

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

Riccardo Omenti<sup>1</sup> Monica Alexander<sup>2,3</sup> Nicola Barban<sup>1</sup>

<sup>1</sup>Department of Statistical Sciences, University of Bologna

<sup>2</sup>Department of Statistical Sciences, University of Toronto

<sup>3</sup>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

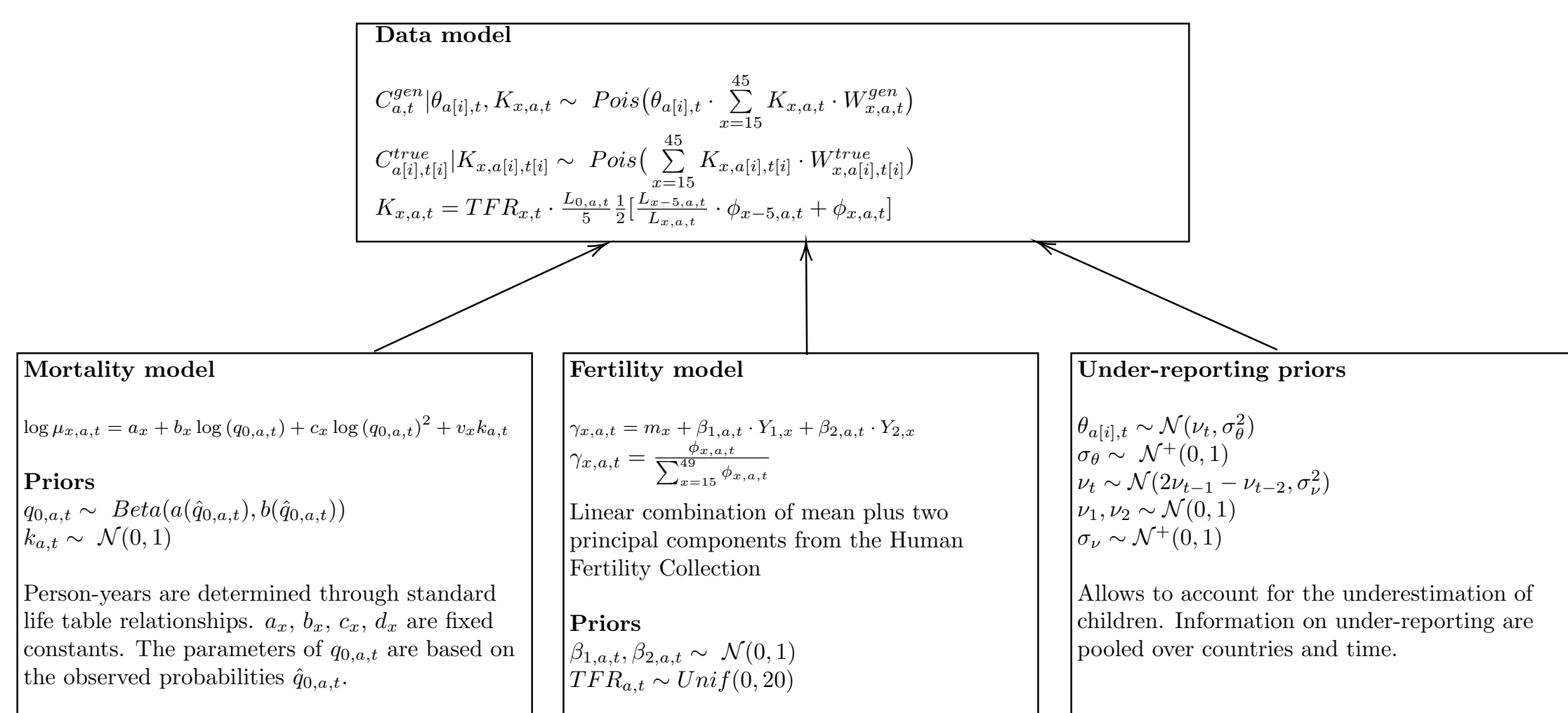


Figure 1. Graphical summary of the proposed Bayesian model

- Only native individuals from online genealogies with reasonable age at death ( $\geq 0$  and  $\leq 110$ ) 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 obtained from multiple sources, including Human Mortality Database, INED and historical demography studies.

