A new inference strategy for general population mortality tables

Alexandre Boumezoued, Marc Hoffmann, Paulien Jeunesse

A new inference strategy for general population mortality tables

Alexandre Boumezoued[1], Marc Hoffmann[2], Paulien Jeunesse[3]

April 22, 2018

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 links 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 (2008), Cairns et al. (2016) and Boumezoued (2016).

Keywords: Mortality tables, general population, statistical inference, population dynamics, cohort effect.

JEL: 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

[1]Milliman R&D, 14 Avenue de la Grande Armée, 75017 Paris, France.
Email: alexandre.boumezoued@milliman.com

[2]CEREMADE, CNRS-UMR 7534, Universite Paris Dauphine, Place du maréchal de Lattre de Tassigny 75016 Paris, France. Email: hoffmann@ceremade.dauphine.fr

[3]CEREMADE, CNRS-UMR 7534, Universite Paris Dauphine, Place du maréchal de Lattre de Tassigny 75016 Paris, France. Email: jeunesse@ceremade.dauphine.fr

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 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 (non-homogeneous 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 with 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 births 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 (2016) 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 non-

parametric perspective by Beran (1981), Dabrowska (1987), Keiding (1990), McKeague and Utikal (1990), Nielsen and Linton (1995), Brunel et al. (2008), 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 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), 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 one-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 Human Mortality Database, whereas the Human Fertility Database is a public source providing in particular number of births by months for several countries. Such population, death and fertility data allows 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) has to be reached by means of national institutes.

The paper is organized as follows. In Section 2, we present the non-homogeneous birth-death model and derive the inference strategy - the related interpretations and link with existing estimators is discussed in Subsection 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 Non-homogeneous 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 $g(a,t)$ the population density at (a,t), a non-negative 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 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 birth 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 $\times$ age coordinates. In an ideal demographic world, two kinds of population estimates are recorded in the one-year age $\times$ 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 Human Mortality Database is given in Figure 1. This can be analysed 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 between birth patterns from one year to the next is of interest, as it introduces some bias in the classical formulas used in practice for death rate estimation.

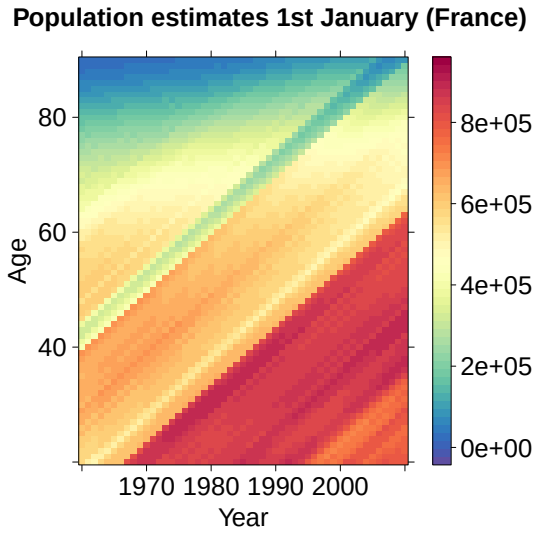

Figure 1: Population estimates for France by year for one-year age classes extracted from the Human Mortality Database

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 $\times$ times 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 death in the Lexis triangles can be written

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

An illustration of death counts in the Lexis triangles (x,t) for the French population extracted form the Human Mortality Database is represented in Figure 2. Variations in number of deaths are closely linked to those of the underlying exposure (Figure 1) but also to the death rate itself, to be estimated.

