ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN

SCIENZE STATISTICHE

Ciclo XXXVII

Settore Concorsuale: 13/D3 - DEMOGRAFIA E STATISTICA SOCIALE

Settore Scientifico Disciplinare: SECS-S/04 DEMOGRAFIA

Blank Space: Demographic Estimation in Data-sparse Contexts

Presentata da: Riccardo Omenti

Coordinatore Dottorato

Prof. Angela Montanari

Supervisore

Prof. Nicola Barban

Co-supervisore

Prof. Monica Alexander

Esame finale anno 2025

Abstract

This thesis explores novel approaches for the estimation of demographic outcomes in contexts where data are limited.

In the first part, we investigate the potential of an emerging non-traditional data source for demographic research: online genealogies. Harnessing FamiLinx, a big genealogical database with over 86 million observations, we show that the availability of accurate and non-missing demographic information in online genealogical data is selective. Our findings reveal that individuals with a non-missing value in a demographic variable are more likely to present non-missing data in the other demographic variables, and to be embedded in family networks, whose members exhibit demographic information of superior quality and completeness.

In the second part, we develop a Bayesian method for estimating the total fertility rate (TFR) indirectly from defective data. By combining online genealogical data from FamiLinx populations with information from more reliable sources, the proposed method allows to obtain TFR estimates for seven European countries and the US during the historical period 1751 – 1910, a time when many of these countries lacked well-functioning civil registration systems.

In the third part, we build a Bayesian model for the estimation of subnational male and female TFRs. Using real data from the United States and simulated data from Australia, we demonstrate that the proposed method can produce reasonably accurate TFR estimates in contexts, such as small areas, where data tend to be sparse and highly variable.

Throughout the second and third parts of this thesis, we leverage indirect estimation techniques within a flexible Bayesian modeling framework. This approach allows to incorporate multiple data sources, to capture regularities in demographic trends over time and space, and to account for uncertainty.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Starting from the early 1990s, massive technological improvements in information storage and computing have led to a 'data revolution' (Hilbert and López, 2011). The emergence of the 'big data' era has radically reshaped demography as a discipline (Kashyap, 2021). Since its inception with the seminal study 'Natural and Political Observations Made upon the Bills of Mortality' by Graunt (1662), the field of demography has been devoted to the employment of large-scale population data and the repurposing of imperfect data. Nonetheless, the explosion in the volume and in the variety of data, encouraged by the unprecedented technological progress of the last three decades, has given rise to unprecedented developments in the field of demography along three distinct dimensions. First, demographic data have become increasingly granular and diverse, with novel opportunities for data linkage that have strengthened the capabilities of traditional big population data sources, such as censuses, administrative records and surveys (Kashyap, 2021). Second, there has been a growing interest in the analysis of population processes using new 'big data', such as digital trace data generated through Internet and social media (Billari and Zagheni, 2017; Alburez-Gutierrez et al., 2019; Kashyap and Zagheni, 2023). Third, the emergence of sophisticated computational methods has enabled not only to obtain more accurate estimates of demographic processes compared to traditional demographic methods but also to quantify their uncertainty. In this regard, machine learning and Bayesian statistics have acquired unprecedented prominence in demographic

research. The former has provided population scientists with statistical approaches for linking data sets by common variables (Abramitzky et al., 2020, 2021) and for producing predictions of demographic outcomes (Arpino et al., 2022). The latter has flourished because of its ability to integrate multiple data sources and to handle uncertainty in a coherent manner (Bijak and Bryant, 2016). Applications of Bayesian modeling in demographic studies include the estimation of demographic outcomes in presence of limited data (Alexander et al., 2017, 2020; Rampazzo et al., 2021; Alexander and Alkema, 2022; Chong and Alexander, 2024), demographic forecasting (Raftery et al., 2012, 2013; Yu et al., 2023) and the employment of sophisticated and structured models for the examination of complex population dynamics (Poole and Raftery, 2000; Wheldon et al., 2013).

In light of these reflections, a number of research questions arise.

(a) *What are the promises and pitfalls of using non-traditional data sources in demography?*

(b) *How can we measure the quality and the completeness of their demographic information?*

(c) *How can we measure demographic outcomes, such as fertility, accurately in situations where data are limited?*

(d) *How can we coherently handle uncertainty in demographic estimates originating from defective data?*

This thesis is a collection of three papers that have the overarching aim to address these research questions[1]. The first paper offers a detailed examination of an emerging non-traditional data source for demographic research: online genealogies. This data source originates from websites where a transnational network of users reconstructs their own family trees, often including essential demographic information about their ancestors, such as birth and death dates and locations. Due to the inclusion of individuals, who lived in distinct historical periods and countries, online genealogies could serve as an alternative

[1] Although the papers share some similarities, they are designed to be read and interpreted independently.

historical census. However, unlike official population censuses, online genealogies are not a representative picture of the general population, as they are not primarily designed for demographic research. Hence, to accurately measure demographic outcomes from online genealogical data, it is essential to employ appropriate bias-correcting statistical methods. Building on the methodological framework by Schmertmann and Hauer (2019), the second paper develops a Bayesian method for estimating historical fertility patterns in seven European countries and the United States, combining online genealogical data with more trustworthy data sources, such as censuses and population registers. The third paper extends the same modeling framework by Schmertmann and Hauer (2019) to estimate subnational male and female fertility in contexts with small populations and highly variable data. Broadly speaking, both the second and third papers tackle fertility estimation in situations where the available data are imperfect. In this regard, Bayesian modeling proves to be an invaluable tool due to its ability of combining the limited available data with plausible information coming from prior distributions, while taking uncertainty into account.

The remainder of the introduction is structured as follows. First, the research objectives of the thesis are described. Next, we provide a brief summary of the main topics of the thesis, including online genealogies, fertility estimation and Bayesian modeling for demographic estimation. Finally, a summary of the main contributions of the three papers is provided.

1.2 Research Aim

The overarching goal of this thesis is to address the challenges associated with the use of non-traditional and imperfect data sources in demography. As a growing number of demographers turn to digital and other non-traditional data sources for their own research, a careful assessment of the demographic information that these sources provide is needed. Demographic variables, such as birth and death locations and dates, are vital for measuring demographic processes, making it crucial to investigate both their share of missing values and their accuracy. The employment of defective data with a high share of

missing values in demographic research can lead to biased demographic estimates that fail to capture the true population trends. Hence, it is imperative to develop ad-hoc statistical methods that enable the estimation of accurate demographic indicators from inherently defective data sources while accounting for their limitations.

One objective of this thesis is to evaluate the potential of non-traditional data sources in demography, using online genealogical data as a case study. This investigation has led to the development of indicators designed to measure the completeness and quality of demographic information derived from such data. In addition, since individuals from this data source are embedded in family networks, the concept of completeness and quality are also investigated not only at an individual level but also at a family level.

Another key goal of this thesis is to develop statistical methods for the estimation of demographic outcomes in data-sparse contexts. In this thesis, we aim to develop a Bayesian methodological framework that produces fertility estimates in settings where the available data are limited. By allowing for the incorporation of multiple data sources and for their uncertainty, Bayesian modeling provides a valuable methodological framework for the estimation of demographic outcomes with imperfect data.

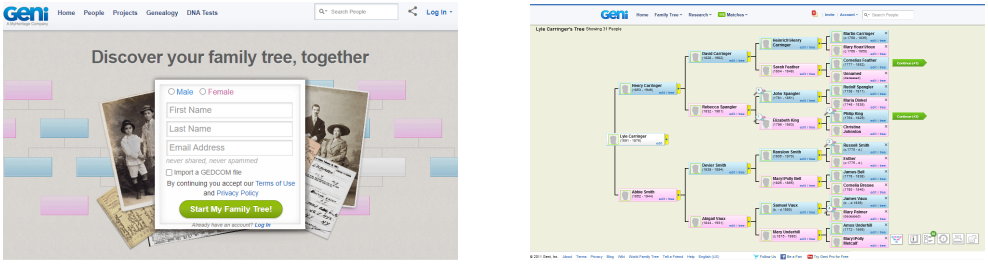
In particular, we extend an existing Bayesian method by Schmertmann and Hauer (2019) to estimate fertility from online genealogical data in combination with more trustworthy data sources in the second paper of this thesis. We further adapt this method to produce subnational fertility estimates for both men and women in the third paper of this thesis. Broadly speaking, this thesis provides a detailed analysis of the promises and pitfalls of using online genealogies for demographic research and introduces novel statistical methods for fertility estimation in contexts where data are limited.

1.3 Key Concepts

1.3.1 Online Genealogies

Since its inception, one of the defining features of demography has been the repurposing of non-traditional data sources for the study of demographic dynamics (Kashyap,

2021), with genealogies representing a prime example. Genealogies are constructed retrospectively, as individuals trace back their own ancestors, generally by compiling the data themselves. After incorporating the descendants and the collateral relatives of the ancestors into the family tree, a genealogy can be continuously updated to mirror births, marriages and deaths occurring among the present and future members of the family (Hollingsworth and Hollingsworth, 1976).



