Leveraging online crowdsourced genealogical data to measure fertility in Europe and North America during the First Demographic Transition

Riccardo Omenti

BSPS Conference 2023

September 11th, 2023

What are online crowd-sourced genealogies?

- Web sites that allow a decentralized network of users to reconstruct their own family tree.

- bottom-up user-generated content.

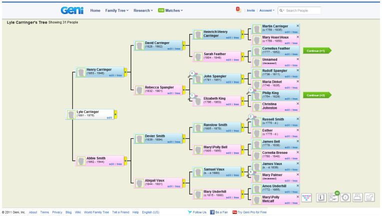

Figure: family tree on geni.com.

Figure: geni.com home page.

Online genealogies for Demographic Research

- Network of profiles with **life courses unfolding across different centuries** and with **transnational kin ties**.

- Unique opportunity to gain new insights about the **evolution of long-term demographic dynamics** (Chong et al., 2022), the **intergenerational transmission of demographic behaviors** (Kolk, 2014; Minardi et al., 2023) as well as the **study of demographic change from kin's perspective** (Murphy, 2011).

- Several potential biases (Alburez-Gutierrez et al., 2022): **bias due to the bottom-up construction of the genealogical tree**, **selection bias**, **selective-remembering**.

Objectives

- Development of a **Bayesian Hierarchical Model** and **Indirect Estimation Indicators** to examine fertility patterns in Europe and North America (1751-1900).

- Providing new estimates of fertility levels for historical periods lacking ground-truth data.

- **Critical Analysis** of the potential of online genealogical data for demographic research.

FamiLinx

- A huge data set curated by Kaplanis et al. (2018) consisting of **86 million** individuals over the last 400 years.

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: Abstract of the article by Kaplanis et al. (2018)

Limitations in FamiLinx

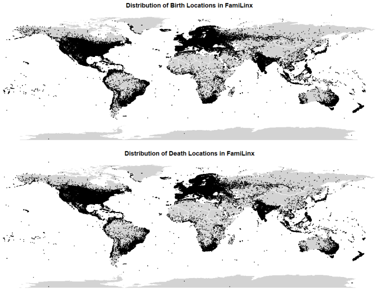

Figure: Distribution of profiles by countries of birth and death.

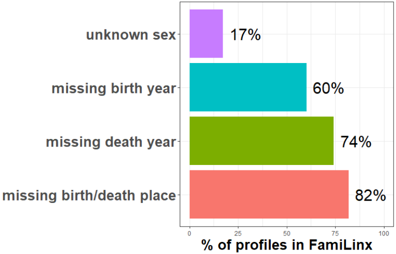

Figure: Percentage of missing data in key demographic variables.

Country selection

- Country selection procedures: **exact matching using the country code**, **regular expression matching** and **inferred coordinates**.

- We focus on two countries:

  - **Sweden** $\rightarrow$ accurate time series of national demographic rates dating back to the middle of the 18^{th} century.

  - **United States of America** $\rightarrow$ country with the highest number of vital events.

Sample Selection

1. Initial sample of **86 million** observations.
2. Selection of approximately **1.5 million** profiles born and/or died in one of the two previous countries.
3. Inclusion of profiles with the same country of birth and death, death year≥ 1741, birth year≤ 1900, age at death≥ 0 and ≤ 110.
4. A final sample of $\mathbf{987,188}$ individuals and **48,901,405** person-years is selected.

Fertility estimation based on Population Pyramid

Following Schmertmann & Hauer (2019, 2020), the following factorization for the Total Fertility Rate (TFR) is proposed.

Proposed Factorization of TFR

$$TFR = \underbrace{\frac{1}{r}}_{\textbf{under-reporting}} \times \underbrace{\underbrace{\frac{1}{s}}_{\textbf{survival multiplier}} \times \underbrace{\frac{1}{p}}_{\textbf{age multiplier}} \times \underbrace{\frac{C_{0-4}}{W_{15-49}}}_{\textbf{CW ratio}}}_{\textbf{Hauer \& Schmertmann (2019, 2020)}}$$

