A NEW INFERENCE STRATEGY FOR GENERAL POPULATION MORTALITY TABLES

BY

ALEXANDRE BOUMEZOUED, MARC HOFFMANN AND PAULIEN JEUNESSE

ABSTRACT

We propose a new inference strategy for general population mortality tables based on annual population and death estimates, completed by monthly birth counts. We rely on a deterministic population dynamics model and establish formulas that link the death rates to be estimated with the observables at hand. The inference algorithm takes the form of a recursive and implicit scheme for computing death rate estimates. This paper demonstrates both theoretically and numerically the efficiency of using additional monthly birth counts for appropriately computing annual mortality tables. As a main result, the improved mortality estimators show better features, including the fact that previous anomalies in the form of isolated cohort effects disappear, which confirms from a mathematical perspective the previous contributions by Richards, Cairns *et al*., and Boumezoued.

KEYWORDS

Mortality tables, general population, death rate inference, population dynamics, cohort effect.

JEL codes: C020.
MSC (2010): 92D25, 62P05, 62N02.

1. INTRODUCTION

General population mortality tables are crucial inputs for actuarial studies as they provide estimates of mortality rates for several age classes at several periods in time. Since the publication of the first mortality tables (attributed to John Graunt in 1662), the mathematical problem of providing consistent statistical estimates of mortality has fascinated mathematicians – for a brief history, the reader is referred to the well-documented dedicated part of the introduction of Daley and Vere-Jones (2003). Two centuries later, there

Astin Bulletin 50(2), 325-356. doi:10.1017/asb.2020.5

was a huge development of graphical formalizations of life trajectories within a population by Lexis (1875) and his contemporaries. These first demographers showed that it is crucial to address simultaneously two components: (1) consider the fact that the death rate depends on both age and time (nonhomogeneous setting) and (2) understand the mortality rate as an aggregate quantity which depends on an underlying population dynamics.

Recently, several papers and publications paid attention to data quality issues in the way we usually build mortality tables, especially in relation to the "discrete time" nature of population estimates provided by national censuses. To our knowledge, the first insights have been suggested by Richards (2008); his conjecture was focused on the 1919 birth cohort for England & Wales, for which he suggested that errors occurred in the computation of mortality rates due to shocks in the birth series. The ONS methodology has then been studied by Cairns *et al*. (2016) in several directions, who confirmed the conjecture by Richards (2008) and proposed an approach to illustrate and correct mortality tables, applied to the data for England & Wales; the *Convexity Adjustment Ratio* introduced in their work has then been adapted by Boumezoued (2020) who focused on the Human Mortality Database, HMD (2018) – which provides mortality tables for more than 30 countries and regions worldwide – and showed that these anomalies are universal while using the "population dynamics" point of view to properly define mortality estimates. To build new mortality tables for several countries, a link with the Human Fertility Database (HFD 2018, the HMD counterpart for fertility) has been made to correct such errors in a systematic way.

However, all precedent contributions did not succeed to introduce a proper mathematical setting for computing mortality rates based on information extracted from censuses. In this paper, we aim at performing a first step in this direction by deriving an inference strategy from a deterministic population dynamics model. The derivation of a consistent theory in the stochastic setting is in parallel provided in a companion theoretical paper, see Boumezoued *et al*. (2018).

The main difficulty in establishing a consistent theory to estimate mortality rates lies in points (1) and (2) mentioned above, which can be summarized as follows: inferring an age and time-dependent mortality rate based on a population dynamics model. In the literature, we argue that each point is treated separately.

The inference of a time-dependent death rate also depending on a time-dependent covariate (possibly age), which relates to point (1), has been addressed from a nonparametric perspective by Beran (1981), Dabrowska (1987), Keiding (1990), McKeague and Utikal (1990), Nielsen and Linton (1995), Brunel *et al*. (2008), and Comte *et al*. (2011). From Keiding (1990), *"One way of understanding the difficulties in establishing an Aalen theory in the Lexis diagram is that although the diagram is two-dimensional, all movements are in the same direction (slope 1) and in the fully non-parametric model the diagram disintegrates into a continuum of life lines of slope 1 with freely varying intensities*

https://doi.org/10.1017/asb.2020.5 Published online by Cambridge University Press

across lines. The cumulation trick from Aalen's estimator (generalizing ordinary empirical distribution functions and Kaplan & Meier's (1958) non-parametric empirical distribution function from censored data) does not help us here." This explains why data aggregation and smoothing is required to derive an estimate with two crossing dimensions, age and time.

On the other side, the inference of an age-dependent death rate in an homogeneous birth-death model (or similar) – point (2) – has been addressed by Clémençon et al. (2008), Doumic *et al.* (2015), and Hoffmann and Olivier (2016). To our knowledge, no statistical method deals with the usual problem faced by demographers related to the construction of a mortality table based on population estimates and death counts.

In this paper, we rely on a deterministic age-structured population model and derive exact formulas in the so-called Lexis diagram, allowing to build new and improved mortality estimates. The inference problem is summarized as follows:

- The death rate depends on both age and time and is to be estimated,
- The population evolves as an age-structured and time inhomogeneous birth–death dynamics,
- The following observables are available in the Lexis diagram:
  - The number of individuals in each 1-year age-class assumed to be recorded at each beginning of year,
  - The number of deaths in annual Lexis triangles,
  - The number of births available each month (or more generally at some intra-year frequency).

Note that the practical availability of annual population estimates as well as death counts in the Lexis triangle can be achieved according to the HMD, whereas the HFD is a public source providing in a particular number of births by month for several countries. Such population, death, and fertility data allow at this date the method proposed in this paper to be applied to around 10 countries. For other countries, the data (especially number of births by month) have to be reached by means of national institutes.

The paper is organized as follows. In Section 2, we present the nonhomogeneous birth–death model and derive the inference strategy – the related interpretations and link with existing estimators is discussed in Section 2.6. In Section 3, we compute mortality tables according to our method and compare it to those obtained by the usual formulas. The paper ends with some concluding remarks in Section 4.

2. Model and Inference Strategy

2.1. Nonhomogeneous birth–death dynamics

Let us denote by $\mu(a, t)$ the mortality rate at exact age $a \in \mathbb{R}_+ = [0, \infty)$ and exact time $t \in \mathbb{R}_+$, with an arbitrary time origin – let us also denote by $g(a, t)$

https://doi.org/10.1017/asb.2020.5 Published online by Cambridge University Press

the population density at (a, t), a nonnegative real value. In its core definition, the death rate drives the number of living in a closed population. Formally, consider $g(0, \nu)$ the newborn at (exact) time ν (starting number in the cohort born at time ν), then the survivors at some age $a > 0$ in the cohort write

$$g(a, \nu + a) = g(0, \nu) \exp\left(- \int_0^a \mu(s, \nu + s) \mathrm{d}s \right).$$

Changing variables to represent $g(a, t)$, and differentiating by age and time, leads to the transport component of the so-called McKendrick–Von Foerster equation (see McKendrick (1926) and Von Foerster (1959)):

$$(\partial_a + \partial_t) g(a, t) = -\mu(a, t) g(a, t), \tag{1}$$

with notation $\partial_a \equiv \partial / \partial a$. Clearly, at this stage, the population dynamics of $g(a, t)$ is not fully specified as the future path of $g(a, t)$ depends on the quantity $g(0, t - a)$. The McKendrick–Von Foerster equation specifies how births are given in the (asexual) population, based on a birth rate $b(a, t)$, as

$$\text{for each time } \nu > 0, \quad g(0, \nu) = \int_0^\infty g(a, \nu) b(a, \nu) \mathrm{d}a.$$

That is simply, the newborn at each time is given by the total number of births from all parents alive at the same time.

2.2. Observables in the Lexis diagram

We work here in the Lexis diagram – that is, we study lifelines in the time × age coordinates. In an ideal demographic world, two kinds of population estimates are recorded in the 1-year age × time square:

- Population at exact time t, with age x at its last birthday:

$$P(x, t) = \int_x^{x+1} g(a, t) \mathrm{d}a. \tag{2}$$

- Individuals who attained exact age x during the year $[t, t + 1)$:

$$N(x, t) = \int_t^{t+1} g(x, s) \mathrm{d}s.$$

An illustration of population estimates $P(x, t)$ for the French population extracted form the HMD is given in Figure 1. This can be analyzed in the light of a Lexis diagram in several directions. First, the diagonal effects appear clearly showing that generations (or cohorts) are not equally represented: as an example, the generations born between around 1915 and 1920 are less represented (World War I), whereas the generations born after around 1946 are highly represented (Baby Boom). In this work, the impact of the discrepancy

https://doi.org/10.1017/asb.2020.5 Published online by Cambridge University Press

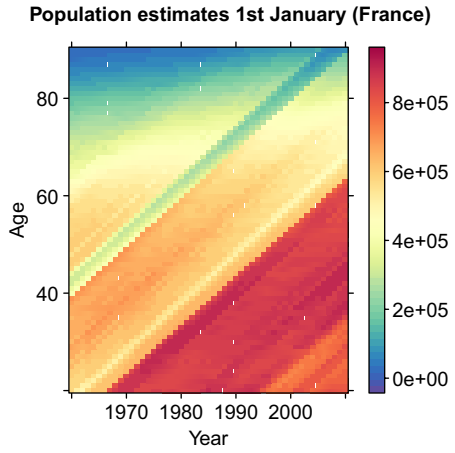

FIGURE 1: Population estimates for France by year for 1-year age classes extracted from the Human Mortality Database.

between birth patterns from 1 year to the next is of interest, as it introduces some bias in the classical formulas used in practice for death rate estimation.

Also, death counts are provided on the upper and lower triangles of the Lexis diagram, as defined below.

Definition 1. *The upper (U) and lower (L) triangles for each age range x and observation year t are the age × time sets defined by*

$$T_U(x,t) = \{(a,s) : a \in [x, x+1) \text{ and } s \in [t, t-x+a)\}, \tag{3}$$

and

$$T_L(x,t) = \{(a,s) : a \in [x, x+1) \text{ and } s \in [t-x+a, t+1)\}. \tag{4}$$

Based on this definition, the number of deaths in the Lexis triangles can be written as

$$D_U(x,t) = \iint_{T_U(x,t)} \mu(a,s)g(a,s)\mathrm{d}a\mathrm{d}s \text{ and } D_L(x,t) = \iint_{T_L(x,t)} \mu(a,s)g(a,s)\mathrm{d}a\mathrm{d}s. \tag{5}$$

An illustration of death counts in the Lexis triangles (x, t) for the French population extracted from the HMD is represented in Figure 2. Variations in number of deaths are closely linked not only to those of the underlying exposure (Figure 1) but also to the death rate itself, which is to be estimated.

Assuming that the population is closed, the following fundamental relations apply (which can be proved by integration by parts):

$$\begin{aligned} N(x+1,t) &= P(x,t) - D_U(x,t), \\ P(x,t+1) &= N(x,t) - D_L(x,t). \end{aligned} \tag{6}$$

https://doi.org/10.1017/asb.2020.5 Published online by Cambridge University Press

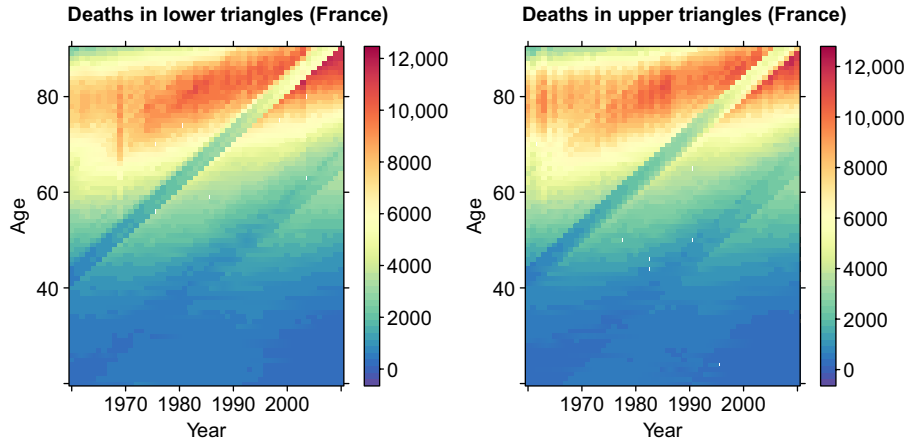

FIGURE 2: Death counts in Lexis triangles extracted from the Human Mortality Database.