## Estimated Population Pyramids

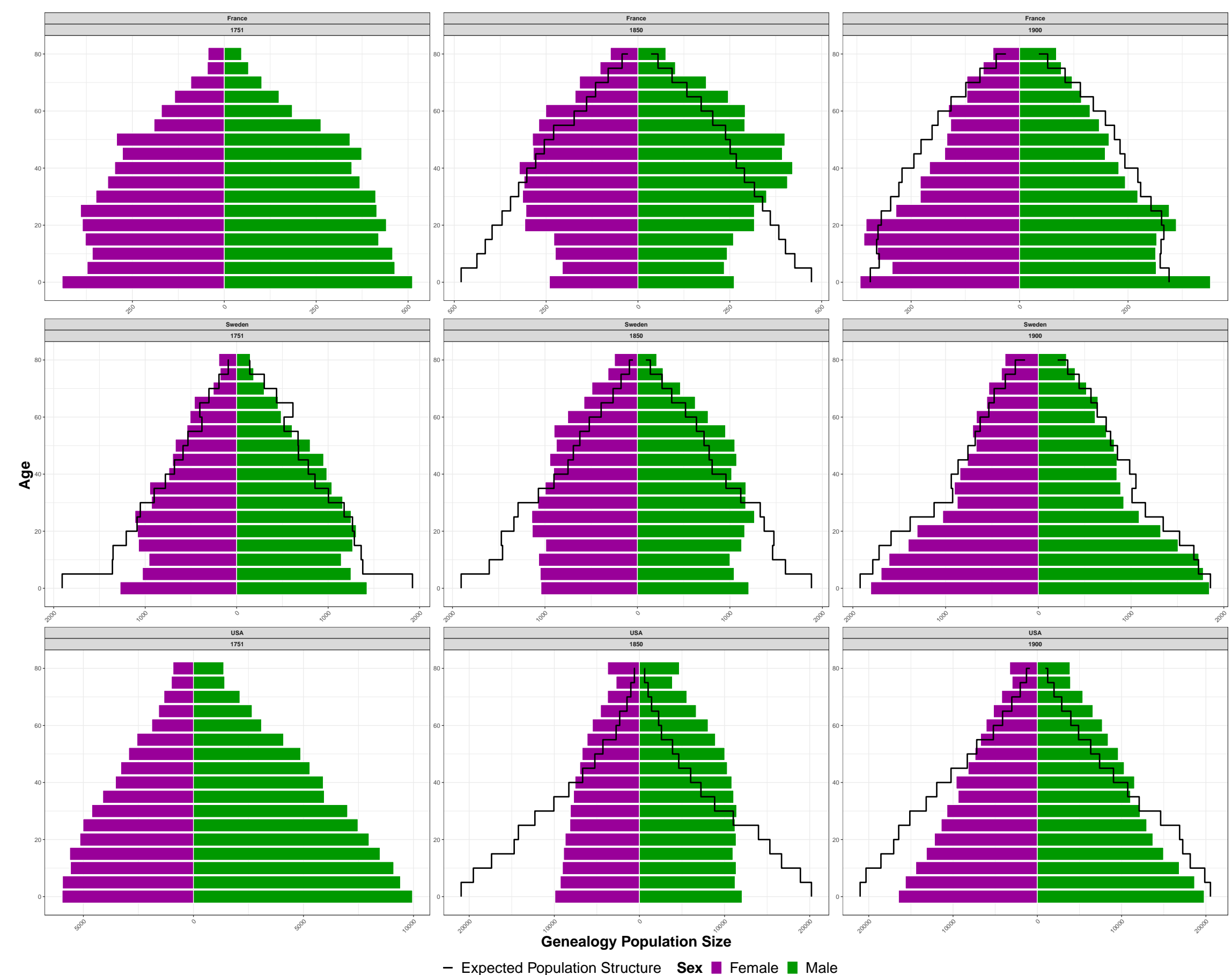


Figure 2. Age-sex distribution for online genealogical populations of Sweden, France and USA in 1751, 1850, 1900.