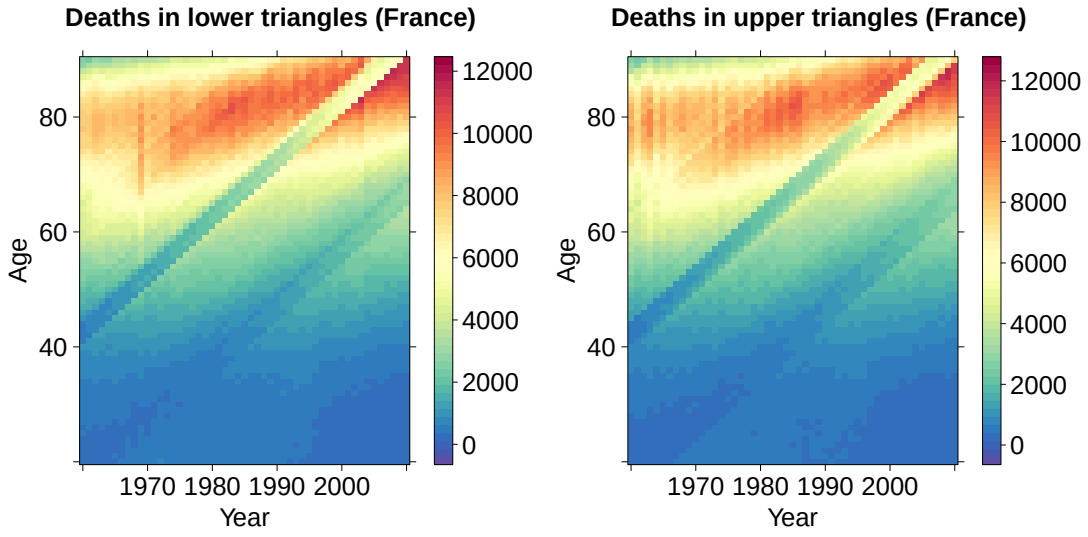

Figure 2: Death counts in Lexis triangles extracted from the Human Mortality Database

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

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

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

In addition to population estimates and death counts, as analyzed by Cairns et al. (2016) and Boumezoued (2016), we aim at including birth counts by month in the inference process - these can be extracted from the Human Fertility Database 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 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.

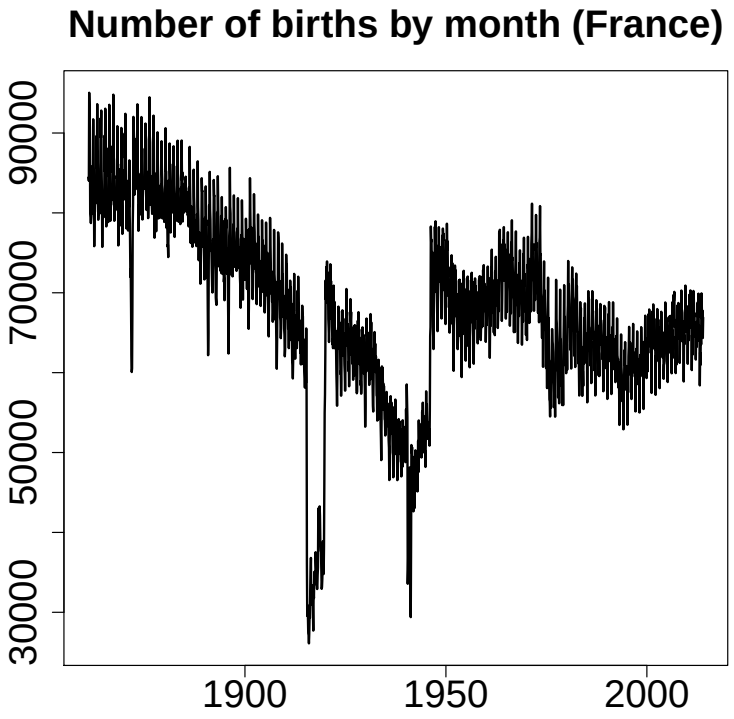

Figure 3: Number of birth by month extracted from the Human Fertility Database

2.3 Death rate inference

When two time-dependent dimensions are involved (here age and calendar time), the natural generalization of classical non-parametric estimates of the death rate is not direct (see again the discussion in Keiding (1990)), therefore smoothing is required - see e.g. 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 to use aggregate information by year and age-class to derive (approximate) estimators.

In the classical demographic and actuarial practice, it is considered two versions of general population mortality tables: period and cohort. We propose here a brief discussion of these two versions and refer the reader to Boumezoued (2016) 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.

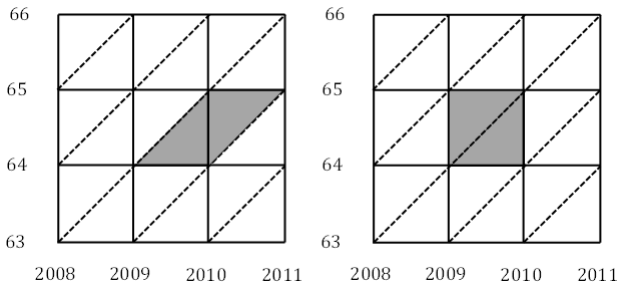

Figure 4: Population used (in grey) for the computation of the cohort death rate (left) and period death rate (right) for age 64 and year 2009.

Overall, period and cohort tables provide complementary information and their use is driven by the underlying objective. In this paper, we illustrate our 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 of 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 the death rate to be 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 equal to μ_T, it follows that:

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

As the left hand side is the opposite of the number of deaths as introduced in Equation (5), it follows from the previous equation that 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

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

are the so-called 'exposures-to-risk' in the lower and upper triangle respectively.