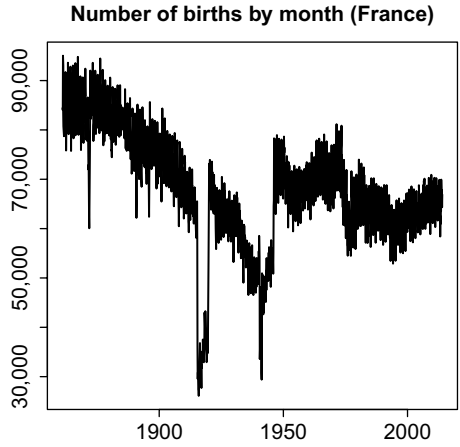

FIGURE 3: Number of birth by month extracted from the Human Fertility Database.

The assumption of closed-population is further discussed in Section 2.6.

In addition to population estimates and death counts, we aim at including birth counts by month in the inference process, as also used by Cairns *et al.* (2016) and Boumezoued (2020) in their estimation procedures – these can be extracted from the HFD for a variety of countries. The dynamics of number of births by month in France is illustrated in Figure 3. The interpretation of this dynamics can be linked to that of Figures 1 (population estimates, see (2)) and 2 (death counts in Lexis triangles, as defined in (5)). Indeed, a similar information arises as the number of births are low in the period 1915–1920, which explains in particular the diagonal effect in Figure 1. Even more importantly, the dynamics at the monthly scale gives insight on what happens inside each year, then can be used to assess how the population is distributed inside

https://doi.org/10.1017/asb.2020.5 Published online by Cambridge University Press

a given age band. This is of great interest as the population distribution appears classically in the form of an "exposure-to-risk", and more precisely the formulas we exhibit in order to estimate the death rate rely explicitly on the births distribution – as such, number of births by month are the key inputs for the inference strategy proposed here as it refines standard annual estimates. This is developed in the following.

2.3. Death rate inference

When two time-dependent dimensions are involved (here age and calendar time), the natural generalization of classical nonparametric estimates of the death rate is not direct (see again the discussion in Keiding (1990)); therefore, smoothing is required – see, for example, McKeague and Utikal (1990) and Nielsen and Linton (1995) for the analysis of such two-dimensional kernel estimator based on continuous observation. Unfortunately, for building national mortality tables, one does not observe continuously the living population (only possibly the date of death through death certificates); therefore, standard kernel smoothing techniques are neither applicable here. This leads to define some geometry on which the death rate is assumed to be piecewise constant, which allows us to use aggregate information by year and age-class to derive (approximate) estimators.

In the classical demographic and actuarial practice, two versions of general population mortality tables are considered: period and cohort. We propose here a brief discussion of these two versions and refer the reader to Boumezoued (2020) for more details (and a study dedicated to period mortality tables). The two versions are illustrated in Figure 4.

- The period table provides death rate estimates based on the assumption that it is piecewise constant on squares in the Lexis diagram; each square (x, t) is equal to the region $T_U(x, t) \cup T_L(x, t)$, where the Lexis triangles T_U and T_L have been defined in Equations (3) and (4). The key advantage of period tables is that they provide an estimate of death rate by using information of a single year; the related drawback is that two generations (cohorts) are merged for a given death rate at (x, t): the lifelines crossing the triangle $T_L(x, t)$ are born in year $t - x$, whereas those crossing $T_U(x, t)$ are born in year $t - x - 1$. This way, the period tables do not strictly reflect the mortality of single cohorts.
- The cohort table is based on the assumption that the death rate is constant on parallelograms $T_L(x, t) \cup T_U(x, t + 1)$, with the advantage that a given death rate at (x, t) relates to lifelines arising from a single cohort: that of people born in year $t - x$. However, the information provided by this death rate reflects conditions of the two consecutive years t and $t + 1$, as illustrated in Figure 4.

Overall, period and cohort tables provide complementary information and their use is driven by the underlying objective. In this paper, we illustrate our

https://doi.org/10.1017/asb.2020.5 Published online by Cambridge University Press

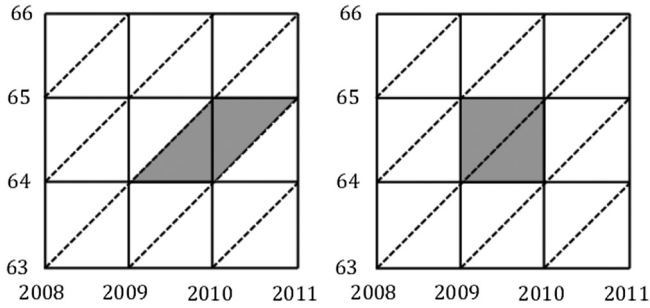

FIGURE 4: Population used (in gray) for the computation of the cohort death rate (left) and period death rate (right) for age 64 and year 2009.

method on the computation of triangle-based mortality tables, which generalize period and cohort mortality tables in a natural way as the death rate is assumed to be piecewise constant on Lexis triangles, instead of squares or parallelograms. This will allow us to draw analyses at a more granular scale compared to the two versions available in practice.

2.4. Main result

In the derivation of the inference formulas, we assume that the death rate is piecewise constant on Lexis triangles:

Assumption 1. *The death rate is piecewise constant on Lexis triangles, that is, for each integer x and t,*

$$\forall (a,s) \in T_L(x,t),\ \mu(a,s) = \mu_L(x,t),$$
$$\forall (a,s) \in T_U(x,t),\ \mu(a,s) = \mu_U(x,t).$$

From the transport component described in Equation (1), for any upper or lower triangle which we denote T, and on which the death rate is constant and equal to μ_T, it follows that:

$$\iint_T (\partial_a + \partial_s) g(a,s) \mathrm{d}a\mathrm{d}s = -\iint_T \mu(a,s) g(a,s) \mathrm{d}a\mathrm{d}s = -\mu_T \iint_T g(a,s) \mathrm{d}a\mathrm{d}s.$$

Note that the middle equation is the negative of the number of deaths (5); therefore, the death rate can be written as the ratio:

$$\mu_L(x,t) = \frac{D_L(x,t)}{E_L(x,t)} \text{ and } \mu_U(x,t) = \frac{D_U(x,t)}{E_U(x,t)},$$

where

$$E_L(x,t) = \iint_{T_L(x,t)} g(a,s) \mathrm{d}a\mathrm{d}s \text{ and } E_U(x,t) = \iint_{T_U(x,t)} g(a,s) \mathrm{d}a\mathrm{d}s,$$

https://doi.org/10.1017/asb.2020.5 Published online by Cambridge University Press

are the so-called "exposures-to-risk" in the lower and upper triangles, respectively.

Now, the number of deaths in Lexis triangles being observed (as provided by the HMD), it remains to appropriately compute the exposure-to-risk. In the literature dedicated to longevity studies, this quantity is approximated by annual observables, see, for example, Brown (1997) and Pitacco *et al.* (2009) Section 2.3.4, as well as the Version 5 Methods Protocol of the Human Mortality Database, see Wilmoth *et al.* (2007). The recent update of the HMD methodology, which allows the inclusion of monthly birth data, is further discussed in Section 2.6. The standard annual approximation can be illustrated for period tables (see Section 2.3) for which the exposure-to-risk writes

$$E(x,t) = \int_t^{t+1} \int_x^{x+1} g(a,s)\mathrm{d}a\mathrm{d}s = \int_t^{t+1} P(x,s)\mathrm{d}s.$$

A possible approximation is therefore given by the trapezoid rule as

$$E(x,t) \approx \frac{1}{2}\left[P(x,t) + P(x,t+1)\right].$$

On the other hand, the exposure-to-risk (period table) can also be written as $E(x,t) = \int_x^{x+1} N(a,t)\mathrm{d}a$ and then approximated by $\frac{1}{2}[N(x,t) + N(x+1,t)] = \frac{1}{2}[P(x,t) + P(x+1,t)] + \frac{1}{2}[D_L(x,t) - D_U(x,t)]$, which leads to another possible approximation. Note that the Version 5 estimates of the HMD rely on a demographic reasoning leading to an approximation in between the two previous ones – see the analysis in Boumezoued (2020) for more details.

Overall, classical approximations have the advantage of being based on observables only, leading to a closed-form for the death rate estimate. The counterpart of this feature is that the validity of the underlying approximation can be put into question for years in which the population curve $s \mapsto P(s,x)$ appears far from linear.

We now detail the recursive and implicit scheme for computing death rate estimates, based on equations linking the death rate with the observables in the Lexis diagram introduced in Section 2.2. Before stating the main result, we introduce two key quantities: first, the Laplace transform of the random variable "date of birth in year y," introduced as:

$$L_y(\theta) = \frac{\int_0^1 g(0,y+v)\exp(-\theta v)\mathrm{d}v}{\int_0^1 g(0,y+v)\mathrm{d}v}, \tag{7}$$

and second, the cumulative gain in longevity at age x last birthday within the same cohort born in year $t-x$ (a diagonal in the Lexis diagram), that is, between those born at exact time $t-x$ and those born at the end of the year $[t-x, t-x+1)$, defined by:

https://doi.org/10.1017/asb.2020.5 Published online by Cambridge University Press

$$H(x,t)=\sum_{y=0}^{x-1}\left\{\mu_U(y,t-x+y+1)-\mu_L(y,t-x+y)\right\},\ x\in\mathbb{N}^*. \tag{8}$$

The result at the core of the inference strategy is stated below:

Proposition 1. *Consider the transport Equation (1). Under Assumption 1, the following equalities hold:*

$$\exp\left(-\mu_L(x,t)\right)L_{t-x}\big(H(x,t)-\mu_L(x,t)\big)=\left(1-\frac{D_L(x,t)}{N(x,t)}\right)L_{t-x}\big(H(x,t)\big), \tag{9}$$

and

$$\begin{aligned}&L_{t-x-1}\big(H(x,t-1)-\mu_L(x,t-1)\big)\\&=\left(1+\frac{D_U(x,t)}{N(x+1,t)}\right)L_{t-x-1}\big(H(x,t-1)-\mu_L(x,t-1)+\mu_U(x,t)\big).\end{aligned} \tag{10}$$

The proof is detailed in the next part, along with a detailed discussion in Section 2.6. The resulting algorithm is described in Section 3.