Class of Indirect TFR estimates

Adjusted for maternal age

- $$iTFR_{t,c} = 7 \cdot \frac{C_{0-4,t,c}}{W_{15-49,t,c}}$$
  $$xTFR_{t,c} = \left(10.65 - 12.55\pi_{25-34,t,c}\right) \cdot \frac{C_{0-4,t,c}}{W_{15-49,t,c}}$$

Adjusted for maternal age and infant mortality

- $$iTFR^{+}_{t,c} = \left(\frac{1}{1-0.75q_{5,t,c}}\right) \cdot 7 \cdot \frac{C_{0-4,t,c}}{W_{15-49,t,c}}$$
  $$xTFR^{+}_{t,c} = \left(\frac{1}{1-0.75q_{5,t,c}}\right) \cdot \left(10.65 - 12.55\pi_{25-34,t,c}\right) \cdot \frac{C_{0-4,t,c}}{W_{15-49,t,c}}$$

Adjusted for maternal age, infant mortality and under-registration

- $$iTFR^{*}_{t,c} = \frac{1}{r_{t,c^*}} \cdot \left(\frac{1}{1-0.75q_{5,t,c}}\right) \cdot 7 \cdot \frac{C_{0-4,t,c}}{W_{15-49,t,c}}$$
  $$xTFR^{*}_{t,c} = \frac{1}{r_{t,c^*}} \cdot \left(\frac{1}{1-0.75q_{5,t,c}}\right) \cdot \left(10.65 - 12.55\pi_{25-34,t,c}\right) \cdot \frac{C_{0-4,t,c}}{W_{15-49,t,c}}$$

Strategy for the estimation of r

- Pick a test country with a complete time series of ground truth TFRs over the whole historical period (Sweden)
- Estimate the under-reporting multiplier.

$$\frac{1}{r_{t,c^*}} = TFR^{\text{true}}_{t,c^*} \times \frac{W_{15-49,t,c^*}}{C_{0-4,t,c^*}} \times s_{t,c^*} \times p_{t,c^*}$$

- Use the previous multiplier to calculate $iTFR^*_{t,c}$ and $xTFR^*_{t,c}$
- Note that $\frac{1}{r_{t,c^*}}$ does not vary across countries.

Bayesian Estimation

- Incorporation of **expert knowledge** (priors) about:
  - child mortality
  - maternal age distribution
  - bias in the observed CW ratios

- Credible intervals of TFR estimates

- **TFR Estimates** → **median** of the conditional posterior distribution of TFR given the observed data and the other parameters.

- **Assumption** → the bias in the Child-Woman ratios does not change significantly in the two countries.

Extented Bayesian TFR ($bTFR^*$)

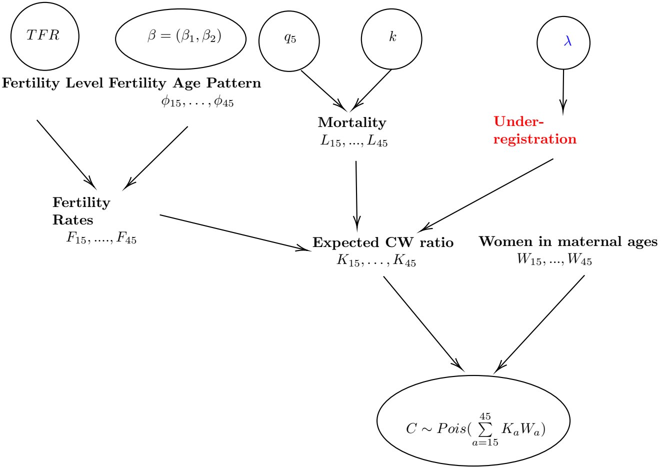

Figure: Proposed Hierarchical Bayesian Model

Estimating TFR using online genealogies

1. Development of country- and period- specific population pyramids.
2. Smooth the genealogy-based counts of children in the age class $0 - 4$ and of women in maternal ages $(15 - 49)$ through a 10-year moving average.
3. Employ the smooth counts to estimate the country- and period-specific TFRs.

