussion Paper, n°PI-1410, The Pensions Institute, Cass Business School.

Delwarde A. et Denuit M. (2005), Construction de tables de mortalité périodiques et prospectives, Paris, Economica (Assurance Audit Actuariat), 428 p.

Guibert Q. et Planchet F. (2016), « Non-Parametric Inference of Transition Probabilities Based on Aalen-Johansen Integral Estimators for Acyclic Multi-State Models: Application to LTC Insurance », Working paper.

Guilloux A. (2007), « Nonparametric estimation for censored lifetimes suffering from unknown selection bias », Mathematical Methods of Statistics, vol. 16, n°3, pp. 202-216.

Hoem J.M. (1971), « Point Estimation of Forces of Transition in Demographic Models », Journal of the Royal Statistical Society. Series B (Methodological), vol. 33, n°2, pp. 275-289.

Human Mortality Database (2016), University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany).

Kaplan E.L. et Meier P. (1958), « Nonparametric Estimation from Incomplete Observations », Journal of the American Statistical Association, vol. 53, n°282, pp. 457-481.

Keiding N. (1990), « Statistical Inference in the Lexis Diagram », Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, vol. 332, n°1627, pp. 487-509.

Keiding N. (1991), « Age-Specific Incidence and Prevalence: A Statistical Perspective », Journal of the

Royal Statistical Society. Series A (Statistics in Society), vol. 154, n°3, pp. 371-412.

Lee R.D. et Carter L.R. (1992), « Modeling and Forecasting U. S. Mortality », Journal of the American Statistical Association, vol. 87, n°419, pp. 659-671.

Lund J. (2000), « Sampling Bias in Population Studies—How to Use the Lexis Diagram », Scandinavian Journal of Statistics, vol. 27, n°4, pp. 589-604.

Mandel M. (2007), « Nonparametric Estimation of a Distribution Function under Biased Sampling and Censoring », Lecture Notes-Monograph Series, vol. 54, pp. 224-238.

Meira-Machado L., Uña-Álvarez J. de et Cadarso-Suárez C. (2006), « Nonparametric estimation of transition probabilities in a non-Markov illness–death model », Lifetime Data Analysis, vol. 12, n°3, pp. 325-344.

Planchet F. et Thérond P.-E. (2011), Modélisation statistique des phénomènes de durée - Applications actuarielles, 2e édition, Paris, Economica (Assurance Audit Actuariat).

Stute W. (1993), « Consistent Estimation Under Random Censorship When Covariables Are Present », Journal of Multivariate Analysis, vol. 45, n°1, pp. 89-103.

Stute W. (1996), « Distributional convergence under random censorship when covariables are present », Scandinavian journal of statistics, vol. 23, n°4, pp. 461-471.

Tomas J. et Planchet F. (2014), « Constructing entity specific projected mortality table: adjustment to a reference », European Actuarial Journal, vol. 4, n°2, pp. 247-279.

Uña-Álvarez J. de et Meira-Machado L. (2015), « Nonparametric estimation of transition probabilities in the non-Markov illness-death model: A comparative study », Biometrics.