2.5. Proof of Proposition 1

To prove (9), let us first focus on the exposure-to-risk in the lower triangle $E_L(x,t)=\int_t^{t+1}\int_x^{x+s-t}g(a,s)\mathrm{d}a\mathrm{d}s$. According to the transport equation (1), the population density in the lower triangle can be expressed as

$$\begin{aligned}g(a,s)&=g(x,s-a+x)\exp\left(-\int_x^a\mu(u,s-a+u)\mathrm{d}u\right)\\&=g(x,s-a+x)\exp\left(-(a-x)\mu_L(x,t)\right).\end{aligned}$$

where the last equality comes from the assumption of a piecewise constant death rate on Lexis triangles. By the change of variable $v\leftarrow s-a+x-t$, the exposure-to-risk can then be rewritten as

$$\begin{aligned}E_L(x,t)&=\int_t^{t+1}\int_x^{x+s-t}g(x,s-a+x)\exp\left(-(a-x)\mu_L(x,t)\right)\mathrm{d}a\mathrm{d}s\\&=\int_0^1\int_{t+v}^{t+1}g(x,t+v)\exp\left(-(s-v-t)\mu_L(x,t)\right)\mathrm{d}s\mathrm{d}v.\end{aligned}$$

By straightforward computation, one finally gets the following expression for the exposure-to-risk in the lower triangle:

$$E_L(x,t)=\int_0^1 g(x,t+v)\frac{1-\exp\left((v-1)\mu_L(x,t)\right)}{\mu_L(x,t)}\mathrm{d}v. \tag{11}$$

https://doi.org/10.1017/asb.2020.5 Published online by Cambridge University Press

Also note that $D_L(x,t) = \mu_L(x,t)E_L(x,t) = \int_0^1 g(x,t+v)(1-\exp((v-1)\mu_L(x,t)))\mathrm{d}v$ and $N(x,t) = \int_0^1 g(x,t+v)\mathrm{d}v$ so that

$$N(x,t) - D_L(x,t) = \int_0^1 g(x,t+v)\exp((v-1)\mu_L(x,t))\,\mathrm{d}v.$$

Let us now derive the population density at exact age x, for any $v \in [0,1)$,

$$\begin{aligned}
g(x,t+v) &= g(0,t-x+v)\exp\left(-\int_0^x \mu(u,t-x+v+u)\mathrm{d}u\right)\\
&= g(0,t-x+v)\exp\left(-\sum_{y=0}^{x-1}\int_y^{y+1}\mu(u,t-x+v+u)\mathrm{d}u\right)\\
&= g(0,t-x+v)\exp\left(-\sum_{y=0}^{x-1}\int_y^{y+1-v}\mu(u,t-x+v+u)\mathrm{d}u\right.\\
&\quad\left.-\sum_{y=0}^{x-1}\int_{y+1-v}^{y+1}\mu(u,t-x+v+u)\mathrm{d}u\right)\\
&= g(0,t-x+v)\exp\left(-(1-v)\sum_{y=0}^{x-1}\mu_L(y,t-x+y)\right.\\
&\quad\left.-v\sum_{y=0}^{x-1}\mu_U(y,t-x+y+1)\right)\\
&= S(x,t)g(0,t-x+v)\exp\left(-vH(x,t)\right),
\end{aligned} \tag{12}$$

where $S(x,t) = \exp\left(-\sum_{y=0}^{x-1}\mu_L(y,t-x+y)\right)$ is the survival function at age x for individuals which attained (exact) age x at (exact) time t, and where the cumulative death rate differential within the cohort $H(x,t)$ has been introduced in Equation (8). Let us now combine the previous results to get

$$N(x,t) - D_L(x,t) = S(x,t)e^{-\mu_L(x,t)}\int_0^1 g(0,t-x+v)e^{-v(H(x,t)-\mu_L(x,t))}\mathrm{d}v,$$

and finally, let us apply some renormalization of the right-hand side, first by $N(x,t)$ and second by $\int_0^1 g(0,t-x+v)\mathrm{d}v$ to get the following formula, which reduces to Equation (9):

$$1 - \frac{D_L(x,t)}{N(x,t)} = \frac{S(x,t)e^{-\mu_L(x,t)}\int_0^1 \tilde{g}(0,t-x+v)e^{-v(H(x,t)-\mu_L(x,t))}\mathrm{d}v}{S(x,t)\int_0^1 \tilde{g}(0,t-x+v)e^{-vH(x,t)}\mathrm{d}v}.$$

where $\tilde{g}(0,t-x+v) = \frac{g(0,t-x+v)}{\int_0^1 g(0,t-x+v)\mathrm{d}v}$.

https://doi.org/10.1017/asb.2020.5 Published online by Cambridge University Press

The proof of (10) follows similarly. Since $E_U(x,t) = \int_t^{t+1} \int_{x+s-t}^{x+1} g(a,s)\mathrm{d}a\mathrm{d}s$ and $g(a,s) = g(x+1, s+x+1-a)\exp\left((x+1-a)\mu_U(x,t)\right)$, then by changing variables, one gets $E_U(x,t) = \int_0^1 g(x+1,t+v)\frac{\exp(v\mu_U(x,t))-1}{\mu_U(x,t)}\mathrm{d}v$, so that

$$N(x+1,t) + D_U(x,t) = \int_0^1 g(x+1,t+v)\exp\left(v\mu_U(x,t)\right)\mathrm{d}v.$$

Then as $g(x+1,t+v) = g(0,t-x-1+v)S(x+1,t)\exp\left(-vH(x+1,t)\right)$, one finally obtains

$$\left(1 + \frac{D_U(x,t)}{N(x+1,t)}\right) L_{t-x-1}\left(H(x+1,t)\right) = L_{t-x-1}\left(H(x+1,t) - \mu_U(x,t)\right),$$

which leads to the result, as the following equality is verified from the definition in Equation (8):

$$H(x+1,t) = H(x,t-1) + \mu_U(x,t) - \mu_L(x,t-1).$$

2.6. Discussion

Exposure-to-risk interpretation. The equality (11) can be interpreted as follows: for each individual attaining exact age x at time $t+v$, its contribution to the exposure-to-risk in the lower triangle is $\frac{1-\exp((v-1)\mu_L(x,t))}{\mu_L(x,t)}$, which depends on the unobserved death rate to be estimated. This contrasts with classical methods which compute approximations of the exposure-to-risk based on observables. At first order, assuming $\mu_L(x,t) << 1$, one recovers that $E_L(x,t) \approx \int_0^1 g(x,t+v)(1-v)\mathrm{d}v$ and the related interpretation that the contribution of any individual which attained exact age x at time $t+v$ and living through the lower triangle is simply $1-v$ as it can be measured in the Lexis diagram.

Biased birthday density. The formula derived in (12) shows that the birthdays density at some age x is exponentially biased through $H(x,t)$ compared to the initial birthdays distribution (at age zero). This is true in general in the triangle model for the piecewise constant death rate (Assumption 1), as well as in the period table for which the cumulative death rate difference matrix reduces to $H(x,t) = \sum_{y=0}^{x-1}\left\{\mu(y,t-x+y+1) - \mu(y,t-x+y)\right\}$ where $\mu(x,t)$ denotes the period death rate for the square (x,t). Moreover, as one expects in general some mortality improvement over the years, age being fixed, one may be interested in interpreting the case $H(x,t) < 0$ – in this situation, one sees that the initial birthdays distribution is distorted to the highest birthdays (youngest individuals) in the cohort as age goes. This demonstrates how even in a discrete time specification, individuals in the same cohort may experience different death rates over life (more precisely they pass through the same rates but do not "spend the same time" in each triangle or square, so that the resulting survival functions are different). However, it is interesting to note that for the cohort table, which by definition assumes that $\mu_U(y,t-x+y+1) = \mu_L(y,t-x+y)$, the H matrix

https://doi.org/10.1017/asb.2020.5 Published online by Cambridge University Press

vanishes, so that the initial birthdays distribution perfectly propagates toward highest ages.

Closed population assumption. The main result in Proposition 1 is obtained using the assumption on the population being closed, that is, no migration flows are considered. This is of course a limit as migration flows do exist, which may distort (a) the time of birth distribution, (b) the population counts, and (c) aggregate mortality; indeed, mortality rates of immigrants, emigrants, and local population can have different levels and dynamics. As for point (a), it could be argued that the main features of the birthdays distribution are driven by shocks, as of interest in this paper; moreover, one could expect that emigrants or immigrants birthdays distribution is more uniform; this is left for further investigation.

For point (b), note that Equations (9) and (10) make use of the actual population count $N(x, t)$ of individuals attaining exact age x in calendar year t; therefore, the population counts used in the estimation procedure of the death rates at age x and calendar year t are only assumed to come from a population closed within calendar year t; in other words, the estimation for year t restarts from a revised estimate $N(x, t)$ so it does not assume that the population count at time t is the result of a pure decrement due to deaths of an initial population $N(0, t - x)$ in a closed cohort.

The problem (c) is probably the most difficult topic to handle, as it refers to building coherent estimation procedures for an heterogeneous population made of several subpopulations with different mortality rates. This is left for further research.

Finally, it is worth mentioning that the modeling framework could be extended to open populations as follows. Assuming that an individual emigration rate $e(a, t)$ and a total immigration rate $I(a, t)$, the transport component (1) would rewrite into:

$$(\partial_a + \partial_t)g(a, t) = -(\mu(a, t) + e(a, t))\, g(a, t) + I(a, t).$$

The estimation of the emigration and immigration rates from the data remains the core challenging issue and is beyond the scope of the present paper.

Link with estimates of the Human Mortality Database. It is worth mentioning that at the time of writing, the HMD released an update on February 2018, including in particular a revision of exposure calculation based on monthly birth counts. We now make the link with both the new Version 6 and the old Version 5 of the Methods Protocol.

From (11), it can be shown by performing a first-order expansion in $\mu_L(x, t)$ that

$$E_L(x, t) \approx E_L^{(1)}(x, t) - \mu_L(x, t)E_L^{(2)}(x, t),$$

https://doi.org/10.1017/asb.2020.5 Published online by Cambridge University Press

where

$$E_L^{(1)}(x,t) = N(x,t)\left(1 + \frac{L'_{t-x}(H(x,t))}{L_{t-x}(H(x,t))}\right),$$

and

$$E_L^{(2)}(x,t) = \frac{1}{2}N(x,t)\left[1 + \frac{2L'_{t-x}(H(x,t)) + L''_{t-x}(H(x,t))}{L_{t-x}(H(x,t))}\right].$$

Let us denote by B_{t-x} the random variable with values in $[0,1]$ that represents the time of birth in the year $t-x$, with mean $m_{t-x} := \mathbb{E}[B_{t-x}]$ and variance $\sigma^2_{t-x} := \mathrm{Var}(B_{t-x})$. Note that the corresponding density is

$$[0,1] \ni v \mapsto \tilde{g}(0, t-x+v) = \frac{g(0,t-x+v)}{\int_0^1 g(0,t-x+v)\mathrm{d}v},$$

so that the mean and variance write

$$m_{t-x} = \int_0^1 v\tilde{g}(0,t-x+v)\mathrm{d}v \text{ and } \sigma^2_{t-x} = \int_0^1 v^2\tilde{g}(0,t-x+v)\mathrm{d}v - m^2_{t-x}.$$

Under the assumption $H(x,t) = 0$, that is, no mortality improvement between the youngest and oldest individuals within the same cohort, one can write

$$E_L(x,t) \approx N(x,t)\,(1-m_{t-x}) - \frac{1}{2}\mu_L(x,t)N(x,t)\left((1-m_{t-x})^2 + \sigma^2_{t-x}\right).$$

Note again that the assumption $H(x,t) = 0$ is not consistent with the piecewise constant death rate assumption on Lexis triangles, nor with the framework underlying the period tables.

Now, if one uses (6) and replaces $\mu_L(x,t) = \frac{D_L(x,t)}{E_L(x,t)}$ by its zero-order approximation

$$\mu_L(x,t) \approx \frac{D_L(x,t)}{N(x,t)\,(1-m_{t-x})},$$

one finally obtains the formula (51) displayed in the Version 6 in the methods protocol, see the Appendix for a derivation:

$$E_L(x,t) \approx P(x,t+1)\,(1-m_{t-x}) + \frac{D_L(x,t)}{2(1-m_{t-x})}\left((1-m_{t-x})^2 - \sigma^2_{t-x}\right). \qquad (13)$$

Finally, if one assumes births to be uniformly distributed, then $m_{t-x} = \frac{1}{2}$ and $\sigma^2_{t-x} = 1/12$ so that the classical formula in Version 5 methods protocol is recovered (see Appendix E therein for the original derivation and again the

https://doi.org/10.1017/asb.2020.5 Published online by Cambridge University Press

Appendix in the present paper):

$$E_L(x,t) \approx \frac{1}{2}P(x,t+1) + \frac{1}{6}D_L(x,t).$$

3. Implementation and Numerical Results

3.1. Implementation

The purpose of this subsection is to detail the step-by-step procedure that allows to apply the method provided in Proposition 1, especially how it is made use of the monthly birth records.

