Improving HMD mortality estimates with HFD fertility data

Alexandre Boumezoued

Improving HMD mortality estimates with HFD fertility data[1]

Alexandre Boumezoued[2]

February 8, 2016

Abstract

Following the work of [6], we aim at correcting mortality estimates based on fertility data. As already conjectured by [13], the computation of exposure to risk can suffer from errors for cohorts born in years in which births are fluctuating. In this context, we first point our attention to the Human Mortality Database [8], the reference mortality data provider. While comparing period and cohort mortality tables, we highlight the presence of anomalies in period ones in the form of isolated cohort effects. Our investigation of the HMD methodology exhibits a strong assumption of uniform distribution of births that is specific to period tables, therefore likely to be at the core of the asymmetry between both. Based on the idea of [6] regarding the construction of kind of "data quality indicator", we make a new and intensive exploitation of the Human Fertility Database [7], which is from our point of view a crucial source as it represents the perfect counterpart of the HMD in terms of fertility. This indicator is then used to construct corrected period mortality tables for several countries, which we analyze on both an historical and prospective point of view. Our main conclusions relate to the reduction of volatility of mortality improvement rates, the impact in the use of cohort parameters in stochastic mortality models, as well as a better fit of corrected tables by classical mortality models.

Keywords: Human Mortality Database, Cohort effect, Death rate, Exposure to risk, Human Fertility Database, Births by month.

1 Introduction

Since its launch in 2002, the Human Mortality Database [8] has become the reference provider of mortality estimates given in an homogenous format for several

[1]This work has been funded by, and carried out at Milliman (Paris Office).

[2]Milliman, 14 rue Pergolèse, 75016 Paris, France
& Laboratoire de Probabilités et Modèles Aléatoires (LPMA), UMR CNRS 7599, Universités Paris 6 & 7, France.
Email: alexandre.boumezoued@milliman.com

countries. These national indicators are extensively used by researchers as demographers, as well as practitioners in insurance companies, as an input to get insights on uncertain future mortality. A huge amount of literature has been dedicated to the sophistication of stochastic mortality models in the past decades, see e.g. [11] and [4]. In their philosophy, such models consider that future mortality rates are random, and the analysis of past mortality rates (often taken from HMD) in terms of their age, time and possibly cohort directions will help to extract the time series driving the mortality pattern, which can be then (randomly) extrapolated.

In comparison, few contributions focused on the reliability of demographic data, and particularly of mortality estimates taken as inputs for historical analysis or predictive forecasts. To our knowledge, the first insights have been suggested by [13], and from our point of view [6] proposed a founding work on this direction. The conjecture of [13] was focused on the 1919 birth cohort for England & Wales for which he suggested the possibility of errors due to erratic number of births. Such conjectures took a concrete form as the ONS (Office for National Statistics) produced corrected tables, in fact a mortality increase for this 1919 cohort, particularly at high ages. The ONS methodology has then been studied by [6] in several directions, who proposed an approach to illustrate and correct mortality tables, applied to the data for England & Wales; the *Convexity Adjustement Ratio* introduced in their work will be of particular interest in this paper.

The common characteristics of such contributions are interesting to highlight, as they emerge from a joint questioning of
A) the demographic causes of *cohort effects* observed as some generations present particular mortality levels/improvements, and
B) the link between mortality and fertility, considering observed aggregate mortality as the result of a whole demographic process.
On this topic, let us also mention other contributions that address the joint impact of mortality and fertility on agregate demographic quantities, such as [2], [1] and [3] and that are source of inspiration for the present study.

While studying the construction of mortality tables, one faces the two components that are at the core of death rate computation: the number of deaths, often reliable, divided by the so-called *exposure to risk* which represents the *quantity* of individuals at risk of death, in other words the total time lived by the population in the period considered. The exposure to risk component is usually approximated based on annual population estimates, as it is done in the Human Mortality Database, and this requires different assumptions depending on the kind of table considered. Indeed, in the HMD mortality tables are provided both on a *cohort* or *period* basis. As cohort tables provide death rates each time computed using individuals of the *same generation* (i.e. born the same year), period tables provide death rates com-

puted using individuals observed the *same year*. In particular, period tables are well suited to understand the time behavior of mortality, and are the structural inputs of stochastic mortality models. The counterpart is that a single death rate is computed while mixing two cohorts; linked to this structure, an additional assumption is required for period tables to approximate the exposure to risk, which takes the form in HMD in assuming a uniform distribution of births.

In this paper, we are interested in the robustness of such assumption and investigate the interaction between the exposure to risk and fertility data in this context. Coming back to the conjecture of [13] and the work on [6] for England & Wales, we want to correct the observed anomalies in HMD period mortality tables based on fertility data at a refined time scale.

From our point of view, the Human Fertility Database [7], started in 2009, is the suitable candidate to address this issue since it represents in its structure the perfect counterpart of the HMD in terms of fertility. The data of interest here is the number of births by month, and we select five countries according to their particularly deep fertility histories: France, Switzerland, Finland, Sweden and Austria. The time origin of such fertility histories is crucial as it represents the first generation (year of birth) that can be corrected. Also, as we will see, the most important anomalies concern four groups of cohorts: around 1915, 1920, 1940 and 1945. As already mentioned, this corresponds to periods in which the number of births is erratic. As these groups are detected for almost the five countries considered, we argue that we are facing some universal issue regarding the construction of mortality tables; the whole correction process for other countries with limited fertility histories is a challenging topic that is left for further research.

The paper is organized as follows. In Section 2, we compare period and cohort mortality tables for specific generations, and we highlight the presence of anomalies in period ones in the form of several isolated cohort effects. We then detail the HMD methodology to compute period and cohort death rates, and we highlight the strong assumption of uniform distribution of births that is specific to period tables, therefore likely to be at the core of the asymmetry between both. In Section 3, we use a rigorous mathematical population framework with continuous age and time in order to properly define and analyze population estimates at stake, as well as to locate the HMD method in the range of integral approximations. This Section 3 is at the core of the paper as it then presents the Human Fertility Database used for the construction of the correction ratio which aims at detecting cohorts that present anomalies as well as to provide corrected period mortality tables. Finally, we analyze in Section 4 the corrected mortality tables and draw their main features; in particular, they do not present the initial anomalies, and in consequence offer lower levels of empirical volatility computed on mortality improvements. We also

discuss the way corrected tables change our use of classical stochastic mortality models, in particular regarding the cohort parameter, and we illustrate how these models are now fitting and reproducing better the data that is corrected. The paper ends with some concluding remarks.

2 The Human Mortality Database: data and methods

The Human Mortality Database (see [8]) provides mortality estimates given in an homogenous format for several countries. These national indicators are seen as references and extensively used by researchers as demographers, as well as practitioners in insurance companies. HMD provides mortality tables of estimated death rates by age and time. In this paper, we focus on such tables for the finest time and age scale available, that is given each year and by one-year age classes. Two kinds of mortality tables are available, giving either cohort death rates or period death rates that are described and shown in the following.

2.1 Cohort and period mortality tables

The estimation of death rates by age and time is a statistical challenge when the two crossing components are continuous, see the formalism and discussion in Section 3. In practice, individuals are regrouped by several blocks depending on their age, time and time of birth on the basis of a space partition: the so-called Lexis diagram (see [12]), represented in a simplified version in Figure 1. Note that on this Figure, the numbers represent the exact ages or times; as an example, if we refer to the *year 2008*, this corresponds to the line between 2008 and 2009. Three degrees of freedom are at the core of this representation:

- the time component: time at which an individual is observed,
- the age component: age of the individual,
- the cohort component: time at which the individual is born.

The difference between cohort and period death rates for a given year t and a given one-year age-class x, denoted respectively $\mu_C(x,t)$ and $\mu_P(x,t)$, relies on the choice of the two degrees of freedom to be fixed among the three described above. For the period death rate $\mu_P(x,t)$, one regroups the individuals whose ages lie in the age-class x at any time of the year t, assuming the death rate to be constant on a square. In this case, two cohorts are mixed: those born in year $t-x$, as well as those born in year $t-x-1$, therefore the individuals belong to two distinct

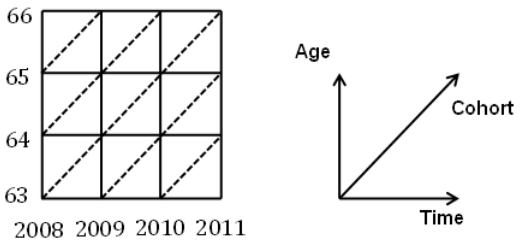

Figure 1: Simplified version of the Lexis diagram (left) and the associated three degress of freedom (right)

generations. On the contrary, cohort mortality rates $\mu_C(x,t)$ are computed so that the individuals concerned belong to the same cohort whereas their age still lies in the one-year age-class x; the obverse is that the times at which they are observed will exceed year t. In this case, the death rate is assumed to be constant on a parallelogram. This is illustrated in Figure 2 for the computation of the death rate for year 2009 and age-class 64.

On the whole, period death rates are particularly interesting to study the dynamics of mortality over time, whereas cohort death rates are better designed to study the age pattern of mortality of the same generation.

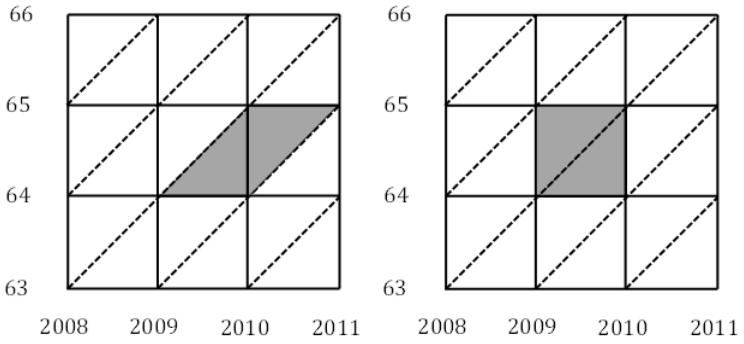

Figure 2: Population used (in grey) for the computation of cohort death rates (left) and period death rates (right) in the Lexis diagram

2.2 Looking at mortality tables from the HMD

Let us first focus on the surface of period death rates, here displayed for France in Figure 3. We represent here the rate for the total population (male and female), which will be the case of interest in this paper; but note that the discussions, results and methodology can be duplicated in a separated analysis of male and female population. Several aspects regarding the dynamics of mortality rates can be described from such graph. First, one notices the level of infant mortality (age zero), then the increase of mortality from intermediate ages to high ages in an exponential shape,

as well as the reduction of the level of mortality for the several age classes over time. Lastly, one can observe the impact of catastrophic events as the First World War combined with the Spanish flu, and the Second World War.

