A new inference strategy for general population mortality tables

Alexandre Boumezoued (Milliman R&D)
joint work with Marc Hoffmann and Paulien Jeunesse (Paris Dauphine University)

September 6, 2018

Agenda

Motivation

The problem

Population dynamics

Inference strategies

Numerical results

Motivation (1/3): an history of demographics

- The first **mortality table** appeared in 1662 by John Graunt
  - He estimated death probabilities as a function of **age**
- Two centuries later, there was a huge development of graphical formalizations of life trajectories **within a population** by Lexis (1875) and his contemporaries

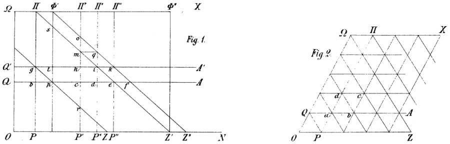
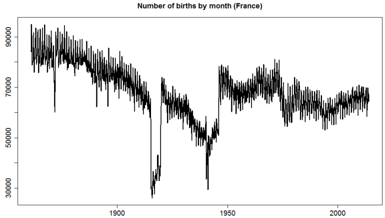


Figure: Examples of the so-called 'Lexis Diagram'

- These first demographers showed that it is crucial to address simulataneously two components:
  1. Consider the **non-homogeneous** case in which the death rate depends on both age and **time**
  2. Understand the mortality rate as an aggregate quantity which depends on an underlying **population dynamics**

Motivation (2/3): recent awareness about anomalies

- The analysis of **cohort effects** has long fascinated demographers
  - these effects correspond to the observation that specific generations can have longevity characteristics different from those of the previous and the following ones
- It is through the study of such cohort effects that Richards (2008) suggested that these could be **anomalies in the calculation of death rates due to shocks in birth patterns**
  - **Cairns, Blake, Dowd & Kessler (2016)** confirmed the conjecture by Richards on the example of England and Wales, and used monthly fertility data to detect and correct the anomalies
  - **B. (2016)** focused on the Human Mortality Database (V5), showed that these anomalies are universal and proposed to link it with the Human Fertility Database to correct such errors



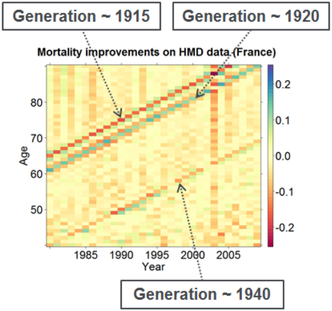


Figure: LEFT: births by month in France. RIGHT: **False "Cohort effects"** in mortality improvements from crude tables of the V5 Human Mortality Database (now V6)

Motivation (3/3): improving mortality estimates with monthly fertility data

- Using **fertility data** at a refined time scale (monthly), it is possible to **refine the traditional death rate estimates**
  - Example below extracted from B. (2016)

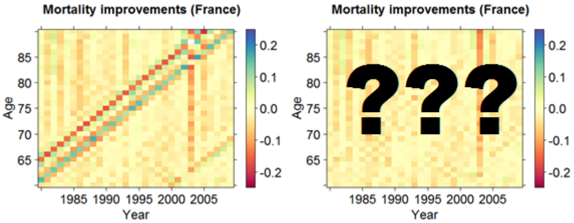


Figure: Mortality improvement rates before (left) and after (right) correction based on monthly fertility data

- **Aim of our project:** build on the previous empirical work and propose a mathematically-founded construction of mortality tables based on traditional census estimates while taking advantage of monthly fertility data

Agenda

Motivation

The problem

Population dynamics

Inference strategies

Numerical results

Non-parametric inference from one to two dimensions

- The Nelson-Aalen estimate in one dimension writes

$$\hat{\beta}(t) = \int_0^t \frac{I\left(Y(s) > 0\right)}{Y(s)} \mathrm{d}N_s$$

- Generalization of one-dimensional non-parametric estimators is not straightforward! Indeed, one would like to define

$$\hat{\beta}(t) = \int_0^\infty \frac{I\left(Y(t,a) > 0\right)}{Y(t,a)} N(t, \mathrm{d}a)$$

where $Y(t,a)$ is the (stochastic) number of living with exact age a at exact time t. **Issue:** $Y(t,a) = 0$ or 1

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 I 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."*

