Development of Bayesian and Formal Demography approaches to unveil historical fertility patterns using online genealogical data

Riccardo Omenti [1] Monica Alexander [2] Nicola Barban [1]

[1]University of Bologna

[2]University of Toronto

April 19th, 2024

Objectives

- **Goal** $\rightarrow$ combine **online genealogical data** with **traditional data sources** to examine **fertility patterns** during the historical period $1751 - 1910$ in 7 European countries and US by:
  - **Bayesian modeling framework**
  - **Indirect estimation techniques**

Online genealogies and FamiLinx

- Web sites that allow users to reconstruct their own family tree from bottom up
- Focus on **FamiLinx** → big genealogical database by Kaplanis et al. (2018) with **86 million** individuals
- Recorded variables: **birth** and **death locations** and **dates**, and **kin relationships**

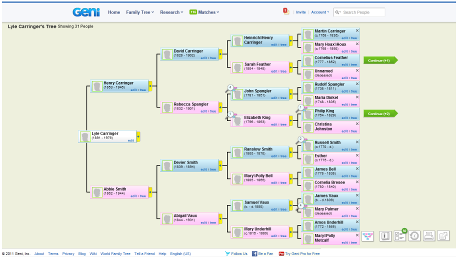
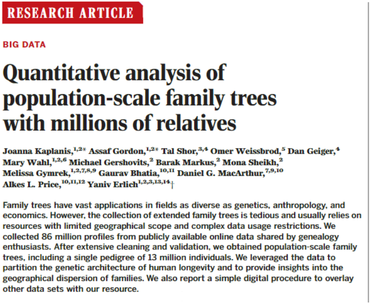


Figure: family tree on geni.com.



RESEARCH ARTICLE

BIG DATA

Quantitative analysis of population-scale family trees with millions of relatives

Joanna Kaplanis,[1,2*] Assaf Gordon,[1,2*] Tal Shor,[3,4] Omer Weissbrod,[5] Dan Geiger,[4] Mary Wahl,[1,2,6] Michael Gershovits,[2] Barak Markus,[2] Mona Sheikh,[2] Melissa Gymrek,[1,2,7,8,9] Gaurav Bhatia,[10,11] Daniel G. MacArthur,[7,9,10] Alkes L. Price,[10,11,12] Yaniv Erlich[1,2,3,13,14†]

Family trees have vast applications in fields as diverse as genetics, anthropology, and economics. However, the collection of extended family trees is tedious and usually relies on resources with limited geographical scope and complex data usage restrictions. We collected 86 million profiles from publicly available online data shared by genealogy enthusiasts. After extensive cleaning and validation, we obtained population-scale family trees, including a single pedigree of 13 million individuals. We leveraged the data to partition the genetic architecture of human longevity and to provide insights into the geographical dispersion of families. We also report a simple digital procedure to overlay other data sets with our resource.

Figure: Kaplanis et al. (2018)

Limitations in FamiLinx

- High percentage of **missing values** in common **demographic variables**
- Most of individuals from **Europe** and **North America**
- Times of other demographic events beyond **births** and **deaths** are not recorded
- Data **representativeness** and **quality** are not consistent across countries and over time

Sample Selection Criteria

1. Countries of birth and death:
   - **Nothern Europe** $\rightarrow$ Denmark, Finland, Norway and **Sweden**
   - **Western Europe** $\rightarrow$ England & Wales, Netherlands and **France**
   - **North America** $\rightarrow$ **United States of America**
2. Place of birth = Place of death
3. Birth Year ≤ 1910 and Death Year ≥ 1751
4. $0 \leq$ Age at death ≤ 110

Methodological Framework

Extend of **Bayesian modeling** and **indirect estimation** by Schmertmann & Hauer (2019, 2020)

Idea → Estimation of period Total Fertility Rate (TFR) without knowledge of births by maternal ages

Minimal Input Data Requirements

- Accurate **counts** of **children under 5** and **women aged 15-49**
- Prior information on
  - **Child mortality**
  - **Age-specific fertility patterns**

