Leveraging crowd-sourced online genealogical data to study the evolution of fertility

Riccardo Omenti

Max Planck Institute for Demographic Research
Rostock, Germany

March 15th, 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.

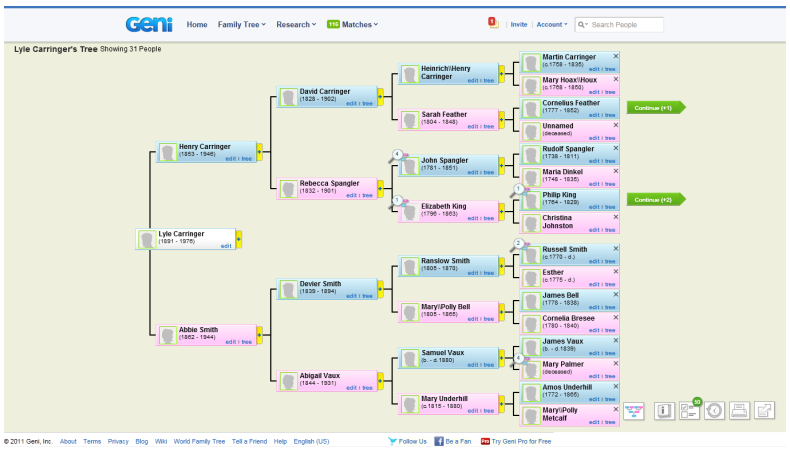


Figure: family tree on geni.com.

Figure: geni.com home page.

Online genealogies for Demographic Research

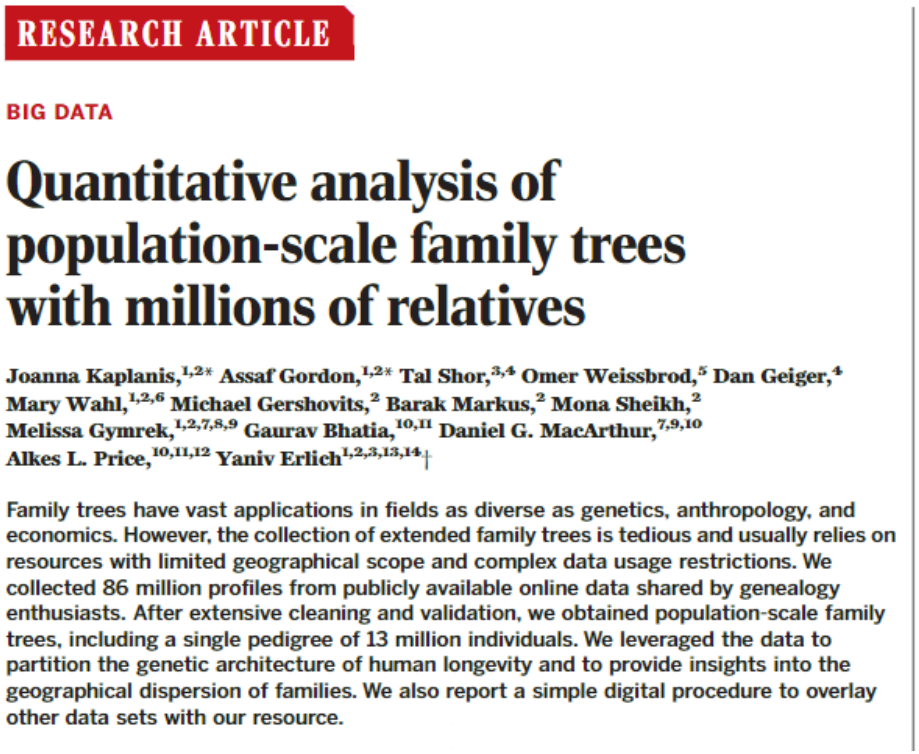
- Network of profiles with **life courses unfolding across different centuries** and with **transnational kin ties**.
- Unique opportunity to gain new insights about the **evolution of long-term demographic dynamics** (Chong et al., 2022), the **intergenerational transmission of demographic behaviors** (Kolk, 2014) 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

- Proposal of new indicators for the evaluation of the quality of online genealogical data.
- Fertility estimation from online genealogical data using population pyramids.
- Critical analysis of the potential of online genealogical data to measure fertility.

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 country of birth (Kaplanis et al., 2018)

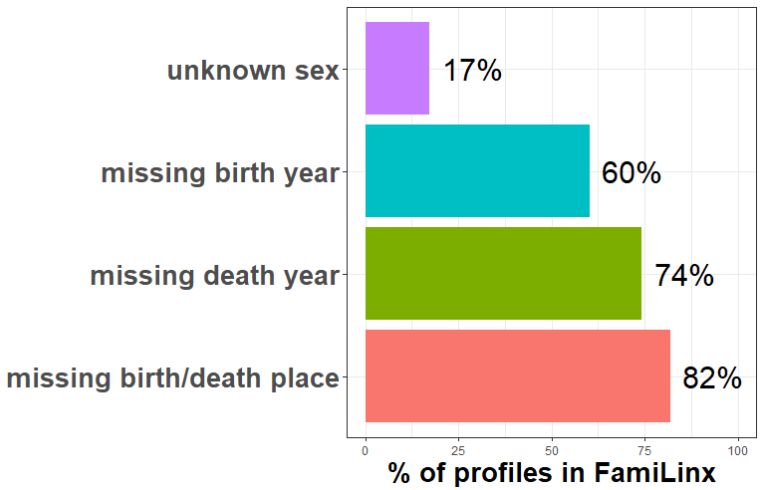


Figure: Percentage of missing data in key demographic variables.

Country selection

- The analysis will cover the historical period **1750 – 1900**.
- The following countries have been selected:
  - **Sweden** → accurate time series of national demographic rates dating back to the middle of the 18^{th} century.
  - France → first country to experience the fertility transition.
  - United States of America and **United Kingdom** → high percentage of profiles.
- Country selection procedures: **exact matching using the country code**, **regular expression matching** and **inferred coordinates**.

Sample Selection

1. Initial sample of **86 million** observations.
2. Selection of approximately **2 million** profiles born and/or died in one of the selected countries with at least one parent or one child.
3. Inclusion of profiles with the same country of birth and death, death year$\geq$ 1750, birth year$\leq$ 1900, age at death$\geq$ 0 and $\leq$ 110.
4. A final sample of **430**, **476** individuals is selected.

Data Quality Assessment

Demographic-information quality indicator

$$I_i^{\text{demo}} = \sum_{j=1}^{k} \frac{x_{ij}}{k}$$

$D_i = \{x_{i1}, x_{i2}, \ldots, x_{ik}\} \rightarrow$ set of discrete variables concerning the completeness of demographic information for an individual i in the genealogy.

Family-network quality indicator

$$I_i^{\text{fam}} = \sum_{j=1}^{p} \frac{y_{ij}}{p}$$

$F_i = \{y_{i1}, y_{i2}, \ldots, y_{ip}\} \rightarrow$ set of variables concerning the presence of parents in the 3 generations preceding an individual i in the genealogy.