Preliminary step. As Equations (9) and (10) rely on the Laplace transform $L_y(\theta)$ of the distribution of times of birth within year of birth y, a preliminary step is to build an estimator of this function, for any θ, by considering the following integral approximation:

$$\begin{aligned} L_y(\theta) &= \frac{\int_0^1 g(0,y+v)\exp(-\theta v)\mathrm{d}v}{\int_0^1 g(0,y+v)\mathrm{d}v} \\ &= \frac{1}{\int_0^1 g(0,y+v)\mathrm{d}v}\sum_{i=0}^{11}\int_{\frac{i}{12}}^{\frac{i+1}{12}} g(0,y+v)\exp(-\theta v)\mathrm{d}v \\ &\approx \frac{1}{\int_0^1 g(0,y+v)\mathrm{d}v}\sum_{i=0}^{11}\exp\left(-\theta\left(\frac{2i+1}{24}\right)\right)\int_{\frac{i}{12}}^{\frac{i+1}{12}} g(0,y+v)\mathrm{d}v, \end{aligned}$$

where we remark that for $i \in \{0,...,11\}$, $\int_{\frac{i}{12}}^{\frac{i+1}{12}} g(0,y+v)\mathrm{d}v$ is the number of births in month $i+1$ within year y (as the continuous sum over the month of individuals with exact age zero) and that the quantity $\int_0^1 g(0,y+v)\mathrm{d}v$ is the number of births within year y. These estimates can therefore be calculated using the monthly birth data provided in the HFD. As a result of this preliminary step, the functions $L_y(.)$ are computed for any year of birth y for which monthly birth records are available.

We now provide the detail of the algorithm for the recursive computation of the death rates, resulting from Proposition 1. In the following, the first three steps are discussed in detail, then the generic algorithm is provided.

First step (age 0, lower triangle). Considering age $x=0$ for a given calendar year t, then Equation (9) simplifies into:

$$\exp(-\mu_L(0,t))\, L_t\big(\mu_L(0,t)\big) = \left(1 - \frac{D_L(0,t)}{N(0,t)}\right).$$

https://doi.org/10.1017/asb.2020.5 Published online by Cambridge University Press

The right-hand side is known, as well as the function $L_t(.)$, see above; therefore, an implicit equation in the lower triangle death rate $\mu_L(0,t)$ appears, which can be solved using a standard optimization routine. As a result, for any calendar year t, the death rates $\mu_L(0,t)$ have been estimated.

Second step (age 0, upper triangle). Switching now to Equation (10), for any year t and still considering age $x=0$, one obtains:

$$L_{t-1}\big(-\mu_L(0,t-1)\big)=\left(1+\frac{D_U(0,t)}{N(1,t)}\right)L_{t-1}\big(-\mu_L(0,t-1)+\mu_U(0,t)\big).$$

First recall that the function $L_{t-1}(.)$ is known from the preliminary step described above, and that the death rate $\mu_L(0,t-1)$ has been estimated from the first step; the remaining single unknown is the death rate $\mu_U(0,t)$ on the right-hand side which is to be estimated relying, again, on any standard optimization method.

Third step (age 1, lower triangle). Finally, we illustrate the third step of the algorithm, which goes back to Equation (9) in order to infer the mortality rate of the lower triangle at age one, leading to:

$$\exp\left(-\mu_L(1,t)\right)L_{t-1}\big(H(1,t)-\mu_L(1,t)\big)=\left(1-\frac{D_L(1,t)}{N(1,t)}\right)L_{t-1}\big(H(1,t)\big),$$

where by Equation (8),

$$H(1,t)=\mu_U(0,t)-\mu_L(0,t-1).$$

Note that $H(1,t)$ is known as the result of the first and the second steps, therefore the only unknown is $\mu_L(1,t)$ in the left-hand side, which can be solved numerically.

Generic algorithm. The general recursive algorithm for computing the death rates is described below in a generic form:

Algorithm 1. *For age x starting at zero:*
(i) *Solve Equation (9) to estimate the death rate $\mu_L(x,t)$ for the lower triangles of any available year t,*
(ii) *Then based on the previous estimates, solve Equation (10) to infer the death rate $\mu_U(x,t)$ for the upper triangles of any available year t,*
(ii) *Compute the value for $H(x+1,t)=H(x,t-1)+\mu_U(x,t)-\mu_L(x,t-1)$ for all possible years t, let $x\leftarrow x+1$ and go to step* (i)

Remark 1. *Note that the method is past-dependent – this is natural as any change in past death rates modifies the future birthdays distribution in the cohort. This way, any revision of past death or population count at (x,t), which may occur in practice, requires the reuse of the methodology which will provide an update of the mortality rates at $(y,t+y-x)$ for $y\geq x$.*

https://doi.org/10.1017/asb.2020.5 Published online by Cambridge University Press

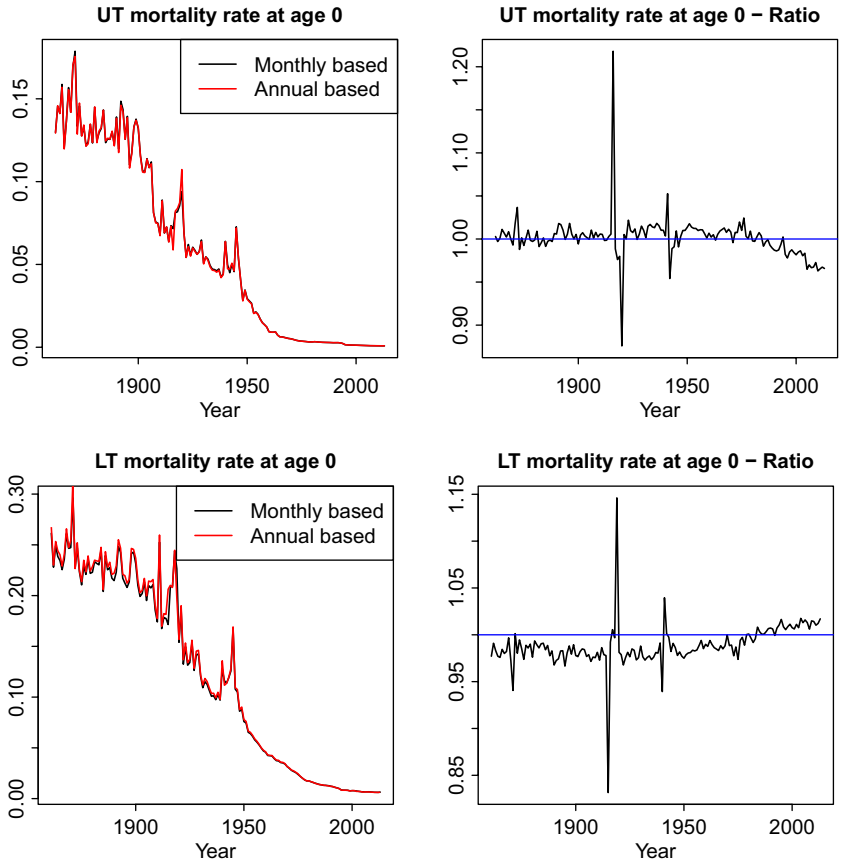

FIGURE 5: Left: death rates estimated based on the new inference method (in black) and compared to estimates using the standard method based on annual population records (in red). Right: ratio between new and old estimates. Top: upper triangle. Bottom: lower triangle.

3.2. Numerical results

In Figures 5, 6, 7, and 8, we depict the death rate estimates obtained with the method developed in this paper applied to French data sourced from the HMD (annual population estimates, Figure 1 and number of deaths in Lexis triangles, Figure 2) and the HFD (births by month, Figure 3). The number of births by month is used to approximate the Laplace transform of the birthdays distribution which is used in the inference process.

The results are compared with estimates as they would be classically computed based on annual observables (see Wilmoth *et al.* (2007) and Boumezoued (2020) for further details):

$$\widehat{\mu_L}(x,t) = \frac{D_L(x,t)}{\frac{1}{2}N(x,t) - \frac{1}{3}D_L(x,t)} \text{ and } \widehat{\mu_U}(x,t) = \frac{D_U(x,t)}{\frac{1}{2}N(x+1,t) + \frac{1}{3}D_U(x,t)}.$$

https://doi.org/10.1017/asb.2020.5 Published online by Cambridge University Press

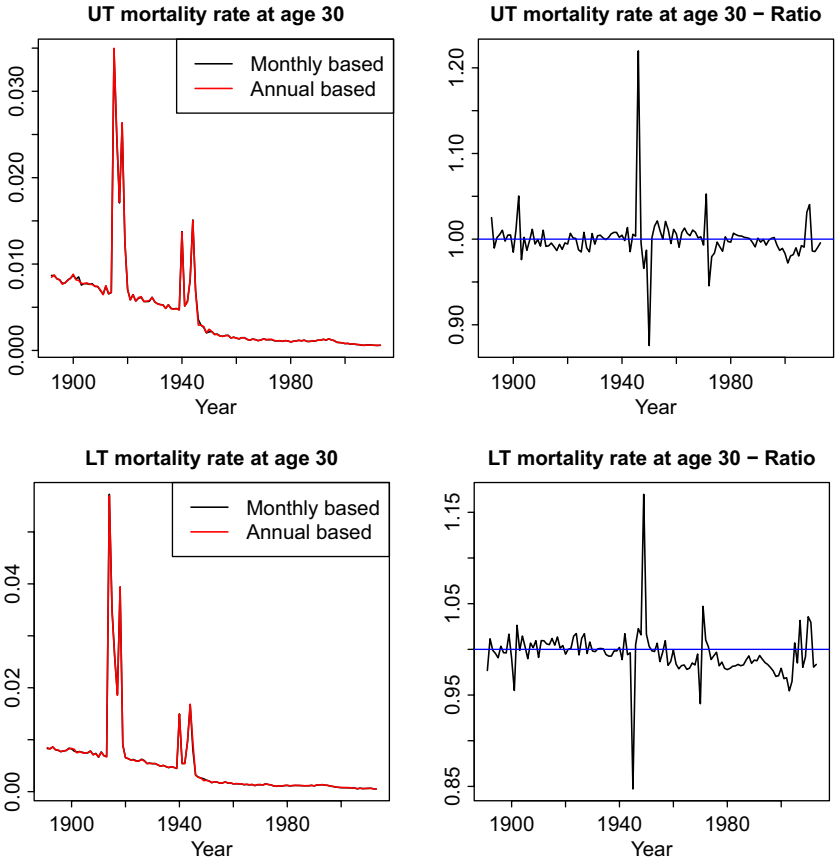

FIGURE 6: Left: death rates estimated based on the new inference method (in black) and compared to estimates using the standard method based on annual population records (in red). Right: ratio between new and old estimates. Top: upper triangle. Bottom: lower triangle.

Each figure includes on the right the ratio between the new and the old estimates, which helps quantify the differences between both. First, the ratio is for several age classes close to one, which indicates that the new estimate does not differ much from the classical one, in other words that the classical approximation is valid. However, one sees strong deviations for specific ages in time, and this translates over time and ages, so that it appears that the anomalies belong to specific generations. As displayed, relative discrepancies between the two estimates can reach up to around +/− 20%.