Dealing with life lines in the Lexis diagram

- Statistical point of view:
  - **Bi-variate smoothing** is required to tackle the life lines issue in the Lexis diagram
  - Non-parametric inference with age x time (no birth-death process)
    - Keiding (1990)
    - McKeague & Utikal (1990)
    - Nielsen & Linton (1995)
    - Brunel, Comte & Guilloux (2008)
    - Comte, Gaiffas & Guilloux (2010)
- Practical demographic point of view:
  - The death rate is assumed to be **piecewise constant** on squares, parallelograms or triangles in the Lexis diagram
    $\Rightarrow$ all life lines crossing the region can be used to estimate the death rate
    $\Rightarrow$ the approach amounts to a smoothing with uniform kernel

Key constraints in the project

The (applied part of the) project must deal with the following constraints

- The death rate depends on both age and time
- The propulation evolves as a stochastic age-structured and time inhomogeneous birth-death process
- Only the following observables are available in the Lexis diagram:
  - Traditional annual census estimates
  - Death counts in annual Lexis triangles
  - Birth counts at the montly scale

Agenda

Motivation

The problem

Population dynamics

Inference strategies

Numerical results

Age pyramid

- Evolves over time due to **several demographic events**:
  - **Deaths**
  - **Births**
  - Migration flows

Age pyramid

- Evolves over time due to **several demographic events**:
  - **Deaths**
  - **Births**
  - Migration flows
- Let $g(a, t)$: number of individuals with exact age a at exact time t
  $\Rightarrow$ Continuous age and time setting

Age pyramid

- Evolves over time due to **several demographic events**:
  - **Deaths**
  - **Births**
  - Migration flows
- Let $g(a, t)$: number of individuals with exact age a at exact time t
  $\Rightarrow$ Continuous age and time setting
- Example: $\int_{a_1}^{a_2} g(a, t)\mathrm{d}a$
  the number of individuals with exact age in $[a_1, a_2)$ at time t

Age pyramid

- Evolves over time due to **several demographic events**:
  - **Deaths**
  - **Births**
  - Migration flows
- Let $g(a, t)$: number of individuals with exact age a at exact time t
  $\Rightarrow$ Continuous age and time setting
- Example: $\int_{a_1}^{a_2} g(a, t) \mathrm{d}a$
  the number of individuals with exact age in $[a_1, a_2)$ at time t
- Example: [intergenerational issues] Dependency ratio

$$r_t = \frac{\int_{65}^{\infty} g(a, t) \mathrm{d}a}{\int_{15}^{65} g(a, t) \mathrm{d}a}.$$

Mortality force & Cohort dynamics

- Let $\mu(a, t) \equiv$ mortality force at exact age a and exact time t
- Drives the time evolution of a given cohort
- Let $g(0, \nu)$ be given (number of newborns at time ν)

Mortality force & Cohort dynamics

- Let $\mu(a,t) \equiv$ mortality force at exact age a and exact time t
- Drives the time evolution of a given cohort
- Let $g(0,\nu)$ be given (number of newborns at time ν)
- The number of survivors at age a in the cohort is

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

Mortality force & Cohort dynamics

- Let $\mu(a, t) \equiv$ mortality force at exact age a and exact time t
- Drives the time evolution of a given cohort
- Let $g(0, \nu)$ be given (number of newborns at time ν)
- The number of survivors at age a in the cohort is

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

- Differentiation (age and time) leads to the...

...transport component of McKendrick-Von Foerster equation

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

Endogeneous births in the renewal component

- People of a birth cohort share the fact that they are born from the same population:

Renewal component of the McKendrick-Von Foerster equation

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

Recall the transport component :

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

Stochastic setting and micro/macro link

- Due to the finite population size, demographic events (individual births and deaths) occur at random times
  $\Rightarrow$ Microscopic point of view
- Need of stochastic modeling to account for idiosyncratic risk

Stochastic setting and micro/macro link

- Due to the finite population size, demographic events (individual births and deaths) occur at random times
  $\Rightarrow$ Microscopic point of view
- Need of stochastic modeling to account for idiosyncratic risk
- $Z_t([a_1, a_2)) \equiv$ the stochastic number of individuals with age in $[a_1, a_2)$ at exact time t

Stochastic setting and micro/macro link

- Due to the finite population size, demographic events (individual births and deaths) occur at random times
  $\Rightarrow$ Microscopic point of view
- Need of stochastic modeling to account for idiosyncratic risk
- $Z_t([a_1, a_2)) \equiv$ the stochastic number of individuals with age in $[a_1, a_2)$ at exact time t