Now, the number of deaths in Lexis triangles being observed (as provided by the Human Mortality Database), it remains to appropriately compute the exposure-to-risk. In the literature dedicated to longevity studies, this quantity is approximated by annual observables, see e.g. 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 Human Mortality Database methodology allowing to include monthly birth data is further discussed in Subsection 2.6. The standard annual approximation can be illustrated for period tables (see Subsection 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 $\frac{1}{2}\left[N(x,t)+N(x+1,t)\right] = \frac{1}{2}\left[P(x,t)+P(x+1,t)\right]+\frac{1}{2}\left[D_L(x,t)-D_U(x,t)\right]$, which leads to another possible approximation. Note that the Version 5 estimates of the Human Mortality Database rely on a demographic reasoning leading to an approximation in between the two previous ones - see the analysis in Boumezoued (2016) 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 Subsection 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:

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

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:

$$H(x,t) = \sum_{y=0}^{x-1} \mu_U(y, t-x+y+1) - \mu_L(y, t-x+y),\ x \in \mathbb{N}^*. \tag{7}$$

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:*

$$\exp\left(-\mu_L(x,t)\right) L_{t-x}\big(H(x,t)-\mu_L(x,t)\big) = \left(1-\frac{D_L(x,t)}{N(x,t)}\right) L_{t-x}\big(H(x,t)\big), \tag{8}$$

and

$$\begin{aligned} &L_{t-x-1}\big(H(x,t-1)-\mu_L(x,t-1)\big) \\ &= \left(1+\frac{D_U(x,t)}{N(x+1,t)}\right) L_{t-x-1}\big(H(x,t-1)-\mu_L(x,t-1)+\mu_U(x,t)\big). \end{aligned} \tag{9}$$

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

2.5 Proof of Proposition 1

To prove (8), 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:

$$E_L(x,t) = \int_0^1 g(x,t+v)\frac{1-\exp\left((v-1)\mu_L(x,t)\right)}{\mu_L(x,t)}\mathrm{d}v. \tag{10}$$

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

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

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

$$\begin{aligned} &g(x,t+v) = g(0,t-x+v)\exp\left(-\int_0^x \mu(u,t-x+v+u)\mathrm{d}u\right) \\ &= g(0,t-x+v)\exp\left(-\sum_{y=0}^{x-1}\int_y^{y+1}\mu(u,t-x+v+u)\mathrm{d}u\right) \\ &= g(0,t-x+v)\exp\left(-\sum_{y=0}^{x-1}\int_y^{y+1-v}\mu(u,t-x+v+u)\mathrm{d}u - \sum_{y=0}^{x-1}\int_{y+1-v}^{y+1}\mu(u,t-x+v+u)\mathrm{d}u\right) \\ &= g(0,t-x+v)\exp\left(-(1-v)\sum_{y=0}^{x-1}\mu_L(y,t-x+y) - v\sum_{y=0}^{x-1}\mu_U(y,t-x+y+1)\right) \\ &= S(x,t)g(0,t-x+v)\exp\left(-vH(x,t)\right), \end{aligned} \tag{11}$$

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 (7). 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 (8):

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

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

The proof of (9) 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 (7):

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

2.6 Discussion

Exposure-to-risk interpretation. The equality (10) 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 (11) 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 $H(x,t) = \sum_{y=0}^{x-1} \mu(y, t-x+y+1) - \mu(y, t-x+y)$ 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 $\mu(y, t-x+y+1) = \mu(y, t-x+y)$, the H matrix vanishes, so that the initial birthdays distribution perfectly propagates towards highest ages.

Closed population assumption. Due to the renormalization in the final result (8), the death rate relates to the closest annual population estimate; therefore, the assumption that the population is closed is only local in terms of population count, as the population estimate N may include population flow effects. Also, the assumption of a closed population implies here that the birthdays distribution at some age is obtained as a transformation of the initial birth distribution - to this extent the assumption applies globally in each cohort.

Link with estimates of the Human Mortality Database. It is worth mentioning that at the time of writing, the Human Mortality Database 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 HMD Methods Protocol.

From (10), it can be shown by performing a first order expansion in $\mu_L(x,t)$ that

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

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 $\sigma^2_{t-x} := \mathrm{Var}(B_{t-x})$.

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 HMD methods protocol:

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

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):

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

3 Numerical results

Based on Proposition 1, one can exhibit a recursive and implicit scheme for computing the death rates, as described below.

Algorithm 1. *For age x starting at zero:*
(i) *Solve Equation (8) 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 (9) 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 modify 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 re-use of the methodology which will provide an update of the mortality rates at $(y,t+y-x)$ for $y \geq x$.

In Figures 5 to 8, we depict the death rate estimates obtained with the method developed in this paper applied to French data sourced from the Human Mortality Database (annual population estimates, Figure 1 and number of deaths in Lexis triangles, Figure 2) and the Human Fertility Database (births by months, Figure 3). The number of births by month are 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 (2016) for further details):

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

Each figure includes on the right the ratio between the new and the old estimate, 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%. To assess this specificity, 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 (2016).