(a) Home page **(b)** Family tree

Figure 1.1: (a) Home page of the website geni.com and (b) Example of a family tree from geni.com

During the last three decades, the digital revolution has facilitated the spread of online platforms, known as online genealogies, where a transnational network of users can reconstruct their own family trees and upload demographic information for each of their ancestors. Popular websites include familysearch.org, geni.com and ancestry.com.
Online genealogies represent a milestone in family history and historical demography, incentivizing the process of 'democratizing' the access to genealogical information. Once a privilege of the wealthy, online genealogies became more accessible with the proliferation of modern censuses and population registers in the late 18th and early 19th centuries, which allowed for the systematic recording of basic demographic information about individuals from a broader range of social classes. Nowadays, online genealogies enable people of all backgrounds to reconstruct their family history and provide researchers with an unprecedented wealth of demographic information about individuals across different countries and historical periods (Stelter and Alburez-Gutierrez, 2022; Colasurdo and Omenti, 2024).
However, online genealogies are not exempted from limitations. First, individuals recorded in online genealogies tend to experience higher survival rates compared to the general pop-

ulations. As online genealogies are generally constructed retrospectively, more longevous individuals have a higher probability of being recorded. On the contrary, as noted by Hollingsworth and Hollingsworth (1976), women, childless individuals, and individuals dying at young ages are generally underrepresented. Second, online genealogies tend to over-represent individuals coming from families of higher historical relevance and socio-economic status. It is well-known that individuals, whose ancestors were historically relevant figures or part of wealthy families, are facilitated in the reconstruction of their family history (Hollingsworth and Hollingsworth, 1976). For instance, harnessing a big genealogical database, FamiLinx, which was created by Kaplanis et al. (2018) by scraping data from the website geni.com, Stelter and Alburez-Gutierrez (2022) showed that male mortality patterns before the end of the 19th century in the Netherlands and Germany resembled those of elite groups in both territories. Third, since online genealogies were not primarily designed for demographic research, the accuracy of their demographic information depends on the knowledge of the user compiling the family tree (Colasurdo and Omenti, 2024). Fourth, the ascending construction of the family trees reduces the probability of including more distant ancestors (Calderón Bernal et al., 2023). Despite these limitations, the vast majority of the previous demographic studies using online genealogies have operated under the assumption that this data source is representative (Hsu et al., 2021; Cozzani et al., 2023; Pojman et al., 2023; Blanc, 2024a,b; Corti et al., 2024). The first and second papers of this thesis address crucial gaps in the use of online genealogies for demographic research. The first paper investigates the extent to which populations scientists can employ online genealogies for demographic research by analyzing the completeness and the accuracy of their demographic variables at both individual and family levels. The second paper develops a Bayesian model for fertility estimation by combing online genealogical data with more trustworthy data sources.

1.3.2 Fertility Estimation

Fertility estimation is an essential part of this thesis, as both the second and third papers aim to produce fertility indicators in data-limited contexts. In this subsection, we provide a brief introduction of fertility estimation in demography.

In demography, fertility data are usually provided in the form of live births occurring to women across distinct reproductive age groups (usually covering the reproductive age span $15-49$) and the corresponding population of women at risk of giving birth.

The focus of the estimation has generally been towards the calculation of age-specific fertility rates (ASFR), which denote the number of live births per women in a certain age-group per time. One of the most widely recognized fertility indicators is the TFR that is obtained by summing over the age-specific fertility rates. The TFR is interpreted as the average number of children that are born to a woman over her lifetime, if they were to experience the exact current age-specific fertility rates (ASFRs) throughout their lifetime, and they were to live from birth until the end of their reproductive life (Wachter, 2014)[2].

In human populations, fertility is a complex and dynamic process, shaped by a wide range of social, economic and biological factors (Balbo et al., 2013). Human fertility is influenced by two major components: tempo (i.e., timing of childbearing) and quantum (i.e., total number of children). The TFR is the most commonly used measure of the fertility quantum, whereas the mean age at childbearing (MAB) is one of the most important measures of fertility timing. The MAB is the average age of mothers at the birth of their children, assuming women were to experience throughout their lives the age-specific fertility rates observed in a given year.

[2]These fertility indicators are period-based measures. However, they can be equivalently defined from a cohort perspective.

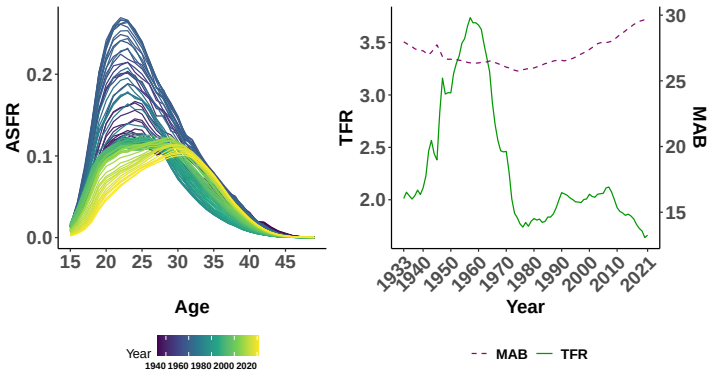


Figure 1.2: ASFR, TFR and MAB in the US during the period 1933-2021. Data are from the Human Fertility Database.

As a woman can bear more than one child during her reproductive ages, a substantial effort in fertility research has been devoted to the estimation of transitions from lower-order births to higher-order births. The most common demographic indicator for estimating such transitions is represented by the parity progression ratio (PPR), which measures the proportion of women with a certain number of children who go on to have another child. Unlike the TFR and the MAB, PPR is generally measured by cohort rather than by period.

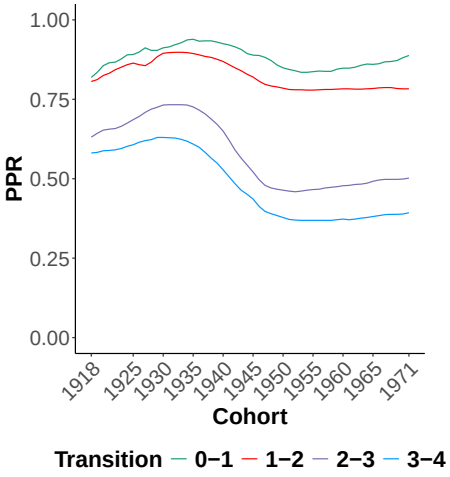


Figure 1.3: Parity Progression Ratios for US women born between 1918 and 1971. Data are from the Human Fertility Database.

Nonetheless, the calculation of the previous indicators via standard demographic techniques generally require either live births stratified by maternal ages (TFR, MAB) or live births classified by maternal ages and parity (PPR). In case unavailability of such de-

tailed information, demographers must opt for alternative measurement techniques such as indirect estimation.

1.3.3 Indirect Fertility Estimation

Indirect estimation is also a crucial component of this thesis. In this regard, the Bayesian methodological framework of the second and third papers builds on the idea that we can estimate the TFR indirectly without the need of knowledge about the number of live births classified by parental ages. This subsection provides a brief overview of indirect estimation for fertility estimation in demography.

Despite the unprecedented improvements in data collection over the past three decades, the availability of high-quality birth data is not always guaranteed. Many developing countries still lack high-quality vital registration systems and timely nationally-representative surveys. Meanwhile, register and census data from developed countries may not always provide direct access to data needed for the calculation of specific demographic indicators, especially at a subnational level. For instance, if a country does not disclose information on the number of births classified by maternal ages within its regions, it becomes unfeasible to estimate subnational TFRs using standard demographic techniques.

In the contexts where data are limited or lacking, demographers have often relied on indirect estimation methods. As stated by the United Nations in their 1983 manual (UN-DESA, 1983), indirect estimation refers to 'techniques suited for analysis of incomplete or defective demographic data'. Several indirect estimation methods have been developed to obtain fertility indicators from incomplete data. One of the most well-known is the own-children method (OCM), which was first introduced in the 1960s by Grabill and Cho (1965). The OCM consists in reconstructing women's fertility behavior from censuses in absence of retrospective birth histories. It relies on the principle of reverse survival, meaning that children aged x at time t are survivors of births that occurred x to x+1 years before. Similarly, exposed women are back-survived to time when children of age x were born. This method has been widely applied in historical demography to study fertility dynamics in historical populations (Breschi et al., 2003; Garrett et al., 2010; Scalone

and Dribe, 2017; Reid et al., 2020) as well as for the estimation of fertility experienced by minority groups in contemporary populations (Abbasi-Shavazi and McDonald, 2000; Dubuc, 2009).

More recently, Hauer et al. (2013) developed an indirect estimation method that allows to approximate the TFR with minimal input, namely, the number of children aged 0-4 and of women aged 15-49. This method produced accurate TFR estimates, especially in countries with the low child mortality. A refined version of this indirect estimation method, which also accounts for child mortality, was introduced by Hauer and Schmertmann (2020).