Micro-macro consistency*

$$\mathbb{E}\left[Z_t([a_1, a_2))\right] = \int_{a_1}^{a_2} g(a, t)\mathrm{d}a \quad \text{[Linear model]}$$

Stochastic setting and micro/macro link

- Due to the finite population size, demographic events (individual births and deaths) occur at random times
  $\Rightarrow$ Microscopic point of view
- Need of stochastic modeling to account for idiosyncratic risk
- $Z_t([a_1, a_2)) \equiv$ the stochastic number of individuals with age in $[a_1, a_2)$ at exact time t

Micro-macro consistency*

$$\mathbb{E}\left[Z_t([a_1, a_2))\right] = \int_{a_1}^{a_2} g(a, t)\mathrm{d}a \quad \text{[Linear model]}$$

- Simulation by means of the *Thinning algorithm*

* Convergence of sequence of renormalized population processes (large number effect) also holds

Agenda

Motivation

The problem

Population dynamics

Inference strategies

Numerical results

Period and cohort tables

- Three directions of analysis in the Lexis diagram; age, period and cohort
  - The difference between **cohort** and **period** tables lies on the choice of the two degrees of freedom to be fixed among the three described above

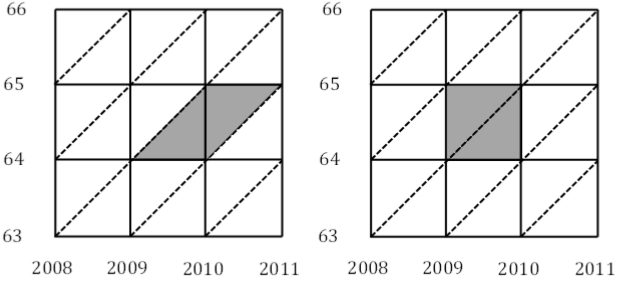


Figure: Population used (in grey) for the computation of cohort death rates (left) and period death rates (right) in the Lexis diagram

Observables in the Lexis diagram - population counts

In an ideal demographic world, two kinds of population estimates are recorded in the one-year age $\times$ time square:

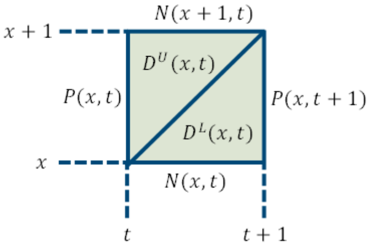
(deterministic or stochastic setting)

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

$$P(t,x) = \int_x^{x+1} g(a,t)\mathrm{d}a \quad \text{or} \quad Z_t([x, x+1))$$

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

$$N(t,x) = \int_t^{t+1} g(x,s)\mathrm{d}s \quad \text{or} \quad \int_t^{t+1} Z_s(\{x\})\mathrm{d}s$$



Observables in the Lexis diagram - death counts

- **Death counts** Also, number of deaths are provided on the upper and lower triangles of the Lexis diagram. Let us first introduce such upper (U) and lower (L) triangles for each age range x and observation year t as

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

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

If we denote $\Gamma(\mathrm{d}t, \mathrm{d}a)$ the point process of deaths then the number of deaths provided write $D_U(t,x) = \Gamma(T_U(t,x))$ and $D_L(t,x) = \Gamma(T_L(t,x))$.

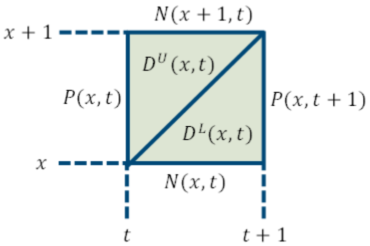


Figure: Observables in the Lexis diagram

Observables in the Lexis diagram - relations

- Fundamental relations in a closed population (integration by parts):

$$N(t, x+1) = P(t, x) - D_U(t, x),$$
$$P(t+1, x) = N(t, x) - D_L(t, x).$$

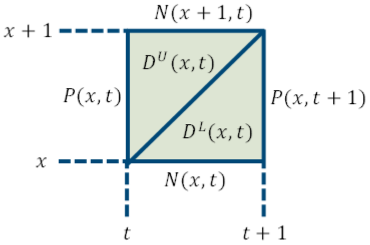
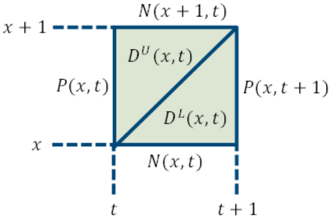