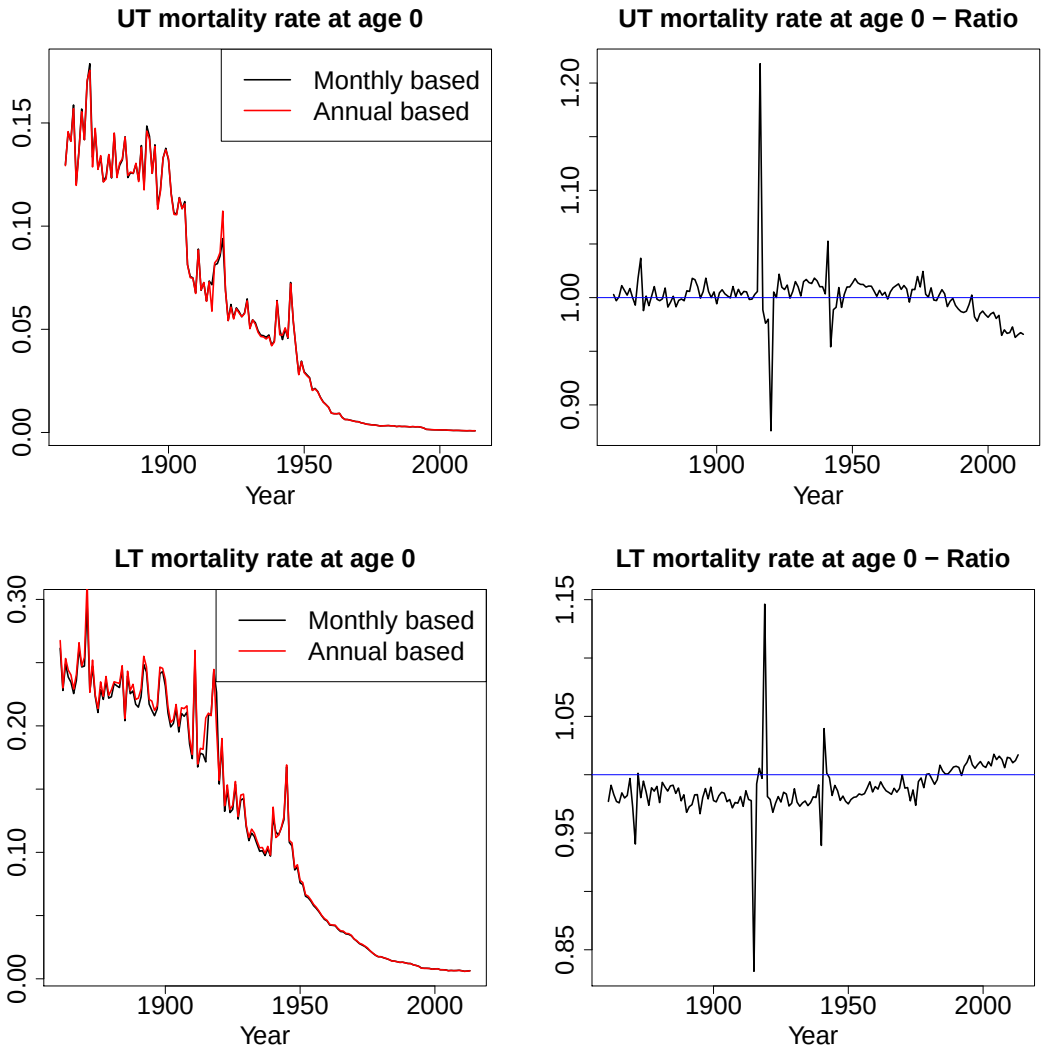

Figure 5: Left: death rates estimated based on the new inference method (in black), and compared to estimates using the standard method based on annual population records (in red). Right: ratio between new and old estimates. Top: Upper triangle. Bottom: Lower triangle.