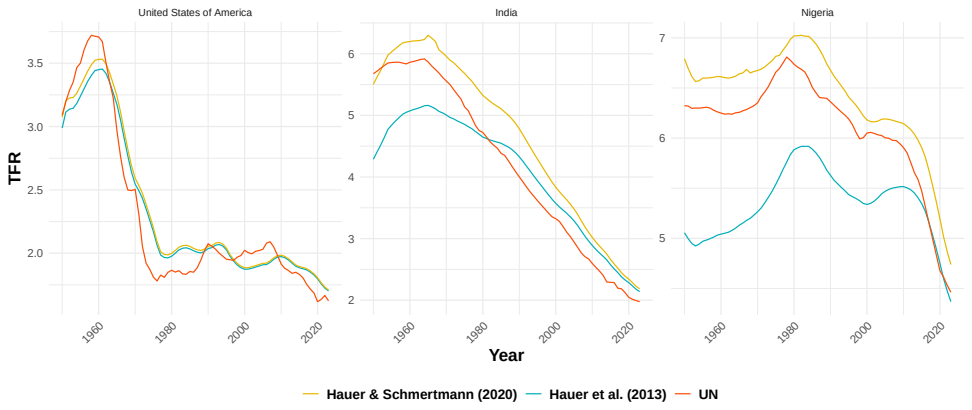


Figure 1.4: Comparison between the indirect TFR estimates based on methods from Hauer et al. (2013) and Hauer and Schmertmann (2020), and TFR estimates taken from the United Nation World Population Prospects (UNDESA, 2024) for the period 1950 – 2023.

One of the major drawbacks of indirect estimation is its inability to account for the reliability of the data sources and for the stochastic nature of demographic processes. While obtaining numerical estimates of demographic trends from incomplete data is essential, it is equally important to assess the uncertainty surrounding these estimates. To overcome this challenge, Bayesian methods have shown much promise as they allow for the use of indirect estimation methods while accounting for the uncertainty coming from the different sources as well as the randomness of the demographic processes (Bijak and

Bryant, 2016).

In the context of Bayesian indirect estimation, Schmertmann and Hauer (2019) proposed a Bayesian version of the indirect method by Hauer et al. (2013). This approach provides accurate TFR estimates along with their associated uncertainty while accounting for mortality experienced by children under 5 and women aged 15-49. The second and third papers of this thesis extend the methodological framework by Schmertmann and Hauer (2019) and consistently compare the new estimates with those obtained from other indirect methods.

1.3.4 Male Fertility Estimation

The third paper of this dissertation deals with the estimation of male fertility. Unlike female fertility, male fertility has received much less attention in the context of demographic research (Coleman, 1995). A lack of focus on male fertility in data collection efforts has been attributed to two major factors. First, the definition of the female population at risk of childbearing is more precise and limited within a specific age range, generally women aged 15-49 (Coleman, 1995). Second, in many settings, women have shown a greater availability to respond to surveys (Greene and Biddlecom, 2000). Nonetheless, limiting the fertility research to women can lead research scientists to ignore the specificities of male reproductive behavior or to assume that men and women display similar fertility patterns (Schoumaker, 2019).

Previous research (Dudel and Klüsener, 2016; Schoumaker, 2017, 2019) has shed new light on the distinct fertility patterns experienced by men in comparison to women. Specifically, these studies have underlined that the age-specific fertility curves of men and women tend to be fairly similar, even though the reproductive age period is larger among men. The total level of fertility has been shown to vary by sex. In low-fertility countries, male and female TFRs tend to be very close to one another, with women experiencing slightly higher fertility levels (Dudel and Klüsener, 2016, 2021; Dudel et al., 2023). In high-fertility settings, Schoumaker (2019) demonstrated that the male TFR can be up to 50% higher than the female TFR. In addition, non-negligible differences between men and women

have been also observed in terms of age at childbearing. For instance, in high-income countries, the age at fatherhood tends to be 3-4 years higher than the age at motherhood (Dudel and Klüsener, 2016, 2021). This difference can be as high as 12 years in certain developing countries as noted by Schoumaker (2017).

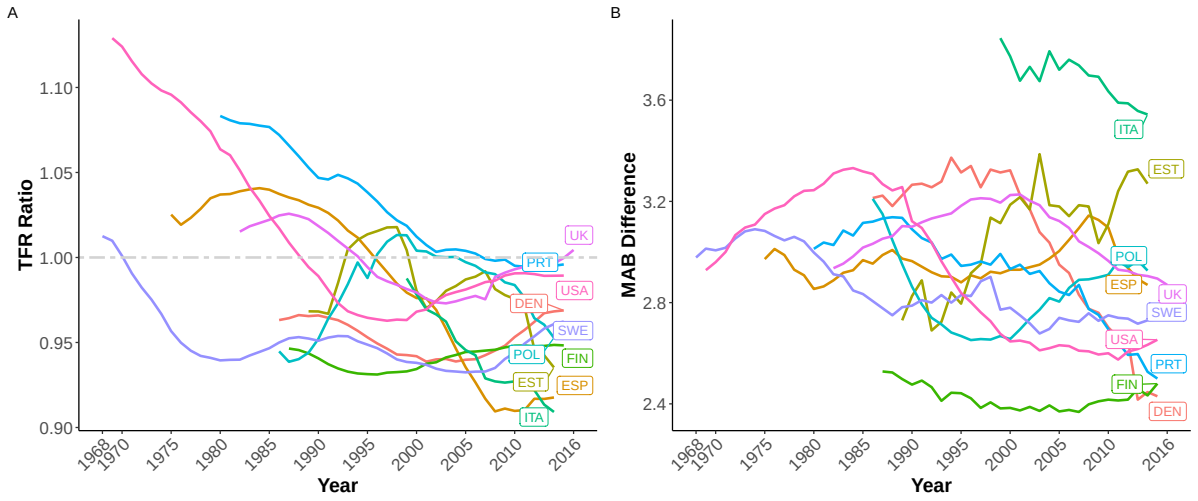


Figure 1.5: (A) Ratio of male TFR to female TFR during period 1968-2016 for 10 selected countries (B) Difference between mean age at fatherhood and mean age at motherhood during period 1968-2016 for 10 selected countries. The indicators were computed using data from the Human Fertility Collection.

Concerning the measurement of male fertility, previous research has developed methods to compute common male fertility indicators, that is TFR and MAB, from either birth registers (Dudel and Klüsener, 2016) or nationally representative surveys (Schoumaker, 2017). Dudel and Klüsener (2016) relied on data from birth registers of developed countries, imputed the missing ages at fatherhood using common imputation methods and computed the previous fertility indicators via standard demographic techniques. Schoumaker (2017) relied on Demographic and Health Surveys (DHS) data and used imputation techniques to infer the missing paternal ages. To compute the male fertility, Schoumaker (2017) developed a "masculine" version of the OCM method. Overall, both of these methods produce accurate male fertility indicators if accurate information on parental ages for a certain country are available. For this reason, it is essential to develop statistical methods that allows for the estimation male fertility indicators in contexts,

where detailed birth data are not available, while also accounting for the incompleteness of the data sources being used.

1.3.5 Bayesian Modeling

Bayesian methods for demographic estimation represent another fundamental topic of this dissertation. In the second and third papers, we develop Bayesian methods to produce TFR estimates and their corresponding uncertainty in contexts with limited data. This subsection provides a brief overview of Bayesian modeling and of its usefulness for demographic estimation.

Over the last three decades, the increasing computational power has determined a surge in the employment of Bayesian methods in statistical modeling. In traditional frequentist methods, the observed data $\mathbf{x}$ are modelled using a likelihood function $f(\mathbf{x}|\theta)$ that depends on some fixed parameters θ. The goal is to estimate the parameters θ so that the observed data are the most probable. Conversely, Bayesian methods treat the parameters θ as random variables by assigning them prior probability distributions. The objective of Bayesian methods is to estimate the posterior probabilities for the parameters θ using a combination of the observed data and prior beliefs about the parameters' values. More formally, according to the Bayes rule, the posterior probability distribution of the parameters θ given the observed data $\mathbf{x}$, $f(\theta|\mathbf{x})$, is expressed as follows

$$f(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta) \cdot f(\theta)}{f(\mathbf{x})} \tag{1.1}$$

where $f(\mathbf{x};\theta)$ is the likelihood function, $f(\theta)$ is the prior distribution of the parameters θ and $f(\mathbf{x})$ is the marginal probability of the data. In most Bayesian models, priors are usually hierarchical, i.e., the parameters governing the priors are themselves given hyper-priors governed by hyper-parameters (Gelman et al., 1995).

A key advantage of Bayesian modeling is that the outcome of analysis is a probability distribution, which can be summarized in more intuitively meaningful ways (Bijak and Bryant, 2016). For example, 95% Bayesian credible intervals represents a 95% probabil-

ity of containing the true value of the parameter; a more straightforward probabilistic interpretation than frequentist confidence intervals, which refers to the proportion of hypothetical intervals that would contain the true value if the study were to be repeated many times.

The posterior estimates of the parameters θ, which are summary statistics, usually means or medians, from their corresponding posterior distributions, are a combination of the information in the prior and what it is observed in the data. The less informative the priors assigned to the parameters θ are the more dependent the posterior estimates are on the observed data. When the observed data are sparse or limited, the Bayesian estimates tend to be more heavily influenced by the information in the priors.