Figure: Observables in the Lexis diagram

Monthly fertility records

- Monthly fertility records are available in the Human Fertility Database
  - **Deterministic setting:** The number of births in the month $[t, t+1/12)$ is
  $$\int_t^{t+1/12} g(0,s)\mathrm{d}s$$
  - **Stochastic setting:** To properly define such estimates, one can construt the counting process related to births events as
  $$N_b(dt) = \int_{(i,\theta)\in\mathbb{N}^\star\times[0,\infty)} \mathbf{1}_{i\leq\langle Z^N_{t-},1\rangle}\mathbf{1}_{0<\theta\leq m1(i,t)}Q(dt,di,d\theta).$$
  The available estimates are then $N_b([t, t+1/12))$ for each $t\in\frac{1}{12}\mathbb{N}$.

Demographic reasoning



Several assumptions underly the classical formulas, **in particular**:

- (H1) Uniform distribution of births within each cohort
- (H2) Uniform distribution of deaths within each triangle

The classical demographic reasoning is split in two main steps:

- **Step 1:** computation of the total exposure under the assumption that **no deaths occur**, gives under (H1):

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

- **Step 2:** adjust the main component to **death occurrences** in the triangle, under (H2) - this corresponds to add a second order term of the form

$$\frac{1}{3}\left[D_U(x,t) - D_L(x,t)\right]$$

Closed forms at first order (1/3)

Notations used:

- $S(x,t) := e^{-\sum_{y=0}^{x-1} \mu_L(y,t-x+y)}$ is the base survival function to age x
- $H(x,t) := \sum_{y=0}^{x-1} \{\mu_U(y,t-x+y+1) - \mu_L(y,t-x+y)\}$ quantifies the gain in longevity **within the same cohort**

$$E_L(x,t) = S(x,t) \int_t^{t+1} \int_x^{x+s-t} g(0,s-a) e^{-(t-x-s+a)H(x,t)} e^{-(a-x)\mu_L(x,t)} \mathrm{d}a\mathrm{d}s$$

$$\approx S(x,t) \int_t^{t+1} \int_x^{x+s-t} g(0,s-a) e^{-(t-x-s+a)H(x,t)} (1-\mu_L(x,t)(a-x)) \mathrm{d}a\mathrm{d}s$$

$$= E_L^1(x,t) - \mu_L(x,t) E_L^2(x,t)$$

where the 'if no deaths occur' exposure is

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

- $L_{t-x}(.)$ is the Laplace transform of the r.v. B_{t-x} "date of birth in the year $t-x$", taking values in $[0,1]$
- If **no improvement in mortality** within the cohort, then $H(x,t)=0$, and $E_L^1(x,t) = N(x,t)\left(1-\mathbb{E}\left[B_{t-x}\right]\right)$
- If additionally **births are uniformly distributed** within the year, then $E_L^1(x,t) = \frac{1}{2}N(x,t)$ = **Classical main component of the exposure-to-risk**

Closed forms at first order (2/3)

$$E_L(x,t) = S(x,t)\int_t^{t+1}\int_x^{x+s-t} g(0,s-a)e^{-(t-x-s+a)H(x,t)}e^{-(a-x)\mu_L(x,t)}\mathrm{d}a\mathrm{d}s$$

$$\approx S(x,t)\int_t^{t+1}\int_x^{x+s-t} g(0,s-a)e^{-(t-x-s+a)H(x,t)}(1-\mu_L(x,t)(a-x))\mathrm{d}a\mathrm{d}s$$

$$= E_L^1(x,t) - \mu_L(x,t)E_L^2(x,t)$$

where the 'if we correct for deaths' component writes

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

- $L_{t-x}(.)$ is the Laplace transform of the r.v. B_{t-x} "date of birth in the year $t-x$", taking values in $[0,1]$
- If **no improvement in mortality** within the cohort, then $H(x,t)=0$, and $E_L^2(x,t) = \frac{1}{2}N(x,t)\left[1-2\mathbb{E}\left[B_{t-x}\right] + Var(B_{t-x})\right]$
- If additionally **births are uniformly distributed** within the year, then $E_L^2(x,t) = \frac{1}{24}N(x,t)$, therefore $\mu_L(x,t)E_L^2(x,t) \approx \frac{1}{12}D_L(x,t)$ $\approx$ **classical second order correction of the exposure-to-risk**