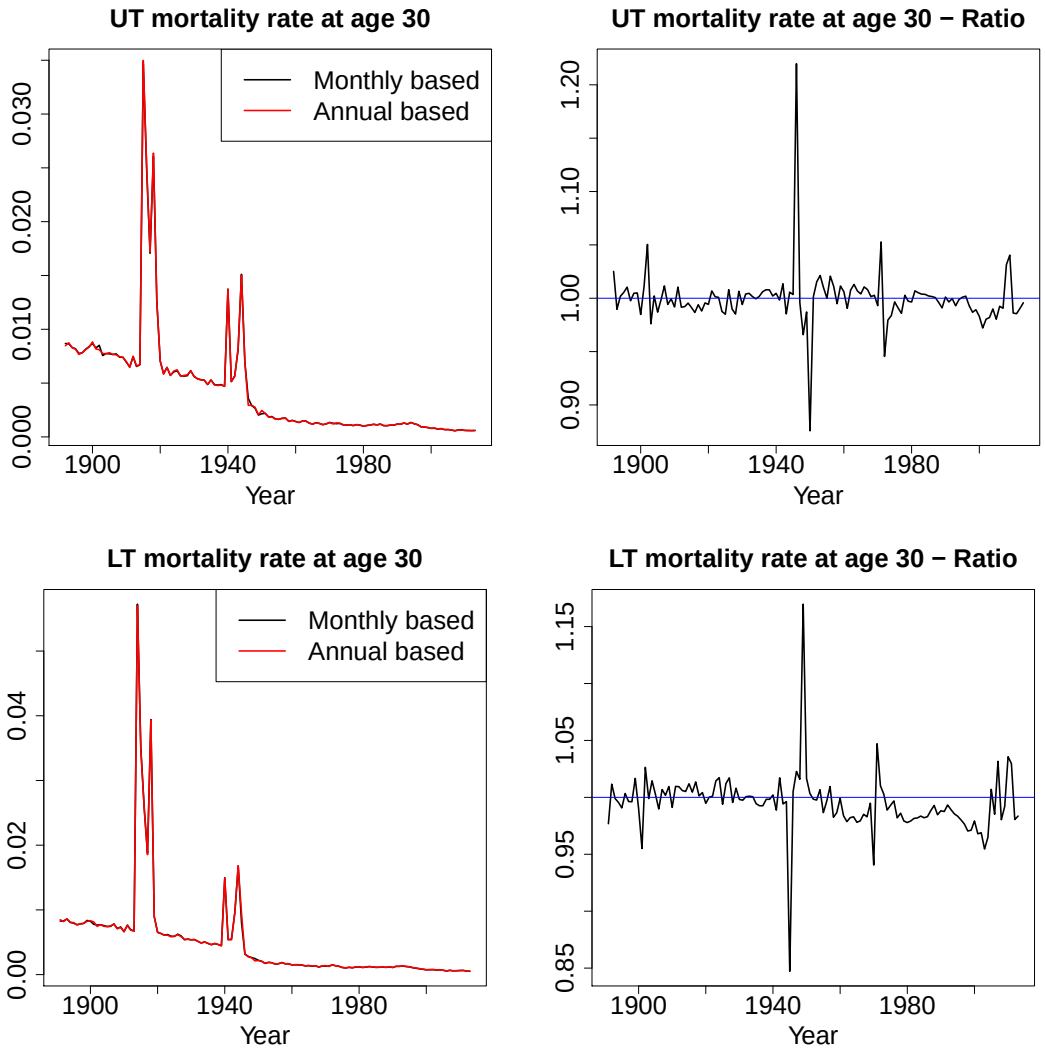

Figure 6: Left: death rates estimated based on the new inference method (in black), and compared to estimates using the standard method based on annual population records (in red). Right: ratio between new and old estimates. Top: Upper triangle. Bottom: Lower triangle.