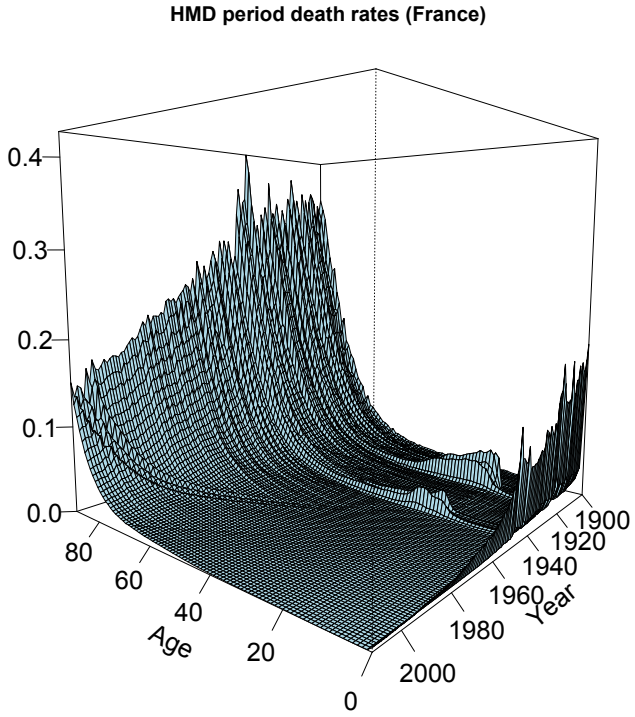

Figure 3: Period death rates for France for year 1900 to 2010 and age from 0 to 90

To better understand the dynamics of mortality rates over time for the several age classes, one often looks at mortality improvement rates. As discussed before, period death rates are well designed to study the time pattern of mortality, therefore improvement rates are computed based on period mortality tables. For any year t and age-class x, the mortality improvement rate $r(x,t)$ is computed from the period mortality rate $\mu^p(x,t)$ as

$$r(x,t) = \frac{\mu_P(x,t+1) - \mu_P(x,t)}{\mu_P(x,t)}. \tag{1}$$

Therefore, the improvement rate $r(x,t)$ measures the evolution in time of the age-dependent mortality, and is as such often negative since mortality is generally decreasing over time. These are depicted for France in Figure 4. Other crucial mortality patterns appear based on the observation of mortality improvement rates in the diagonal, that is while following a given cohort. Mortality improvements appear to be particularly low (red) or high (green/blue) for specific generations, in order from the lowest to the highest diagonal:

- Individuals aged 40 around year 1980, that is born around 1940,

- Individuals aged 40 around year 1960, that is born around 1920,
- Individuals aged around 45 in year 1960, that is born around 1915.

Note that for other countries, such phenomenon will also be detected for the generations born around 1945. Several demographic contributions have been dedicated to the explanations of so-called *cohort effects*, that is the fact that some generations present particularly high/low morality improvement rates. Aim of this paper is not to provide a list of such contributions, rather to show that such effects represent anomalies in the computations of death rates due to erratic fertility patterns at year of birth, as already suggested by [13], and also worked by [6] for England & Wales based on quarterly birth data.

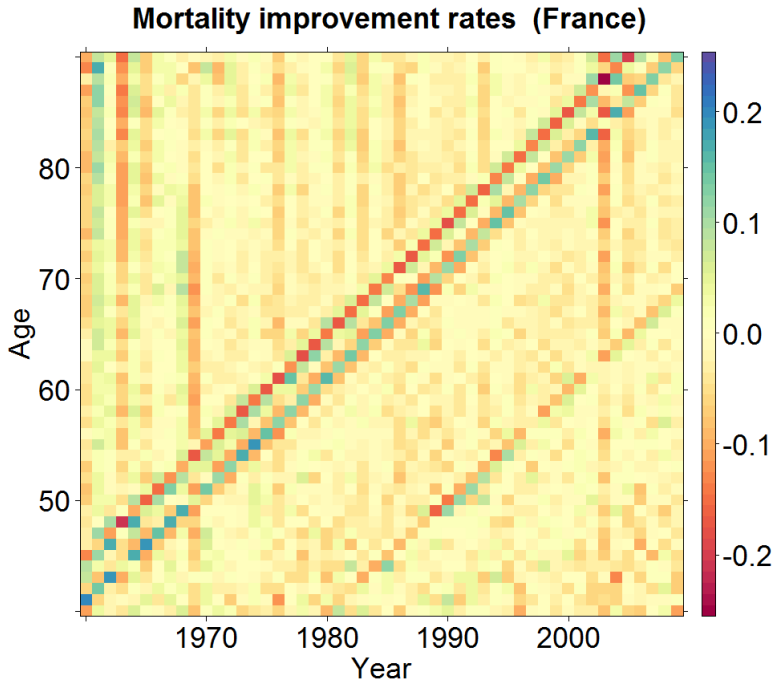

Figure 4: Period death rates for France for year 1960 to 2010 and ages from 40 to 90

Focus on the 1919-1920 birth cohorts A way to better understand the issues regarding mortality improvements is to focus on specific cohort concerned, as 1919-1920. In the following, we compare the mortality rates of such cohorts for both period and cohort HMD mortality tables. This comparison is depicted in Figure 5. Recall that for cohort data this corresponds to real generations, whereas for period data we represent a diagonal starting with age zero at the year of birth considered, see again Figure 2. First, the observation of the right panel (period mortality rates) in Figure 5 explains the shape of the mortality improvement rates displayed in Figure 4: a downward jump from 1918 to 1919, then an increase from 1919 to 1920, then

again a downward jump from 1920 to 1921. Note that these orders of magnitude are not intuitive regarding demographic insights. On this topic, the comparison with cohort data is instructive, see now the left panel in Figure 5. Indeed, for cohort data some regular mortality improvement is observed from 1918 to 1922. This comparison is depicted for the other countries we consider in this paper in Figure 6: Switzeland, Finland, Sweden and Austria. Let us recall that this choice is due to their special fertility records, see the discussions in Sections 1 and 3. Based on these observations, noting also that cohort and period data share common sub-populations in their computation, see Figure 2, we argue at this stage that some structural anomaly is at the origin of such difference. To address this issue, there is a need to go into the details of the construction of cohort and period mortality tables in the Human Mortality Database. This is investigated in the next subsection.

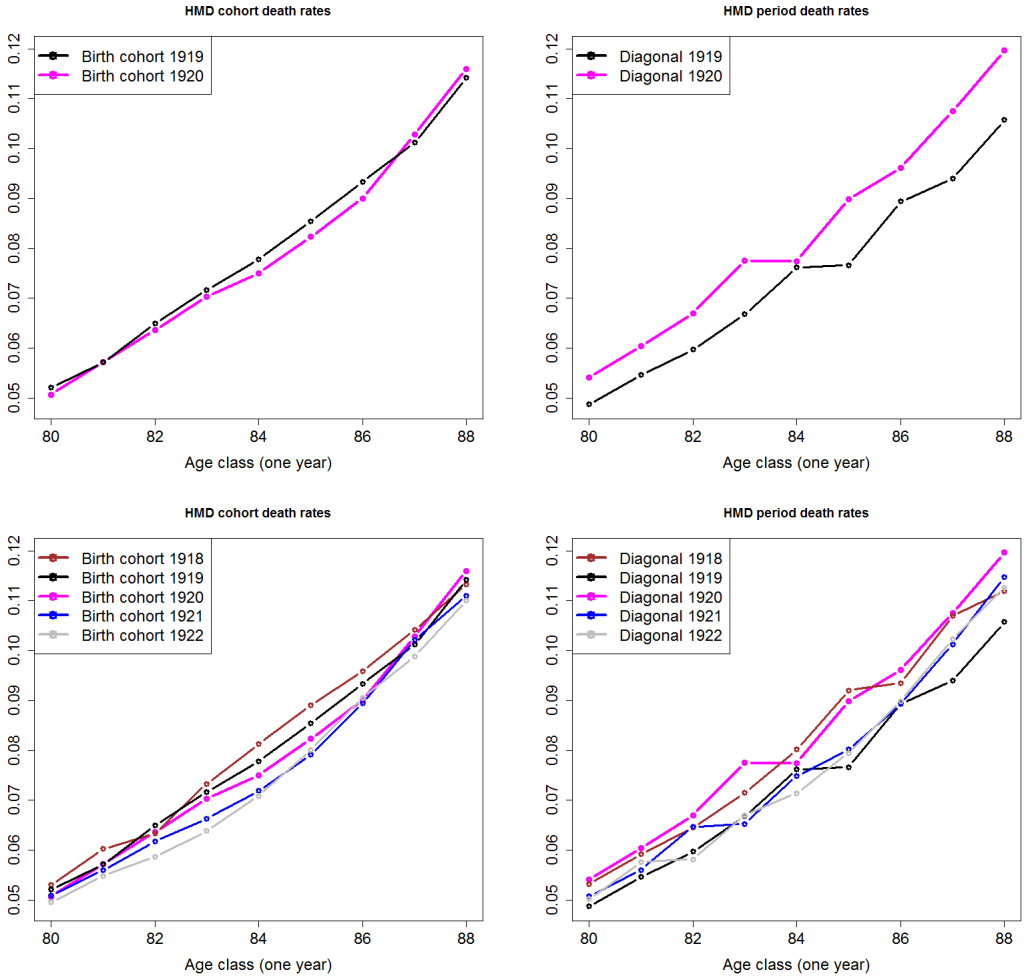

Figure 5: Left: mortality rates for the birth cohorts 1918 to 1922 from HMD cohort data. Right: mortality rates for diagonals starting between 1918 and 1922 from HMD period data.

2.3 HMD methodology to compute cohort and period death rates

The aim of this part is to detail the HMD methodology as it is described in the associated technical note, see [16]. The mathematical interpretation based on the full continuous age and time formalism will be detailed in Section 3.

At the beginning, the aim is to estimate the mortality rate, assumed to be constant in each square (period setting) or parallelogram (cohort setting). As classical, the corresponding estimators are computed as

$$\widehat{\mu_P}(x,t) = \frac{D_P(x,t)}{E_P(x,t)} \text{ and } \widehat{\mu_C}(x,t) = \frac{D_C(x,t)}{E_C(x,t)},$$

where $D_P(x,t)$ (resp. $D_C(x,t)$) is the number of deaths in the square (resp. parallelogram), and $E_P(x,t)$ (resp. $E_C(x,t)$) is the famous *exposure to risk*, that is the total time lived by individuals in the square (resp. parallelogram). While observing a given population, counting the number of deaths, and measuring the total time lived, it is worth mentioning that these numbers only allow to compute an *estimator* $\widehat{\mu}(x,t)$ of the *true* death rate $\mu(x,t)$, and also that this estimator is computed under the assumption of a constant death rate in a square or parallelogram.

From censuses, one can compute the number of deaths, and it is reasonable to think that these numbers are accurate, although several approximation methods have to be used to recompose the number of deaths in the case where the information is not available between several years, see the discussion in [16]. In all this paper, we assume that the number of deaths is computed carefully with no errors and we focus on the computation of exposures to risk.

To compute the exposure to risk on a surface (square or parallelogram), one has to measure the total time lived by individuals. Unfortunately, this can not be measured exactly since the population is not *continuously* observed, see the discussion in Section 3. Therefore, the HMD methodology relies on fundamental quantities that are measured annually (or approximated, but as well we assume that HMD numbers regarding population estimates are accurate).

In the one-year age $\times$ time square, two kind of population estimates are recorded, see Figure 7:

- The number of individuals at *exact* time t, with age x last birthday, denoted $P(x,t)$,
- The number of individuals who attained *exact* age x in the year $[t, t+1)$, denoted $N(x,t)$.