Contribution

- Extension of the methods in **contexts** with **imperfect data** by incorporating prior information on **non-representativeness** of **women** and **children**

Bayesian modeling framework

Data model:

$$C^{\text{gen}}_{a,t} | K_{x,a,t}, \tau_{x,a,t} \sim \text{Pois}\left(\sum_{x=15}^{45} K_{x,a,t} \cdot W^{\text{gen}}_{x,a,t} \cdot \tau_{x,a,t} \right)$$

$$C^{\text{true}}_{a,t} | K_{x,a,t} \sim \text{Pois}\left(\sum_{x=15}^{45} K_{x,a,t} \cdot W^{\text{true}}_{x,a,t} \right)$$

$$K_{x,a,t} = TFR_{a,t} \cdot \frac{L_{0,a,t}}{5} \cdot \frac{1}{2} \cdot \left[\frac{L_{x-5,a,t}}{L_{x,a,t}} \cdot \phi_{x-5,a,t} + \phi_{x,a,t} \right]$$

- **Overall fertility** ($TFR_{a,t}$) $\rightarrow$ non-informative prior
- **Age-specific fertility proportions** ($\phi_{x,a,t}$) $\rightarrow$ linear model based on a set of standard schedules
- **Age-specific person-years** ($L_{x,a,t}$) $\rightarrow$ log-quadratic mortality model on child mortality

Bias-adjustment process and TFR estimation

$\tau_{x,a,t}$ → bias-adjustment parameters

Interpretation → extent to which the **age-specific child-woman ratios** $(K_{x,a,t})$ are biased in **FamiLinx**.

Problem → bias patterns are not available for all countries and years.

Solution → borrowing information from countries with more reliable data.

- **information pooling** → $\log(\tau_{x,a,t}) \sim \mathcal{N}(\nu_{x,t}, \sigma_\tau^2)$
- **smoothing over time** → $\nu_{x,t} \sim \mathcal{N}(2\nu_{x,t-1} - \nu_{x,t-2}, \sigma_\nu^2)$

Final Goal: draw estimates from the **marginal posterior distribution** $TFR_{t,a}|$ data, other parameters

Indirect Estimation

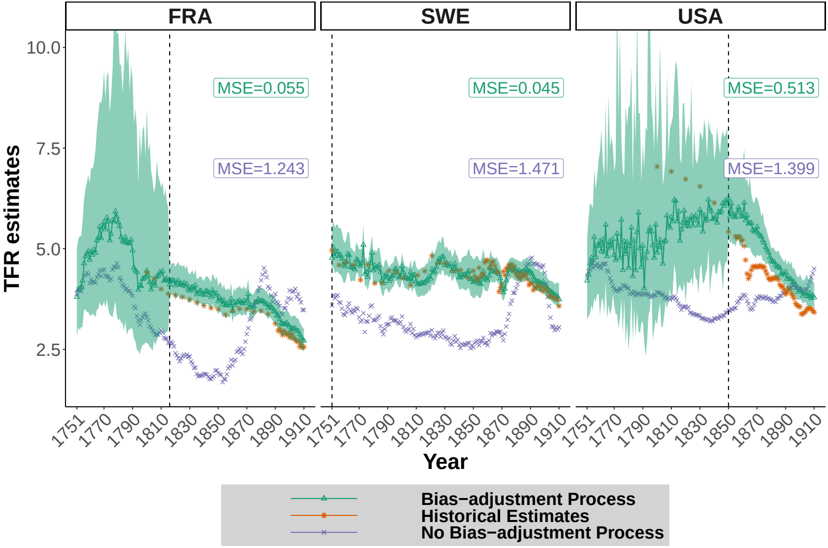
Extend the TFR decomposition by Hauer & Schmertmann (2020) through the addition of a bias-adjustment factor

$$TFR_{a,t} = \underbrace{r_{a,t}}_{\text{bias-adjustment multiplier}} \cdot \underbrace{\frac{1}{p_{a,t}}}_{\text{age multiplier}} \cdot \underbrace{\frac{1}{1 - 0.75 \cdot q_{0,a,t}}}_{\text{survival multiplier}} \cdot \underbrace{\frac{C_{a,t}}{W_{a,t}}}_{\text{CW ratio}}$$