Performance of the indicators

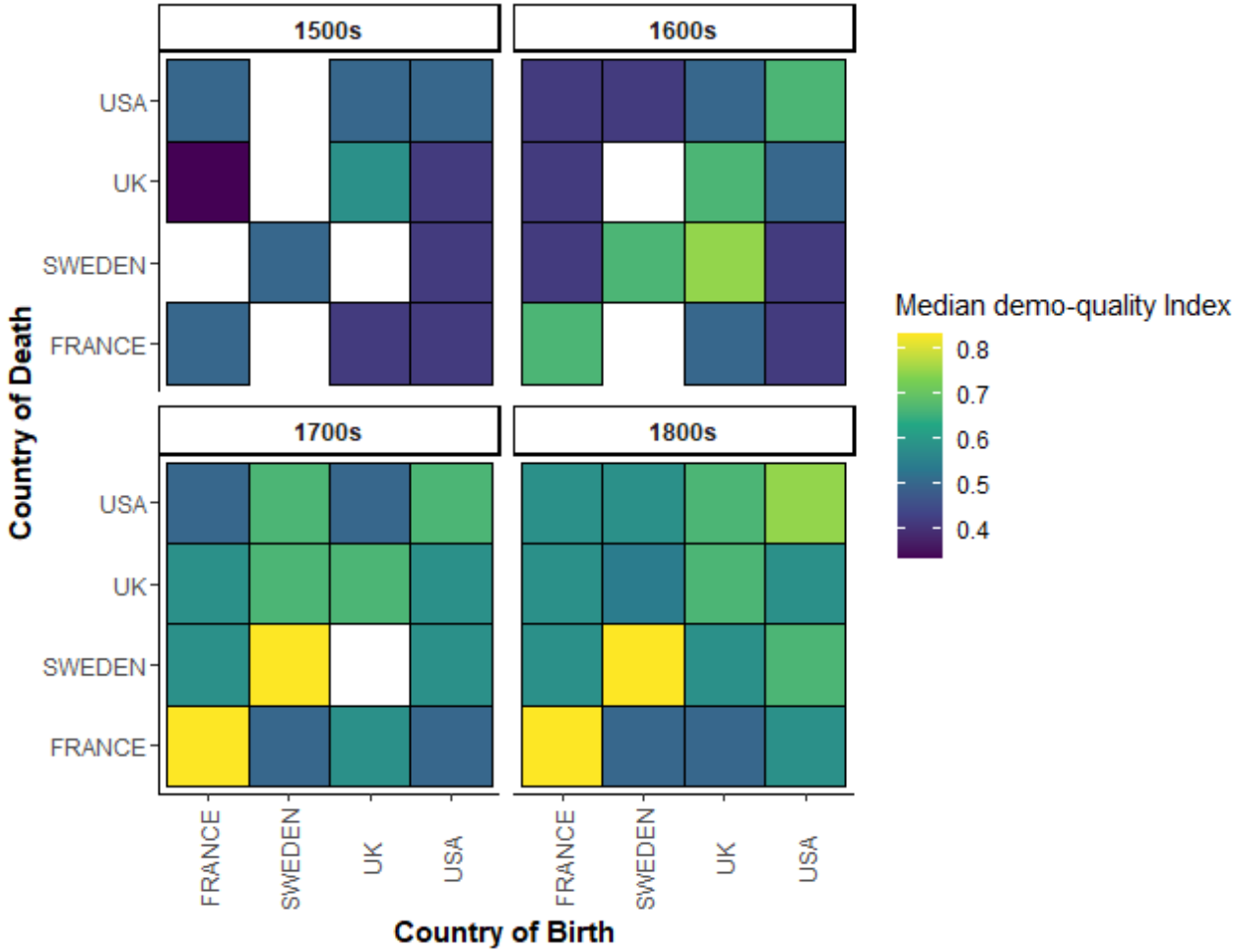


Figure: Median of demographic information quality index by country of birth and death across multiple centuries.

Performance of the indicators

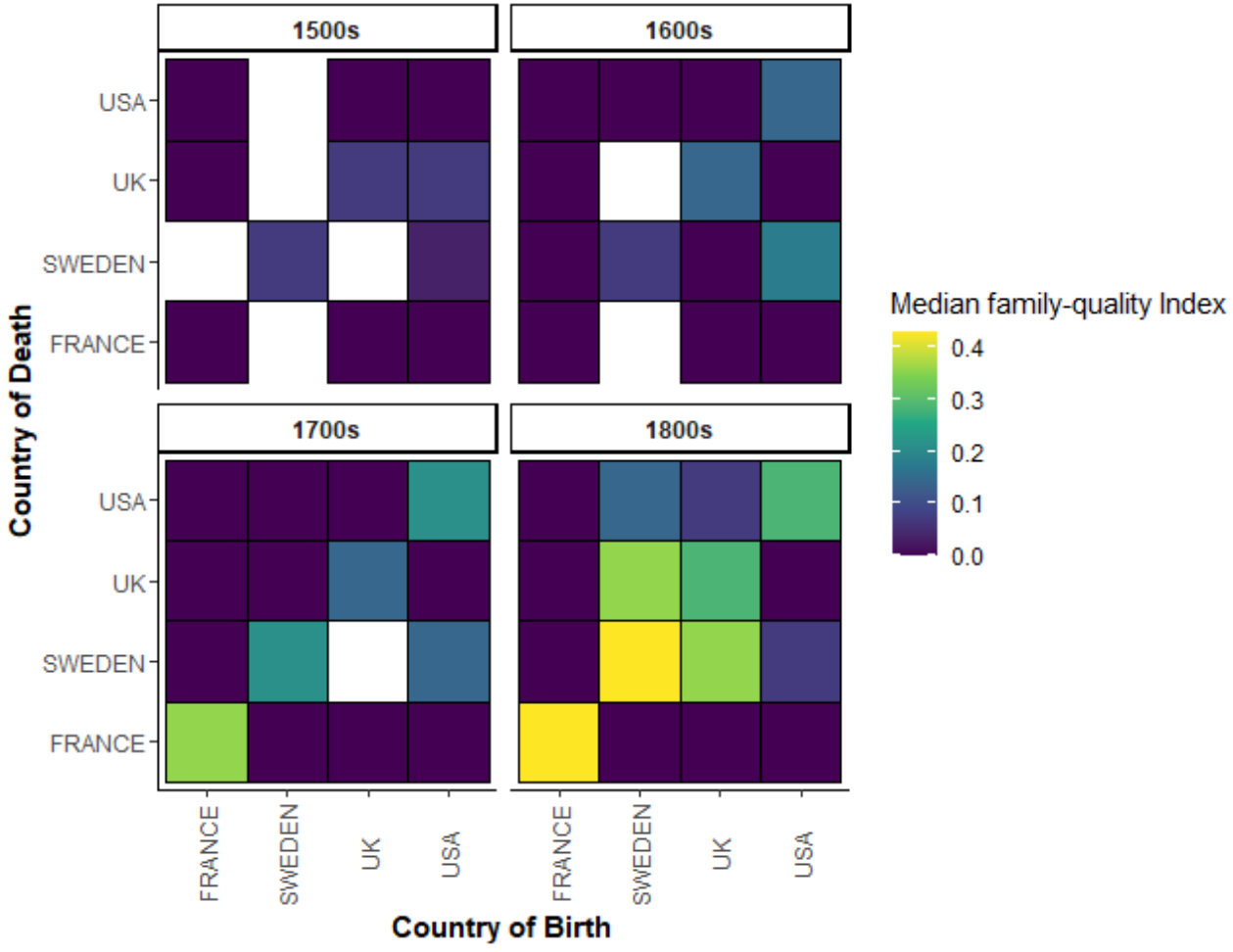


Figure: Median of family network quality index by country of birth and death across multiple centuries.

Overall quality indicator

Each subjected is assigned an overall quality indicator:

$$I_i^{\text{Overall}} = \frac{(I_i^{\text{demo}} + I_i^{\text{fam}})}{2}$$

Identification of multiple tresholds from the distribution of the overall quality indicator for the selection of country-specific sub-samples.