Swedish and US Population Pyramids

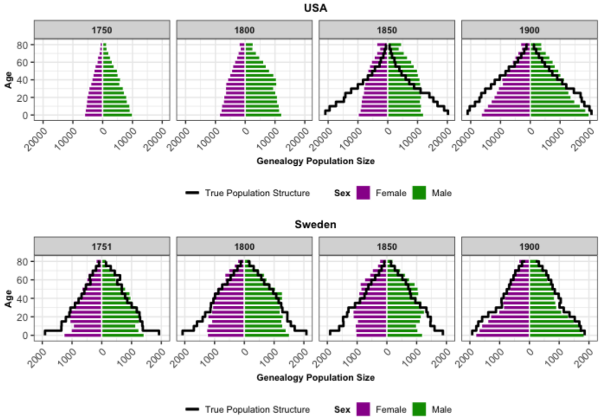

Figure: Genealogy-based Swedish and US population pyramids for calendar years 1750, 1800, 1850 and 1900 for different sub-samples.

Infant mortality in the US and Sweden

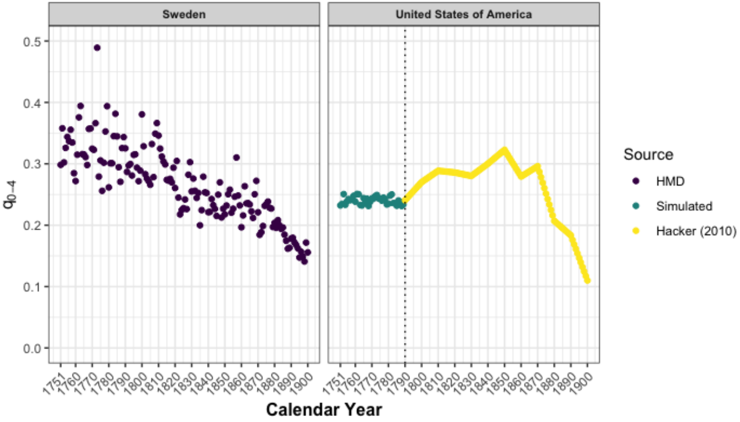

Figure: Probability of death under age 5 (q_{0-4}) in Sweden and in the US during the historical period $1750 - 1900$.

$iTFR$, $iTFR^+$ and $iTFR^*$ in Sweden and the US

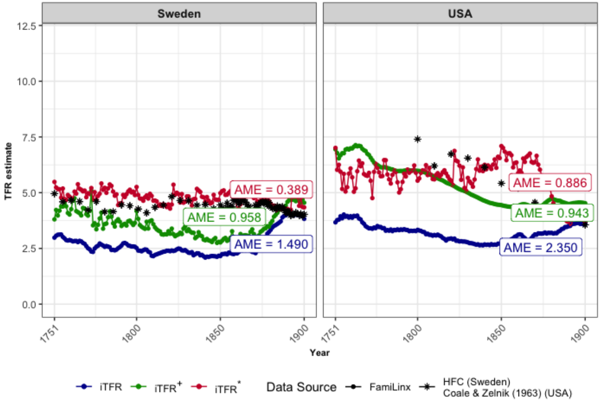

Figure: Time series of TFR estimates in Sweden for the historical period 1751-1900.

$bTFR$ and $bTFR^*$ in Sweden and the US

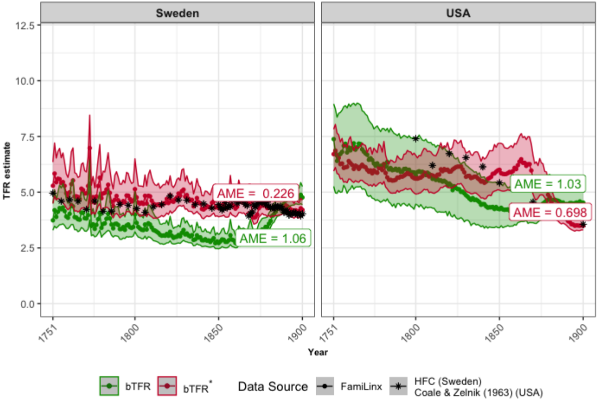

Figure: Time series of TFR estimates in the US for the historical period 1751-1900.

Main Limitations

- Lack of ground truth historical infant mortality rates for the US.
- Difficult to draw appropriate conclusions about the actual timing of the First Demographic Transition.
- Bias in the Child-Woman ratio is assumed to be the same in the considered countries.