The **bias multiplier** is defined to mimic **information sharing** across countries in absence of accurate data

$$r_{a,t} = \begin{cases} \frac{\text{True CW ratio in country} \quad a}{\text{Genealogical CW in country} \quad a} & \text{if} \quad a \in \mathcal{T}_a^{\text{true}} \\ \frac{\text{True CW ratio in country} \quad a^*}{\text{Genealogical CW in country} \quad a^*} & \text{if} \quad a \notin \mathcal{T}_a^{\text{true}} \end{cases}$$

Bayesian model results



Conclusions

- Combining **online genealogical data** with **more reliable data sources** allows to obtain relatively accurate TFR estimates
- Possibility to apply the proposed methods for fertility measurement in countries and populations with imperfect data
- Estimated fertility patterns seem to align with those observed in previous historical studies

Any Questions??

Looking forward to your feedback!

Contact: `riccardo.omenti2@unibo.it`

romenti.github.io @OmentiRiccardo

Essential Bibliography

Robert Stelter and Diego Alburez-Gutierrez.
Representativeness is crucial for inferring demographic processes from online genealogies: Evidence from lifespan dynamics.
Proceedings of the National Academy of Sciences, 119(10):e2120455119, 2022.

Joanna Kaplanis, Assaf Gordon, Tal Shor, Omer Weissbrod, Dan Geiger, Mary Wahl, Michael Gershovits, Barak Markus, Mona Sheikh, Melissa Gymrek, et al.
Quantitative analysis of population-scale family trees with millions of relatives.
Science, 360(6385):171–175, 2018.

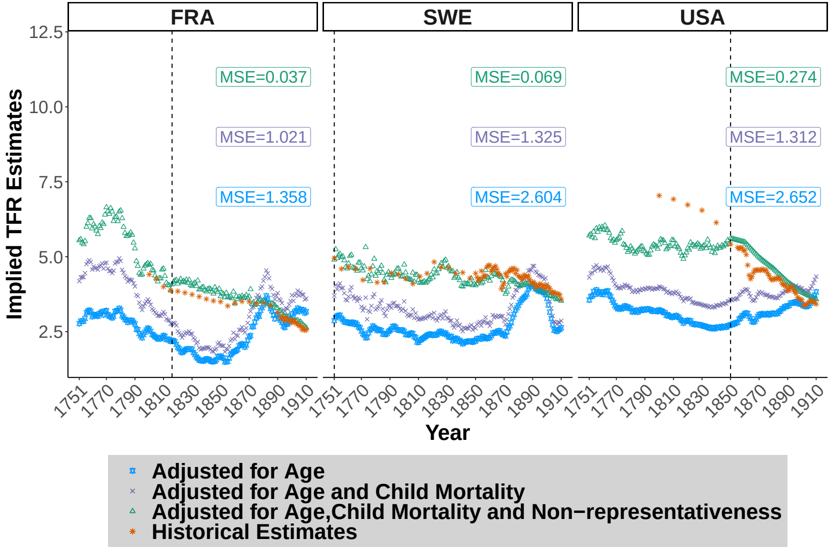
Saverio Minardi, Giulia Corti, and Nicola Barban.
Historical patterns in the intergenerational transmission of lifespan and longevity: Evidence from the united states, 1700-1900.
SocArXiv. doi, 10, 2023.

Mathew E Hauer and Carl P Schmertmann.
Population pyramids yield accurate estimates of total fertility rates.
Demography, 57(1):221–241, 2020.

Carl P Schmertmann and Mathew E Hauer.
Bayesian estimation of total fertility from a population's age–sex structure.
Statistical Modelling, 19(3):225–247, 2019.

Michael Chong, Diego Alburez-Gutierrez, Emanuele Del Fava, Monica Alexander, Emilio Zagheni, et al.
Identifying and correcting bias in big crowd-sourced online genealogies.
Max Planck Institute for Demographic Research, 2022.