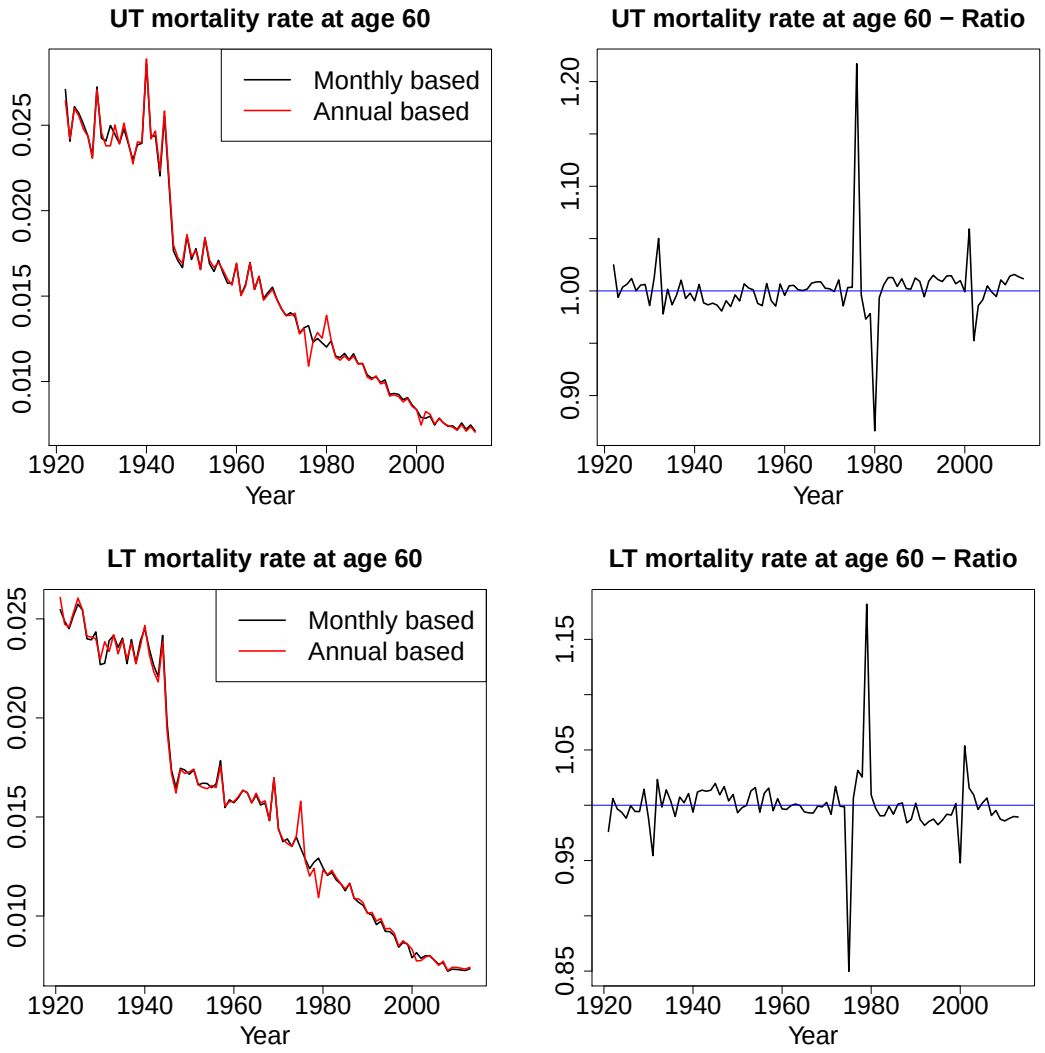

Figure 7: Left: death rates estimated based on the new inference method (in black), and compared to estimates using the standard method based on annual population records (in red). Right: ratio between new and old estimates. Top: Upper triangle. Bottom: Lower triangle.