Conclusions

- Reconstruction of historical TFR estimates for historical periods and countries without ground-truth data.
- Additional adjustment for sample under-registration provides more precise TFR estimates.
- Strong potential for fertility estimation in data-sparse settings.
- Better representativeness of genealogies towards the end of the 19^{th} century. (Stelter & Alburez-Gutierrez, 2022).

What comes next?

- Deeper investigation into the potential of online genealogical data for the examination of historical fertility patterns.
- Improving the role of the under-reporting multiplier as a prior in the Bayesian Model.
- Incorporation of a dependence structure among the parameters to allow information to be shared across calendar years.

Thank you!

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.

Diego Alburez-Gutierrez, Nicola Barban, Hal Caswell, Martin Kolk, Rachel Margolis, Emily Smith-Greenaway, Xi Song, Ashton M Verdery, and Emilio Zagheni.
Kinship, demography, and inequality: Review and key areas for future development.
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.

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.

Guillaume Blanc.
Demographic change and development using crowdsourced genealogies.
2022.

$bTFR$: Bayesian Total Fertility Rate

The objective is to obtain the posterior distribution TFR after observing the number of children under age 5 and the distribution of women by childbearing age group.

The point estimates of $bTFR$ are given by the median of the conditional distribution $TFR|C$.

$$P(TFR|C) \propto \int L(C|TFR, \beta, {}_5q_0, k) \cdot f_\beta(\beta) \cdot f_{{}_5q_0}({}_5q_0) \cdot f_k(k) \cdot f_\lambda(\lambda) d\beta d{}_5q_0 dk d\lambda$$

$$C|TFR, \beta, {}_5q_0, k \sim \textsf{Pois}\left(\sum_{x=15}^{45} W_x K_x(TFR, \beta, {}_5q_0, k, \lambda)\right)$$

$$TFR \sim \textsf{Unif}(0, 20)$$

$$\beta \sim \textsf{MVN}_2(\mathbf{0}_2, I_2)$$

$${}_5q_0 \sim \textsf{Beta}(a({}_5\hat{q}_0), b({}_5\hat{q}_0))$$

s.t.

$$P({}_5q_0 < 0.5 \cdot {}_5\hat{q}_0) = P({}_5q_0 > 2 \cdot {}_5\hat{q}_0) = 0.05$$

$$k \sim N(0, 1)$$

$$\lambda \sim N(\log(r), 10^{-3})$$

Expected Child-Woman Ratios: K_a

The relationship between the Expected Child-Woman ratio and the demographic parameters comes from Formal Demography.

From the first row of a Leslie Matrix, we can calculate the expected number of children per woman in the age group $[a, a+5)$ denoted by K_a.

$$K_a = \left[\frac{L_{a-5}}{L_a} \cdot F_{a-5} + F_a\right] \cdot \frac{L_0}{2} \cdot exp(\lambda)$$

$$K_a \cdot exp(-\lambda) = TFR \cdot \underbrace{\frac{L_0}{5}}_{s} \cdot \underbrace{\frac{1}{2} \cdot \left[\frac{L_{a-5}}{L_a} \cdot \phi_{a-5} + \phi_a\right]}_{p_a} \cdot exp(\lambda)$$

$\frac{L_0}{5}$ $\rightarrow$ expected number of children still alive in the past five years.
p_a $\rightarrow$ average of the fertility proportions in the maternal age groups a and $a-5$ with year weight on the age group $a-5$ to account for maternal mortality.

$$C = \sum_{a=15}^{45} K_a \cdot W_a$$

$$\frac{C}{W} = TFR \cdot s \cdot exp(\lambda) \cdot \sum_{a=15}^{45} \frac{W_a}{W} p_a$$

$$\frac{C}{W} = TFR \cdot s \cdot exp(\lambda) \cdot \cdot \bar{p} \rightarrow TFR = \frac{1}{s} \cdot \frac{1}{\bar{p}} \cdot \frac{C}{W}$$

Parameters: Fertility

Apply the following transformation to the the proportion of lifetime fertility that occurs in age group a
$\gamma_a = \ln\left(\frac{\phi_a}{\phi_{15}}\right) \; \forall a \in \{15, \ldots, 45\}$ and $\phi_a = \frac{5}{TFR} \cdot F_a$

Model the transformed parameters as

$$\underbrace{\boldsymbol{\gamma}}_{7\times 1} = \underbrace{\mathbf{m}}_{7\times 1} + \underbrace{\mathbf{X}}_{7\times 2} \cdot \underbrace{\boldsymbol{\beta}}_{2\times 1}$$

The values in $\mathbf{m}$ are the averages of the transformed parameter γ_a for each age group of interest based upon all fertility schedules up to the year 1900 from the Human Fertility Collection (collected in a rectangular matrix of dimension).