Also, the number of deaths $D_P(x,t)$ in the square is split between the upper (U) and the lower (L) triangle, respectively denoted $D^U(x,t)$ and $D^L(x,t)$, so that

$$D_P(x,t) = D^U(x,t) + D^L(x,t).$$

From the observation of the square, two fundamental relations appear, that can be proved rigorously with the mathematical formalism introduced in Section 3:

$$\begin{aligned} N(x+1,t) &= P(x,t) - D^U(x,t), \\ P(x,t+1) &= N(x,t) - D^L(x,t). \end{aligned} \tag{2}$$

Remark 1. *The fundamental equalities of Equation (2) are valid for a closed population, that is without any migration flow. This will be used by HMD in the reasoning to establish the final formulas for the approximation of the exposure to risk. Note however that straightfoward consequences of the fundamental relations as* $P(x,t+1) = P(x-1,t) - D^U(x-1,t) - D^L(x,t)$ *are not verified numerically based on population estimates and death counts in HMD. This comes from the fact that although the reasoning is performed for a closed population, the population estimates take into account to some extent the way population numbers fluctuate. In this paper, we also derive reasonings without migration flows, but note that the analysis of this aspect in the HMD methodology could be of interest as well.*

At this stage, the main question arises: as the exact exposure can not be computed, how to approximate it by means of these quantities? To address this issue, let us first go back to the cohort case. We will then exhibit the additional assumption needed for the period case to be likely at the source of the asymmetry between period and cohort mortality tables. Note that the quantities introduced previously, namely P, N, D^U and D^L are common to both the cohort and the period setting, see Figure 8.

We now describe the HMD approximation for the *cohort* exposure to risk as detailed in [16], see Figure 8. The reasoning for the approximation of the exposure to risk is made in two steps:

1. Let us first assume that no deaths occur in the parallelogram. In this case, it is easy to compute the exposure to risk: each individual of the cohort lives one year while aging in diagonal through the parallelogram. As in the case individuals reach the vertical barrier, they are numbered as $P(x,t+1)$, this quantity is the main component of the exposure to risk.

2. In practice, deaths occur so there is a need for accounting of two adjustments: first, individuals who died in the lower triangle are missed by the population estimate, so there is a need to add their contribution to the exposure to risk. Second, individuals who died in the upper triangle are counted for 1 at this stage so it is needed to substract the time from death until the end of the period. In HMD, the assumption is made that deaths are uniformly spread in each triangle, and a straightforward computation (see again [16]) show that

such positive or negative contribution is equal to 1/3, so that in the end the exposure to risk is approximated as:

$$\widehat{E_C}(x,t) = P(x,t+1) + \frac{1}{3}\left(D^L(x,t) - D^U(x,t+1)\right). \tag{3}$$

Let us now focus on the HMD approximation for the *period* exposure to risk, and reproduce the two main steps of the reasoning, see Figure 7.

1. The problem arising here is that in the first step, assuming no deaths, the time lived by individuals is not one year in general: this strongly depends on the distribution of individuals of each cohort over the year. For example, let us focus on the oldest cohort, that is the cohort of individuals going through the upper triangle. Assume that all these individuals are born at exact time t and nowhere else; in this extreme case, all individuals from this cohort live the total diagonal that is one year. On the contrary, let us assume that this generation is born at exact time t; in this case, its contribution to the exposure to risk is zero. In between these extreme cases, one has to know about the distribution of births for each cohort as the contribution to the exposure to risk strongly depends on it. In HMD, see again [16], an assumption of uniform distribution of births for every cohort is made. In this case, the *average* contribution of an individual from any cohort is 1/2, which, from our interpretation, is nothing but the area of each triangle. Then, if the horizontal barriers are considered as references (see Remark 2 below), the main component of the exposure to risk thus writes
$$\frac{1}{2}\left(N(x,t) + N(x+1,t)\right).$$

2. As in the cohort setting, one has to correct the main component from the deaths in the square. More precisely, for the oldest cohort one has to add the contribution of those died in the upper triangle; also, for the youngest cohort one has to substract some time due to deaths in the lower triangle. With the same elementary contribution 1/3, this leads to
$$\widehat{E_P}(x,t) = \frac{1}{2}\left(N(x,t) + N(x+1,t)\right) + \frac{1}{3}\left(D^U(x,t) - D^L(x,t+1)\right). \tag{4}$$

A last step is made by HMD to convert the N quantities into P quantities; from the fundamental relations in Equation (2), one gets:

$$\begin{aligned}
\widehat{E_P}(x,t) &= \frac{1}{2}\left(P(x,t+1) + D^L(x,t) + P(x,t) - D^U(x,t)\right) + \frac{1}{3}\left(D^U(x,t) - D^L(x,t)\right) \\
\widehat{E_P}(x,t) &= \frac{1}{2}\left(P(x,t+1) + P(x,t)\right) + \frac{1}{6}\left(D^L(x,t) - D^U(x,t)\right).
\end{aligned} \tag{5}$$

This is the final formula used in HMD for the approximation, as we can numerically check starting from population estimates and Lexis death counts as given in the database. A rigorous mathematical interpretation will be discussed in Section 3.

Remark 2. *It is interesting to note here the sensitivity of the formulas derived depending on the reasoning considered. Indeed, instead of using population estimates N in step 1, let us proceed similarly with population estimates P. In this case, see Figure 7, the first step of the reasoning leads to a main component of the exposure to risk equal to $(P(x,t) + P(x,t+1))/2$, while assuming no deaths and uniform births. In the second step, we adjust this estimate to the deaths in both triangles: this remains to add the average contribution to the exposure per death in the lower triangle and to substract the average lost exposure in the upper triangle, which is in each case equal to $1/3$ under the assumption of uniform distribution of deaths, see again [16]. In the end, the obtained formula is*

$$\frac{1}{2}\left(P(x,t+1) + P(x,t)\right) + \frac{1}{3}\left(D^L(x,t) - D^U(x,t)\right).$$

Note that this is slightly different from the formula derived in Equation (5); both will be interpreted in terms of integral approximation in Section 3.

At this stage, the asymetry between period and cohort exposure to risk computation appears: this corresponds to the assumption of uniform distribution of births in the period setting, whereas the cohort framework does not need such additional assumption. Based on the observations in Subsection 2.2, as well as on the previous insights by [13] and [6], we argue that such assumption is likely to be the source of errors for several diagonals in period mortality tables. In fact, this issue is already reported in the HMD technical note itself, see again [16], which we reproduce below:

"This assumption [of uniform distribution of births] is violated most severely in situations where there are rapid changes in the size of successive cohorts, owing to fluctuations in the birth series many years before. The worst situation is when a sharp discontinuity in births occurs in the middle of one calendar year, creating a cohort that is "heavy" at one end and "light" at the other. We have not attempted to correct our mortality estimates for the error introduced by such occurrences, which may result in artificially elevated or depressed levels of mortality along a diagonal of the Lexis diagram that follows the cohort(s) in question. The user should be aware of this possibility and not misinterpret the data."

3 Correcting population exposure with fertility data

This section is dedicated to the mathematical formalism as well as the diagnose/ correction methodology relying on the Human *Fertility* Database. We first detail in

Subsection 3.1 the underlying contiuous age and time population framework; this allows us to highlight the impact of births in the whole dynamics, to rigorously define the population estimates as well as the death counts at stake, and give an integral approximation interpretation of the formula used in the HMD. Although we think that this theoretical part is crucial to get a whole understanding of the issue as well as for further research on this topic, we allow the (hurried) reader to directly go to Subsections 3.2 and 3.3 which present the fertility data from the HFD as well as the associated methodology to diagnose and correct period mortality estimates. The corrected tables will be analyzed in Section 4.

3.1 Continuous framework and population estimates

Continuous age-time population framework Whereas we deal with one year age classes x and years t to characterize mortality rates, the mortality surface described by the Lexis diagram is structured into continuous age and time axes. To make the difference, we denote a a continuous age variable and s a continuous time variable, with $a \in \mathbb{R}_+$ and $s \in \mathbb{R}_+$. At this stage, and in the rest of this paper, we deal with deterministic population densities, that is $g(a,s) \in \mathbb{R}_+$ denotes the quantity of individuals with exact age a at exact time s. This can be seen as the density of individuals in an infinite population, or the average number of individuals in a given (finite) population evolving stochastically over time; see [2] for further discussions on this aspect, especially the link between deterministic and stochastic population dynamics. In this framework, for any fixed time $s \in \mathbb{R}_+$, the trajectory $(g(a, s+a))_{a\geq 0}$, or the map $a \mapsto g(a, s+a)$, is the rigorous representation of what is called the birth cohort s, that is the evolution of the population of all individuals born at exact time s. Note also that with fixed time s as well, the component $(g(a,s))_{a\geq 0}$ represents the well known *age pyramid* as it gives the repartition of individuals by age. Finally, now with fixed age $a \in \mathbb{R}_+$, the component $(g(a,s))_{s\geq 0}$ characterizes the evolution of populations at the same age over time.

Let us introduce the *true* mortality rate $\mu(a,s)$ for exact age a and exact time s; it drives the evolution of the birth cohort s by: for each $a \in \mathbb{R}_+$,

$$g(a, s+a) = g(0,s) \exp\left(-\int_0^a \mu(u, s+u)\mathrm{d}u\right), \tag{6}$$

where $g(0,s)$ represents the individuals with age zero (born) at time s, and with $S(a,s) := \exp\left(-\int_0^a \mu(u,s+u)\mathrm{d}u\right)$ the survival function at age a for the individuals born at time s. From this equation, it is clear that the cohort dynamics depends on the number of newborn in a crucial way.

Let us write the number of deaths of individuals aged a at time s in any small

time interval with length Δu as

$$g(a,s) - g(a+\Delta u, s+\Delta u) \approx -(\partial_a + \partial_s)g(a,s)\Delta u.$$

From Equation (6), one can then prove that

$$(\partial_a + \partial_s)g(a,s) = -\mu(a,s)g(a,s). \tag{7}$$

In this form, the equation is the first basic component of the classical population dynamics model of Mc Kendrick and Von Foerster (see [10] and [14]). This is called the transport component in the field of partial differential equations as is states that age is translated along the time axis, in other words that individuals are ageing.

Theoretically, another component is needed, since in itself Equation (6) embeds some unknown quantity $g(0,s)$. But in fact, such estimates of the number of newborn in a given country can be found for a number of countries, see Section 3, and this *transport* component will be sufficient in itself four our study. Nevertheless, we mention the second component in the following remark.

Remark 3. *The second component of the Mc Kendrick-Von Foerster population model is referred to as the renewal component: given a birth rate $b(a,s)$ for age a and exact time s, the number of newborn is computed as the sum over all possible parents as, for each $s \in \mathbb{R}_+$,*

$$g(0,s) = \int_0^\infty g(a,s)b(a,s)\mathrm{d}a. \tag{8}$$

With Equations (7) and (8) one is able to analyze theoretically the (deterministic) population dynamics, which is beyond the scope of the present paper.

In the following, we express the well known exposure to risk in terms of the underlying continuous age and time population, and we discuss its approximation at a refined time scale.

