Development of a Bayesian model to unveil historical fertility patterns using online genealogical data

Riccardo Omenti [1] Monica Alexander [2,3] Nicola Barban [1]

[1]Department of Statistical Sciences, University of Bologna [2]Department of Statistical Sciences, University of Toronto [3]Department of Sociology, University of Toronto

Background

- Online genealogical populations derive from digital trees constructed by a transnational network of users willing to trace back their ancestors.
- Harnessing online genealogies for demographic research offer much promise but it also raises important methodological challenges.
- This project relies on the database **FamiLinx** created by Kaplanis et al. (2018).

Main contribution

- Proposal of a statistical framework for producing fertility estimates (TFR) in historical populations by combining online genealogical data with more traditional data sources (e.g, Censuses, Population registers).
- Generate time series of historical TFR estimates for Sweden, USA and France during the time period 1751-1910.

Motivation

- Online genealogical data are not primarily designed for demographic research.
- Combining data from online genealogical populations with more reliable data sources is essential to measure demographic estimates accurately.
- Extension of the modelling framework by Schmertmann and Hauer (2019) to estimate TFRs using counts from population pyramids while adjusting for the non-representativeness of the population of interest and for child mortality.

Bayesian Model summary

Data model

$$C^{gen}_{x,t} | \theta_{a[i],t}, K_{x,a,t} \sim \; Pois(\theta_{a[i],t} \cdot \sum_{x=15}^{45} K_{x,a,t} \cdot W^{gen}_{x,a,t})$$

$$C^{true}_{a[i],t[i]} | K_{x,a[i],t[i]} \sim \; Pois(\sum_{x=15}^{45} K_{x,a[i],t[i]} \cdot W^{true}_{x,a[i],t[i]})$$

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

Mortality model

$$\log \mu_{x,a,t} = a_x + b_x \log(q_{0,a,t}) + c_x \log(q_{0,a,t})^2 + v_x k_{a,t}$$

Priors

$q_{0,a,t} \sim \; Beta(a(\hat{q}_{0,a,t}), b(\hat{q}_{0,a,t}))$
$k_{a,t} \sim \; \mathcal{N}(0,1)$

Person-years are determined through standard life table relationships. a_x, b_x, c_x, d_x are fixed constants. The parameters of $q_{0,a,t}$ are based on the observed probabilities $\hat{q}_{0,a,t}$.

Fertility model

$\gamma_{x,a,t} = m_x + \beta_{1,a,t} \cdot Y_{1,x} + \beta_{2,a,t} \cdot Y_{2,x}$
$\gamma_{x,a,t} = \frac{\phi_{x,a,t}}{\sum_{x=15}^{49} \phi_{x,a,t}}$

Linear combination of mean plus two principal components from the Human Fertility Collection

Priors

$\beta_{1,a,t}, \beta_{2,a,t} \sim \; \mathcal{N}(0,1)$
$TFR_{a,t} \sim Unif(0,20)$

Under-reporting priors

$\theta_{a[i],t} \sim \mathcal{N}(\nu_1, \sigma^2_\theta)$
$\sigma_\theta \sim \; \mathcal{N}^+(0,1)$
$\nu_3 \sim \mathcal{N}(2\nu_{3-1} - \nu_{3-2}, \sigma^2_\nu)$
$\nu_1, \nu_2 \sim \mathcal{N}(0,1)$
$\sigma_\nu \sim \mathcal{N}^+(0,1)$

Allows to account for the underestimation of children. Information on under-reporting are pooled over countries and time.

Figure 1. Graphical summary of the proposed Bayesian model

- Only native individuals from online genealogies with reasonable age at death (≥ 0 and ≤ 110) and non-missing sex are included.
- Reliable population estimates are observed only for selected years in France (1816-1910) and the US (1850-1910).
- The under-estimation process in France and the US during the years with missing population estimates is informed from Sweden.
- Child mortality estimates are obtain from multiple sources, including Human Mortality Database, INED and historical demography studies.