Proposed tresholds: **25th percentile**, **50th percentile**, **75th percentile**.

Fertility in Genealogies: Blanc's works

- First attempt by Blanc (2020,2022).
- Fertility estimated as:
  - **total number of children ever-born**.
  - to individuals having one parent in the four preceding generations with a fertility greater than 1.
- **Blanc's conclusions**:
  - representativeness of online genealogies in Europe from the $18th$ century onwards.
  - Evidence of an early fertility decline by France in the mid-18^{th} century.

Fertility in Genealogies: Blanc's works

Main weaknesses:

- Lack of use of a proper demographic fertility measure, i.e. Total Fertility Rate (TFR).
- Unclear sample selection.
- Using Coale & Watkins Fertility Index as gold-standard fertility measure to assess the representativeness of genealogy-based fertility may not be appropriate.

Fertility estimation based on Population Pyramid

Schmertmann & Hauer (2019) developed a series of indicators to estimate Total Fertility Rates from population pyramids.

Proposed Factorization of *TFR*

$$TFR = \underbrace{\frac{1}{s}}_{\textbf{survival multiplier}} \times \underbrace{\frac{1}{p}}_{\textbf{age-distribution multiplier}} \times \underbrace{\frac{C}{W}}_{\textbf{child-woman ratio}}$$

Usefulness of the Population Pyramid approach for online genealogies

- Knowledge of births disaggregated by maternal ages is not required.
- Possibility to include prior information to account for infant mortality and mothers' age distribution.
- No need to define direct links between mothers and their children.

Variants for the estimation of TFR

Five indicators of increasing complexity.

1. $iTFR = 7 \cdot \frac{C}{W}$
2. $xTFR = \left(10.65 - 12.55\pi_{25-34}\right) \cdot \frac{C}{W}$
3. $iTFR^{+} = \left(\frac{1}{1-{}_5q_0}\right) \cdot iTFR$
4. $xTFR^{+} = \left(\frac{1}{1-{}_5q_0}\right) \cdot xTFR$
5. $bTFR$

Bayesian TFR

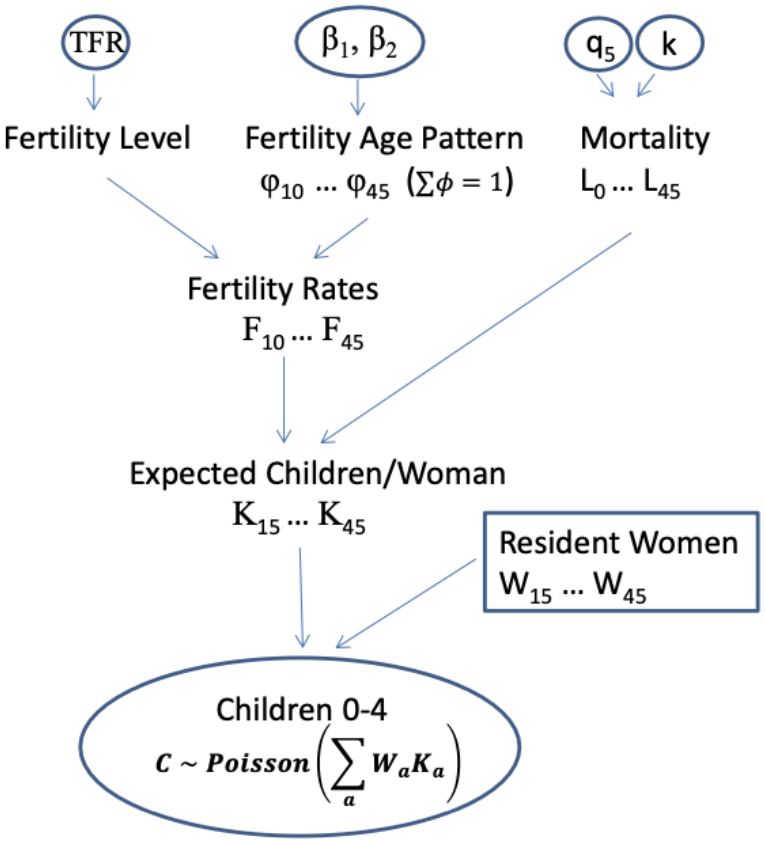


Figure: Graphical Summary of the Bayesian model for the calculation of *bTFR* (Schmertmann & Hauer, 2019)

Estimating TFR using online genealogies

1. Development of country-specific population pyramids considering different sub-samples.

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 Population Pyramid

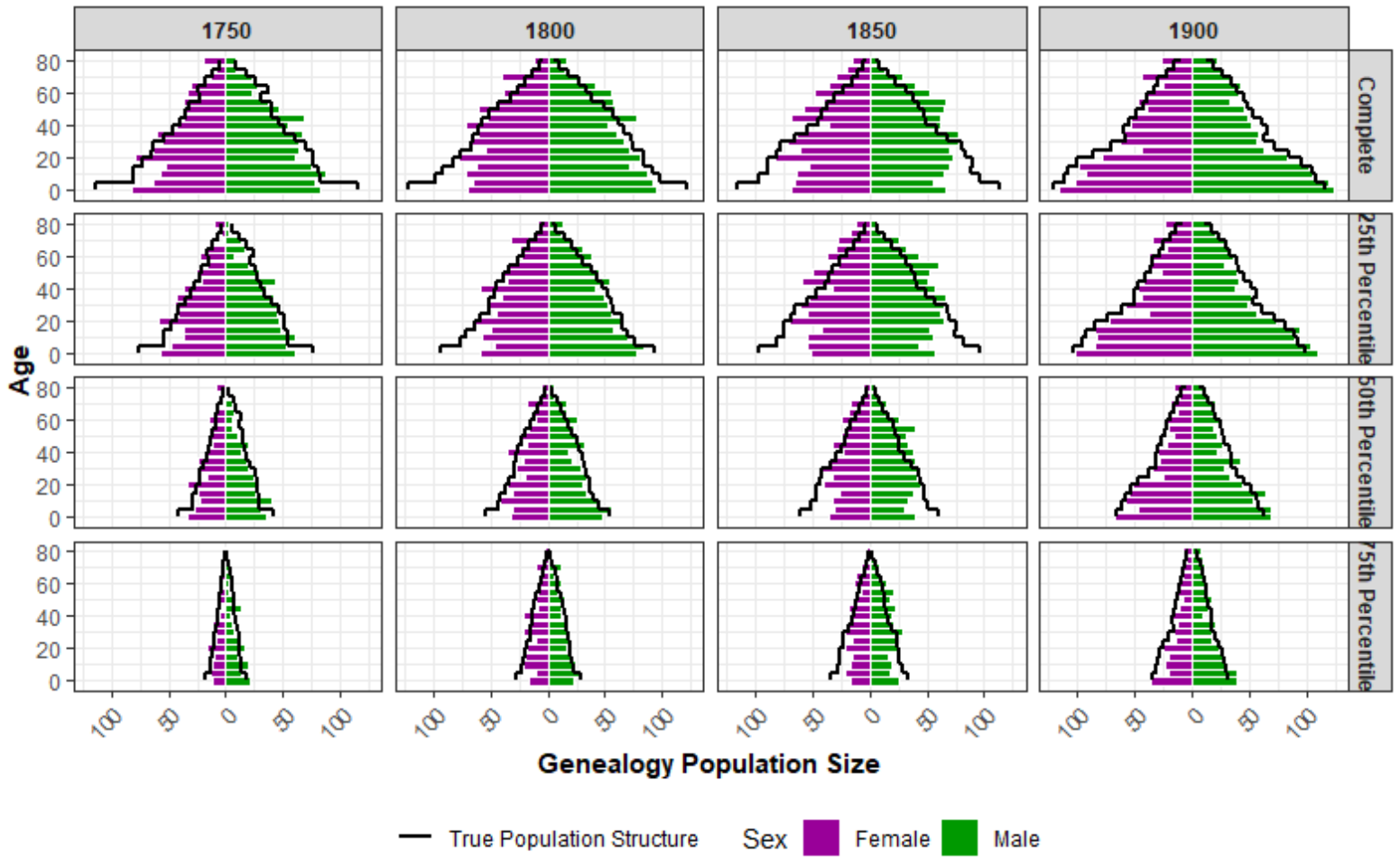


Figure: Genealogy-based Swedish population pyramids for calendar years 1750, 1800, 1850 and 1900 for different sub-samples.