Indirect estimation results



Bayesian model graphical illustration

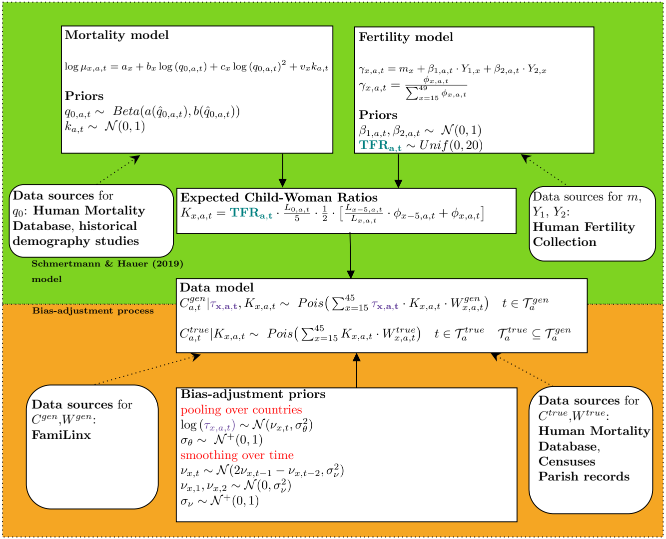


Figure: Proposed Hierarchical Bayesian Model

Population Pyramids

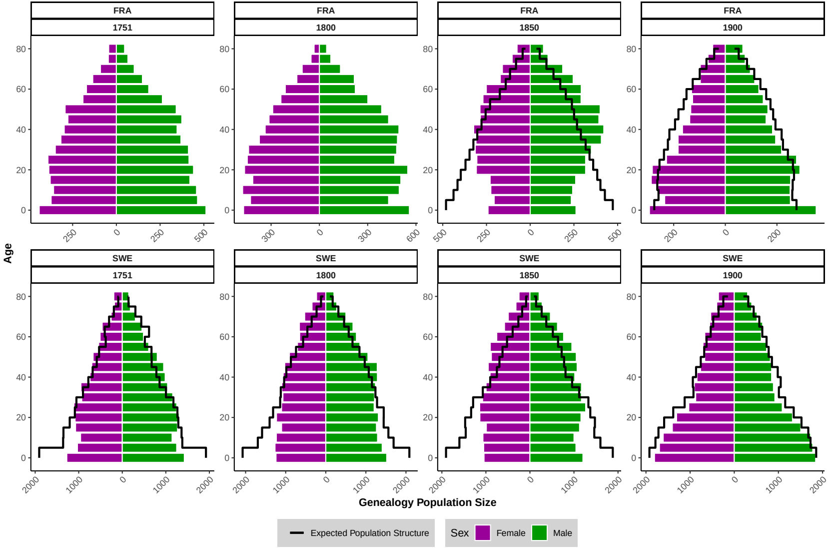


Figure: Population pyramids in France and Sweden in selected years .

Child Mortality Estimates

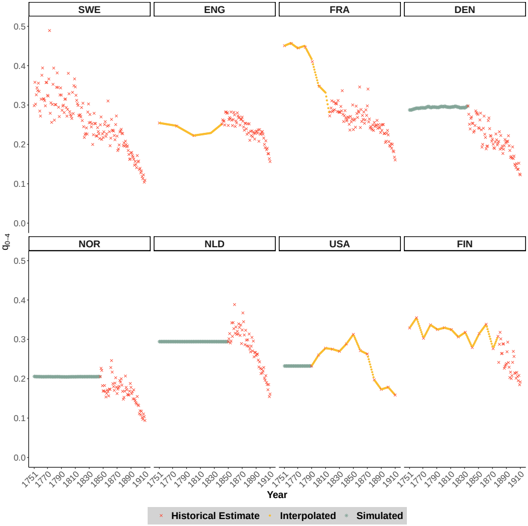


Figure: Child mortality estimates by country in the historical period 1751-1910.