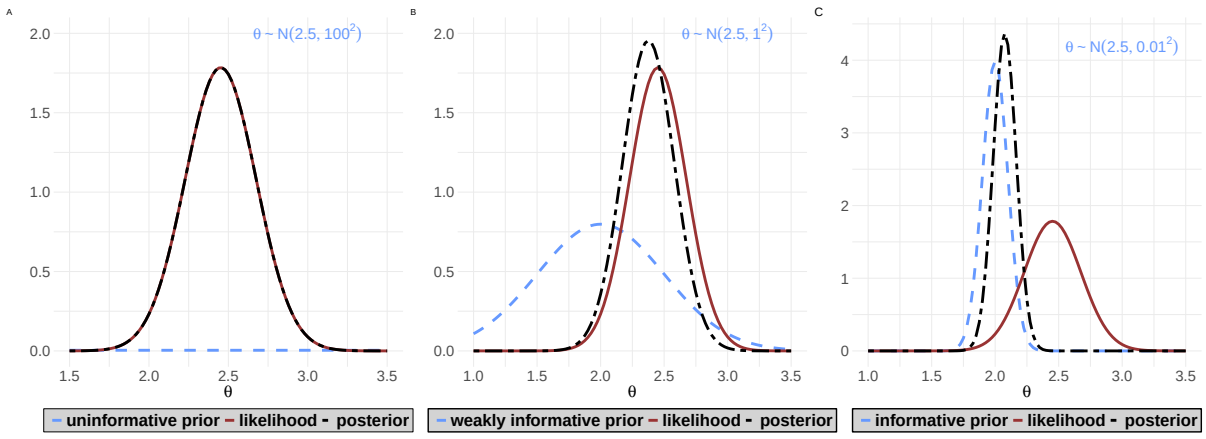


Figure 1.6: Simulated TFR distribution for n =20 randomly selected countries in 2022 with mean θ and known variance $\sigma^2 = 1$ ($TFR_c \sim N(\theta, 1^2)$). Panel (A) shows the posterior distribution for $\theta|TFR$ assuming an informative prior on θ. Panel (B) shows the posterior distribution for $\theta|TFR$ assuming a weakly informative prior on θ. Panel (C) shows the posterior distribution for $\theta|TFR$ assuming an informative prior on θ. TFR estimates are taken from the United Nations World Population Prospects (UNDESA, 2024).

In the context of demography, Bayesian methods were first applied to produce population forecasts with uncertainty (Daponte et al., 1997; Booth, 2006). Further advancements in Bayesian population forecasting were implemented by Alkema et al. (2011) and Raftery

et al. (2012), whose population projection estimates were adopted by the United Nations Population Divisions. In parallel, researchers have increasingly applied Bayesian methods for producing estimates and forecasts of all the major demographic processes, including mortality (Girosi and King, 2008; Alkema and New, 2014; Raftery et al., 2013; Alexander et al., 2017; Schmertmann and Gonzaga, 2018; Chong et al., 2022; Alexander and Root, 2022), fertility (Schmertmann et al., 2014, 2013; Ellison et al., 2020, 2024; Alkema et al., 2011; Schmertmann and Hauer, 2019; Batyra et al., 2023) and migration (Bijak and Wiśniowski, 2010; Disney et al., 2015; Alexander et al., 2020; Rampazzo et al., 2021; Bijak, 2022; Aparicio Castro et al., 2024).

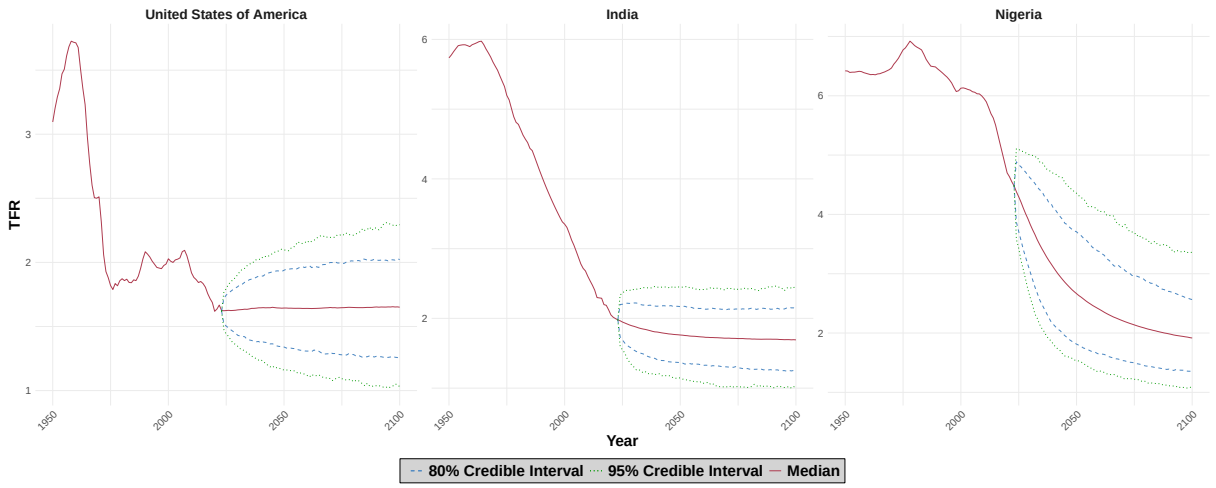


Figure 1.7: Median TFR estimates from United Nations World Population Prospects (UN-DESA, 2024) during the period 1950-2100 in US, India and Nigeria. Uncertainty in the TFR forecasts is incorporated by including the upper and lower limits of the 80% and 95% credible intervals for the forecasting period 2024-2100.

Bayesian methods provide several advantages in demographic estimation. First, multiple sources of uncertainty can be incorporated within a single modeling framework. Possible sampling and non-sampling errors in the observed data as well as the randomness of the demographic process are captured by the likelihood, whereas parameter uncertainty is introduced by using priors. Indeed, Bayesian methods not only allow to combine different

data sources within a single modeling framework but also to account for the reliability of each of these sources.

Second, prior information on the parameters can be incorporated into the model. This feature is especially valuable when the observed data are limited or defective. For instance, if fertility in a country was being estimated from defective data within a Bayesian framework, more robust fertility estimates could be produced by calibrating priors with information from more trustworthy data sources either from the same country or other countries with similar fertility patterns.

Third, the hierarchical nature of many Bayesian models allows to flexibly capture regularities over age, time and space. For example, in the context of fertility estimation, spatial and temporal auto-correlations in fertility trends could be captured by placing time-series and spatial models on the relevant parameters.

Overall, Bayesian methods represent a flexible and powerful framework for performing demographic estimation in various contexts.

1.4 Main Contributions of the Thesis

1.4.1 Online Genealogical Data for Demographic Research

The first paper of this thesis examines the strengths and weaknesses of online genealogical data for demographic research. Online genealogical data arise from the collaborative efforts of users willing to reconstruct their own family trees. Specifically, this project relies on FamiLinx, a big crowd-sourced genealogical database, that was developed by Kaplanis et al. (2018) by scraping data for over 86 million profiles from the website geni.com. This data source records demographic variables, including birth and death dates and locations, and kinship ties for individuals that lived in the last 400 years mostly in the Global North (Kaplanis et al., 2018). In this paper, we show that the share of missing values in the reported demographic variables is selective. Individuals with a non-missing demographic variable are more likely to have non-missing values in the other demographic variables. This finding underlines the importance of a careful sample

selection when using FamiLinx for the study of population dynamics. This thesis also reveals that individuals with more complete and accurate demographic information tend to be embedded within family networks of individuals whose demographic variables are of higher accuracy and completeness. In addition, a comparison with more trustworthy data from Human Mortality Database (Wilmoth et al., 2007), reveals that populations from FamiLinx experience lower mortality than the general population in alignment with previous findings by Stelter and Alburez-Gutierrez (2022).

To conclude, this paper can be regarded as a guide for other researchers interested in harnessing online genealogical data for demographic research.

1.4.2 Correcting Fertility in Online Genealogical Data

The second paper introduces a methodological framework to estimate fertility from defective data using data from FamiLinx as an example.

Building on the Bayesian modeling framework by Schmertmann and Hauer (2019), this research paper combines FamiLinx data with more traditional data sources to compute accurate TFR estimates in seven European countries and the United States of America during the historical period 1751 – 1910.

The original model by Schmertmann and Hauer (2019) allows to estimate the TFR using a minimal data input: the number of children under 5, the number of women aged 15-49 and the probability of death under age 5. The proposed extension consists in adding a bias-adjustment model into the Bayesian methodological framework to account for the fact that the proportions of children aged 0-4 and women aged 15-49 generally do not mirror those observed in the general population. The bias-adjustment model is specified to flexibly capture temporal patterns in the non-representativeness of online genealogical populations. The findings reveal the ability of the proposed Bayesian method to produce fairly accurate TFR estimates in countries and historical periods lacking well-functioning national civil registration systems. In addition, in the majority of the countries, the proposed Bayesian method outperforms a similar indirect estimation based on the work by Hauer and Schmertmann (2020) that we have extended to account for the

non-representativeness of children and women in online genealogical populations.