UK Population Pyramids

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

xTFR and *iTFR* in Sweden and the UK

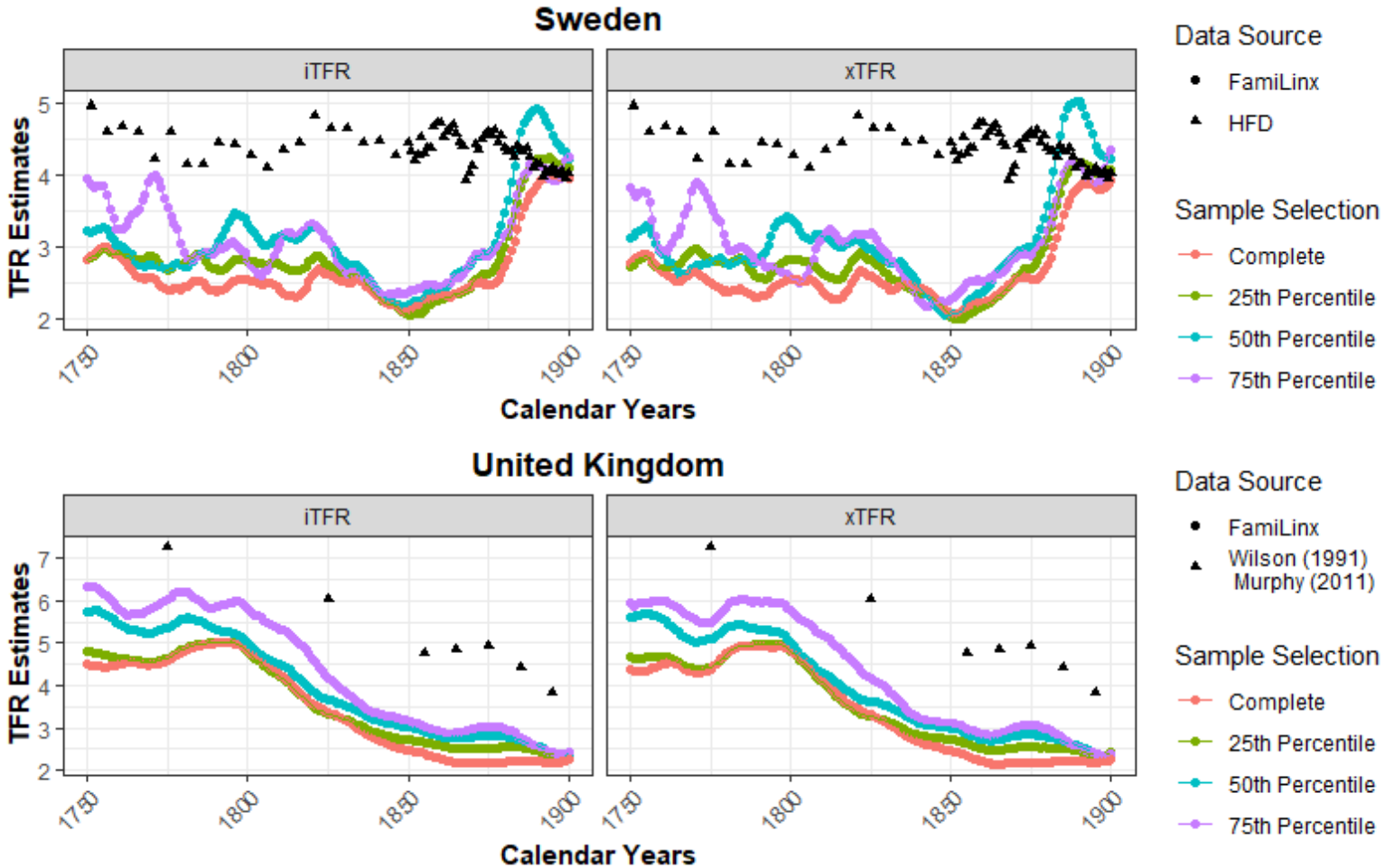


Figure: Time series of TFR estimates (*iTFR* and *xTFR*) in the selected countries for different sub-samples during the historical period $1750 - 1900$.