Period and cohort mortality estimates The estimation of death rates with two crossing continuous dimensions (here age and time) is a statistical challenge. In particular, there is no standard non-parametric technique to tackle this problem without making assumption of a piecewise constant mortality rate, see e.g. [9] for a discussion on this topic. In the period framework, the assumption is made that the mortality rate is constant equal to some $\mu_P(x,t)$ in each square with left-lower point $(x,t) \in \mathbb{R}_+^2$, that is:

$$\text{for each } (a,s) \in [x,x+1)\times[t,t+1),\ \mu(a,s) = \mu_P(x,t).$$

In the cohort framework however, this assumption is made on each parallelogramm with left-lower point $(x,t) \in \mathbb{R}_+^2$, which can be rigorously written as

$$\text{for each } a \in [x,x+1), \text{for each } s \in [t+a-x, t+a-x+1),\ \mu(a,s) = \mu_C(x,t),$$

since individuals have their time of birth in the year $t-x$, i.e. mathematically $s-a \in [t-x, t-x+1)$. Let us remark the difference between the two assumptions, especially the dependence in age of the time interval in the cohort setting.

Now, let us address the implication of each assumption on Equation (7). Let us introduce the regions "square (x,t)" $R_P(x,t)$ and "parallelogramm (x,t)" $R_C(x,t)$ in the period and cohort settings respectively as:

$$R_P(x,t) = [x, x+1)\times[t, t+1) \text{ and } R_C(x,t) = \{(a,s) : a \in [x, x+1),\ s-a \in [t-x, t-x+1)\}$$

Let us integrate Equation (7) for ages and times in the regions $R_P(x,t)$ and $R_C(x,t)$ respectively[1], and let us first focus on the period setting. On the left hand side, we obtain $D_P(x,t) = \int_{(a,s)\in R_P(x,t)} -(\partial_a + \partial_s)g(a,s)\mathrm{d}a\mathrm{d}s$ which is nothing but the total number of deaths in the square. On the right hand side, we get, using the assumption of a constant death rate on the square:

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

therefore equalling the left and right hand sides in the period setting it follows that

$$D_P(x,t) = \mu_P(x,t)E_P(x,t),$$

with $E_P(x,t) = \int_x^{x+1}\int_t^{t+1} g(a,s)\mathrm{d}s\mathrm{d}a$ the famous exposure to risk in the period setting.

With the same reasoning, in the cohort setting, with the corresponding assumption, we get

$$D_C(x,t) = \mu_C(x,t)E_C(x,t),$$

with $E_C(x,t) = \int_x^{x+1}\int_{t+a-x}^{t+a-x+1} g(a,s)\mathrm{d}s\mathrm{d}a$ and $D_C(x,t) = \int_{(a,s)\in R_C(x,t)} -(\partial_a+\partial_s)g(a,s)\mathrm{d}a\mathrm{d}s$.

Integral approximation of the exposure to risk Let us now investigate the approximation of the exposure to risk in the period setting, which is by construction sensitive to the repartition of births in successive years. From a mathematical point of view, this problem reduces to the approximation of the two-dimensional integral

$$E_P(x,t) = \int_x^{x+1}\int_t^{t+1} g(a,s)\mathrm{d}s\mathrm{d}a. \tag{9}$$

Traditionally, this problem is tackled by using values of the two-dimensional function at some collection of ages and times (a_i, s_i), and weights w_i so that e.g.

$$E_P(x,t) \approx \sum_i w_i g(a_i, s_i)(s_{i+1}-s_i)(a_{i+1}-a_i).$$

[1] Both sides are multiplied by -1 to get the (positive) number of deaths.

However, in practice, demographic information for joint isolated times and ages is not available. Instead, one has access to *population estimates* as described in Subsection 2.3, such as the population estimate with age in a one year age class $[x, x+1)$ at an *exact* time t, which we denoted $P(x,t)$, or the estimation of the number of individuals reaching *exact* age x in a year t, which we denoted $N(x,t)$, all notations in accordance with the description of the HMD methodology.

Let us express these quantities in terms of the underlying population. First, the quantity $P(x,t)$ can be obtained by summing all ages in the age class considered, that is rigorously:

$$P(x,t) = \int_x^{x+1} g(a,t)\mathrm{d}a. \tag{10}$$

Also, as $g(x,s)$ is the number of individuals reaching exact age x at exact time s, we deduce that

$$N(x,t) = \int_t^{t+1} g(x,s)\mathrm{d}s. \tag{11}$$

Recall that the number of deaths can be split into upper and lower triangles. As in Subsection 2.3, for any square with left-lower point (x,t), we denote $D^L(x,t)$ the number of deaths in the lower triangle and $D^U(x,t)$ the number of deaths in the upper triangle. Let us finally rigorously define such quantitites. First, deaths in the upper triangle concern individuals aged a at time s so that $a \in [x, x+1)$, $s \in [t, t+1)$ and that belong to the cohort born in year $t-x-1$, i.e. such that $s - a \in [t-x-1, t-x)$, leading to

$$D^U(x,t) = \int_x^{x+1} \int_t^{t-x+a} -(\partial_a + \partial_s)g(a,s)\mathrm{d}s\mathrm{d}a.$$

Analogously, as individuals dying in the lower triangle are born in the year $t-x$, one gets

$$D^L(x,t) = \int_x^{x+1} \int_{t-x+a}^{t+1} -(\partial_a + \partial_s)g(a,s)\mathrm{d}s\mathrm{d}a.$$

It is clear that $D_P(x,t) = D^U(x,t) + D^L(x,t)$, and starting from all quantities introduced above, it is possible to prove the fundamental relations introduced in Equation (2); this proof is left to the reader.

In order to make the link between the HMD formula and classical integral approximation methods, let us go back to the double integral representation of the exposure to risk. Two representations follow from Equations (9), (10) and (11):

$$E_P(x,t) = \int_x^{x+1} N(a,t)\mathrm{d}a \text{ and } E_P(x,t) = \int_t^{t+1} P(x,s)\mathrm{d}s.$$

This leads to two *naïve* one-dimensional integral approximations:

$$\widehat{E_P}^1(x,t) = \frac{1}{2}\left(N(x,t) + N(x+1,t)\right) \text{ and } \widehat{E_P}^0(x,t) = \frac{1}{2}\left(P(x,t) + P(x,t+1)\right),$$

the first one being rewritten, according to the fundamental relations in Equation (2) as $\widehat{E_P}^1(x,t) = \frac{1}{2}\left(P(x,t) + P(x,t+1)\right) + \frac{1}{2}\left(D^L(x,t) - D^U(x,t)\right)$. Finally, let us introduce the weighted average approximation, for $\alpha \in [0,1]$,

$$\widehat{E_P}^{\alpha}(x,t) = \alpha \widehat{E_P}^1(x,t) + (1-\alpha)\widehat{E_P}^0(x,t).$$

We conclude that the HMD approximation, see Equation (5), is equal to $\widehat{E_P}^{\alpha}(x,t)$ for $\alpha = \frac{1}{3}$. The interesting point here is that a demographic reasoning combined with several assumptions (uniform deaths and births) leads to a formula which is similar to some weighted average of simple one-dimensional integral approximations.

3.2 The Human Fertility Database

In order to detect the cohorts that are sensitive to the assumption of uniform distribution of births made by the HMD, there is a need to study the time behavior of fertility at a refined time scale. Moreover, since the analysis of several countries in the HMD is at stake, there is a need for a comparable database in terms of fertility, with as the HMD an homogenous data structure between different countries.

As a suitable candidate, the Human Fertility Database [7] has been launched in 2009 with the aim to be the HMD counterpart in terms of fertility. Many kinds of fertility data are available for more that 20 countries, such as fertility rates by age and/or parity[2] in the standard form of fertility tables. For our purpose, we are rather interested in number of births over time, given at a monthly time scale.

It is now time to recall the set of countries we will focus on in this paper. We perform our analysis on countries for which deep historical records are available in the HFD, that are the following countries in order from the deepest historical record[3]: France (1861), Switzerland (1871), Finland (1900), Sweden (1911) and Austria (1914). Although for the other countries the proposed methodology is achievable, the correction of old cohorts (as 1919-1920) can not be performed with the current histories, therefore advanced statistical methodology is required, see the discussion in the concluding Section 5.

The data collection of number of births by month is depicted in Figure 9 for the five countries considered. The trajectories are interesting as they make appear several upward or downward *shocks*. These shocks are of interest as they will impact the robustness of the assumption of uniform distribution of births used in HMD. It is not the purpose of the present paper to provide detailed demographic insights on these fluctuations, rather to understand the dynamics from a modelling and *data* point of view. Let us still mention that the main shocks likely to be due to the world

[2]Number of children already born to a woman.

[3]We do not consider here HFD data that is stated as "preliminary release", see http://www.humanfertility.org/, as it is not fully processed and checked.

wars will be of interest as the main anomalies will be detected for the generations born at the beginning and the end of the two world wars, that is around 1915, 1920, 1940 and 1945.

Let us note the complex shape of these numbers depicted in Figure 9, explained by the fact that they depend on both the underlying population and the level of birth rates in a non-trivial way. Indeed, mathematically, for a time u on a monthly grid $\frac{1}{12}\mathbb{N}$, the number of births in the month u is

$$N(u) = \int_u^{u+\frac{1}{12}} g(0,s)\mathrm{d}s,$$

then the renewal Equation (8) leads to

$$N(u) = \int_u^{u+\frac{1}{12}} \int_0^{\infty} g(a,s)b(a,s)\mathrm{d}a\mathrm{d}s.$$

It is interesting to note here that the database already provides the previous quantity as a whole, therefore it is not needed to go into the details of a *population dynamics* analysis and simulation, although we think that it is a powerfull tool in other contexts of application, see the references cited in the Introduction focusing on the interaction between mortality and fertility.

3.3 Quality indicator and corrected mortality tables

Following the work of [6], we focus on the computation of the exposure to risk at age zero with two methods: an annually estimate, in line with HMD method, and an (annual) exposure to risk computation based on monthly birth data. The deviation between both will be summarized in some *correction ratio* (close to that called *Convexity Adjustment Ratio* in the work of [6]), which will then be used to correct HMD period tables assuming that birthdays are uniformly spread at each age. Moreover, this correction ratio will not depend on deaths. An advantage of such approach is that this ratio does not depend on HMD data, therefore only on HFD data. This helps avoiding any endogeneity in the methodology, while considering HFD as an external benchmarking database.