Note that one can also observe that the fluctuations of the ratio are mostly characterized by opposite adjustments for consecutive years. This phenomenon has been analyzed in Cairns *et al*. (2016), where it has been first shown how sudden break in the birth series in a year creates some convexity then concavity of the population curve, leading to consecutive over- and under-estimation of the

https://doi.org/10.1017/asb.2020.5 Published online by Cambridge University Press

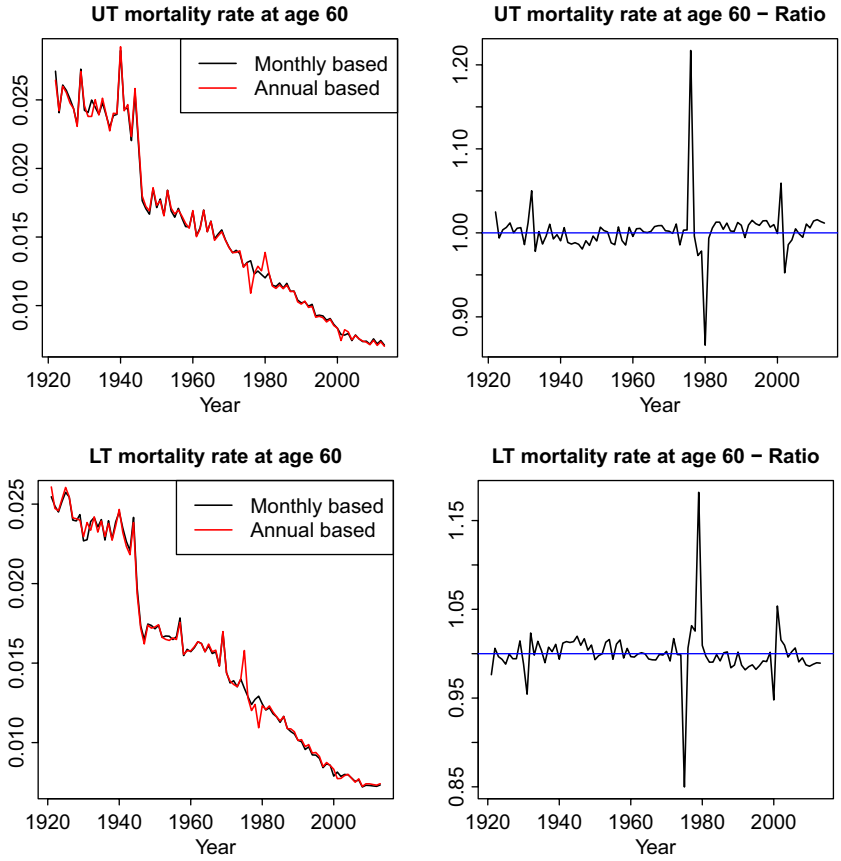

FIGURE 7: Left: death rates estimated based on the new inference method (in black) and compared to estimates using the standard method based on annual population records (in red). Right: ratio between new and old estimates. Top: upper triangle. Bottom: lower triangle.

exposure-to-risk. This can be particularly observed for the periods characterized by major birth fluctuations, see, for example, cohorts 1940 and 1941 in Figure 5 bottom right panel.

To further analyze these discrepancies, we depict in Figure 9, mortality improvement rates separated between upper and lower triangles as

$$\frac{\mu_L(x,t+1)-\mu_L(x,t)}{\mu_L(x,t)} \text{ and } \frac{\mu_U(x,t+1)-\mu_U(x,t)}{\mu_U(x,t)}.$$

Clearly, the isolated cohort effects disappear in the new mortality tables: mainly the diagonals around 1915 and 1920 and to a lower extent those born around 1940; note that this indeed corresponds to the shocks in birth numbers as illustrated in Figure 3, which confirms from a mathematical perspective the previous contributions by Richards (2008), Cairns *et al.* (2016), and Boumezoued (2020).

https://doi.org/10.1017/asb.2020.5 Published online by Cambridge University Press

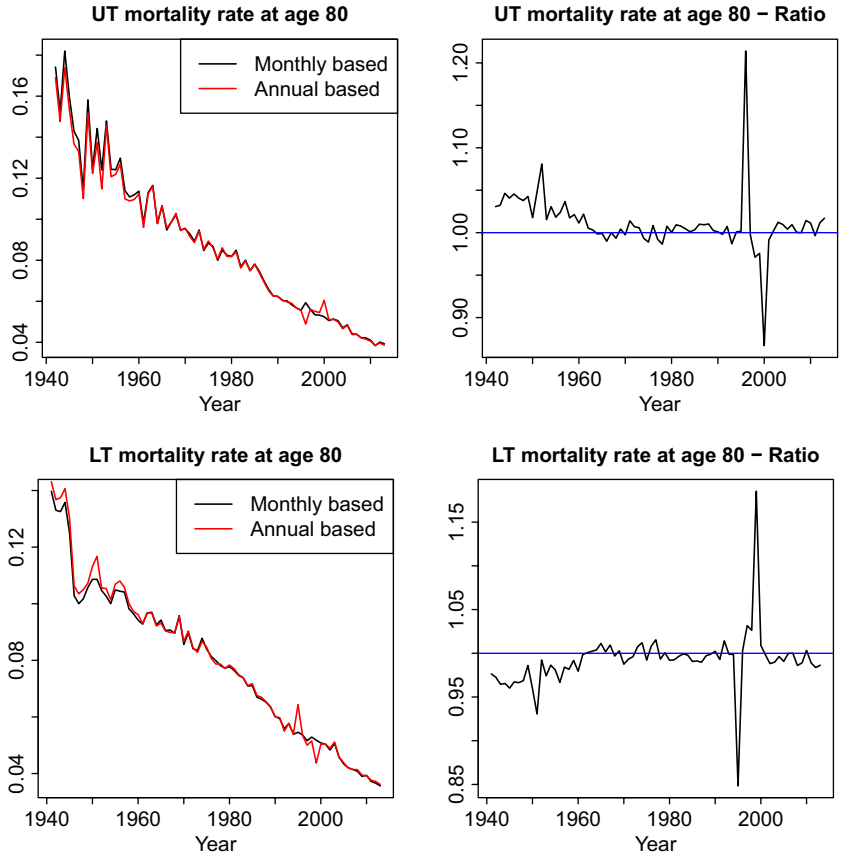

FIGURE 8: Left: death rates estimated based on the new inference method (in black) and compared to estimates using the standard method based on annual population records (in red). Right: ratio between new and old estimates. Top: upper triangle. Bottom: lower triangle.

Further comparison with previous work on data correction. Recall that Cairns *et al.* (2016) and Boumezoued (2020) proposed methods to detect and correct death rate estimates in period tables based on monthly birth counts. The empirical method applied by Boumezoued (2020) to the HMD Version 5 relied on a so-called correction indicator for each cohort, as a ratio between two estimates of the exposure-to-risk at age zero, based on either monthly or annual birth counts; let us emphasize that these ratios were obtained at age zero and did not account for deaths; they were then applied to the uncorrected mortality rates at all ages within each cohort to provide new estimates which apparently removed the strong isolated cohort effects.

It is worth mentioning that such previous works relied on a set of empirical choices and approximations, as it is also the case for the Version 5 and 6 estimates, see the discussion in Section 2.6 as well as the detail of the underlying reasoning in Appendix. Also, we recall that the focus of Boumezoued (2020) is on period tables, whereas we develop a model to infer mortality rates in each

https://doi.org/10.1017/asb.2020.5 Published online by Cambridge University Press

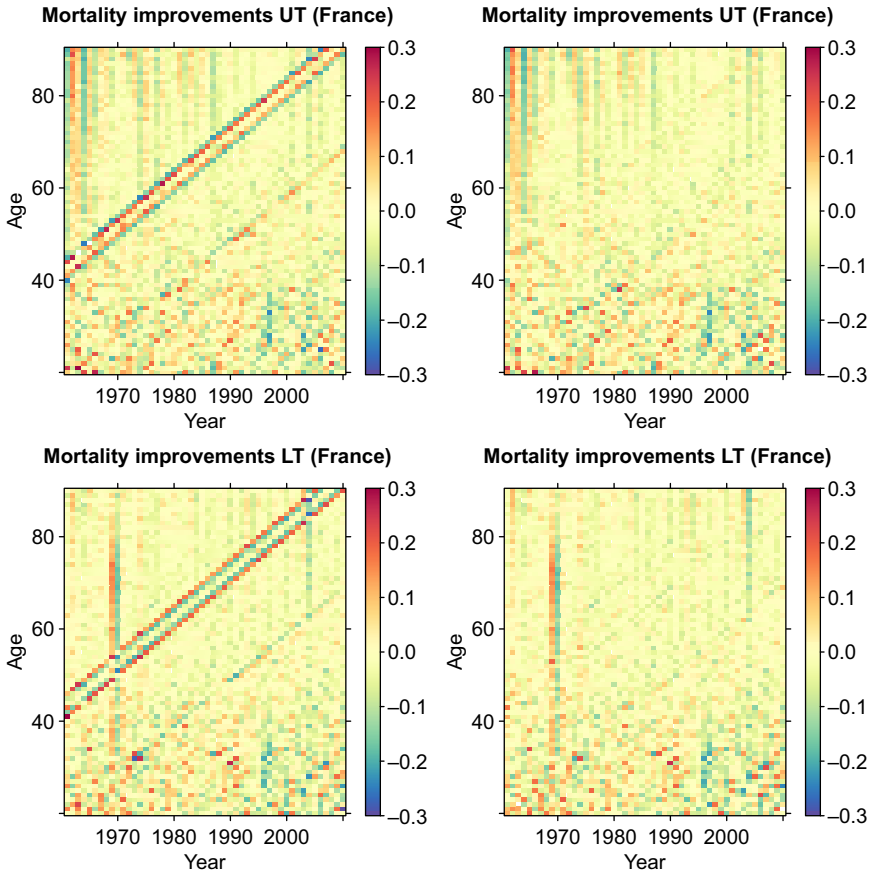

FIGURE 9: Left: mortality improvement rates using the standard method based on annual population records. Right: mortality improvement rates using the new inference method. Top: upper triangles. Bottom: lower triangles.

triangle of the Lexis diagram; at this granularity, this fully illustrates how previous estimators provided biased values (more than 15% deviation, see Figures 5, 6, 7, and 8); moreover, we think that the inference at Lexis triangle scale can be leveraged for further research in the field of stochastic mortality modeling, as it duplicates the number of observed points while allowing for a refined analysis of the age and period effects which may differ, as illustrated in Figure 9.

Therefore, we claim that the core advantage of the method proposed in this paper lies in the mathematical basis and the weaker assumptions needed to derive the estimates; from such viewpoint, the estimates provided are more reliable. From a numerical standpoint, although the algorithm is new, since the implementation is done the derivation estimates are output with insignificant computational cost.