Infant mortality in the UK and Sweden

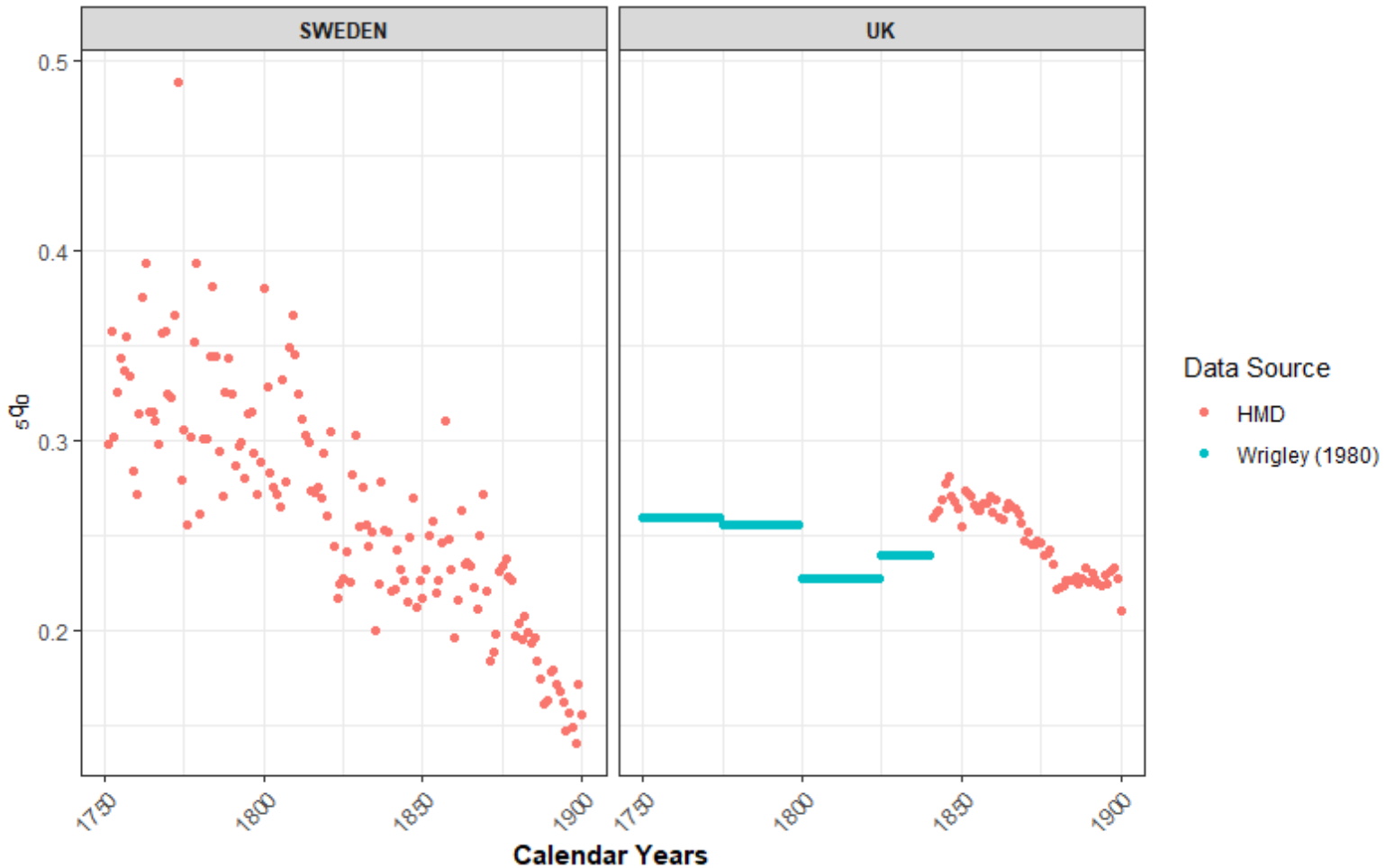


Figure: Probability of death under age 5 ($_5q_0$) in Sweden and in the UK during the historical period $1750 - 1900$.

$xTFR^+$, $iTFR^+$ and $bTFR$ in the UK and Sweden

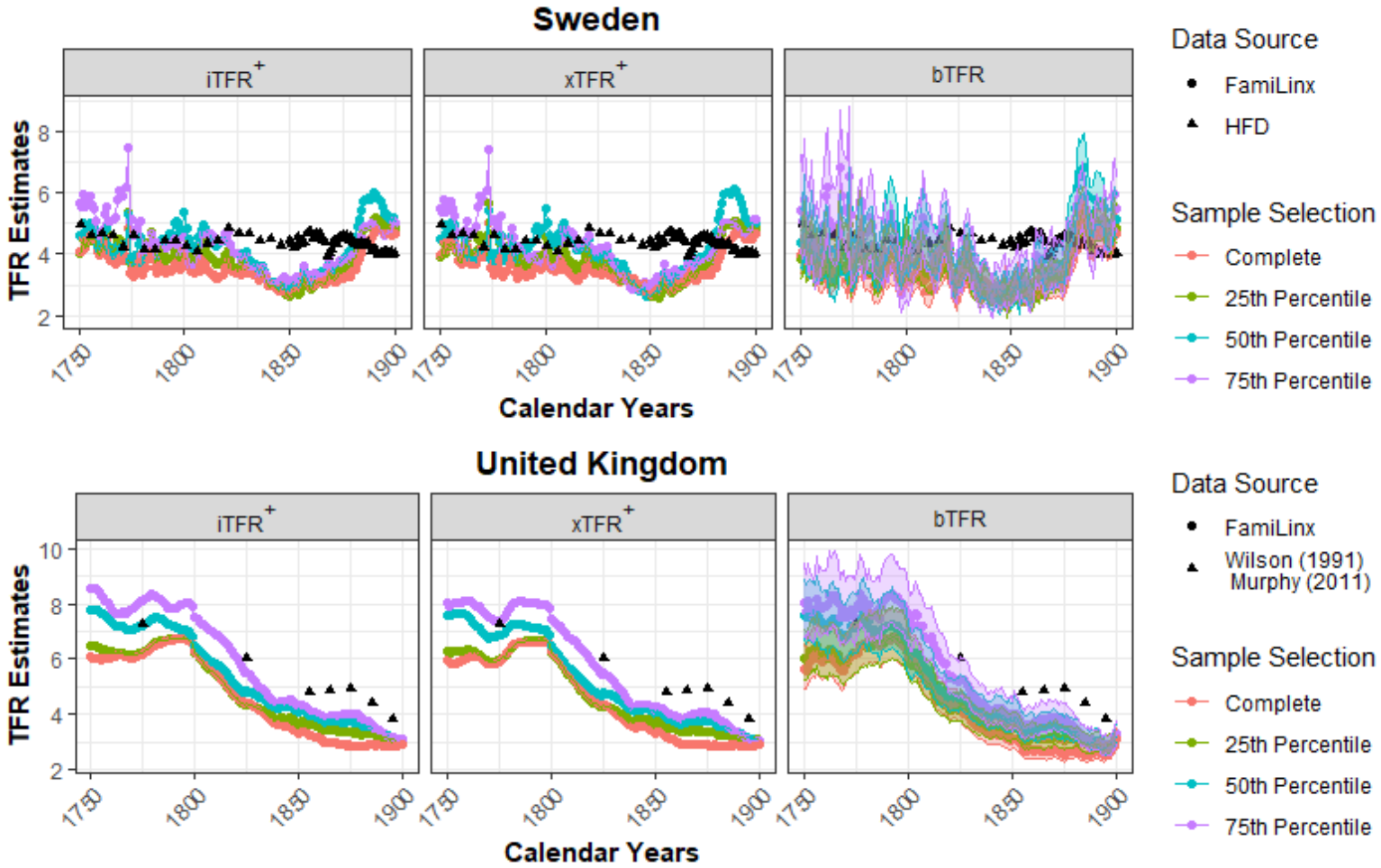


Figure: Time series of TFR estimates in Sweden and UK for different sub-samples.