In conclusion, this paper provides a major contribution to the development of Bayesian methods and indirect estimation techniques for fertility estimation in settings with deficient data.

1.4.3 Estimating male and female fertility at a subnational level

The third paper of this thesis extends the modeling framework by Schmertmann and Hauer (2019) for the estimation of male and female fertility at a subnational level. Male fertility measurement has proven to be challenging due to the more uncertain and broader reproductive age interval, as well as the lower information quality in surveys (Greene and Biddlecom, 2000) and birth registers (Dudel and Klüsener, 2016). Additionally, in small-area populations, information on births stratified by the parental ages may not be recorded or may be restricted due to privacy concerns. Hence, the model by Schmertmann and Hauer (2019) represents a suitable alternative as it does not require births to be disaggregated by parental ages.

This study develops a Bayesian model, extending the original framework by Schmertmann and Hauer (2019) in two important ways. First, the model is extended to allow for the estimation of male fertility by considering men aged 15 – 59 as exposed to the risk of having children and by incorporating prior knowledge on national age-specific male fertility patterns as well as subnational mortality patterns. Second, spatial and temporal components are added to account for potential dependencies in fertility patterns observed among adjacent areas and in consecutive years. As a data example, the model is applied to simulated data from Australia and to real population data from US counties during the period 1982 – 2019.

The analysis reveals a substantial heterogeneity in the fertility patterns. While the majority of US counties have shown an overall decline in the TFR, considerable disparities persist, with some counties experiencing fertility levels well above the replacement value of 2.1 and with others presenting extremely low fertility levels below 1.3. Regarding sex-

specific differences, the analysis displays that in the majority of US counties women and men tend to display fairly similar fertility levels, with female fertility being slightly higher. This result aligns with previous findings by Dudel and Klüsener (2021) and Dudel et al. (2023).

To conclude, this paper provides a novel statistical framework for subnational male and female fertility estimation in absence of detailed birth data, paving the way for new applications to other countries.

Chapter 2

Using Online Genealogical Data for Demographic Research: An Empirical Examination of the FamiLinx Database

Riccardo Omenti

Andrea Colasurdo

Using Online Genealogical Data for Demographic Research: An Empirical Examination of the FamiLinx Database

Andrea Colasurdo[1,2,†] and Riccardo Omenti[3,‡]

[1] *Kinship Inequalities Research Group, Max Planck Institute for Demography Research, Rostock, Germany*

[2] *Faculty of Spatial Sciences, University of Groningen, Groningen, Netherlands*

[3] *Department of Statistical Sciences, University of Bologna, Bologna, Italy*

[†] *Corresponding author:* colasurdo@demogr.mpg.de

[‡] *Corresponding author:* riccardo.omenti2@unibo.it

This is a post-peer-review, pre-copyedit version of an article that has been recently published in Demographic Research.

Submitted: February 2024; Published: November 2024

Abstract

Background: Online genealogies are promising data sources for demographic research, but their limitations are understudied. This paper takes a critical approach to evaluating the potential strengths and weaknesses of using online genealogical data for population studies. We focus on the FamiLinx dataset, which contains demographic information and kinship ties across multiple countries and centuries.

Objective: We propose novel measures to assess the completeness and the quality of demographic variables in the FamiLinx data at both the individual and the familial level over the 1600-1900 period. Utilizing Sweden as a test country, we investigate how the age-sex distribution and the mortality levels of the digital population extracted from FamiLinx diverge from the registered population.

Method: We employ descriptive statistics, negative binomial regression modeling, and standard life table techniques for our measures of completeness and quality.

Results: Missing values and accuracy in demographic information from FamiLinx

are selective. When one demographic variable is available, researchers can effectively anticipate the availability of other demographic information. The completeness and quality of demographic variables within kinship networks are markedly higher for individuals with more complete and accurate demographic information. Populations from FamiLinx display lower mortality levels compared to the registered population and their representativeness improves towards the end of the 19th century.

Contribution: This study sheds new light on the opportunities and challenges of harnessing online genealogies for demographic research. Although this data source offers much promise, its usability in population studies is dependent on the quality and completeness of its recorded demographic information and their selectivity.

Keywords: · Online Genealogies · Digital Data · Data Quality · Kinship networks · Completeness · FamiLinx

2.1 Introduction

The digital revolution has provided researchers with access to an unprecedented wealth of non-traditional data sources that can be used in population studies (Cesare et al., 2018; Kashyap, 2021). Among these emerging sources, online genealogical data have garnered significant attention (Gavrilova and Gavrilov, 2007; Hsu et al., 2021; Stelter and Alburez-Gutierrez, 2022; Cozzani et al., 2023; Blanc, 2024a,b; Minardi et al., 2024; Corti et al., 2024). These data sources present themselves as vast repositories of information from genealogical websites that enable users to reconstruct their own family trees. Online genealogical data are micro-level data that a) are scraped from digital family tree information stored in genealogical websites; b) link individuals not only to their parents, but also to more distant relatives; and c) provide additional details on the demographic characteristics of individuals, such as their dates and locations of birth and death (Song and Campbell, 2017).

Although online genealogical data were not primarily designed for use in social science research, they hold significant potential. First, they serve as exclusive repositories of data on extended family networks that span multiple centuries and cross-national borders. These data allow researchers to link individuals not only to their parents, but also to their more distant ancestors. Additionally, the kinship structure of these data sources enables researchers to examine multi-generational processes, and thus to go beyond the traditional two-generation approach that primarily focuses on parent-offspring associations (Mare, 2011; Song and Campbell, 2017). Second, the large volume of demographic information in these data sources, including details about birth and death locations and dates, permits researchers to investigate long-term population dynamics in regions and historical periods for which official population data may be scarce or unavailable (Stelter and Alburez-Gutierrez, 2022).

The use of genealogies in demography has emerged in response to the lack of historical data on the demographic experiences of kin groups (Post et al., 1997). Scholars have turned to genealogical data to advance the field of historical demography, to analyze historical trends in key demographic behaviors over time, and to study past mortality patterns

and the influence of heredity and family dynamics (Zhao, 2001; Otterstrom and Bunker, 2013). Louis Henry is recognized as the pioneer of family reconstitution. His work, which identified genealogies as rich sources for demographic research, has enabled researchers to pose a broader range of questions about family history, and to trace ancestors and more distant kin further back in time (Henry, 1968; Hollingsworth and Hollingsworth, 1976; Wrigley, 1981). Early genealogies and existing historical studies relying on genealogical data mainly focused on the ancestors and descendants of specific social groups living in specific areas (Henry, 1968; Otterstrom and Bunker, 2013). Furthermore, most genealogical reconstitution efforts suffered from incomplete location specificity and family networks (Kasakoff and Adams, 1995; Post et al., 1997). More recently, thanks to the digital revolution, genealogies have become powerful resources for tracing multiple generations of relatives over time using online platforms (Otterstrom and Bunker, 2013).

While their inherent structure makes the use of online genealogical data in population studies appealing, we should be critical of the claims that have been made about their merits. We contend that a thorough explanation of their limitations, including issues of data quality and potential biases, is imperative to ensure the responsible use of these data in population studies. The presence of individuals in a genealogy typically hinges on genealogists' knowledge of their relatives or their decisions about whom to include in their family trees (Calderón Bernal et al., 2023). Hence, these databases generally overrepresent the family networks of individuals who experienced more favorable demographic conditions than the general population, including higher fertility, lower mortality, and higher nuptiality (Zhao, 2001). Conversely, individuals with matrilineal and extinct lineages are often neglected. In genealogies, certain subpopulations are consistently underrepresented, including children who passed away at an early age and childless women. Online genealogies are also affected by selective remembering and the inclusion of individuals in online genealogical trees is contingent upon having a living descendant interested in tracing their family history (Zhao, 2001; Chong et al., 2022; Cozzani et al., 2023; Minardi et al., 2024). Genealogy users are more inclined to remember ancestors with important roles in their family history (Chong et al., 2022) and may tend to downplay relatives

who dishonored the family (Zhao, 2001). These problems combine to create considerable demographic selectivity issues, including the underestimation of mortality and the overestimation of fertility (Calderón Bernal et al., 2023). At the same time, the underreporting of individuals dying at young ages might result in an underestimation of fertility levels (Calderón Bernal et al., 2023; Hollingsworth and Hollingsworth, 1976). When genealogies exhibit inadequate coverage and representativeness, particularly when recording only a few generations or closer relatives, biases become more pronounced, consequently reducing the accuracy of estimations (Zhao, 2001; Calderón Bernal et al., 2023).

Additionally, online genealogical data may suffer from a high prevalence of missing values for essential demographic variables, such as birth and death locations and dates. This is not unexpected, as users of genealogical websites are more focused on tracing their ancestors than on meticulously recording precise information about the locations and the dates of their relatives' vital events. Limited familiarity with one's own relatives may also contribute to imprecise or missing information. Furthermore, certain subpopulations within genealogies are typically underrepresented and more likely to feature missing information. Examples include children who passed away at a young age and childless women. Genealogies often commence with a patriarch documenting his family history, with women typically recorded solely as wives or daughters, resulting in their information being less comprehensive compared to that of males (Zhao, 2001). In light of these issues, we argue that a comprehensive examination of missing value patterns in crowdsourced genealogies is warranted.