In order to refine our analysis, let us focus on another core assumption in the previous work by Boumezoued (2020) who suggested that the ratio between corrected and crude estimates is the same for each age within a given cohort.

https://doi.org/10.1017/asb.2020.5 Published online by Cambridge University Press

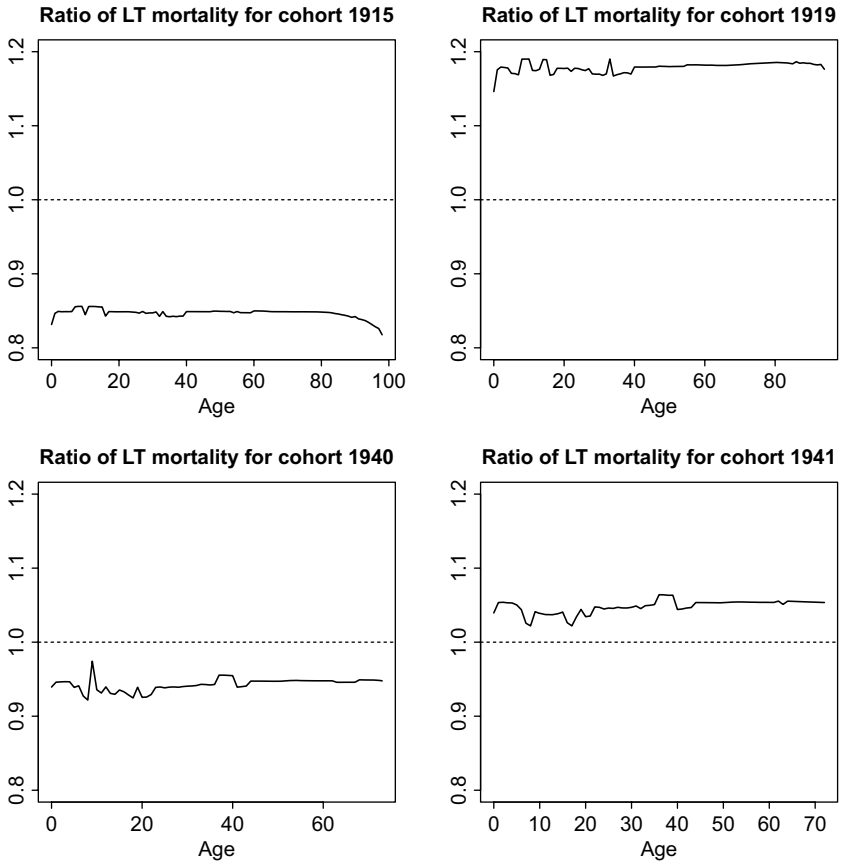

FIGURE 10: Ratio of corrected and crude mortality rates of the lower triangles. Each figure corresponds to a specific cohort: 1915, 1919, 1940, and 1941.

We analyze in our case the value of this ratio for cohorts 1915, 1919, 1940, and 1941, focusing on the lower triangles for illustration; these results are depicted in Figure 10. At first sight, the figures illustrate that the ratio between corrected and crude mortality estimates is roughly stable over ages within a given cohort; note that this would have been already observable in Figures 5, 6, 7, and 8: when fixing a cohort, one observes that the magnitude of the correction remains in the same order at the several ages. Overall stability is observed after age around 50 for cohorts 1919, 1940, and 1941, although this general remark does not hold for cohort 1915, which shows at high ages an increasing correction magnitude. In more details, it appears that the variations can reach several percentages between – as an example, the cohort 1919 shows a correction at age zero which seems lower by at least 2% than that at the other ages.

Impact on stochastic mortality modeling. The correction of the mortality data in general, and the removal of isolated cohort effects in particular, has already been discussed in the context of stochastic mortality modeling by Cairns *et al.*

https://doi.org/10.1017/asb.2020.5 Published online by Cambridge University Press

(2016) and Boumezoued (2020). The purpose of this subsection is to show how previous conclusions still hold with the proposed new method, as well as to discuss additional insights. We focus on the classical Lee and Carter (1992) as well as on some extension with cohort component, derived from Renshaw and Haberman (2006). In the following, we again focus on the lower triangle mortality rates; similar conclusions can be drawn from the analysis of the upper triangles. The experiment is carried out using the StMoMo R package, see Villegas *et al.* (2018), using the standard log-Poisson calibration. The age range for model inference is 20–90, and the period is 1954–2013.

The Lee–Carter model applied to lower triangle mortality rates that can be written as follows:

$$\ln \mu_L(x,t) = \alpha(x) + \beta(x)\kappa(t).$$

In this model, $\alpha(x)$ captures the static age structure of mortality rates in the log-scale, $\kappa(t)$ represents the time dynamics, whereas $\beta(x)$ is the sensitivity of age class x to mortality yearly variations.

The results of the model fitting are depicted in Figure 11. It can be seen that the parameters order of magnitude and shape are rather unchanged, although the absolute differences exhibit changes in the age parameters $\alpha(x)$ and $\beta(x)$, especially at high ages, and in the period parameter $\kappa(t)$ which seems in particular under-estimated in the crude data for the period 1960–1990 and overestimated for the period 1990–2013. This results in a slight adjustment in the mortality rate forecasts, which are lower using new mortality estimates for ages 20, 30, and 60, as illustrated in Figure 12 assuming a standard random walk with drift for the period parameter $\kappa(t)$. Note however that for age 90 the predicted mortality confidence interval is higher and shows a clear reduction in volatility, which can be explained since the $\beta(x)$ parameter is significantly reduced at this age using the new mortality estimates.

As expected, the major difference in the two model inference exercises is captured in the residuals, as it can be seen in Figure 13, where the isolated diagonal appears in the fitting using the crude data. This translates into an improvement of the log-likelihood from $-33,559$ (raw mortality table) to $-29,405$ (new mortality table). It is worth mentioning that "true" cohort effects do remain in the new table, as large diagonal trends can still be observed in the residuals. This observation typically leads to consider stochastic mortality models including a so-called cohort component, as we present in the following.

In order to assess the impact on mortality models including cohort parameters, we rely on a simplification of the model by Renshaw and Haberman (2006), which reduces to an extension to the Lee–Carter model to account for cohort effects:

$$\ln \mu_L(x,t) = \alpha(x) + \beta(x)\kappa(t) + \gamma(t-x).$$

The reason for illustrating the results using this simplification lies in the challenge to infer properly the general model parameters, which can distort the

https://doi.org/10.1017/asb.2020.5 Published online by Cambridge University Press

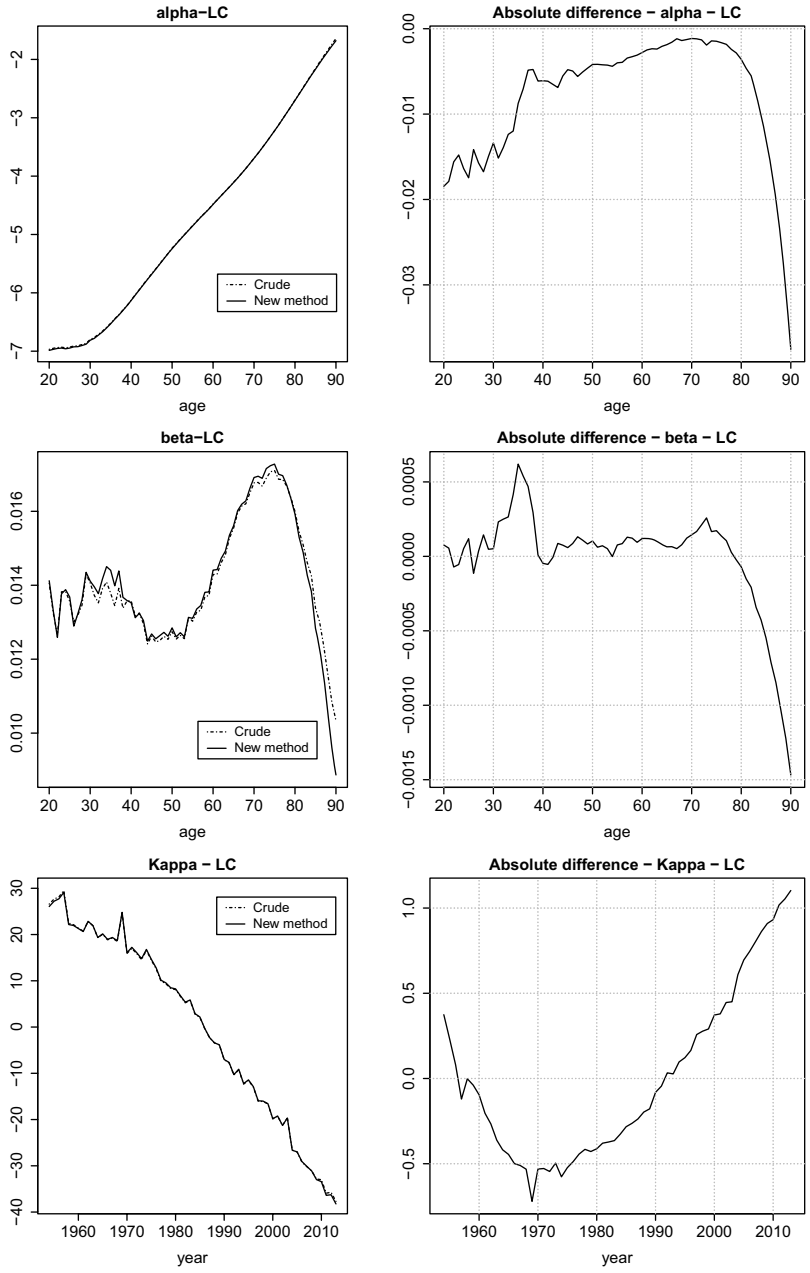

FIGURE 11: Comparison of fitting results for the Lee–Carter between crude estimates and proposed method. Left: fitted values; right: absolute differences.

https://doi.org/10.1017/asb.2020.5 Published online by Cambridge University Press

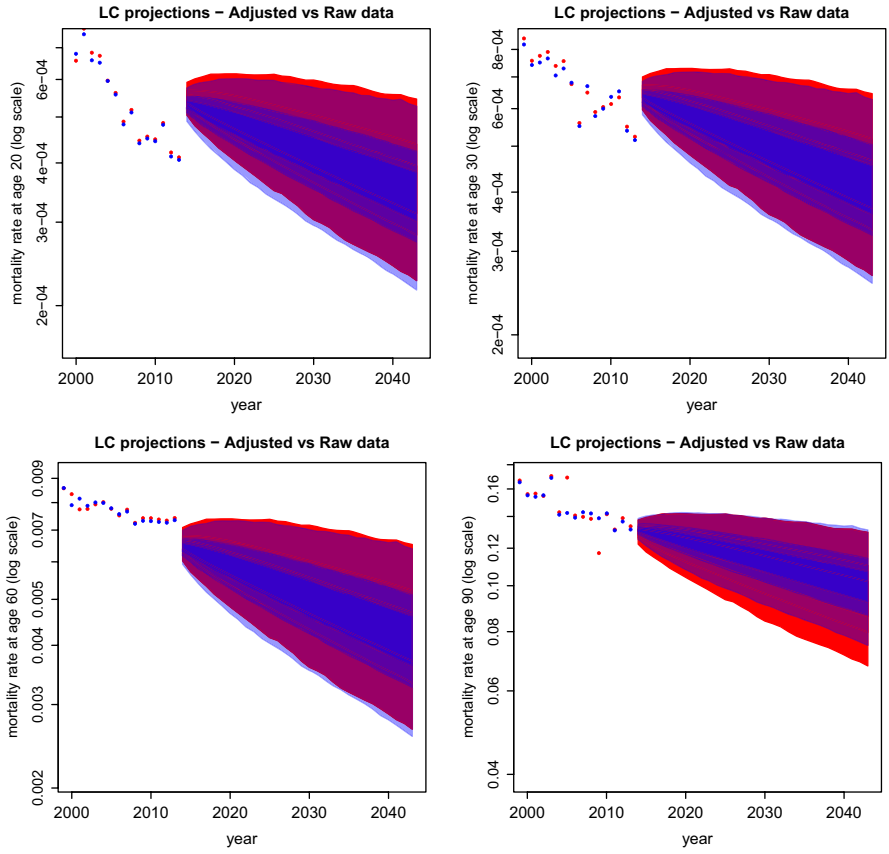

FIGURE 12: Comparison of 99% confidence intervals forecasts by the Lee–Carter model between crude estimates (red) or proposed method (blue). From left to right: age class 20, 30, 60, and 90.