bTFR in Sweden

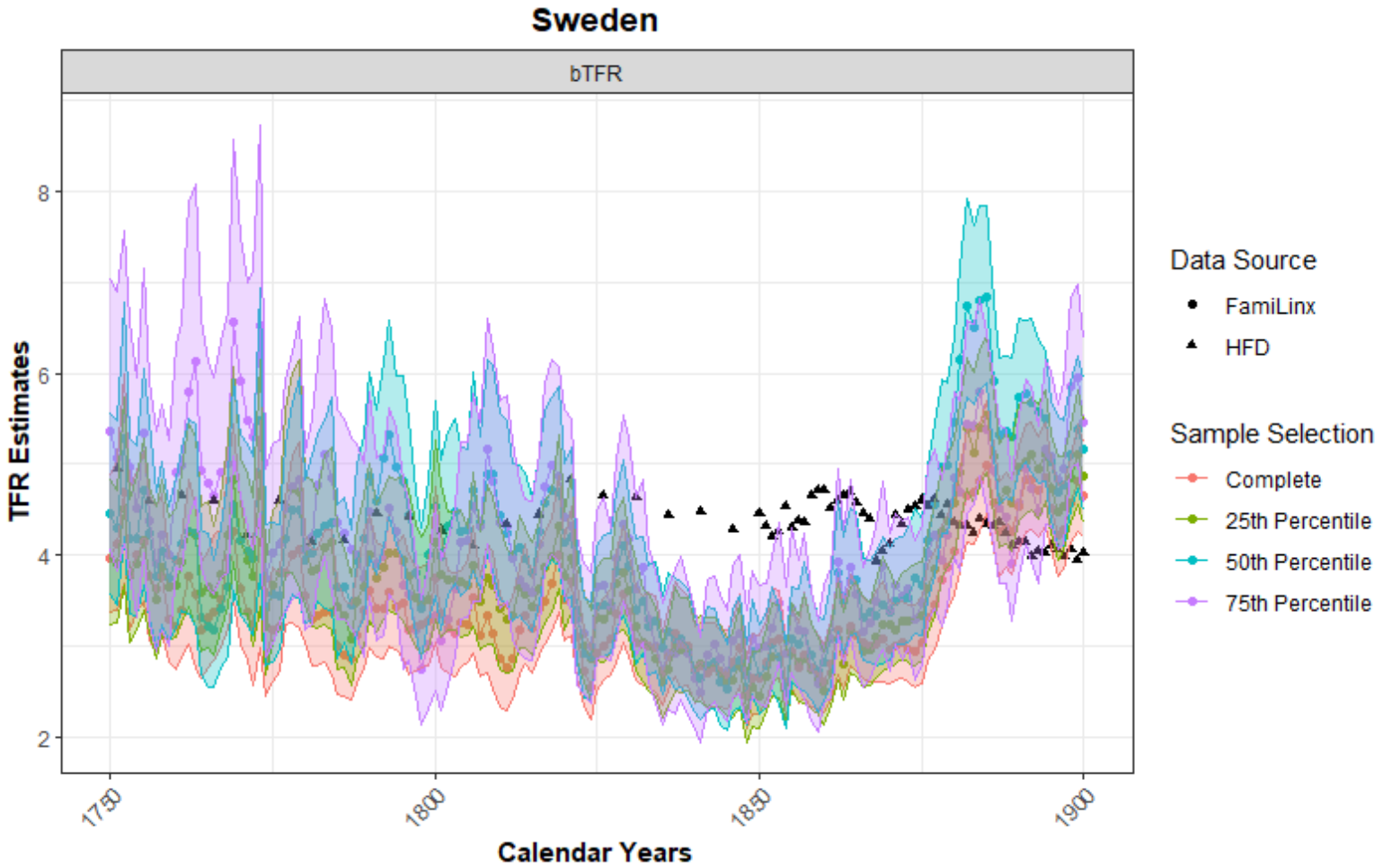


Figure: Time series of TFR estimates in Sweden for different sub-samples.

bTFR in the UK

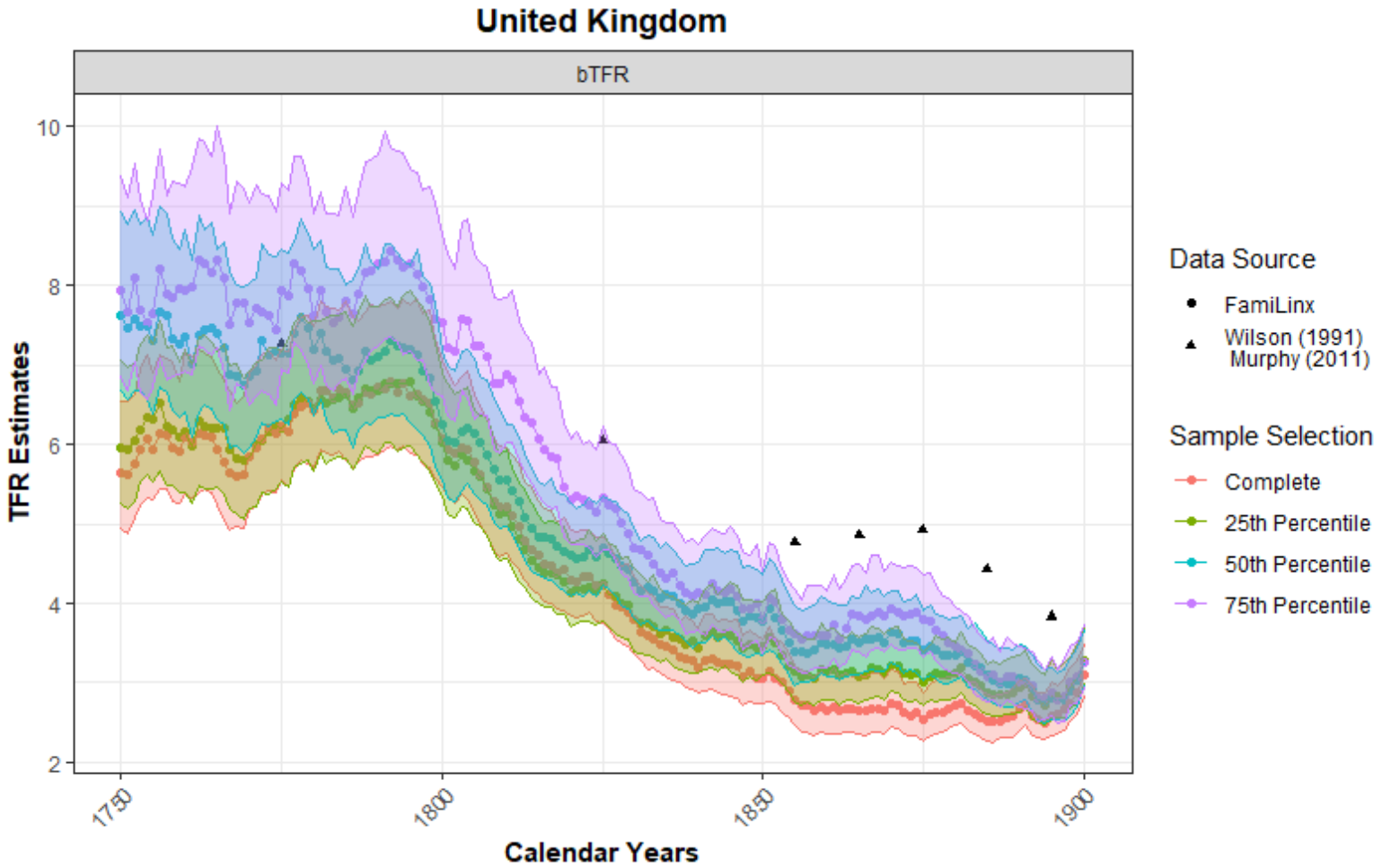


Figure: Time series of TFR estimates in the UK for different sub-samples.

Main Limitations

- Downward biased TFRs.
- Under-estimation of infant deaths.
- Skewed Population Pyramids (over-representation of male individuals).
- Missing ground-truth data covering early historical periods (before 1800s) for most of the considered countries.
- Selecting sub-samples of profiles with more accurate information does not translate to an improvement in the *TFR* estimates.

Conclusions

- Improvement of the historical TFR estimates through infant mortality adjustment.
- Development of data-quality indicators as a guide to the sample selection of the FamiLinx data set.
- Better representativeness of genealogies towards the end of the 19^{th} century. (Stelter & Alburez-Gutierrez, 2022)

What comes next?

- Improvement of the genealogy-based data-quality indicators.
- Deeper investigation into the potential of online genealogical data for the examination of historical fertility patterns.
- Integration of online genealogical data and models from the Formal Demography of Kinship (see Caswell, 2019) to study the evolution of kinship networks.

Thank you!

Looking forward to your feedback!

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.