Despite the previously mentioned limitations of crowdsourced genealogical databases (as shown by Stelter and Alburez-Gutierrez (2022), Chong et al. (2022) and Calderón Bernal et al. (2023)), the majority of population studies relying on these data sources have operated under the assumption that their selected individual samples accurately mirror the broader population (Hsu et al., 2021; Cozzani et al., 2023; Blanc, 2024a,b; Minardi et al., 2024). Prior research has attempted to illustrate the biases stemming from ascending genealogies and their impact on crucial demographic measures, such as life expectancy at birth and the total fertility rate, by means of simulations (see Zhao (2001) and

Calderón Bernal et al. (2023)). To the best of our knowledge, our study represents the first attempt to evaluate the accuracy and the completeness of demographic variables at both the individual and the family network level in a big genealogical digital database, and to analyze the implications for the use of this database in population studies. We look at the age structure of the population drawn from the genealogical data and a key demographic measure, life expectancy, to illustrate how the quality of the reported demographic information can vary. Our aim is to offer scholars a more comprehensive understanding of the dataset's strengths and limitations, thus enabling them to make more informed decisions when utilizing the FamiLinx data for their research projects.

In this paper, our objective is to investigate the challenges associated with missing information in extensive genealogical data, and to highlight the critical issues that may hinder the usability of these data for demographic research. Specifically, we assess the accuracy and the comprehensiveness of vital demographic variables in online genealogies, including birth and death dates and locations, for individuals and their associated family networks. In our analysis, we rely on the concepts of completeness and quality. Completeness refers to the quantification of the percentage of non-missing values for common demographic variables, while quality indicates the accuracy of the reported demographic information. Further details on the measurement of completeness and quality are provided in the method section. Although we focus on the FamiLinx database, we believe our findings and methods are also applicable to other genealogical databases. In a nutshell, this paper seeks to address the following research questions:

(a) What are the potential advantages and pitfalls of using online genealogies for demographic research?

(b) How do the completeness and the quality of the demographic information in online genealogical data affect their usability? Are completeness and quality clustered within selected kinship networks?

(c) How are the age-sex distributions and the demographic estimates derived from online genealogical populations impacted by the completeness and the quality of the reported demographic information?

2.2 Data

2.2.1 The FamiLinx Database

The analysis relies on the FamiLinx database, which is sourced using publicly available genealogies accessible on the geni.com website. These digital data are derived from family trees that have been constructed by a network of users from multiple countries with a common interest in tracing their own ancestral lineages. Since these genealogies have been built using a bottom-up approach, they are of the ascending type. This means that the genealogist begins the construction of their family tree from the bottom and then traces their lineage backward in time, including their parents, grandparents, great-grandparents, and so on. This process allows for the creation of a family tree that "ascends" through the generations, illustrating the kinship ties between individuals when moving from present relatives to earlier ancestors.

Furthermore, FamiLinx has a passive registration system where only the main vital events, i.e., births and deaths, are recorded and the genealogists are not aware of the individuals' status at all the time points. This limitation hampers the applicability of FamiLinx to examine more complex demographic phenomena, such as migration trends and marriage patterns.

In recent years, scholars have increasingly turned to FamiLinx for population research. Leveraging the dataset's rich information spanning numerous centuries, FamiLinx has primarily served as a tool to investigate historical demographic trends. Existing studies have predominantly delved into historical mortality dynamics (Chong et al., 2022; Stelter and Alburez-Gutierrez, 2022; Cozzani et al., 2023; Minardi et al., 2024), scrutinizing the dataset's biases and representativeness compared to more reliable data sources (Chong et al., 2022; Stelter and Alburez-Gutierrez, 2022) or examining disparities in lifespan (Stelter and Alburez-Gutierrez, 2022; Cozzani et al., 2023; Minardi et al., 2024). Other research initiatives utilizing FamiLinx have centered on historical fertility patterns (Gay et al., 2023; Blanc, 2024a,b) and the correlation between fertility and longevity (Hsu et al., 2021). Blanc (2024a) additionally utilized the dataset to uncover patterns of internal mi-

gration to and from urban centers. Overall, FamiLinx has emerged as a valuable resource for analyzing pivotal historical processes such as demographic transitions, shedding light on the potential of online genealogies in population research while acknowledging their inherent limitations in terms of bias and representativeness.

Our focus on FamiLinx derives from its easy accessibility, which makes it appealing to researchers. All the dataset's records are anonymized, and no formal request to access the information is needed. Additionally, compared to other genealogies, FamiLinx covers more countries and provides quite detailed information about the location of events, surpassing the limited geographic scope of traditional genealogies. The demographic information stored in FamiLinx spans multiple generations of individuals, and thus covers a long period of time. Among the database's strengths is the ease with which the individual profiles can be linked to their family networks, thus facilitating a more comprehensive tracing of both ancestors and collateral kin.

The dataset was curated by Kaplanis et al. (2018), who gathered an extensive collection of 86 million profiles from the `geni.com` website. This social media platform allows users to upload their family trees and establish individual profiles for each member of their familial network. FamiLinx includes a dataset containing anonymized individual-level records for all 86 million individuals, as well as a dataset with information about the kinship ties between children and parents for approximately 43 million of these individuals. By leveraging these two types of records, researchers can identify distinct types of kin beyond parents and children, such as siblings and grandparents. Additionally, Kaplanis et al. (2018) eliminated implausible kinship ties, e.g., individuals with more than two parents. The task of linking the two datasets is made easier by the fact that each individual is assigned a unique identification number. Specifically, the dataset with all the individual-level records incorporates vital demographic variables, including birth and death dates and locations, as well as gender. Each demographic variable is represented by multiple columns. For instance, the demographic information about the dates is presented in separate columns for day, month, and year. The locations of demographic events are documented through a two-digit country code representing the country name of the

vital event, and through the country name itself reported as a string of text. Building on the information contained in the location-based text strings and two-digit country codes, Kaplanis et al. (2018) assigned the latitude and longitude coordinates to profiles with sufficiently detailed location information on the vital events.

Since all the individuals who were still alive as of 2015, when the profiles were scraped from `geni.com` (see Kaplanis et al. (2018) for details), were omitted from the database, the demographic analysis should be restricted to individuals from extinct cohorts (see the appendix of Kaplanis et al. (2018) for details). Additionally, since the records in the database are anonymized, it is not feasible to link them to other micro-level historical data sources, such as parish records or censuses. Moreover, de-anonymization is not allowed under the terms of use of the data.

2.2.2 Analytical Sample

We investigate how the completeness and the quality of the data are manifested within family networks. As the individuals in genealogies are embedded within kinship networks, we believe that it is essential to investigate how the quality and the completeness of the demographic information on individuals in the genealogies are related to those of their kin. To facilitate our analysis, we define a subsample comprising approximately seven million "focal" (or anchor) individuals, which we refer to as the "analytical sample." Based on our selection, we recall that individuals with identifiable kinship networks are inherently a subset of a larger population. To be included in the analytical subsample, individuals must a) have **at least one parent or one child**, as this ensures that the size of the kinship network of the focal individual is non-zero; and b) have **at least one known place of birth or death**, as determined by the following criteria.

2.2.3 Determination of Birth and Death Locations

The locations of the demographic events experienced by focal individuals was determined by a three-tier method, which involved three algorithms ranked in order of preference:

(a) **Exact matching using the country code**: birth and death locations are inferred from the reported two-digit country code.

(b) **Regular expression matching**: birth and death locations are determined by a set of text strings, known as regular expressions, that specify a matching pattern for the name of the country of interest.

(c) **Inferred coordinates**: this method leverages the latitude and longitude coordinates by Kaplanis et al. (2018) to identify the country of the vital event.

The motivation behind the implementation of this approach is that the inferred latitude and longitude coordinates by Kaplanis et al. (2018) may be affected by reporting errors due to historical changes in boundaries between countries.

To establish the definitive birth and death locations, we extract the country names from inferred coordinates harnessing a geo-parsing algorithm available in the R package mapdata (see Becker et al. (2022) for the details). We identify the 20 countries with the highest numbers of vital events.

Subsequently, we select the birth and death countries using the country codes and text strings from the 20 countries according to the methods described above. For instance, if a profile has two different birth locations, one determined by exact matching and the other based on the inferred coordinates, we assign the birth country identified by the exact matching method. Extending our analyses beyond these 20 countries would not affect our results, given the extremely low numbers of reported birth and death events in the remaining countries.

2.3 Method

Our analysis consists of four steps. In the first step, we measure the completeness and the quality of the FamiLinx data. In the second step, we model the association between focal individuals and their kin in terms of the completeness and the quality of the demographic information. In the third step, we aim to generate population pyramids and age-sex distributions to evaluate the representativeness of populations drawn from online

genealogical data. In the fourth step, we calculate life expectancy at age 30 based on the FamiLinx data. The methods applied in each of these steps are outlined.