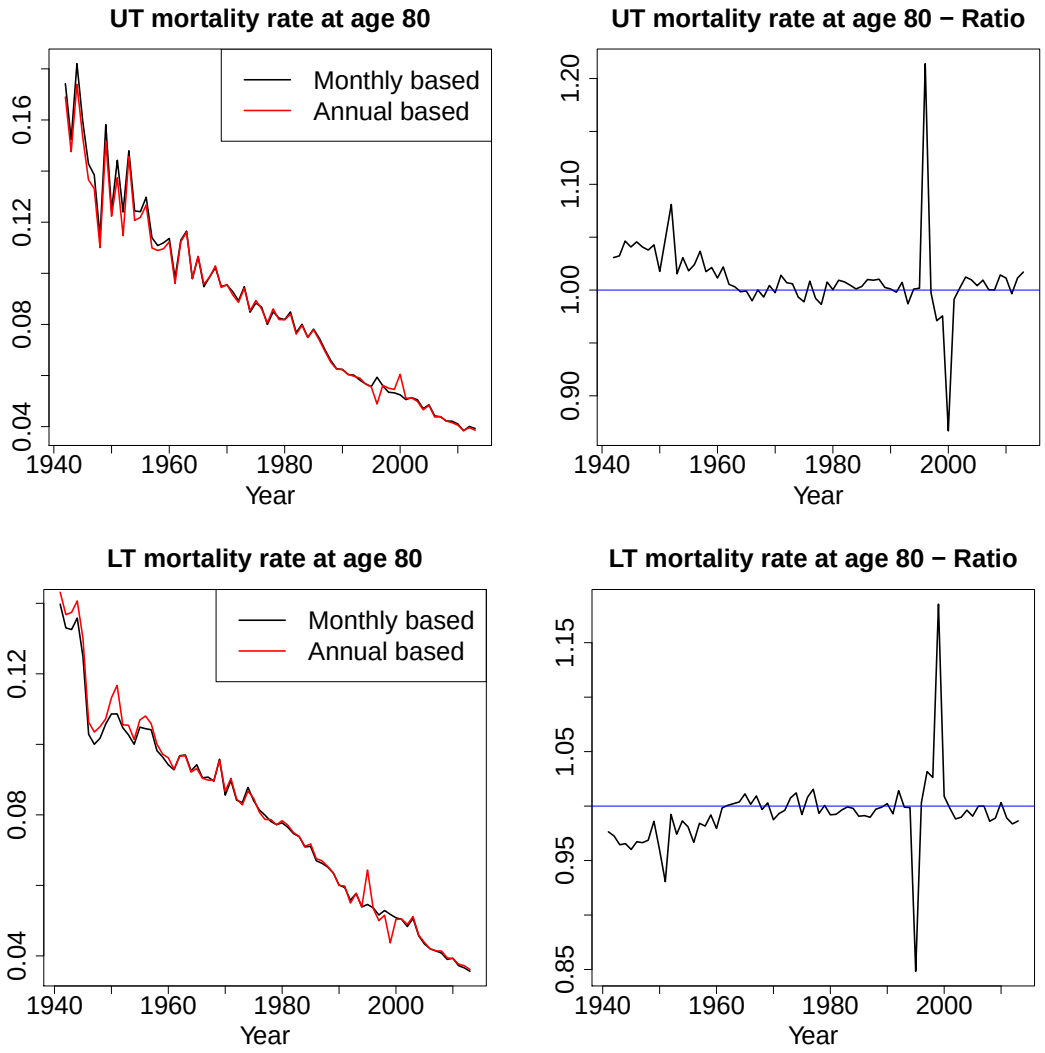

Figure 8: Left: death rates estimated based on the new inference method (in black), and compared to estimates using the standard method based on annual population records (in red). Right: ratio between new and old estimates. Top: Upper triangle. Bottom: Lower triangle.