Let us go back to the continuous age and time framework introduced in Subsection 3.1, and denote $g^*(a,s)$ the population where no deaths occur, with * to be consistent with the notations for population estimates in [6], see below. Our aim is to compute the following exposure to risk at age 0, which does not include deaths, with the two methods previously described:

$$E_P^*(0,t) = \int_t^{t+1} \int_0^1 g^*(a,s)\mathrm{d}a\mathrm{d}s = \int_t^{t+1} P^*(0,s)\mathrm{d}s, \tag{12}$$

where $P^*(0,s) = \int_0^1 g^*(a,s)\mathrm{d}a$ is the population estimate of individuals with age 0 last birthday at exact time s; as no deaths are embedded, we get $g^*(a,s) = g(0,s-a)$,

where we recall that $g(0, s-a)$ is the number of individuals born at exact time $s-a$, therefore the population estimate writes

$$P^*(0,s) = \int_0^1 g(0, s-a)\mathrm{d}s.$$

Through the Human Fertility Database, one has access to monthly number of births: for a time u on a monthly grid $\frac{1}{12}\mathbb{N}$, the number of births in the month u is

$$N(u) = \int_u^{u+\frac{1}{12}} g(0,s)\mathrm{d}s.$$

Therefore the population estimates can be computed since

$$\begin{aligned} P^*(0,s) &= \int_0^1 g(0,s-a)\mathrm{d}s \\ &= \sum_{k=1}^{12} \int_{\frac{k-1}{12}}^{\frac{k}{12}} g(0,s-u)\mathrm{d}u \\ &= \sum_{k=1}^{12} \int_{s-\frac{k}{12}}^{s-\frac{k-1}{12}} g(0,v)\mathrm{d}v \\ &= \sum_{k=1}^{12} N\left(s-\frac{k}{12}\right). \end{aligned} \tag{13}$$

We are now ready to describe the three steps for the computation of the *correction ratio*:

1. For s on a monthly grid $\frac{1}{12}\mathbb{N}$, compute the population estimate $P^*(0,s)$ as

$$P^*(0,s) = \sum_{k=1}^{12} N\left(s-\frac{k}{12}\right),$$

   where we recall that $N(u)$ is the number of births in the month u. Note that this is not an approximation as the integral can be exactly split into the sum of monthly integrals, see Equation (13).

2. Approximate the annual exposure to risk defined in Equation (12) based on the previous monthly estimates as, for annual $t \in \mathbb{N}$,

$$\widehat{E_P^*}(0,t) = \sum_{i=0}^{12} w_i P^*\left(0, t+\frac{i}{12}\right), \tag{14}$$

   where the w_i are chosen in accordance with any integral approximation method. An interesting point here is that the two-dimensional integral approximation problem, see the discussion in Subsection 3.1, reduces to one dimension only.

3. The *correction ratio* is then defined for each annual $t \in \mathbb{N}$ (viewed as a year of birth) as the ratio between the previous annual exposure to risk $\widehat{E^*_P}(0,t)$ (monthly based approximation) and the main component of the HMD exposure to risk approximation in the period setting, see Equation (5), as

$$\widehat{I}(t) = \frac{\widehat{E^*_P}(0,t)}{\frac{1}{2}\left(P^*(0,t) + P^*(0,t+1)\right)}. \tag{15}$$

Note that this ratio is a monthly approximation of the *true* ratio $I(t) = \frac{E^*_P(0,t)}{\frac{1}{2}(P^*(0,t)+P^*(0,t+1))}$. If close to one, the ratio $\widehat{I}(t)$ indicates that the diagonal in the period table starting with age zero in year t is not sensitive to the assumption of uniform distribution of births. On the contrary, if $\widehat{I}(t)$ significatively differs from one, this indicates errors in the period mortatlity table for the "cohort" born in year t (again, note that it is not a real cohort as the diagonal of the period table does not concern a single generation). More precisely: if $\widehat{I}(t)$ is greater than one, we deduce that the uniform distribution of births is likely to under-estimate the exposure to risk, whereas if $\widehat{I}(t)$ is lower than one it is likely that this assumption leads to an over-estimation of the exposure to risk. Two remarks again:

- First, as in [6], all estimates in the previous step do not embed death counts, therefore in our work they only depend on HFD data: we argue that this feature is interesting to avoid any endogeneity of the proposed methodology as well as any interaction between the two databases at this diagnostic stage.
- Second, the exposure to risk computed in step 2 is still an approximation of a one-dimensional integral, but performed at a monthly time scale instead of an annual time scale; in this sense this is considered to be more accurate.

As for the weights w_i in Equation (14), we still make the same choice as in [6] in order to provide comparable results. We fix these weights in accordance to the Simpson approximation method with $w_0 = w_{12} = 1/36$, for $i \in \{2,4,6,9,10\}$, $w_i = 2/36$ and for $i \in \{1,3,5,7,9,11\}$, $w_i = 4/36$. We furthermore performed some sensitivity using uniform weights with no significative difference in the results.

Lastly, the correction of HMD period mortality tables is performed as follows: for any period death rate $\widehat{\mu_P}(x,t)$ from the HMD table, we construct the *corrected period death rate* as:

$$\widetilde{\mu_P}(x,t) = \frac{\widehat{\mu_P}(x,t)}{\widehat{I}(t-x)} = \frac{D_P(x,t)}{\widehat{E_P}(x,t)} \times \frac{\frac{1}{2}\left(P^*(0,t-x) + P^*(0,t-x+1)\right)}{\widehat{E^*_P}(0,t-x)} \tag{16}$$

since $t-x$ is the year at which individuals are aged zero along the diagonal in the Lexis diagram. From this equation, one illustrates the spirit of the methodology: by cross-product, the aim is to correct the deviation due to the assumption of uniform

distribution of births. The assumption underlying this correction lies in the fact that birthdays distribution remain the same at each age. Recall that the ratio on the left hand side comes from HMD, whereas that on the right hand side only depends on HFD.

4 Analysis of corrected period mortality tables

The correction ratios, which have been constructed based on HFD data from Equation (15) for the range of countries considered, are depicted in Figure 10. High or low values indicate a deviation due to the uniform births assumption. Even if historics vary from one country to another, we are able to identify common structural anomalies for specific cohorts, especially those born around 1915-1920 and 1940-1945.

To go further, we now aim at analyzing the corrected mortality tables constructed as detailed in Equation (16). This analysis will be performed in two directions:

- first, an observation of the main features of both original and corrected historical data (retrospective analysis),
- second, the discussion of the impact of such correction on the way we use stochastic mortality models (prospective analysis).

4.1 Retrospective analysis

Let us first focus again on the 1919 and 1920 generations, in order to capture the *new* orders of magnitude between both. The results are depicted in Figure 11 with cohort HMD data, period HMD data, and corrected period data for our five countries of interest (France, Switzerland, Finland, Sweden and Austria). As seen on this graph, the natural order between the two generations is restored, which is coherent with reasonable demographic insights. As well, we note several similarities between cohort and corrected period data, although the comparison can not be made fully since the populations used for their computation are different, see the discussion in Section 2.

In order to analyze the impact on all observed *isolated* cohort effects, it is interesting to recompose the mortality improvement rates, see Equation (1); these are depicted for both HMD data and HFD-corrected data in Figure 12. As crude HMD data exhibit isolated cohort trends, those disappear with the correction using HFD data. This is a main conclusion of this paper, that the use of the ratio linked to the assumption of uniform distribution of births leads to corrected tables that do not present such isolated cohort effects. Let us note that the range of colors is different for each country, and that the corresponding range of values is highly dependent of the order of magnitude of the population size, as smaller populations lead to a

wider range of possible values due to sampling (or "demographic") risk. Therefore, for smaller countries, the isolated cohort trends are less identifiable. Another good feature of the correction ratio here appears as it allows to detect the cohort anomalies even for smaller countries.

Remark 4. *Of course, other kind of cohort effects can be exhibited and are likely to exist from a demographic point of view, but they usually concern generations born in a wider range of years. One can think for example of the Golden Cohort in UK, see in particular demographic explanations in [15], as well as a population modeling analysis in [2] based on a remark of the previously cited paper linking the cohort effect with fluctuating birth patterns in an heterogenous population. It is interesting to note that these other kind of cohort effects have also been studied in the light of fertility issues, which way of taking into account the whole demographic process seems from our point of view a promizing direction in this context as well.*

Finally, let us focus on classical features of mortality improvements: their empirical mean and standard deviation. These are depicted in Figure 13. First, for all countries, the mean of mortality improvements (left column) is preserved in corrected data; this is coherent with the values of the correction ratio which is centered around 1 (see again Figure 10); moreover this shows that the whole methodology amounts to re-distribute part of the population estimate year by year from one cohort to the other. Second, let us now focus on their empirical standard deviation (right column). Another main conclusion is that for all countries, their empirical volatility is reduced, especially for high ages. Note that the impact is lower for higher range of values, and that the impact is particularly high for France; note especially that the shape of volatility of mortality improvements is now closer to that of the other countries. Indeed, as already noticed, the range of values of such mortality improvements depends on the country considered; this drives in particular the level of the mean as well as that of the volatility. As expected, the more sampling risk (smaller countries), the less the *relative* impact of data correction on volatility (with theoretically the same values for the correction ratio). We conclude this retrospective Subsection with Figure 14 which presents a focus on the volatilities for the high age range from 70 to 90. From this graph, it is clear that although the level differs from one country to another, for almost each age the volatility is reduced in the corrected data.

4.2 Impact for stochastic mortality modeling

As new mortality tables are constructed, it is now time to address the impact on stochastic mortality modelling. It is worth mentioning that in the demographic and actuarial literature, a huge amount of contributions have been dedicated to the

increasing sophistication of stochastic mortality models. The stochastic modelling framework for mortality, introduced by [11], has soon been followed in the actuarial litterature, see e.g. [5] and [4], and references therein. In their philosophy, such models consider that future mortality rates are random, and the analysis of the past values in terms of their age, time and possibly cohort directions will help to extract the time series driving the mortality pattern, which can be then (randomly) extrapolated.

For our discussion, we focus on high ages in the range 60-90. In this context, we make choice of a specific model which is detailed in the following; we consider a slightly modified version of the "M7" model described in [4] as it will provide the crucial basic features we want to analyze here:

$$\ln\left(\mu_P(x,t)\right) = \kappa_1(t) + \kappa_2(t)\left(x-\bar{x}\right) + \kappa_3(t)\left((x-\bar{x})^2 - \widehat{\sigma_x}^2\right) + \epsilon(x,t) \qquad (17)$$

where $\bar{x}$ is the mean age over the age range considered, $\widehat{\sigma_x}^2$ is the mean of $(x-\bar{x})^2$, κ_1, κ_2 and κ_3 are the time series driving the mortality dynamics, and the $\epsilon(x,t)$ are centered residuals. Such model will be calibrated with either $\widehat{\mu_P}(x,t)$ (crude HMD data) or $\widetilde{\mu_P}(x,t)$ (HFD-corrected data). Let us emphasize that it is not the purpose of the present paper to perform a selection of the "best" model, and we rather motivate this specific choice based on the following considerations:

- The model embeds no unkown parameters in age so that the whole set of parameters is reasonable; it exploits for each year t the well-known overall linearity of the logarithm in the age range considered through $\kappa_1(t)$ and $\kappa_2(t)$, also known as the Gompertz model, and adds a small adjustment of the curvature with $\kappa_3(t)$.

- The three terms will be able to capture the shape of the mortality surface in a satisfactory way (in the corrected data), and in particular, the replicated mortality improvements are well fitted in mean (for both data).

- The model does not include a cohort component, see the discussion below.

- In its complete form (with cohort term), and tested on uncorrected data, it has proved to provide good results, see again [4].