2.3.1 Measurement of Completeness and Quality in FamiLinx

Our analysis relies on two pivotal concepts that determine the usefulness of FamiLinx for population studies: completeness and quality. These two concepts are investigated by considering the five main demographic variables present in the dataset: gender and birth and death dates and locations.

Following the guidelines laid out by the United Nations (United Nations, 2016), we measure completeness as the extent to which primary demographic variables (birth and death years and countries) display non-missing values. Specifically, this concept is quantified as the proportion of individuals with non-missing values for each of the aforementioned demographic variables. After the completeness of each demographic variable has been computed, we can analyze the variations in the marginal distributions of these variables when one of them is non-missing. Through this approach, we are able to gain novel insights into the overall level of completeness of individual records in FamiLinx. The concept of quality refers to the accuracy of the reported birth and death dates.

Following the guidelines established by Kaplanis et al. (2018) and Minardi et al. (2024), we consider an individual record to exhibit higher quality if the month of the birth and/or death date is not missing. To measure the quality of the dates, we rely on the concept of year heaping. By year heaping, we mean a preference for recording years with a last digit that is either zero or five (see Stockwell and Wicks (1974) for a review on year heaping measurement[1]). Depending on whether we are considering births or deaths, we may use the term birth year heaping or death year heaping. When a sample has year heaping issues, it typically displays a non-uniform distribution of the number of births and deaths with unrealistic spikes in years ending in zero or five. Therefore,

[1]A similar measurement concept for year heaping was employed by Cummins (2017) to assess the accuracy of the birth and death dates when analyzing the lifespan of the Western European nobles from 800 up to 1800.

to examine the quality of the data in FamiLinx, we can only consider individuals with non-missing birth and death years. For this purpose, we define an indicator measuring the level of year heaping in the data. This indicator is calculated separately for the birth events and the death events in the data. Our strategy involves partitioning the selected individuals into two groups: one consisting of individuals with the non-missing month of the vital event and the other consisting of individuals for whom only the year of the vital event is available. Following this primary division, we group these individuals into 25-year intervals, and calculate the proportion whose reported year ends in zero or five. If the proportion in a group is close to 20%, we assume that there is no year heaping. Conversely, if it exceeds this threshold value, we conclude that there are year heaping issues in the data. In our example, if the proportion of individuals with the non-missing month of the vital event is around 20%, we can conclude that the demographic information for this group is relatively accurate Spoorenberg and Dutreuilh (2007) for a review on age-heaping measurement). In our examination of the quality of the data, we restrict our analysis to individuals who were born or died between 1600 and 1900. We do so because the records of individuals with a birth or a death recorded before 1600 are considered unreliable (Kaplanis et al., 2018), whereas the cohorts born after 1900 might include individuals who were still alive as of 2015, which could result in ascertainment bias.

2.3.2 Measurement of Completeness and Quality within Kinship Networks

After examining the completeness and the quality in the whole dataset, we explore how these concepts are applicable within the extended family networks (which include grandchildren, children, siblings, cousins, parents, aunts and uncles, and grandparents). Since researchers may be interested in examining the size and the structure of kin at any time point (Post et al., 1997) or investigating the multigenerational transmission of demographic behaviors, we believe that it is crucial to investigate the quality and the completeness of demographic information not only for focal individuals, but also for their kin. This investigation could provide novel insights that are of value to researchers in-

terested in using FamiLinx to carry out studies in the domains of historical and family demography.

To carry out this analysis, we rely on the individuals in the analytical sample and their respective kinship networks. We consider the demographic variables of birth and death countries and years. We disregard gender in the set of demographic variables due to the high percentage of non-missing values for this variable in the dataset.

To study the association between the completeness of demographic information for a focal individual and that of their kinship network, we use a negative binomial regression model. This model can be seen as a generalization of the Poisson regression model (Hilbe, 2011). In both models the interpretation of the regression coefficients remains the same. However, in the negative binomial regression model the equi-dispersion assumption is not required in that the mean of the response does not need to be equal to its variance. Hence, the modeling approach is appropriate given the over-dispersion present in the data. Concerning our model, for each combination of demographic variable j and type of relative k, we outline the following negative binomial regression model.

$$Y_{ijk}^{\text{completeness}} | \alpha_{0jk}, \alpha_{1jk}, \theta_{jk} \sim \text{NegBinom}\left(\mu_{ijk}, \theta_{jk}\right) \tag{2.1}$$

$$\text{E}\left[Y_{ijk}^{\text{completeness}} | \alpha_{0jk}, \alpha_{1jk}, \theta_{jk}\right] = \mu_{ijk} = \exp\left(\alpha_{0jk} + \alpha_{1jk} z_{ij}^{\text{completeness}} + \psi_{jk} c_{ik}\right) \tag{2.2}$$

$$\text{VAR}\left[Y_{ijk}^{\text{completeness}} | \alpha_{0jk}, \alpha_{1jk}, \theta_{jk}\right] = \mu_{ijk} + \frac{\mu_{ijk}^2}{\theta_{jk}} \tag{2.3}$$

Where the dependent variable $Y_{ijk}^{\text{completeness}}$ denotes the number of relatives of type k of the focal individual i with a non-missing value in the demographic information j and the independent variable $z_{ij}^{\text{completeness}}$ indicates whether focal i has demographic information j non-missing. The parameter μ_{ijk} denotes the mean of the dependent variable and can be interpreted as the expected number of relatives of type k with non-missing values in demographic information j for a focal individual i. The parameter θ_{jk} is the reciprocal

dispersion parameter is included to account for overdispersion in the response variable. c_{ik} denotes the number of relatives of type k of the focal individual i[2].

To evaluate the association between the focal individual and the quality of their kinship network's demographic information, we implement a negative binomial regression model for each type of relative. For the implementation of this set of negative binomial regression models, we selectively include only focal individuals and their kin conditional on possessing non-missing birth (death) years. We aim to assess whether the dates of vital events reported for the relatives of a focal individual are more likely to be accurate when the month of the event for the focal individual is known.

The examination of the quality of the reported dates is based on the following multivariate negative binomial model.

$$Y_{ijk}^{\text{quality}} | \gamma_{0jk}, \gamma_{1jk}, \phi_{jk} \sim \text{NegBinom}\Big(\nu_{ijk}, \phi_{jk}\Big) \tag{2.4}$$

$$\text{E}\Big[Y_{ijk}^{\text{quality}} | \gamma_{0jk}, \gamma_{1jk}, \phi_{jk}\Big] = \nu_{ijk} = \exp\Big(\gamma_{0jk} + \gamma_{1jk} z_{ij}^{\text{quality}} + \beta_{ij}' X_i + \delta_{jk} d_{ijk}\Big) \tag{2.5}$$

$$\text{VAR}\Big[Y_{ijk}^{\text{quality}} | \gamma_{0jk}, \gamma_{1jk}, \phi_{jk}\Big] = \nu_{ijk} + \frac{\nu_{ijk}^2}{\phi_{jk}} \tag{2.6}$$

where the independent covariate z_{ij}^{quality} denotes whether the month of the demographic event j experienced by the focal individual i is non-missing; and the dependent variable Y_{ijk}^{quality} indicates the number of relatives of type k of the focal individual i with a non-missing value in the month of the date for the demographic event j. The parameter ν_{ijk} is the expected value of the outcome variable and is interpretable as the expected number of relatives of type k with non-missing month in demographic information j for a focal individual i. The parameter ϕ_{jk} retains the same meaning as the parameter θ_{jk} in the previous model. X_i denotes a matrix of fixed effects made up of dummies referring

[2]We included the number of relatives of type k as a control variable since having a higher number of relatives of a certain type may increase the probability of having a larger number of relatives with a non-missing value in a demographic variable.

to the period in which the demographic event of interest occurred. We believe that we should account for fixed effects, since the degree of heterogeneity in the quality of the reported demographic information may be higher for individuals with vital events in earlier historical periods. d_{ijk} indicates the number of relatives of type k of the focal individual i with the non-missing year of the demographic event j[3].

To advance our understanding of the representativeness of digital populations drawn from online genealogies, we compare the age-sex distribution extracted from genealogical data with that of the registered population. In this analysis, we identify two samples with distinct quality levels. One sample consists only of individuals with non-missing birth and death months, while the other is made up of individuals with missing birth or death months. This allows us to examine the impact of different sample selections on the age-sex distribution of the genealogical populations. To carry out this comparison, we employ population pyramids, which enables us to visually investigate the extent to which the digital population drawn from online genealogies aligns with the registered population. In addition, we calculate the differences between the genealogy-based age-sex percentages and those based on census data for the same time period.

Finally, we leverage data from online genealogical populations to compute life expectancy at age 30. We aim to compute the demographic estimates from samples with distinct quality levels. This is again motivated by our interest in examining the impact of sample selection in online genealogical populations on the estimation of common demographic indicators, such as life expectancy at age 30. The previous measure is calculated using life tables with mortality rates smoothed over both ages and years. This calculation allows us to examine the ability of online genealogical data to capture historical trends in adult mortality. We smooth our estimates to avoid unrealistic shocks in life expectancy trends due to the small sample sizes. The smoothing is carried out utilizing two-dimensional P-splines implemented through the R package mortalitysmooth developed by Camarda

[3]We added the number of relatives of type k as a control variable since having a higher number of relatives with a non-missing birth or death year may increase the probability of having a higher number of relatives with a non-missing birth or death month.