Bayesian model results

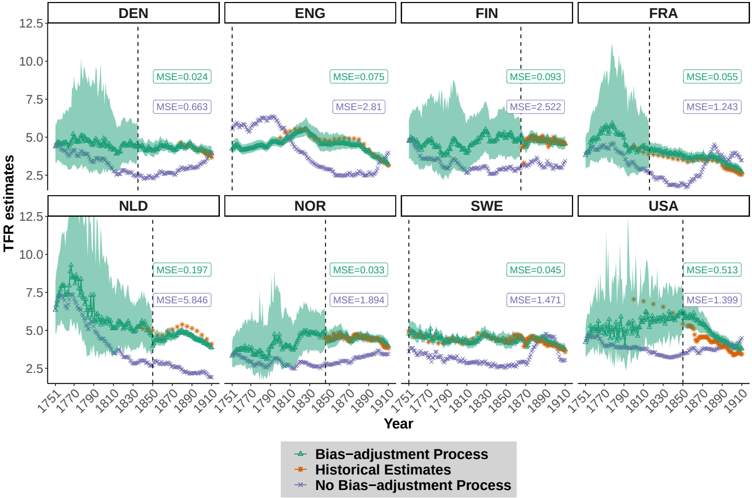


Figure: TFR posterior estimates for the period $1751 - 1910$ in the 8 selected countries with and without the bias-adjustment process. 95% credible intervals are also included.

Extended TFR estimates

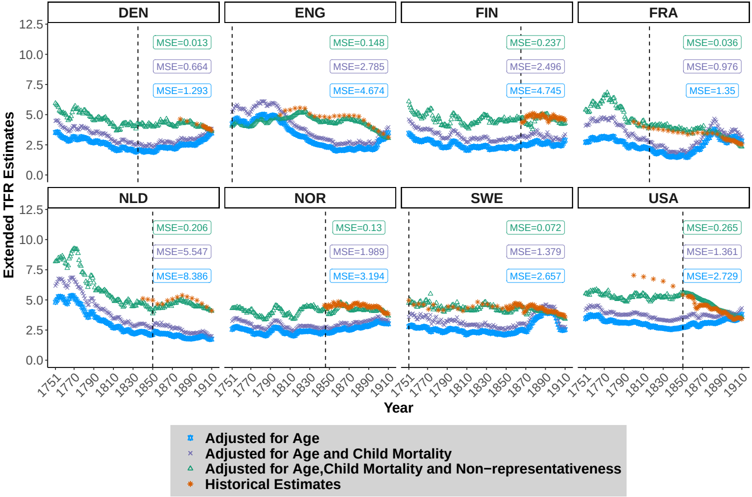


Figure: Extended TFR estimates for the period $1751 - 1910$ in the 8 selected countries with $p_{a,t} = 10.65 - 12.55 \cdot \pi_{25-34}$.

Indirect estimation results

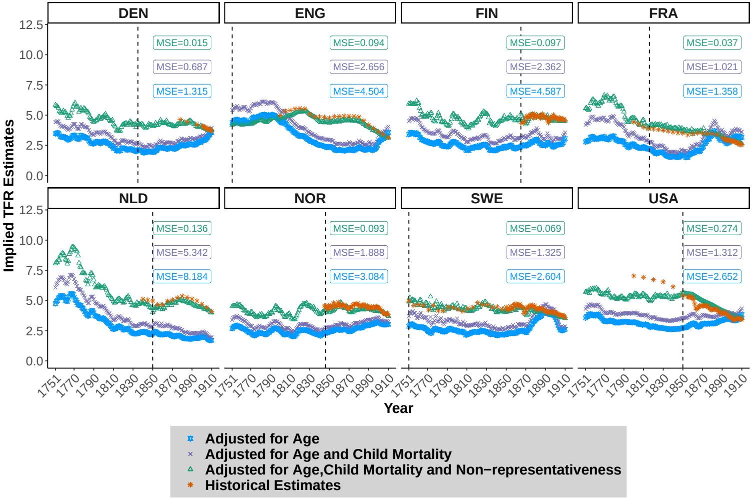


Figure: Time series of indirect TFR estimates with $\frac{1}{p_{a,t}} = 7$ for the historical period 1751-1910 in the 8 selected countries.