Closed forms at first order (3/3)

- The relation $\mu_L(x,t) = \frac{D_L(x,t)}{E_L^1(x,t) - \mu_L(x,t) E_L^2(x,t)}$ leads to (omit dependence in (x,t), and denote $L \equiv L_{t-x}(H(x,t))$ for simplicity):

$$\mu_L = \frac{L + L'}{L + 2L' + L''} \left\{ 1 - \sqrt{1 - \frac{D_L}{N/2} \frac{L(L + 2L' + L'')}{(L + L')^2}} \right\}$$

- Practically, $L_{t-x}(.)$ is estimated based on monthly birth counts, and $H(x,t)$ is estimated recursively based on the mortality table
- Some analysis:
  - Denote $\sigma^2 = Var(B_{t-x})$; if $H \equiv 0$ (no improvement within the cohort), and births are centered ($\mathbb{E}[B_{t-x}] = 1/2$) then

$$\mu_L = \frac{1}{2\sigma^2} \left\{ 1 - \sqrt{1 - \frac{D_L}{N/2} \times 4\sigma^2} \right\} \approx \frac{D_L}{N/2}$$

- Similar reasoning leads to (recursive) closed-forms for the death rate un the upper triangle $\mu_U(x,t)$.

Final estimation method

- **Issues with the closed forms:**
  - The Taylor expansion is not valid for ages below around 5 and above around 60, as death rate values in these ranges are not small
  - The recursive estimation transports the initial bias for low ages to higher ages in each cohort
- **Solution:** keep the untractable formulas to numerically (and recursively) find the death rate estimate as the solution to some inverse problem

Proposition: The following equalities hold:

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

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

Agenda

Motivation

The problem

Population dynamics

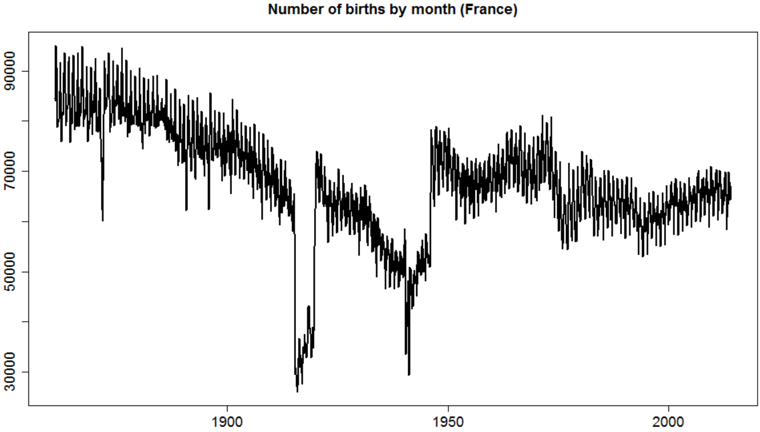
Inference strategies

Numerical results

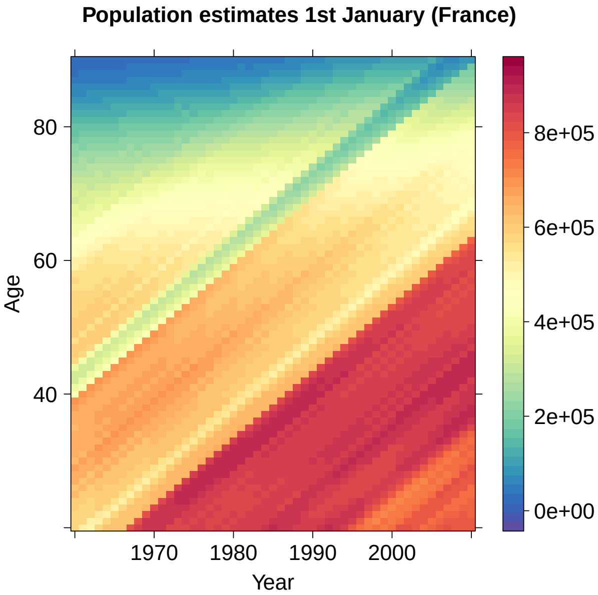
Data & algorithm

- **Initial step:**
  - Start at age zero and estimate the death rate in the lower triangle $\mu_L(0, t)$ for each available year of birth t
    - Only number of births by months and deaths in the lower triangle are required
  - Then compute the death rate in the upper triangle $\mu_U(0, t)$, based on $\mu_L(0, t)$ estimated previously
- Then **Recursive computation** of $\mu_L(x, x + t)$ and then $\mu_U(x, x + t)$ for increasing x.