We propose to calibrate the model on data from France on both crude and corrected tables. The estimated values for κ_1, κ_2 and κ_3 are depicted in Figure 15. Clearly, the parameters are rather similar for both data sets; this is coherent with the fact that such mortality model performs some smoothing of the mortality surface, therefore it is not sensitive to changes from one cohort to the other (having in mind that no cohort parameter is embedded in the mortality model). In Figure 16, we represent the residuals of the stochastic mortality model in Equation (17),

that is the quantities $\epsilon(x,t)$. Here, the difference appears: although for the crude data the residuals show striking cohort trends, leading the user to choose another model including some cohort component, for the corrected data such trends almost disappear. Therefore, as a first main conclusion, the impact of correcting such data is not on the parameter values, rather on the choice of the mortality model itself.

Let us now focus on the mean and volatility of mortality improvements in four configurations, see Figure 17:

- mortality improvements of the period 1980-2010 for crude HMD data,
- mortality improvements of the period 1980-2010 for corrected data,
- next year mortality improvements simulated in the model calibrated on crude HMD data,
- next year mortality improvements simulated in the model calibrated on corrected data.

For the two last analysis based on model-generated mortality improvements, these are computed as:

$$r(x,T) = \frac{\mu_M(x,T+1) - \mu_M(x,T)}{\mu_M(x,T)},$$

where $T = 2010$, $\mu_M(x,T) = \exp\left\{\kappa_1(T) + \kappa_2(T)\,(x-\bar{x}) + \kappa_3(T)\left((x-\bar{x})^2 - \widehat{\sigma_x}^2\right)\right\}$ is the (deterministic) last death rate value fitted by the model, and $\mu_M(x,T+1)$ is the (randomly) simulated death rate for next year, where simple auto regressive model is fitted to the time component $\bar{\kappa}(t) = (\kappa_1(t), \kappa_2(t), \kappa_3(t))$:

$$\bar{\kappa}(t+1) = \bar{\kappa}(t) + \mu + CZ(t),$$

with μ a 3-dimensional vector, C a 3×3 matrix and $Z(t)$ random and independent 3-dimensional vector with centered normal distribution and identity variance-covariance matrix.

Let us first notice that the mean and volatility of mortality improvements generated by the model are really close for each calibration data (corrected or not); this is coherent with the close values fo the corresponding parameters, see again Figure 15. Second, it is interesting to note the ability of the model to reproduce by simulation the quadratic curvature in age of mortality improvements (left graph). Third, and more importantly, with the new volatility levels exhibited in the corrected data, the model shows a far better duplication of historic volatility features. As a second main conclusion, it is therefore interesting to note that as volatility is reduced, the model is now able to give coherent order of magnitude of these, which drive the random possibilities for the future mortality levels.

5 Concluding remarks

In this paper, we exhibited the strong assumption of uniform distribution of births underlying the Human Mortality Database (HMD) methodology to compute period death rates. Based on monthly fertility data, we aimed at detecting the cohorts that are sensitive to such assumption as well as to correct period mortality tables based on a ratio measuring the deviation between annual and monthly approximations of the exposure to risk. We proposed to exploit the Human Fertility Database (HFD) as it represents the perfect couterpart of the HMD, and allows to construct the correction ratio for several countries. We focused on the five countries with the deepest fertility histories available in HFD, namely France, Switzerland, Finland, Sweden and Austria. For all these countries, we constructed corrected period mortality tables which do not present the initial important anomalies in the form of isolated cohort effects for years of birth around the two world wars, that is around 1915, 1920, 1940 and 1945. As these groups are concerned for almost the five countries considered, we argue that we are facing some universal issue regarding the construction of period mortality tables in the HMD; the whole correction process for other countries with limited fertility histories is a challenging topic that is left for further research.

The analysis of corrected mortality tables allowed us to draw several important conclusions. This concerned the way we understand cohort effects, in particular the systematic *pathologic* feature of those concerning isolated years of birth (one to three), the reduction of mortality improvements volatility in the corrected tables, as well as the better ability for classical models to fit the mortality surface (residual analysis) and to reproduce by simulation the mean and volatility of mortality improvements, noting that estimated time parameters remain quite the same when calibrating stochastic mortality models on uncorrected or corrected data.

In the spirit of the cited recent contributions in this direction, we argue that considering the whole demographic dynamics is a promizing direction for further understanding of mortality; this includes linking further HMD and HFD data as well as their associated research communities.

Aknowledgements

The author is grateful to Laurent Devineau for fruitful research directions, as well as for several inputs, discussions and remarks on this work. The author also thanks his collegues at Milliman, in particular Sophie Decupère and Mohamed Talfi for several interesting discussions.

References

[1] S. Arnold, A. Boumezoued, H. Labit Hardy and N. El Karoui. 2015. Cause-of-death mortality: What can be learned from population dynamics? *HAL preprint Id: hal-01157900* 2

[2] H. Bensusan, A. Boumezoued, N. El Karoui and S. Loisel. 2016. Bridging the gap from microsimulation practice to population models: a survey. *Work in progress* 2, 13, 22

[3] A. Boumezoued. 2015. Macroscopic behavior of heterogenous populations with fast random life histories. *HAL preprint Id: hal-01245249* 2

[4] A.J.G. Cairns, D. Blake, K. Dowd, G.D. Coughlan, D. Epstein, A. Ong, I. Balevich. 2009. A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *North American Actuarial Journal* **13**(1) 1–35. 2, 23

[5] A.J.G. Cairns, D. Blake, K. Dowd. 2006. A Two-Factor Model for Stochastic Mortality with Parameter Uncertainty: Theory and Calibration. *Journal of Risk and Insurance* **73**(4) 687–718. 23

[6] A.J.G. Cairns, D. Blake, K. Dowd and A.R. Kessler. 2016. Phantoms Never Die: Living with Unreliable Population Data. *To appear in Journal of the Royal Statistical Society, Series A.* 1, 2, 3, 7, 12, 18, 20

[7] Human Fertility Database. Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria). *Available at www.humanfertility.org (data downloaded on October 2015).* 1, 3, 17

[8] Human Mortality Database. University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). *Available at www.mortality.org or www.humanmortality.de (data downloaded on October 2015).* 1, 4

[9] N. Keiding. 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. 14

[10] A.G. McKendrick. 1926. Application of mathematics to medical problems. *Proc. Edin. Math. Soc.* **54** 98–130. 14

[11] R.D Lee, L.R. Carter. 1992. Modeling and forecasting US mortality. *Journal of the American Statistical Association* **87**(419) 659–671. 2, 23

[12] W. Lexis. 1875. Einleitung in die Theorie der Bevolkerungsstatistik. Strassburg: Triibner. (Pages 5-7 translated to English by N. Keyfitz and printed, with figure 1, in Mathematical Demography (ed. D. Smith & N. Keyfitz). Berlin: Springer (1977).) 4

[13] S.J. Richards. 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, 2008. 1, 2, 3, 7, 12

[14] H. Von Foerster. 1959. *The Kinetics of Cellular Proliferation*. Grune & Stratton. 14

[15] RC. Willets. 2004. *The cohort effect: insights and explanations*. Cambridge Univ Press. 22

[16] J.R. Wilmoth, K. Andreev, D. Jdanov and D.A. Glei. 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, 12

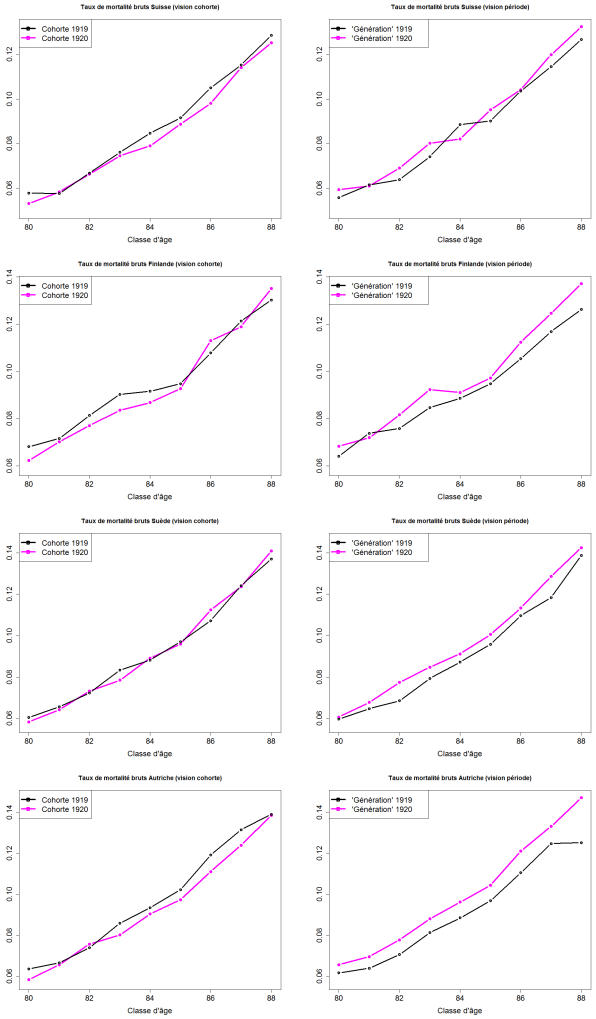

Figure 6: From top to bottom: Switzerland, Finland, Sweden and Austria. Left: mortality rates for the birth cohorts 1919 and 1920 from HMD cohort data. Right: mortality rates for diagonals starting at 1919 and 1920 from HMD period data.

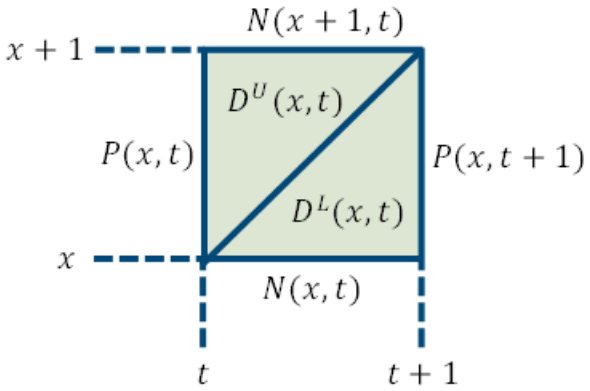

Figure 7: Population estimates and death counts on the Lexis diagram.

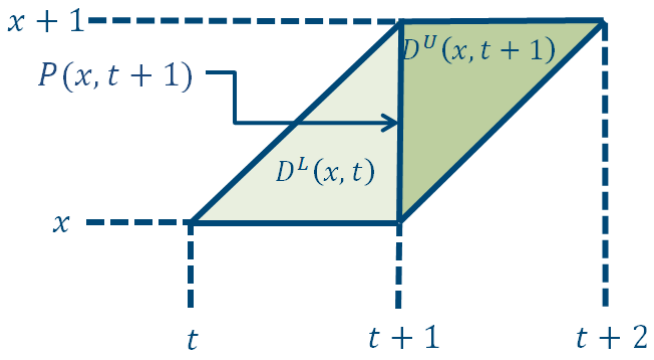

Figure 8: Population estimate and death counts on the Lexis diagram (cohort setting).