List of features for calculating I_i^{Demo}

- x_{i1} (**availability of birth year**) $\rightarrow$ 1 if available, 0 otherwise.
- x_{i2} (**availability of birth month**) $\rightarrow$ 1 if available, 0 otherwise
- x_{i3} (**availability of birth day**) $\rightarrow$ 1 if available, 0 otherwise.
- x_{i4} (**availability of death year**) $\rightarrow$ 1 if available, 0 otherwise.
- x_{i5} (**availability of death month**) $\rightarrow$ 1 if available, 0 otherwise.
- x_{i6} (**availability of death day**) $\rightarrow$ 1 if available, 0 otherwise.
- x_{i7} (**availability of place of birth**) $\rightarrow$ 1 if available, 0 otherwise.
- x_{i8} (**availability of place of death**) $\rightarrow$ 1 if available, 0 otherwise.
- x_{i9} (**accuracy of place of birth**) $\rightarrow$ $\frac{1}{3}$ if inferred from one method, $\frac{2}{3}$ if inferred from two methods, 1 if inferred from all three selection methods, 0 otherwise.
- x_{i10} (**accuracy of place of death**) $\rightarrow$ $\frac{1}{3}$ if inferred from one method, $\frac{2}{3}$ if inferred from two methods, 1 if inferred from all three selection methods, 0 otherwise.

List of features for calculating I_i^{Fam}

- y_{i1} (**knowledge of parents**)$\rightarrow$ 1 if both parents are known, $\frac{1}{2}$ only one known, 0 otherwise.
- y_{i2} (**knowledge of maternal grandparents**) $\rightarrow$ 1 if both maternal grandparents are known, $\frac{1}{2}$ only one known, 0 otherwise.
- y_{i3} (**knowledge of paternal grandparents**) $\rightarrow$ 1 if both paternal grandparents are known, $\frac{1}{2}$ only one known, 0 otherwise.
- y_{i4} (**knowledge of paternal great-grandparents**) $\rightarrow$ 1 if both paternal great-grandparents (paternal grandmother) are known, $\frac{1}{2}$ only one known, 0 otherwise.
- y_{i5} (**knowledge of paternal great-grandparents**) $\rightarrow$ 1 if both paternal great-grandparents (paternal grandfather) are known, $\frac{1}{2}$ only one known, 0 otherwise.
- y_{i6} (**knowledge of maternal great-grandparents**) $\rightarrow$ 1 if both maternal great-grandparents (maternal grandmother) are known, $\frac{1}{2}$ only one known, 0 otherwise.
- y_{i7} (**knowledge of maternal great-grandparents**) $\rightarrow$ 1 if both maternal great-grandparents (maternal grandfather) are known, $\frac{1}{2}$ only one known, 0 otherwise.

Performance of the overall quality indicator

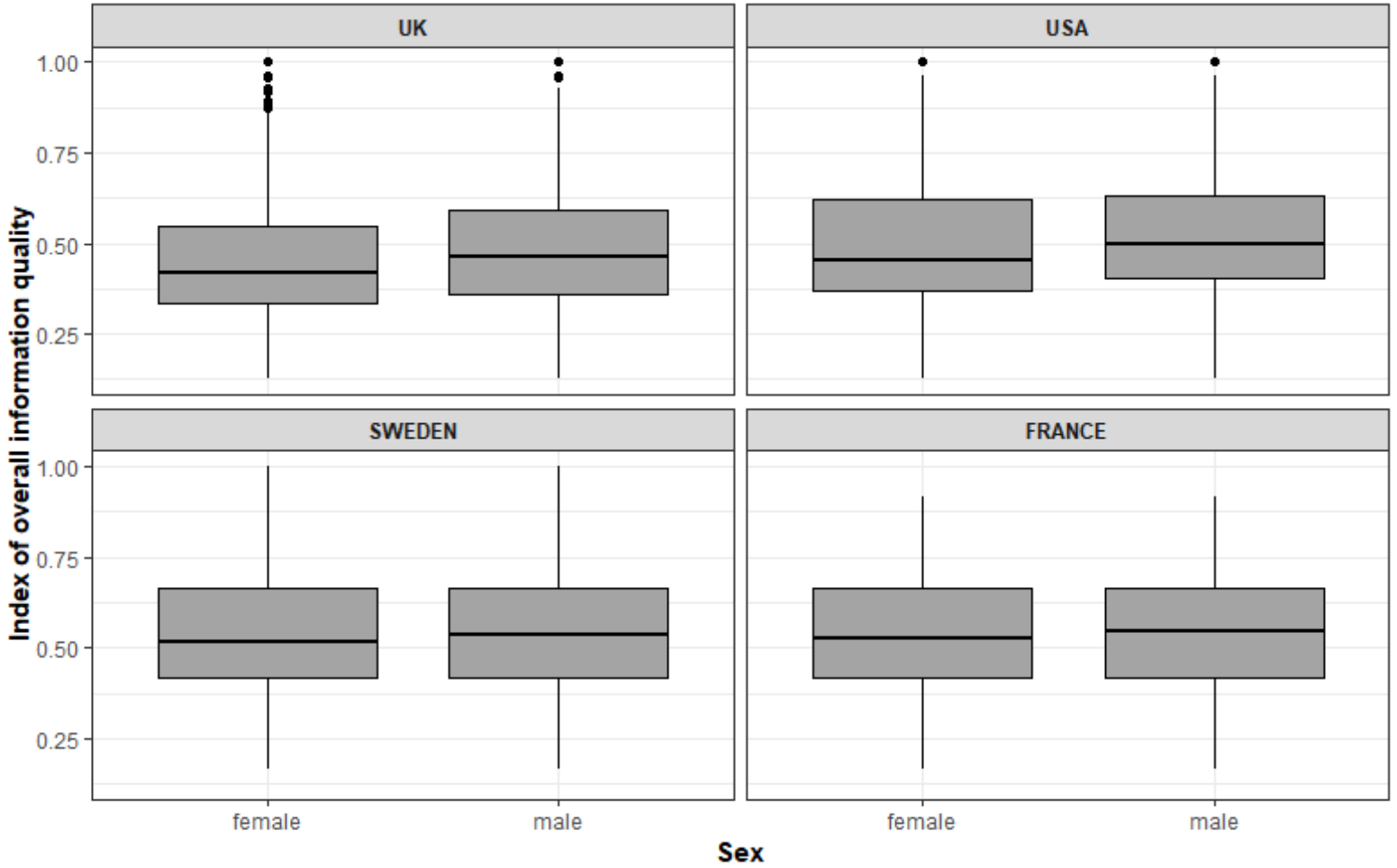


Figure: Distribution of the overall quality index stratified by sex across different countries.