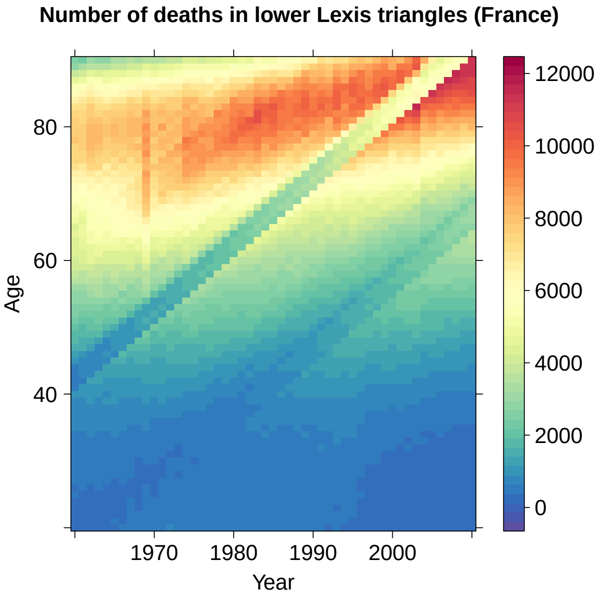
Births distribution



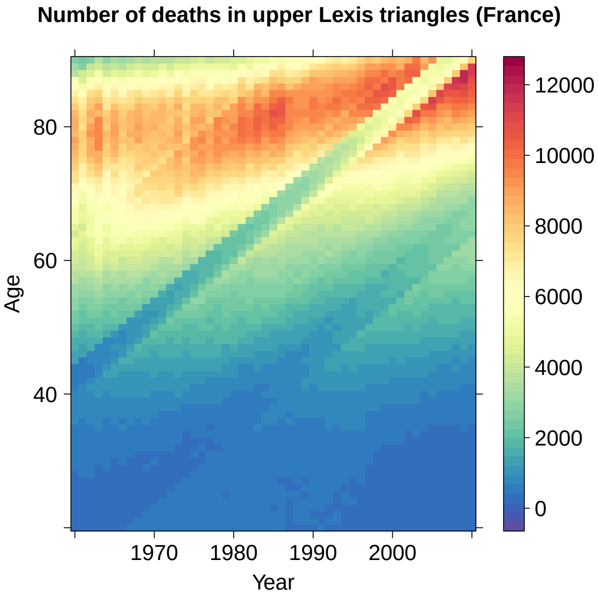
Population counts $P(x, t)$



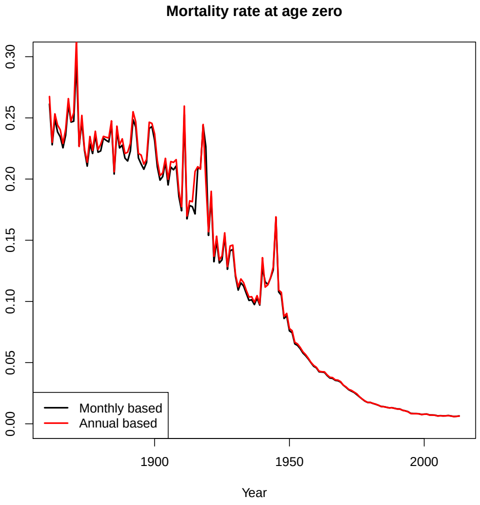
Deaths in lower Lexis triangles: $D_L(x, t)$



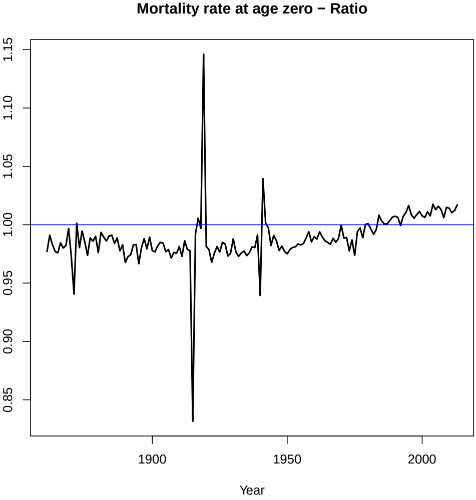
Deaths in upper Lexis triangles: $D_U(x, t)$