## Model performance

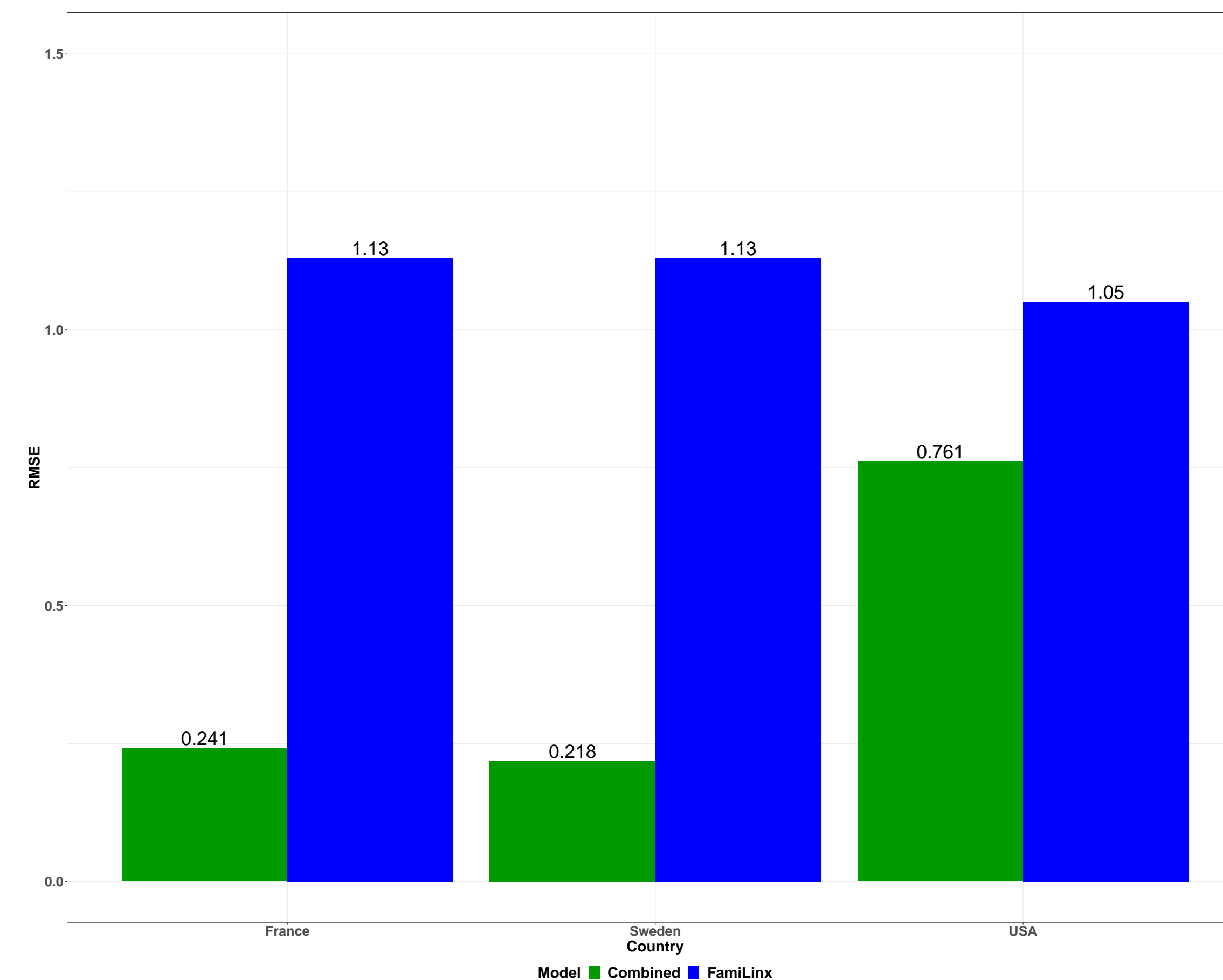


Figure 4. Overall Root Mean Squared Error (RMSE) by type of model and country of interest

## 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

- 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.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 865356