US Population Pyramids

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

French Population Pyramid

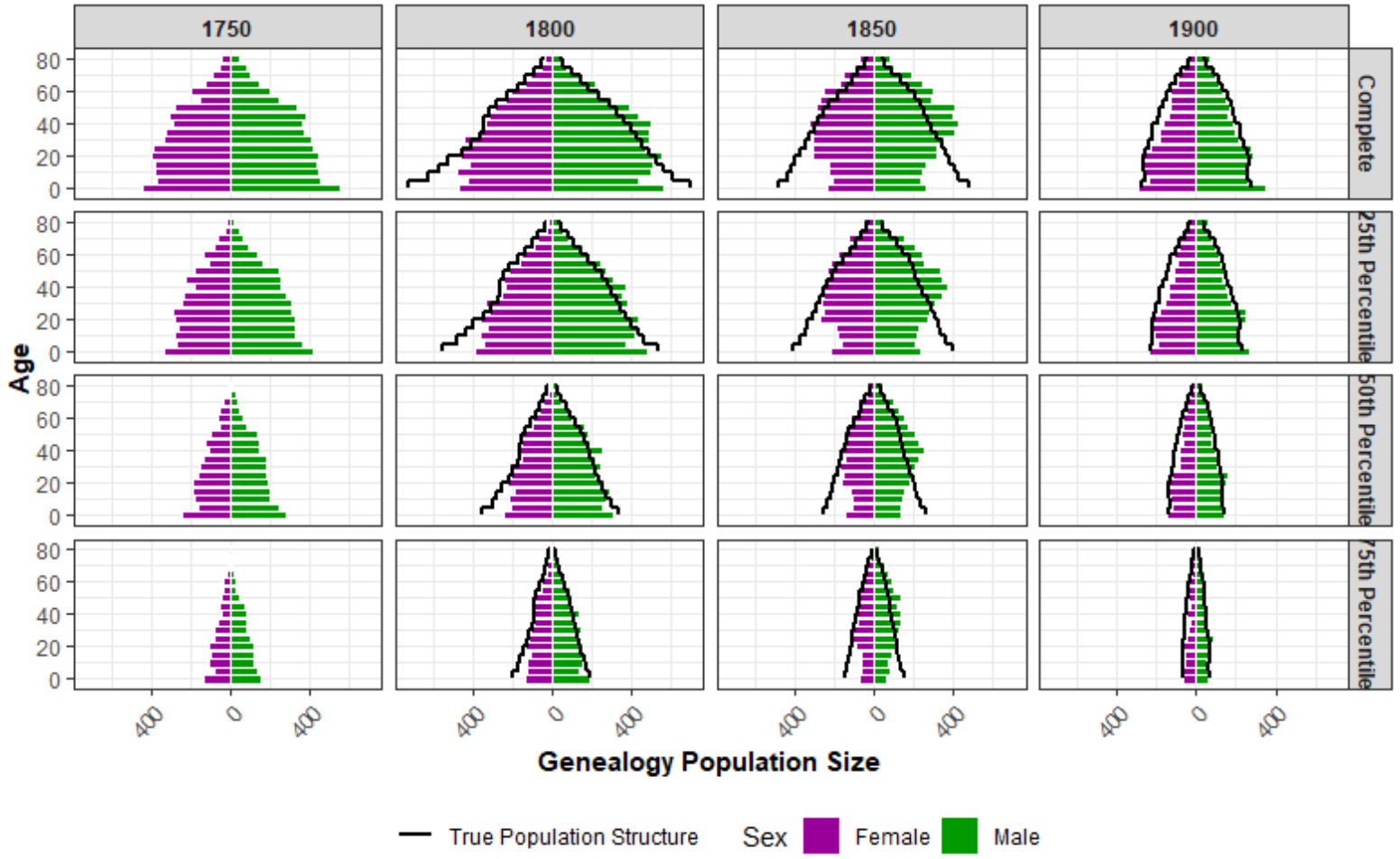


Figure: Genealogy-based French population pyramids for calendar years 1750, 1800, 1850 and 1900 for different sub-samples.

xTFR and *iTFR* in France and the USA

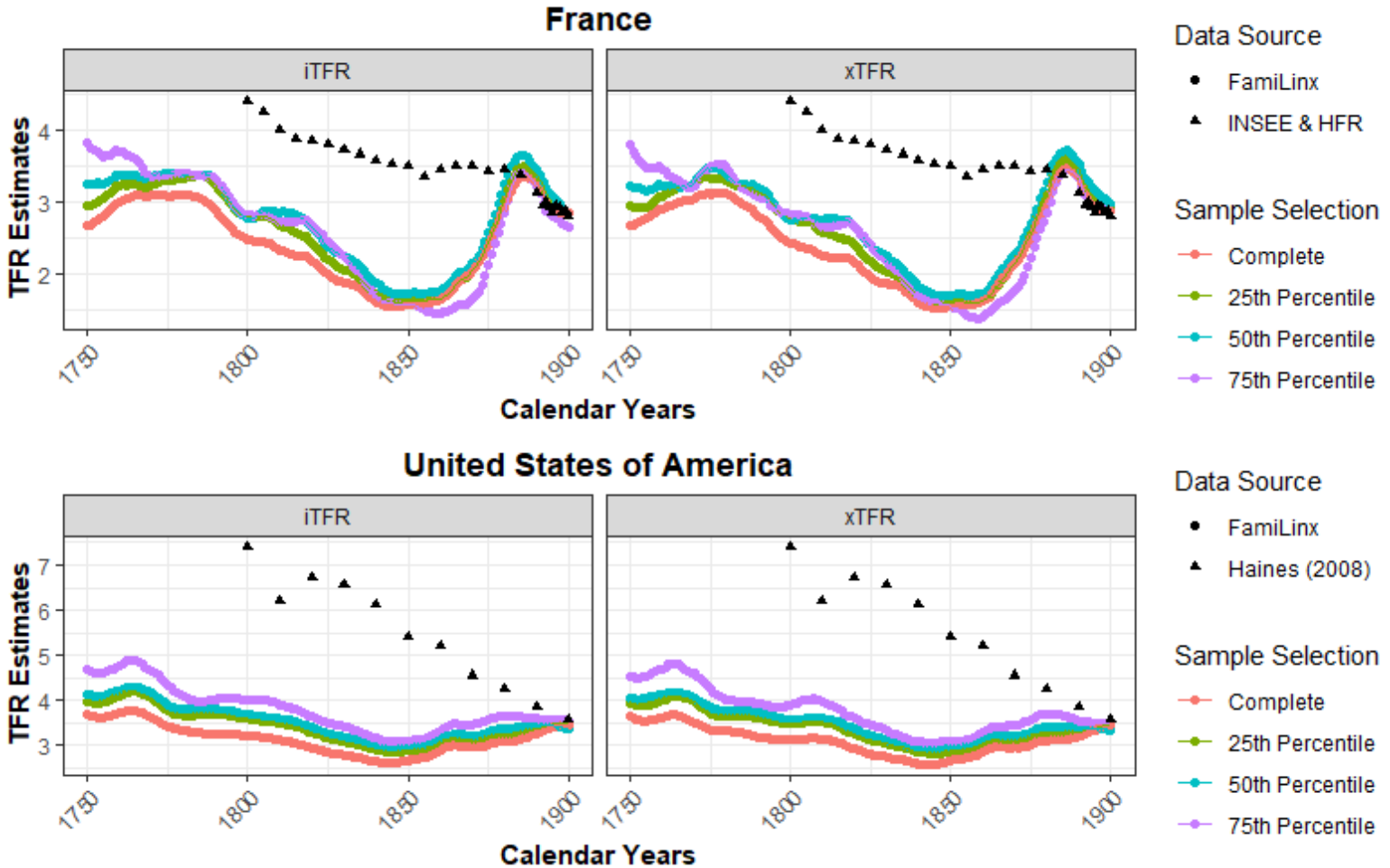


Figure: Time series of TFR estimates (*iTFR* and *xTFR*) in the selected countries for different sub-samples during the historical period $1750 - 1900$.

xTFR: Extended Total Fertility Rate

$$xTFR = \left(10.65 - 12.55\pi_{25-34}\right) \cdot \frac{C}{W}$$

$\pi_{25-34} \rightarrow$ proportion of women aged 25-34 among those in the maternal ages (15-49).

The reported coefficients are estimated by running a simple linear regression model

$$TFR^* \cdot \frac{W}{C} = \beta_0 + \beta_1 \pi_{25-34}$$

The model is trained by using $1,804$ fertility schedules retrieved from the Human Fertility Database.

$TFR^* \rightarrow$ the average TFR over the previous five years.

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 $TFR|C$.

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

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

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

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

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

s.t.

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

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

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 637 fertility schedules from the Human Mortality Database (collected in a rectangular matrix of dimension 7×637).

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 HFD 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.