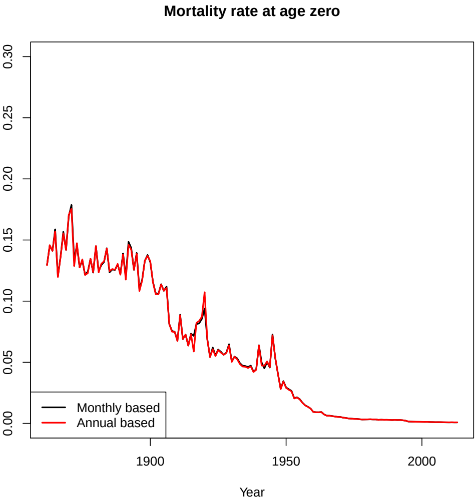
Mortality rate at age zero - lower triangle



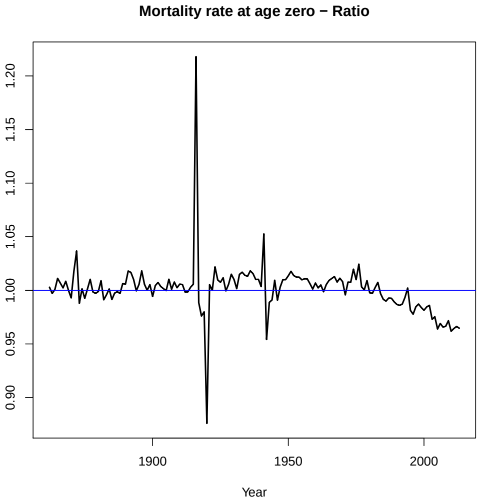
Mortality rate at age zero - lower triangle



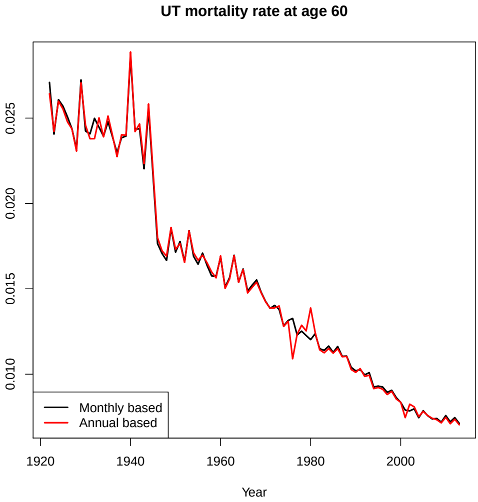
Mortality rate at age zero - upper triangle



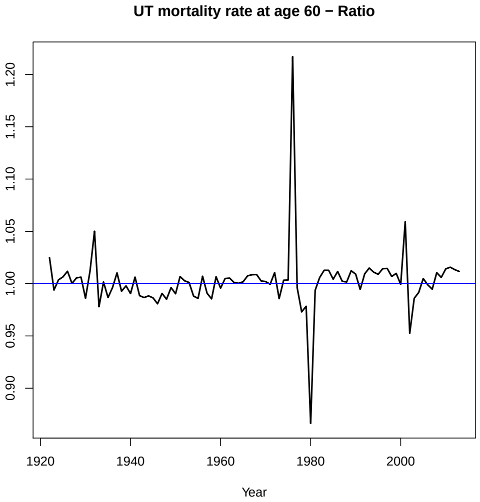
Mortality rate at age zero - upper triangle



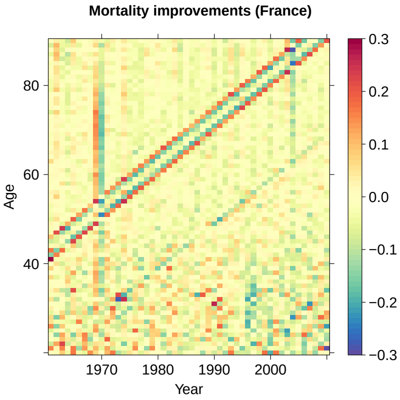
Mortality rate at age 60 - upper triangle