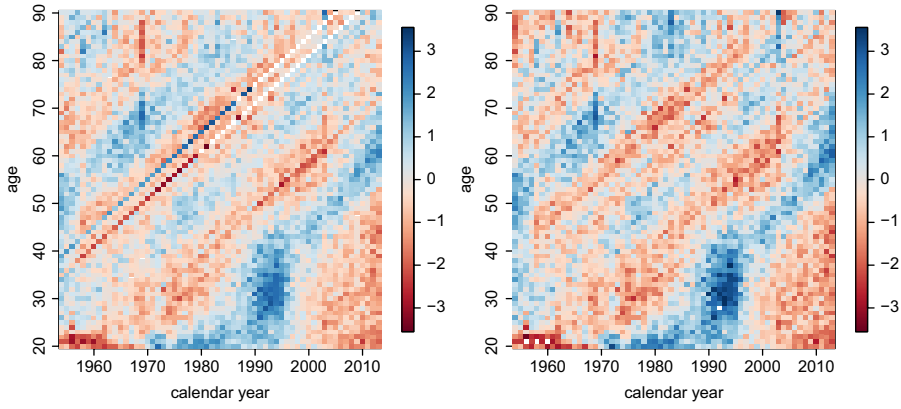

FIGURE 13: Comparison of residuals for the Lee–Carter model using crude estimates (left) or proposed method (right).

https://doi.org/10.1017/asb.2020.5 Published online by Cambridge University Press

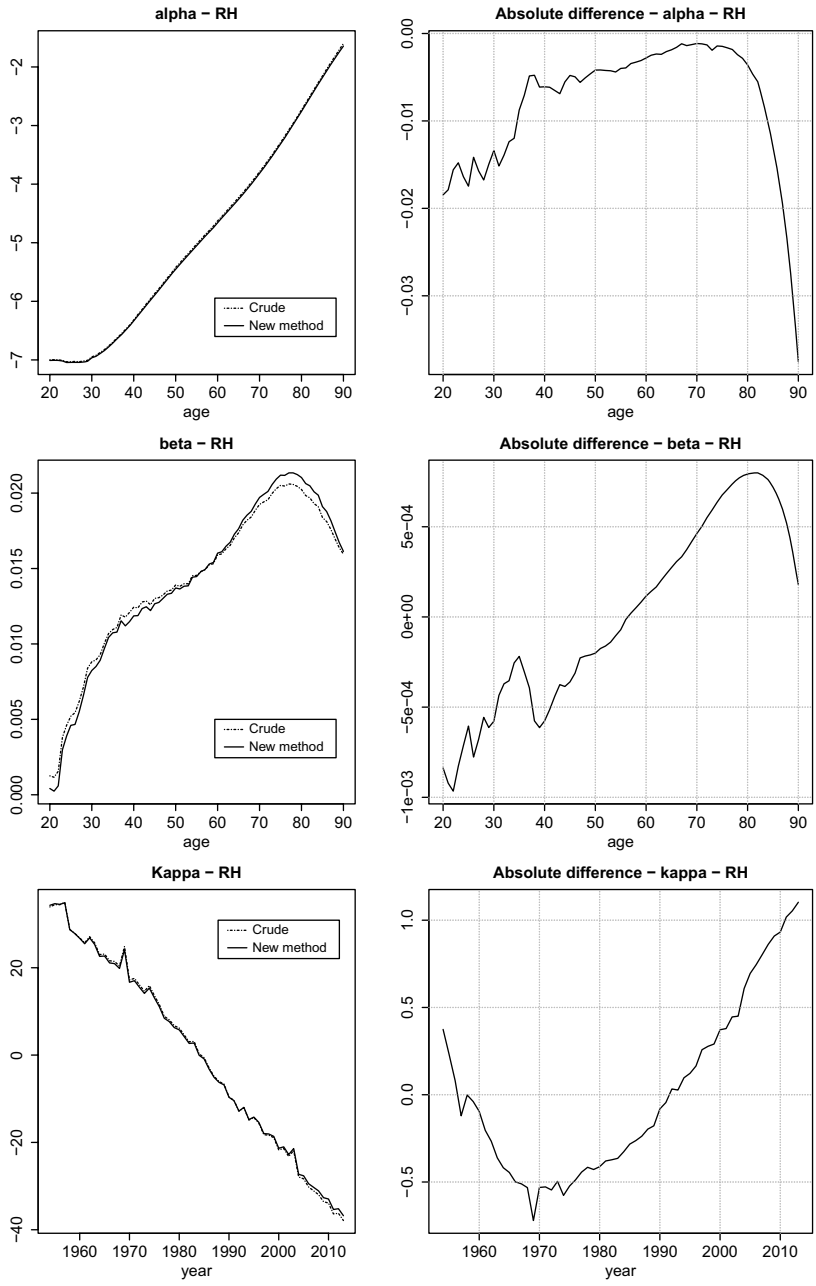

FIGURE 14: Comparison of fitting results for the Renshaw–Haberman model between crude estimates and proposed method. Left: fitted values; right: absolute differences.

comparison between the crude and the new data if the optimization algorithm does not converge. The results of the fitting procedure are depicted in Figure 14. The major differences are observed for the cohort parameter, where

https://doi.org/10.1017/asb.2020.5 Published online by Cambridge University Press

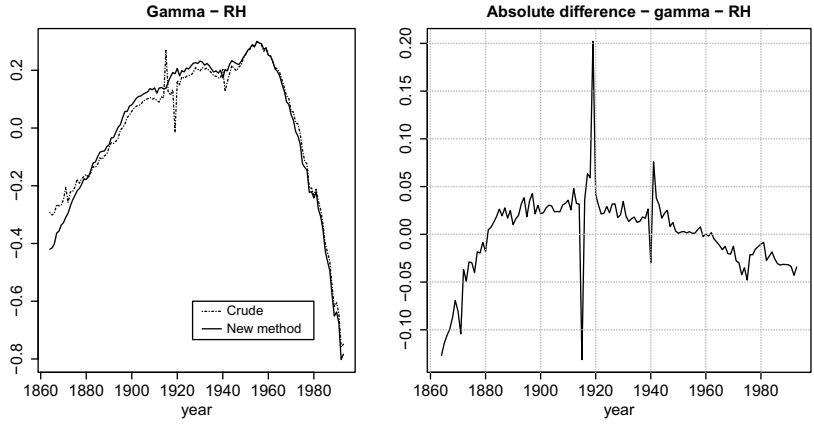

FIGURE 14: Continued.

especially the spikes of the original estimate are removed. The analysis of the forecast, see Figure 15, also demonstrates a reduction in the volatility, with the exception of age 30 starting from 2030, as well as age 20 overall, for which higher possible increase of the mortality rates is especially forecast by the model using the updated mortality estimates. It can also be observed a smoother confidence interval at age 90 for projection years around 2030, which benefits from a more regular cohort parameter around year of birth 1940. Finally, as expected, the residuals of the (simplified) Renshaw–Haberman model are rather unchanged, see Figure 16, since the cohort parameter adjusts to the original data anomalies. Finally, the removal of the abnormal cohort effects tends to less favor this model by decreasing the log-likelihood from $-24,839$ to $-24,954$.

4. CONCLUDING REMARKS

In this paper, we proposed an inference strategy for general population mortality tables based on the derivation of formulas in the Lexis diagram, which relate the death rate to annual observables and the intra-year distribution of birthdays over ages. The method therefore uses monthly birth counts to refine classical mortality estimates. The new mortality tables show better features, including the fact that previous anomalies in the form of isolated cohort effects disappear, which confirms from a mathematical perspective of the previous contributions by Richards (2008), Cairns *et al.* (2016), and Boumezoued (2020).

Several topics remain to be addressed to strengthen the methodology. First, it is of interest to account for population flows which may for several countries deform the closest population count, as well as distort the birthdays distribution over ages. Second, we emphasize that it is of importance to derive confidence intervals for the prediction, by going beyond the classical

https://doi.org/10.1017/asb.2020.5 Published online by Cambridge University Press

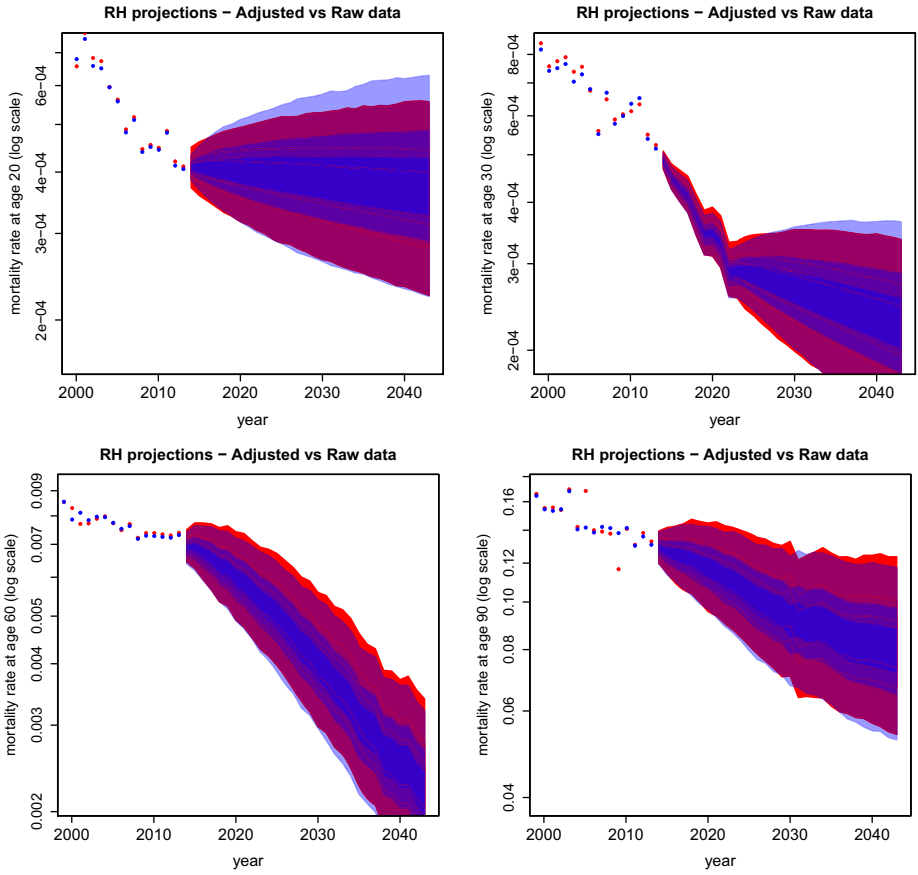

FIGURE 15: Comparison of 99% confidence intervals forecasts by the Renshaw–Haberman model between crude estimates (red) or proposed method (blue). From left to right: age class 20, 30, 60, and 90.

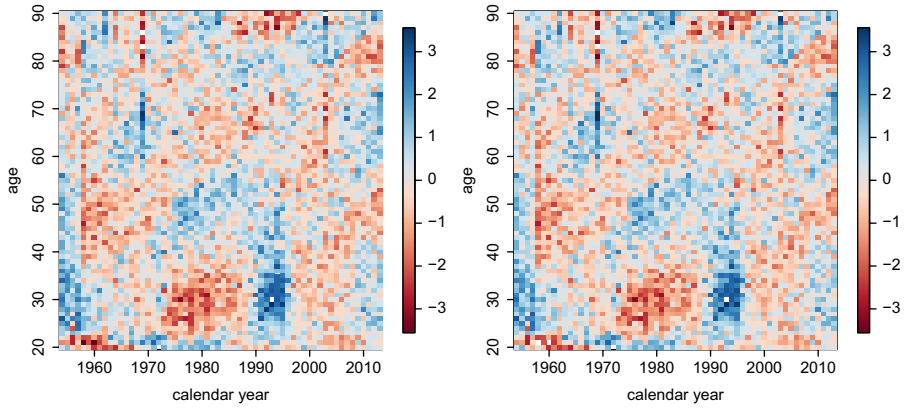

FIGURE 16: Comparison of residuals for the Renshaw–Haberman model using crude estimates (left) or proposed method (right).

https://doi.org/10.1017/asb.2020.5 Published online by Cambridge University Press

Poisson approximation to measure sampling risk. To this extent, a stochastic population dynamics model is required, as well as a dedicated statistical framework.

References

BERAN, R. (1981) *Nonparametric regression with randomly censored survival data*. Technical Report, Univ. California, Berkeley.