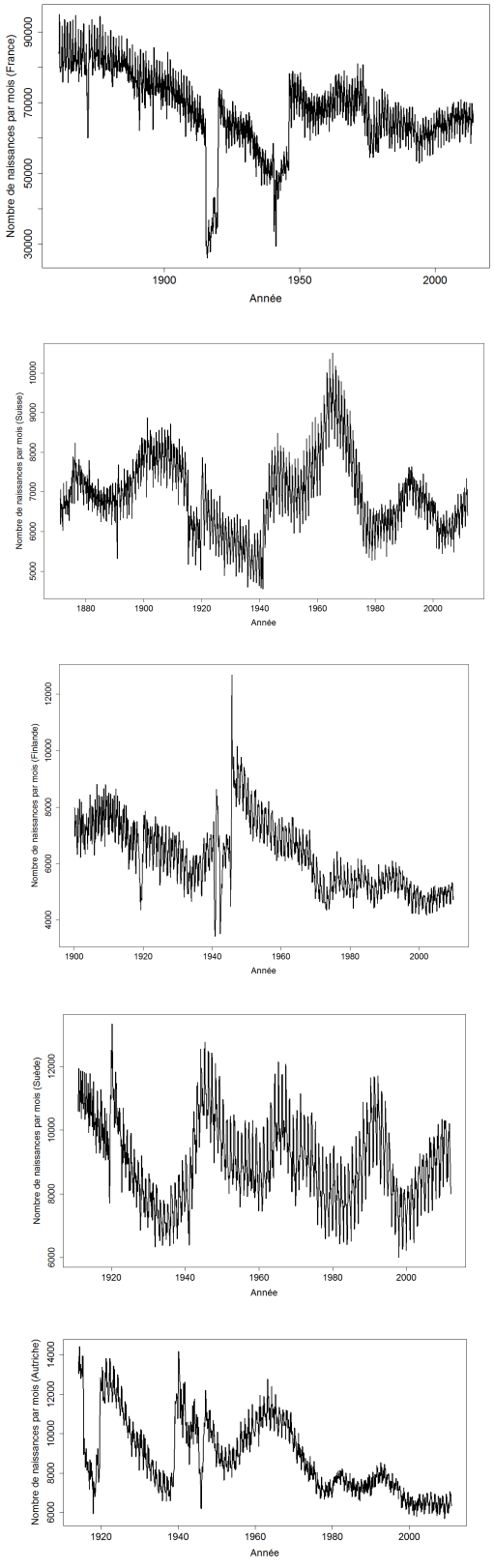

Figure 9: Number of births by month from HFD data. From top to bottom: France, Switzerland, Finland, Sweden and Austria.

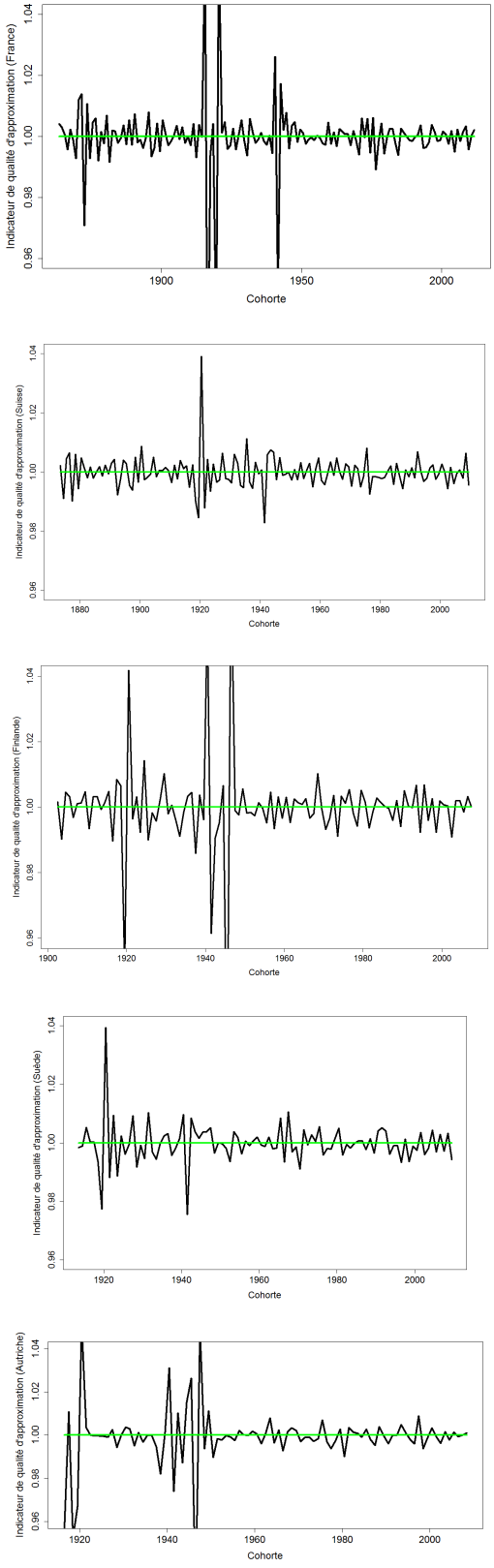

Figure 10: Correction ratio by year of birth. From top to bottom: France, Switzerland, Finland, Sweden and Austria.

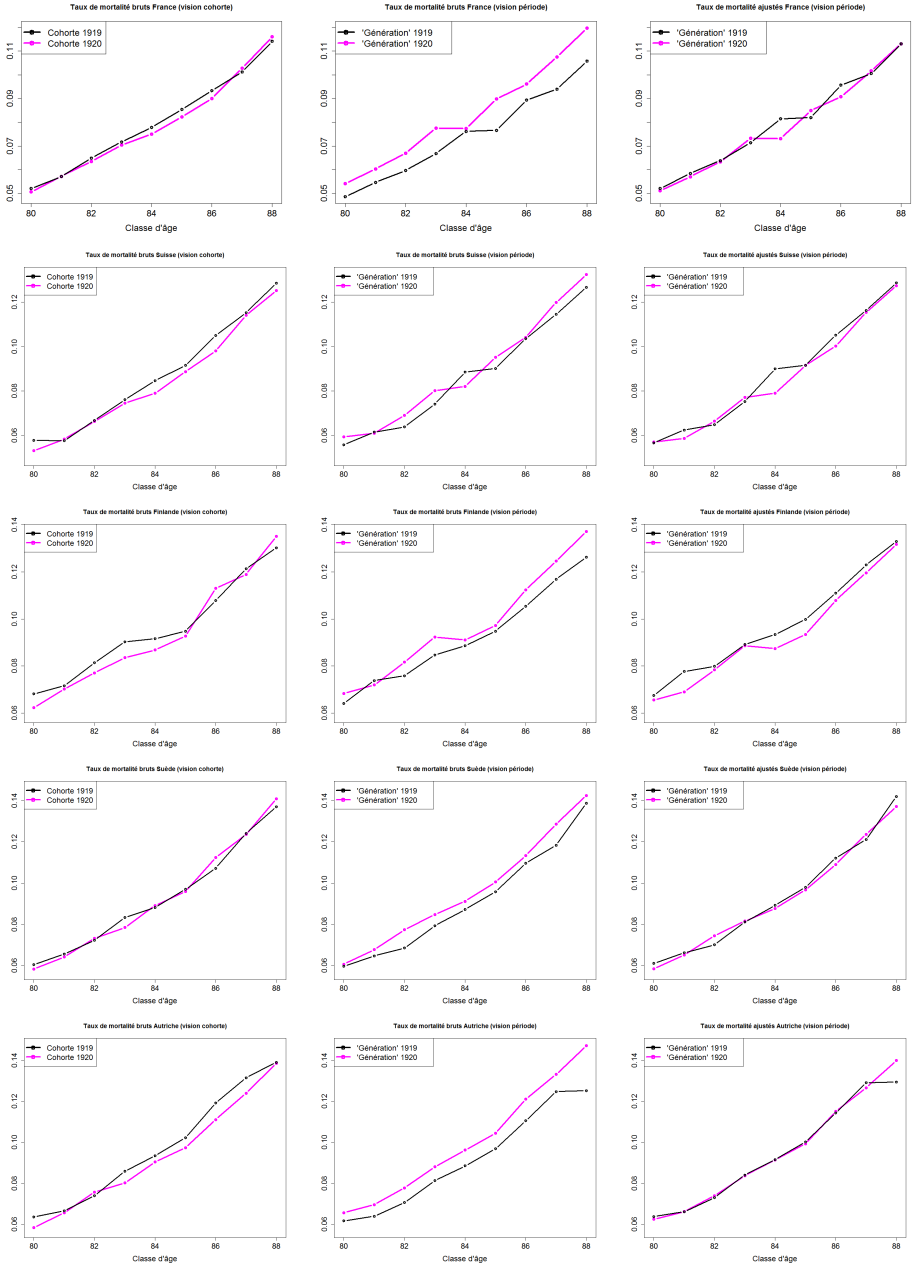

Figure 11: From top to bottom: France, Switzerland, Finland, Sweden and Austria. Left column: mortality rates for the birth cohorts 1919 and 1920 from HMD cohort data. Middle column: mortality rates for diagonals starting at 1919 and 1920 from HMD period data. Right column: mortality rates for diagonals starting at 1919 and 1920 from corrected period data.

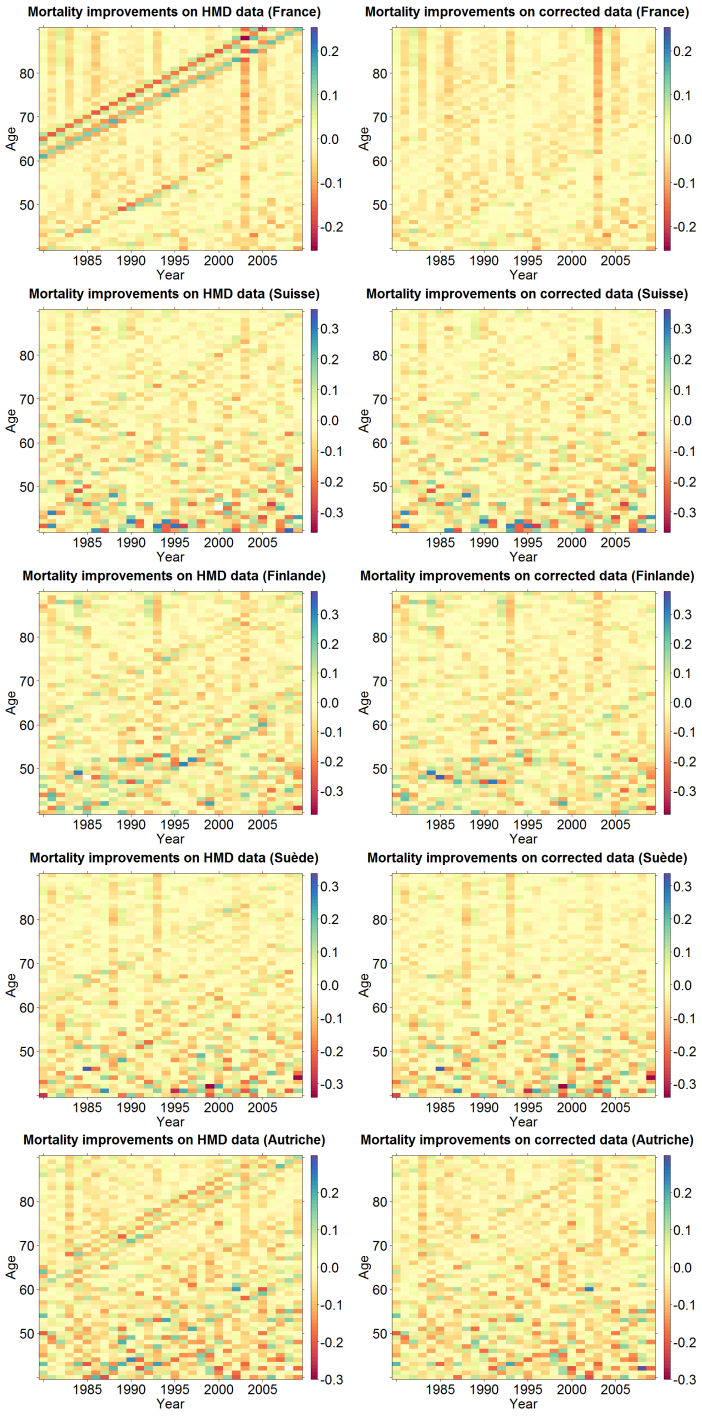

Figure 12: From top to bottom: France, Switzerland, Finland, Sweden and Austria. Left column: mortality improvement rates from HMD period data. Right column: mortality improvement rates from corrected period data.