The two columns in $\mathbf{X}$ are first and second right singular vectors obtained through the Singular Value Decomposition of the matrix of transformed fertility schedules.

$\boldsymbol{\beta}$ is assigned a two-dimensional standard normal distribution to ensure its range is restricted on $[-2, 2] \times [-2, 2]$ to better mimic HFC schedules.

Parameters: Mortality

The estimation of the survival proportions among the mothers, i.e. L_x with $x \in \{0, 5, \ldots, 45\}$, is performed by considering the two-dimensional mortality model developed by Wilmoth et al. (2012).

The model is characterized by a quadratic relationship between the age-specific death rates and the probability of death under age 5.

$$\log(m_x) = a_x + b_x \cdot \log({}_5q_0) + c_x \cdot [\log({}_5q_0)]^2 + v_x \cdot k$$

a_x, b_x, c_x, v_x are age-specific coefficients estimated using information provided by 719 life tables from the Human Mortality Database through the Weighted Least Squares Method.

$k \in [-2, 2]$ denotes the relative excess of adult mortality over one might predict from knowledge of child mortality alone.

Parameters: undercount

The estimation of the undercount parameter is based upon the divergence of the genealogical child-woman ratio from the true child-woman ratio for a test country for which ground-truth data are available.

Let τ be the observed ratio of the true child-woman ratio to the genealogical one.

$$\tau_{t,c^*} = \frac{\frac{C^{\textbf{true}}_{0-4,c^*,t}}{W^{\textbf{true}}_{15-49,c^*,t}}}{\frac{C^{\textbf{gen}}_{0-4,c^*,t}}{W^{\textbf{gen}}_{15-49,c^*,t}}}$$

Assumption: the under-count parameter, denoted by λ, is generated from a normal distribution centered at the log of the observed ratio.

$$\lambda_{t,c} \sim N(\log(\tau_{t,c^*}), 10^{-3})$$

Inclusion of undercount multiplier

Consider the proposed **TFR decomposition** for some test country c^*

$$TFR = \frac{1}{r} \times \frac{1}{s} \times \frac{1}{p} \times \frac{C}{W}$$

where $\frac{1}{r}$ denote the undercount multiplier.

Consider .

Find the multiplier for each calendar year of interest using the inverse formula.

$$\frac{1}{r_{t,c^*}} = TFR^{\textbf{True}}_{c^*,t} \times \frac{W_{c^*,t}}{C_{c^*,t}} \times (1 - 0.75q_{0-4,t,c^*}) \times 7$$

For the years within the period $1751 - 1900$, for which the true TFR is not available, we use linear interpolation to infer the missing values.

Based on the time series of estimated multipliers for the test country c^*, we are able to find the undercount adjusted TFR estimates for the other countries of interest.

$$m_{c^*} = [\frac{1}{r_{1751,c^*}}, \ldots, \frac{1}{r_{1900,c^*}}]$$

$$TFR_{c,t} = \frac{1}{r_{t,c^*}} \times \frac{1}{1 - 0.75q_{0-4,t,c}} \times p \times \frac{C_{t,c}}{W_{t,c}}$$

Performance of the TFR estimates

The performance of the TFR estimates is assessed using the **Root Mean Squared Error** (RMSE)

$$AME_c = \frac{\sum_{t \in T} \left| TFR_{t,c}^{\textbf{gen}} - TFR_{t,c}^{\textbf{True}} \right|}{T}$$

- T refers to the total number of calendar years for which the true TFRs of country c are available.
- $TFR_{t,c}^{\textbf{gen}}$ denotes the genealogy-based TFR in country c during the calendar year t.
- $TFR_{t,c}^{\textbf{True}}$ denotes the true TFR in country c during the calendar year t.