BOUMEZOUED, A. (2020) Improving mortality estimates with HFD fertility data. *North American Actuarial Journal*, 1–25.

BOUMEZOUED, A., HOFFMANN, M. and JEUNESSE, P. (2018) Statistical inference for an in-homogeneous age-structured population process. In revision. Preprint available at http://arxiv.org/abs/1903.00673.

BROWN, R.L. (1997) *Introduction to the Mathematics of Demography*. USA: Society of Actuaries.

BRUNEL, E., COMTE, F. and GUILLOUX, A. (2008) Estimation strategies for censored lifetimes with a lexis-diagram type model. *Scandinavian Journal of Statistics,* **35**(3), 557–576.

CAIRNS, A.J.G., BLAKE, D., DOWD, K. and KESSLER, A.R. (2016) Phantoms never die: Living with unreliable population data. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* **179**(4), 975–1005.

CLÉMENÇON, S., CHI TRAN, V. and DE ARAZOZA, H. (2008) A stochastic SIR model with contact-tracing: Large population limits and statistical inference. *Journal of Biological Dynamics*, **2**(4), 392–414.

COMTE, F., GAÏFFAS, S. and GUILLOUX, A. (2011) Adaptive estimation of the conditional intensity of marker-dependent counting processes. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, vol. 47, pp. 1171–1196. Institut Henri Poincaré.

DABROWSKA, D.M. (1987) Non-parametric regression with censored survival time data. *Scandinavian Journal of Statistics*, **14**(3), 181–197.

DALEY, D.J. and VERE-JONES, D. (2003) *An Introduction to the Theory of Point Processes. Volume I: Elementary Theory and Methods of Probability and Its Applications*. New York: Springer.

DOUMIC, M., HOFFMANN, M., KRELL, N. and ROBERT, L. (2015) Statistical estimation of a growth-fragmentation model observed on a genealogical tree. *Bernoulli,* **21**(3), 1760–1799.

HFD (2018) *The Human Fertility Database*. Germany and Austria: Max Planck Institute for Demographic Research and Vienna Institute of Demography. www.humanfertility.org

HMD (2018) *The Human Mortality Database*. Berkeley, USA and Germany: University of California and Max Planck Institute for Demographic Research. www.mortality.org.

HOFFMANN, M. and OLIVIER, A. (2016) Nonparametric estimation of the division rate of an age dependent branching process. *Stochastic Processes and Their Applications*, **126**(5), 1433–1471.

KEIDING, N. (1990) Statistical inference in the lexis diagram. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences,* **332**(1627), 487–509.

LEE, R.D. and CARTER, L.R. (1992) Modeling and forecasting US mortality. *Journal of the American Statistical Association*, **87**(419), 659–671.

LEXIS, W. (1875) *Einleitung in die Theorie der Bevolkerungsstatistik*. Strassburg: Triibner. Pages 5–7 translated to English by N. Keytz and printed, with gure 1, in Mathematical Demography (ed. D. Smith & N. Keytz). Berlin: Springer (1977).

MCKEAGUE, I.W. and UTIKAL, K.J. (1990) Inference for a nonlinear counting process regression model. *The Annals of Statistics*, **18**(3), 1172–1187.

MCKENDRICK, A.G. (1926) Application of mathematics to medical problems. *Proceedings of the Edinburgh Mathematical Society*, **54**, 98–130.

NIELSEN, J.P. and LINTON, O.B. (1995) Kernel estimation in a nonparametric marker dependent Hazard model. *The Annals of Statistics*, **23**(5), 1735–1748.

PITACCO, E., DENUIT, M. and HABERMAN, S. (2009) *Modelling Longevity Dynamics for Pensions and Annuity Business*. Oxford: Oxford University Press.

RENSHAW, A.E. and HABERMAN, S. (2006) A cohort-based extension to the Lee–Carter model for mortality reduction factors. *Insurance: Mathematics and Economics,* **38**(3), 556–570.

https://doi.org/10.1017/asb.2020.5 Published online by Cambridge University Press

RICHARDS, S.J. (2008) Detecting year-of-birth mortality patterns with limited data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **171**(1), 279–298.
VILLEGAS, A.M., KAISHEV, V.K. and MILLOSSOVICH, P. (2018) StMoMo: An R package for stochastic mortality modeling. *Journal of Statistical Software*, **84**(3), 1–38. doi:10.18637/jss.v084.i03.
VON FOERSTER, H. (1959) *The Kinetics of Cellular Proliferation*. New York: Grune & Stratton.
WILMOTH, J.R., ANDREEV, K., JDANOV, D., GLEI, D.A., BOE, C., BUBENHEIM, M., PHILIPOV, D., SHKOLNIKOV, V. and VACHON, P. (2007) *Methods Protocol for the Human Mortality Database*. Berkeley and Rostock: University of California and Max Planck Institute for Demographic Research. http://mortality.org. [version 31/05/2007], **9**, 10–11.

ALEXANDRE BOUMEZOUED (CORRESPONDING AUTHOR)
Milliman R&D
14 Avenue de la Grande Armée
75017 Paris, France
E-mail: alexandre.boumezoued@milliman.com

MARC HOFFMANN
CEREMADE, CNRS-UMR 7534
Universite Paris Dauphine
Place du maréchal de Lattre de Tassigny
75016 Paris, France
E-mail: hoffmann@ceremade.dauphine.fr

PAULIEN JEUNESSE
CEREMADE, CNRS-UMR 7534
Universite Paris Dauphine
Place du maréchal de Lattre de Tassigny
75016 Paris, France
E-mail: jeunesse@ceremade.dauphine.fr

APPENDICES

The purpose of these appendices is to detail the reasoning underlying the death rate estimate formulas as successively developed in the Versions 5 and 6 of the HMD. We emphasize that the full methodology on building mortality tables is far more complete (including high-ages extrapolation and inter-census estimates), and we only focus here on the death rate estimate formula, which is of interest in this paper.

APPENDIX A. VERSION 5 ESTIMATES

The reasoning for the approximation of the exposure-to-risk is made into two steps:

1. The first step is to assume that no death occurs in the square (x, t). As it is already noticed in Wilmoth *et al.* (2007) (Version 5 Methods Protocol), the distribution of births for

https://doi.org/10.1017/asb.2020.5 Published online by Cambridge University Press

each cohort is key to understand its contribution to the exposure-to-risk. It is therefore assumed that births are uniformly distributed to obtain an average contribution of an individual from any cohort of 1/2. Then, at first order, the exposure-to-risk writes

$$\frac{1}{2}\left(N(x,t)+N(x+1,t)\right).$$

2. In practice, deaths occur so that there is a need for accounting of two adjustments: first, individuals who died in the upper triangle are missed by the population estimate $N(x+1,t)$, so one needs to add their contribution to the exposure-to-risk. Second, individuals who died in the lower triangle are associated with a contribution 1/3, so it is required to subtract the time from death until the end of the period. The assumption is then made that deaths are uniformly spread in each triangle, and a straightforward computation (see again Wilmoth *et al.* (2007)) shows that such positive or negative contribution is equal to 1/3, then the final formula for the exposure-to-risk writes:

$$\widehat{E}(x,t)=\frac{1}{2}\left(N(x,t)+N(x+1,t)\right)+\frac{1}{3}\left(D_U(x,t)-D_L(x,t)\right). \tag{14}$$

APPENDIX B. VERSION 6 ESTIMATES

The release of the new Version 6 in February 2018 has introduced two major methodological changes: the calculation of mortality rates at age zero (more precisely, mean age at death), as well as the exposure adjustment which is of interest in this paper. The following is discussed in the new Version 6 Methods Protocol: *Until now, we used the classic approach which assumes that births are uniformly distributed throughout the calendar year. In the event of a sharp discontinuity in the monthly distribution of births within a calendar year, this assumption results in the incorrect estimation of population exposures and induces false cohort effects on mortality surfaces when these surfaces are based on Lexis squares.* It is worth mentioning that the corrected cohorts are those for which monthly birth counts are available; the development of a correction method in the absence of fertility data appears as an interesting working direction which is left for further research.

In the following, we detail the calculations leading to the estimate given in Equation (13) for the exposure-to-risk $E_L(x,t)$ in the lower triangle (x,t); the steps are similar for the upper triangle formula. This reasoning is, again, performed in two steps:

1. In a first step, it is assumed that no death occurs in the lower triangle and moreover that the density of individuals with exact age x in year t is identical to the birthday density. In other words, it is assumed that the birthdays distribution is similar at any age within the cohort, here neglecting the heterogeneity in the death rate paths of individuals within the same cohort, which likely leads to nonzero cumulative mortality differentials as defined in Equation (8). Under these two assumptions, the representative individual has exact age x at time m_{t-x}, the average time of birth in cohort $t-x$, then is under exposure a time $1-m_{t-x}$ in the lower triangle, leading to the first-order estimate of the exposure-to-risk:

$$N(x,t)(1-m_{t-x}).$$

2. The second step to account for death occurrence in the lower triangle starts first with the characterization of the death distribution, denoted by $\gamma_{x,t}(a,s)$ for $(a,s)\in T_L(x,t)$. Under the assumption that the birthdays distribution remains the same at any age and

https://doi.org/10.1017/asb.2020.5 Published online by Cambridge University Press

moreover that deaths are uniformly distributed, the following formula is obtained under the constraint that $\int_{T_L(x,t)} \gamma_{x,t} = 1$:

$$\gamma_{x,t}(a,s) = \frac{f_{t-x}(s-a)}{1-m_{t-x}},$$

where f_{t-x} is the birthdays distribution density; we also introduce the cumulative distribution function F_{t-x} for the next calculation, such that $f_{t-x} = F'_{t-x}$. Finally, it remains to compute the lost exposure, that is, for all death occurrences, the time from death to the exit from the lower triangle. For a death occurrence at $(x+a, t+s)$ in the lower triangle, the time to exit is $1-s$, leading to the following calculation for the lost exposure:

$$\begin{aligned}
&\int_0^1 \int_0^s \gamma_{x,t}(a,s)(1-s)\mathrm{d}a\mathrm{d}s \\
&= \frac{1}{1-m_{t-x}} \int_0^1 (1-s) \int_0^s f_{t-x}(s-a)\mathrm{d}a\mathrm{d}s \\
&= \frac{1}{1-m_{t-x}} \int_0^1 (1-s) F_{t-x}(s)\mathrm{d}s,
\end{aligned}$$

where we note that $\int_0^1 F_{t-x}(s)\mathrm{d}s = 1-m_{t-x}$ and, by integration by parts,

$$\int_0^1 sF_{t-x}(s)\mathrm{d}s = \frac{1}{2}\left\{1 - \int_0^1 s^2 f_{t-x}(s)\mathrm{d}s\right\} = \frac{1}{2}\left\{1 - \left(\sigma^2_{t-x} + m^2_{t-x}\right)\right\}.$$

By rearranging the two terms, one obtains

$$\int_0^1 \int_0^s \gamma_{x,t}(a,s)(1-s)\mathrm{d}a\mathrm{d}s = \frac{1-m_{t-x}}{2} + \frac{\sigma^2_{t-x}}{2(1-m_{t-x})}.$$

Finally, the total exposure-to-risk can be approximated as follows as a result from steps 1 and 2, and using Equation (6):

$$\begin{aligned}
\widehat{E_L}(x,t) &= N(x,t)(1-m_{t-x}) - D_L(x,t)\left(\frac{1-m_{t-x}}{2} + \frac{\sigma^2_{t-x}}{2(1-m_{t-x})}\right) \\
&= P(x,t+1)(1-m_{t-x}) + D_L(x,t)\left(\frac{1-m_{t-x}}{2} - \frac{\sigma^2_{t-x}}{2(1-m_{t-x})}\right).
\end{aligned}$$

https://doi.org/10.1017/asb.2020.5 Published online by Cambridge University Press