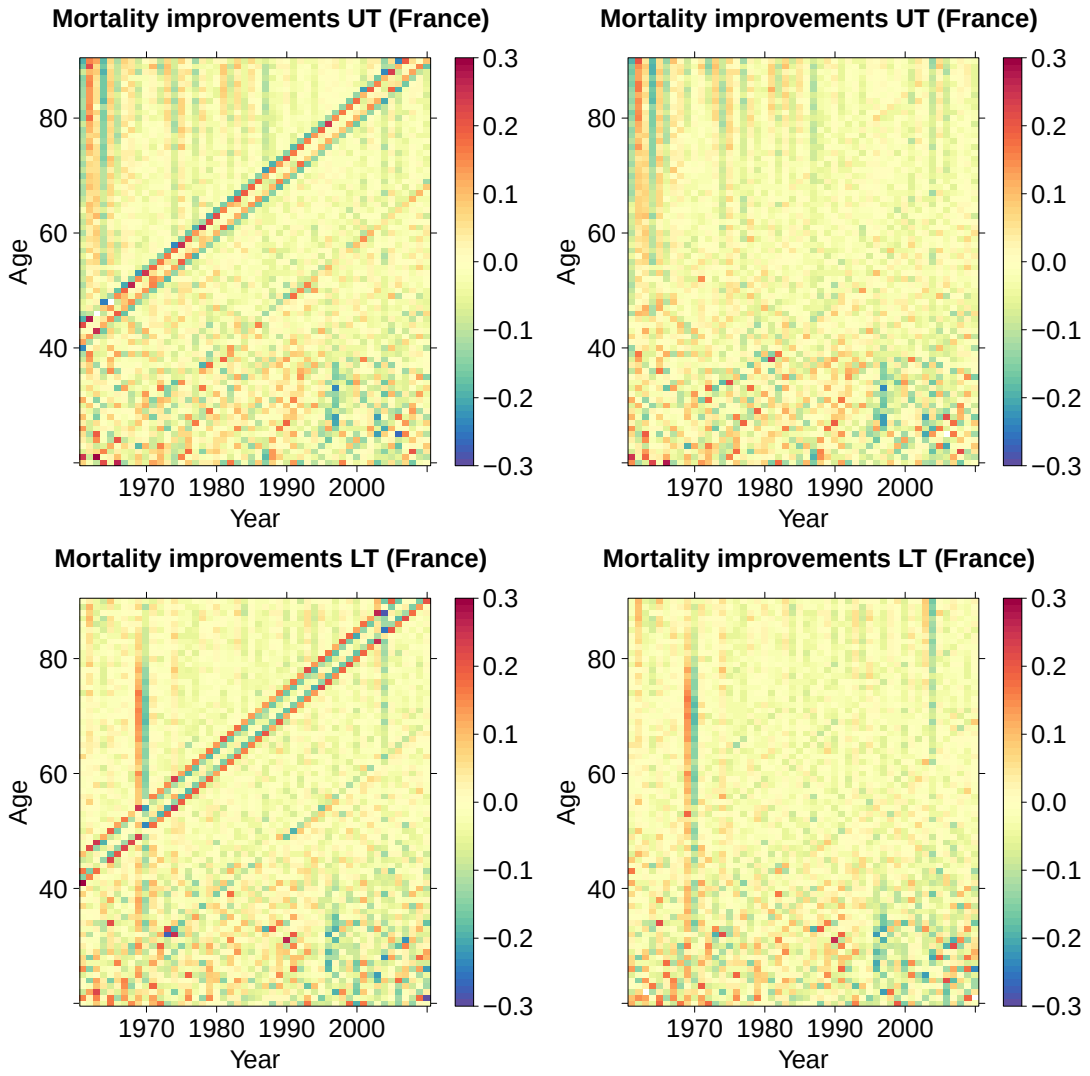

Figure 9: Left: mortality improvement rates using the standard method based on annual population records. Right: mortality improvement rates using the new inference method. Top: upper triangles. Bottom: lower triangles.

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 with 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 the previous contributions by Richards (2008), Cairns et al. (2016) and Boumezoued (2016).

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 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, Rudolf. 1981. Nonparametric regression with randomly censored survival data. Tech. rep., Technical Report, Univ. California, Berkeley. 3

Boumezoued, Alexandre. 2016. Improving HMD mortality estimates with HFD fertility data. *To appear in the North American Actuarial Journal* . 1, 2, 6, 8, 10, 15, 20

Boumezoued, Alexandre, Marc Hoffmann, Paulien Jeunesse. 2018. Statistical inference for an in-homogeneous age-structured population process. *Forthcoming* . 2

Brunel, Elodie, Fabienne Comte, Agathe Guilloux. 2008. Estimation strategies for censored lifetimes with a lexis-diagram type model. *Scandinavian Journal of Statistics* **35**(3) 557–576. 3

Cairns, Andrew JG, David Blake, Kevin Dowd, Amy R Kessler. 2016. Phantoms never die: living with unreliable population data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **179**(4) 975–1005. 1, 2, 6, 15, 20

Clémençon, Stéphan, Viet Chi Tran, Hector De Arazoza. 2008. A stochastic SIR model with contact-tracing: large population limits and statistical inference. *Journal of Biological Dynamics* **2**(4) 392–414. 3

Comte, Fabienne, Stéphane Gaïffas, Agathe Guilloux. 2011. Adaptive estimation of the conditional intensity of marker-dependent counting processes. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, vol. 47. Institut Henri Poincaré, 1171–1196. 3

Dabrowska, Dorota M. 1987. Non-parametric regression with censored survival time data. *Scandinavian Journal of Statistics* 181–197. 3

Daley, DJ, D. Vere-Jones. 2003. *An Introduction to the Theory of Point Processes. Volume I: Elementary Theory and Methods of Probability and its Applications*. Springer, New York. 2

Doumic, Marie, Marc Hoffmann, Nathalie Krell, Lydia Robert. 2015. Statistical estimation of a growth-fragmentation model observed on a genealogical tree. *Bernoulli* **21**(3) 1760–1799. 3

HFD. 2018. The human fertility database. max planck institute for demographic research (germany) and vienna institute of demography (austria). URL `www.humanfertility.org`. 2

HMD. 2018. The human mortality database. university of california, berkeley (usa), and max planck institute for demographic research (germany). URL `www.mortality.org`. 2

Hoffmann, Marc, Adélaïde Olivier. 2016. Nonparametric estimation of the division rate of an age dependent branching process. *Stochastic Processes and their Applications* **126**(5) 1433–1471. 3

Keiding, Niels. 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. 3, 7

McKeague, Ian W, Klaus J Utikal. 1990. Inference for a nonlinear counting process regression model. *The Annals of Statistics* 1172–1187. 3, 7

McKendrick, A.G. 1926. Application of mathematics to medical problems. *Proc. Edin. Math. Soc.* **54** 98–130. 4

Nielsen, Jens P, Oliver B Linton. 1995. Kernel estimation in a nonparametric marker dependent hazard model. *The Annals of Statistics* 1735–1748. 3, 7

Pitacco, Ermanno, Michel Denuit, Steven Haberman. 2009. *Modelling longevity dynamics for pensions and annuity business*. Oxford University Press. 9

Richards, SJ. 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. 1, 2, 15, 20

Von Foerster, H. 1959. *The Kinetics of Cellular Proliferation*. Grune & Stratton. 4

Wilmoth, John R, Kirill Andreev, Dmitri Jdanov, Dana A Glei, C Boe, M Bubenheim, D Philipov, V Shkolnikov, P Vachon. 2007. Methods protocol for the human mortality database. *University of California, Berkeley, and Max Planck Institute for Demographic Research, Rostock. URL: http://mortality. org [version 31/05/2007]* **9** 10–11. 9, 15