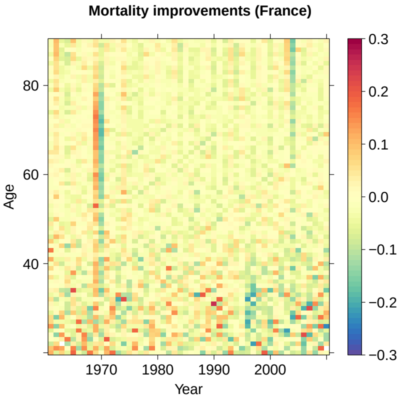


Mortality rate at age 60 - upper triangle

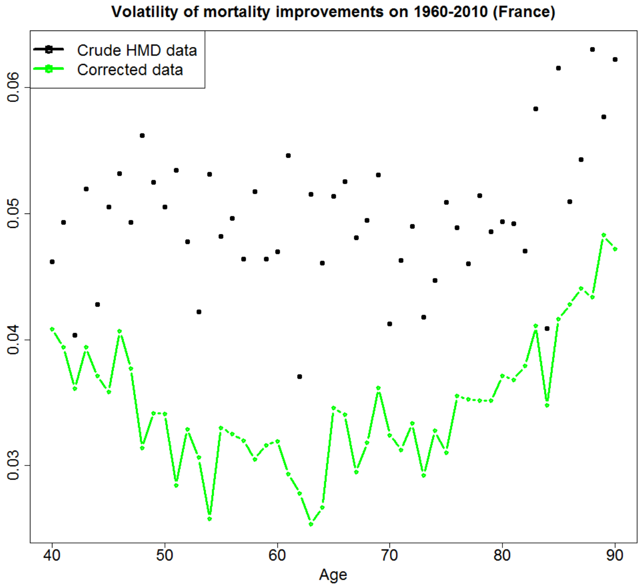


Old and new mortality table (lower triangles)





Consequences for the insurance market?



Stochastic population dynamics - a word

- Based on the **thinning representation** (stochastic equation) for counting processes:

$$N_t = N_0 + \int_0^t \int_{\mathbb{R}_+} \mathbf{1}_{[0,\Psi(N_u, 0 \le u < s)]}(\theta) Q(\mathrm{d}s, \mathrm{d}\theta).$$

- Contruction of a birth-death stochastic age-structured population process
- Statistical setting: We have (i) data Z^N and (ii) a parameter of interest f. Asymptotics are taken as $N \to \infty$.
- Structure of the problem:

$$\mathcal{H}_N(Z^N) = 0 \text{ for some SDE } \mathcal{H}_N,$$

$$Z^N \to \xi \text{ limiting object},$$

$$\mathcal{H}(\xi, f) = 0 \text{ for some PDE } \mathcal{H}.$$

- Here Z^N is a (large) human population evolving through time and $f(t, a)$ the density (or mortality rate, or fertility rate) of the population with age a at time t.

Conclusion & Perspectives

- Summary
  - New tables easy to compute...
  - ...with a slight attention that these are recursive: any revision of past population estimates / death counts will imply to re-compute the following mortality rates... Natural !
- Perspectives
  - statistical analysis of the construction method, based on the stochastic population model
  - dealing with population flows in age $\times$ year squares

References

- With M. Hoffmann and P. Jeunesse (in preparation)
  - **A new inference strategy for general population mortality tables**
  - **Non-parametric inference for in-homogeneous and age-dependent population processes**
- B. 2016. **Improving HMD mortality estimates with HFD fertility data.** To appear in the North American Actuarial Journal.
- B. & L. Devineau. **Enjeux de fiabilité dans la construction des tables de mortalité nationales.** L'Actuariel, janvier 2017.
- **Reliability issues in the construction of national mortality tables for the general population: What you should know**, by A. B., L. Devineau, D. S. Hagstrom
- B. 2016. **Micro-macro analysis of heterogeneous age-structured populations dynamics. Application to self-exciting processes and demography.** Doctoral thesis, available at https://tel.archives-ouvertes.fr/tel-01307921/ (supervision of N. El Karoui and co-supervision of S. Loisel)
- A.J.G. Cairns, D. Blake, K. Dowd and A.R. Kessler. 2016. **Phantoms Never Die: Living with Unreliable Population Data.** Journal of the Royal Statistical Society, Series A. 179(4) 975-1005.
- **Human Mortality Database.** University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org
- **Human Fertility Database.** Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria). Available at www.humanfertility.org
- 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.