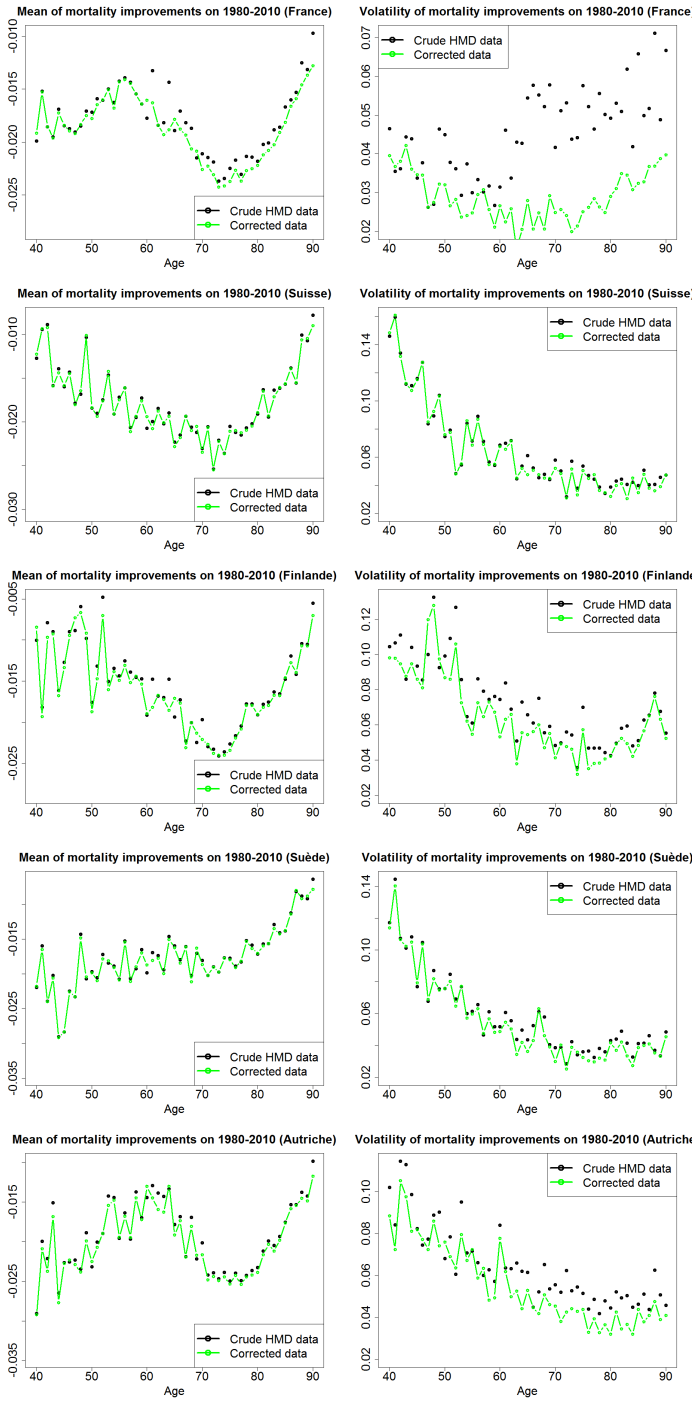

Figure 13: From top to bottom: France, Switzerland, Finland, Sweden and Austria. Comparison of mean and standard deviation of mortality improvements on the period 1980-2010 for both HMD data and corrected data, for the age range 40-90. Left column: mean. Right column: standard deviation.

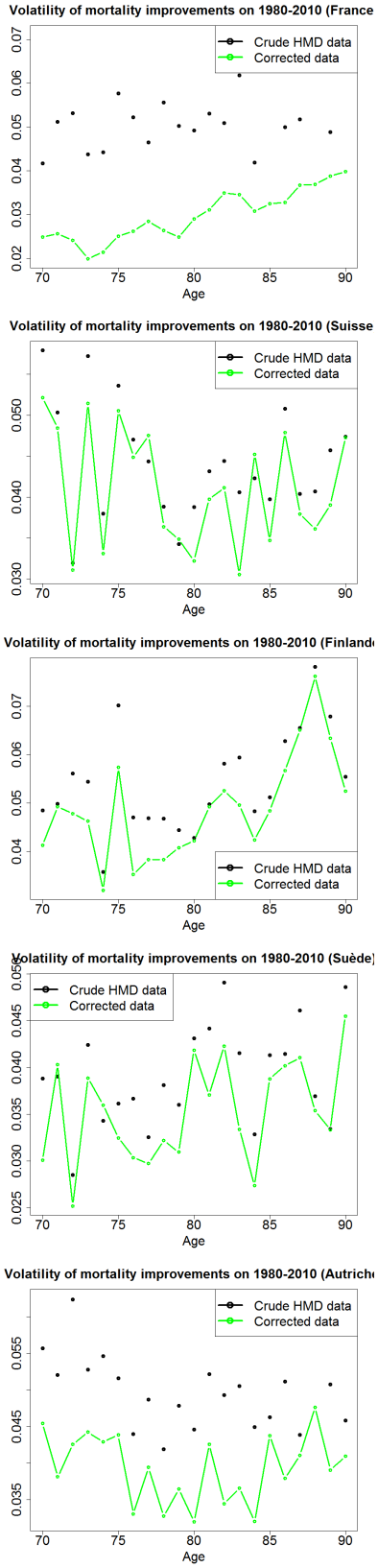

Figure 14: Standard deviation of mortality improvements on the period 1980-2010 for both HMD data and corrected data, for the age range 70-90. From top to bottom: France, Switzerland, Finland, Sweden and Austria.

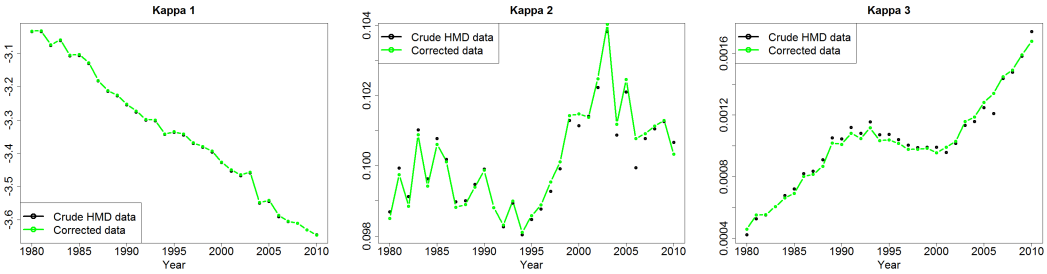

Figure 15: Time parameters in the stochastic mortality model in Equation 17 for the period 1980-2010 calibrated on the age range calibrated on the age range 60-90.

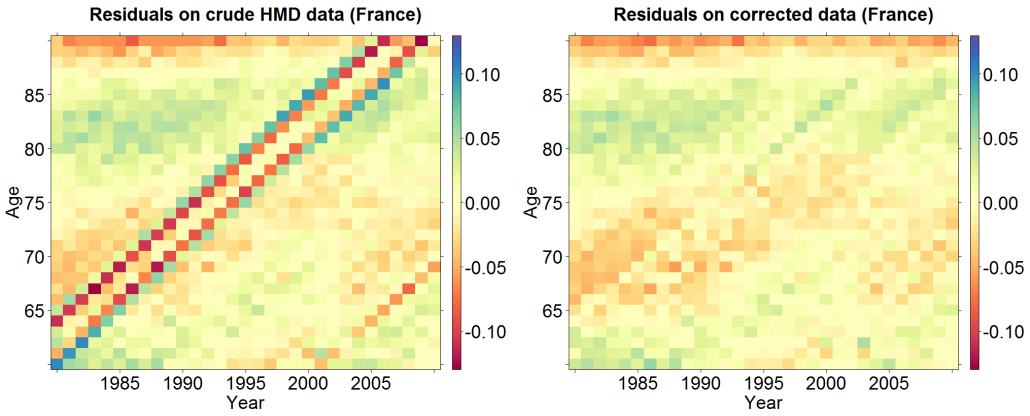

Figure 16: Residuals of the stochastic mortality model in Equation 17 calibrated on the time period 1980-2010 and the age range calibrated on the age range 60-90. Left: crude HMD data. Right: corrected data.

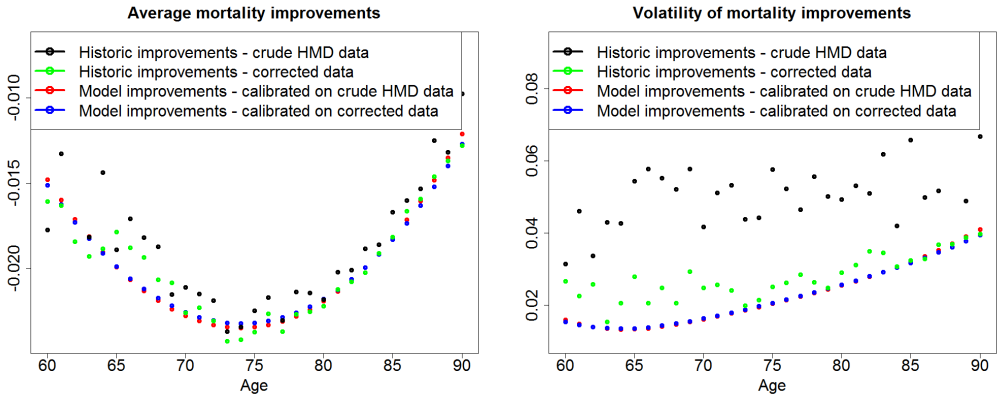

Figure 17: Mean (left) and volatility (right) of mortality improvements from four sources: empirical analysis on 1980-2010 from crude HMD data, empirical analysis on 1980-2010 from corrected data, analysis of one year mortality improvements in the model calibrated either on crude HMD or corrected data.