Estimated Population Pyramids

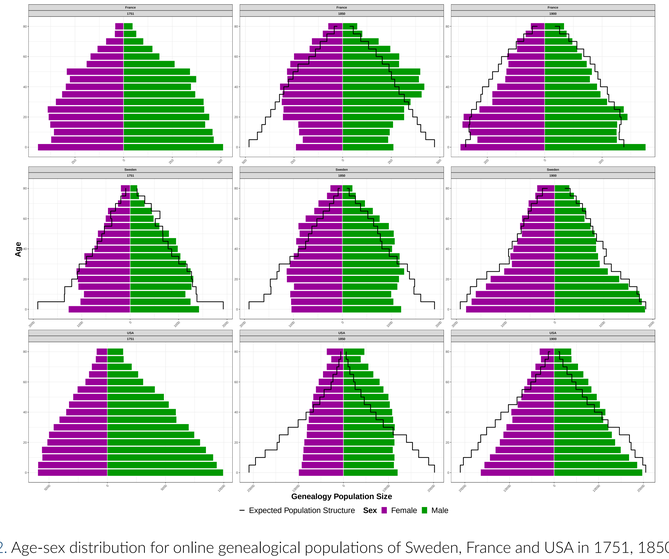

Figure 2. Age-sex distribution for online genealogical populations of Sweden, France and USA in 1751, 1850, 1900.

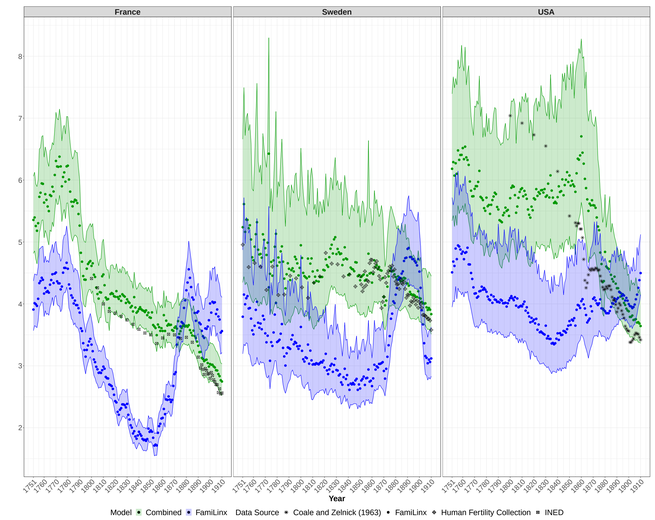

Figure 3. Time series of the TFR estimates by model type.

Model performance

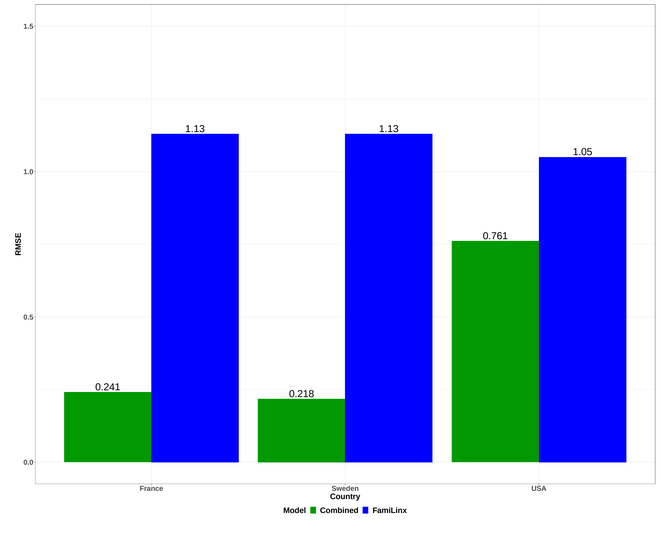

Figure 4. Overall Root Mean Squared Error (RMSE) by type of model and country of interest

Preliminary Conclusions

- By combining digital data with more representative data sources, the fertility estimates of the proposed model closely align the estimates from well-known Demographic Databases or previous historical demography studies.
- The accuracy of the estimates is especially evident for France and Sweden.
- The results for the US must be interpreted carefully as its historical demographic estimates are not as accurate as those of the two previous countries.
- The proposed statistical model could be employed for fertility estimation in other data-sparse settings.

References

[1] 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.

[2] 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.