(2012). For more details on the mortality smoothing and its implementation in R, see the appendix.

2.4 Results

2.4.1 Completeness of Individual Demographic Information in FamiLinx Data

In Figure 1, we present the percentage of individuals with non-missing information for the considered demographic variables (gender and birth and death dates and locations) to describe their availability in the initial dataset and in other subsamples (selected from the initial dataset). The characteristics of the initial full dataset and the subsamples are shown in Table A.2 in the appendix. The radar charts (Figure 2.1) show that in the initial full dataset, most of the observations have missing information for the considered demographic variables, but the presence of at least one available variable considerably reduces the likelihood that other demographic variables are unavailable. The latter condition includes the analytical sample used for this study and several samples of observations, conditioned on having a specific demographic variable available. In the initial full dataset, the year of birth, the year of death, the location of birth, and the location of death are available for only 25% of the individuals, or even less. However, in the analytical sample, the percentage of observations with available demographic information is larger. In particular, while the availability of gender information does not guarantee that other demographic information is available, knowing an individual's place of death increases the probability of having non-missing information for the other variables. Thus, when information on one variable is available, researchers can effectively expect that other demographic information is available, which contributes to a more comprehensive understanding of individual profiles in FamiLinx.

When we look at the completeness of the demographic information for those individuals born in specific countries (Canada, Germany, Sweden, United Kingdom, United States of America) (see Figure A.1 in the appendix), we see that the percentage of individuals

with non-missing information on the selected demographic variables is much higher than that observed in the initial full dataset. In general, the individuals in the genealogies who were born in the UK have more incomplete demographic information, and indeed have the highest percentage of missing information for all the considered variables. While the percentages are quite similar for the other analyzed countries, individuals born in the US seem to have a larger share of non-missing values for the demographic variables, especially those concerning the date of death and the place of death.

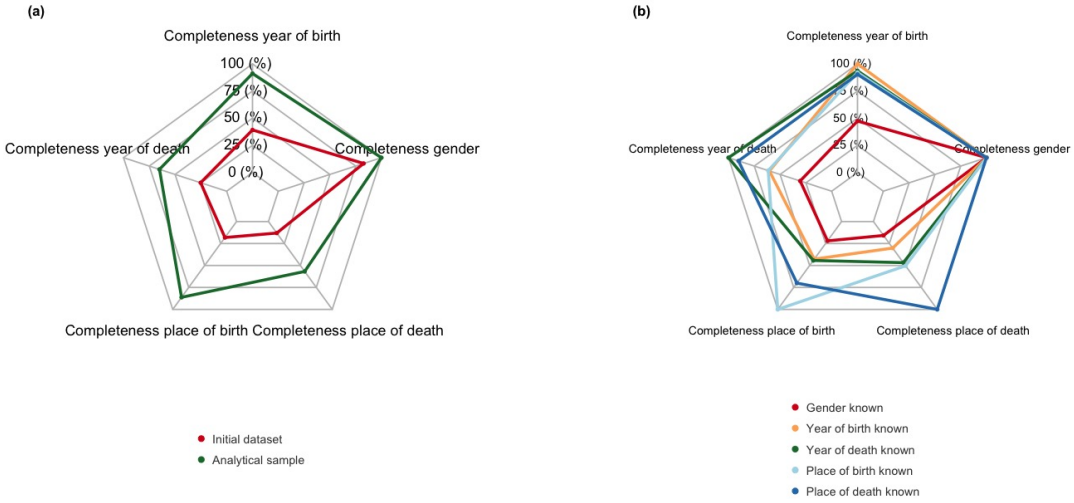


Figure 2.1: (a) Percentage of non-missing values for five demographic variables in the initial full dataset (N=86,124,644) and in the analytical subsample (N=7,618,651). (b) Percentage of non-missing values for five demographic variables in different samples, identified by the availability of specific information.

2.4.2 Completeness of Demographic Information within Kinship Networks

The negative binomial models reveal a positive association between the completeness of the demographic variables for the focal individuals and those for their kin, independent of the size of the kinship network (see Table A.3 in the appendix). This means that the presence of more complete variables for a focal individual is associated with having a higher number of relatives with more complete demographic information. These associations are found across distinct types of relatives and all the considered demographic variables, albeit with heterogeneous degrees of magnitude. Among all the demographic

variables, the strongest association is observed for the birth year. As a robustness check, we ran a logistic regression model using as the response a binary variable equal to 1 if at least one of the relatives of a given type for a focal individual has a non-missing value in a demographic variable. As an additional sensitivity check, we implemented two other regression models: a negative binomial regression model, where the number of relatives is treated as offset, and a binomial regression model. The results of the alternative models, included in the appendix (Table A.5, Table A.7, Table A.9), are consistent with those of the negative binomial model presented in the main text.

Regarding the specific types of relatives, horizontal kin, namely cousins and siblings, tend to exhibit stronger associations for all demographic variables. The associations are weaker for more distant kin, such as grandparents. For instance, the expected number of siblings with a non-missing birth year for a focal individual with non-missing birth year is over three times bigger than that of a focal individual with a missing birth year.

The expected number of children, cousins and parents with non-missing birth year for the same focal individual is approximately twice higher than that of a focal individual with missing birth year. If we focus on the number of grandparents and, aunts and uncles with non-missing birth year for a focal with non-missing birth year, their expected number is more than 50% higher than that of a focal individual with missing birth year.

The expected number of siblings, parents, children and cousins with non-missing values in the death year, birth and death countries increases by at least 50% for a focal individual with non-missing values in the same demographic variables. For more distant kin, such as grandparents, grandchildren and aunts and uncles, these estimates are still above the unit, but are smaller in magnitude.

The observed differences in magnitude can be attributed to the higher proximity between the year of demographic events experienced by focal individuals and those experienced by their horizontal kin. When considering more distant kin, the temporal gap between the demographic events widens. Hence, for genealogists willing to reconstruct their own family trees, knowing the year of a demographic event experienced by the focal individual increases the likelihood of recollecting the same piece of demographic information for

relatives who lived in the same temporal period, e.g. by searching in parish records. Conversely, gathering demographic information for more distant kin proves challenging not only due to the higher temporal distance between the demographic events but also to a more substantial effort to link the focal individual to their more distant relatives.

Overall, these results underscore how the completeness of demographic information tends to be shared among relatives. A focal individual with more complete demographic information has a higher likelihood of being embedded in a kinship network whose members have more complete demographic variables. This finding highlights the potential for studying demographic outcomes (fertility, longevity, etc.) within extended kinship networks beyond the classic parents-focal or children-focal relationships. Consequently, it opens up new opportunities for the exploration of demographic dynamics in the context of extended kinship networks.

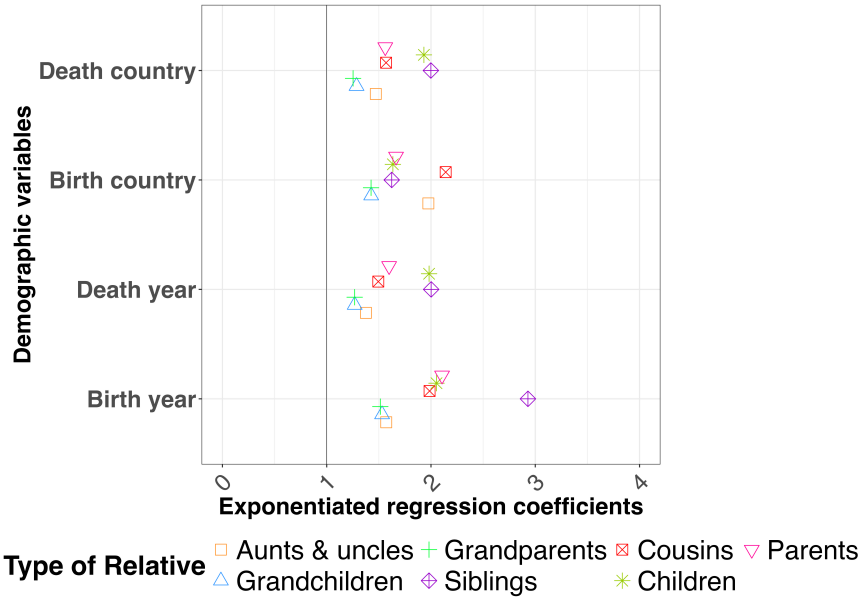


Figure 2.2: Exponentiated coefficients from negative binomial regression measuring the association between a focal individual and their relatives in terms of the completeness of the reported demographic variables.

2.4.3 Quality of Individual Demographic Information in FamiLinx data

Figure 2.3 indicates that observations with complete dates of birth and death (i.e., that specify the years and the months of birth and death) do not seem to show a preference for those years. Individuals for whom only information on the years is available are more prone to year heaping issues. Thus, observations with complete dates of birth and death are of higher quality. We can see an increase in quality over time for birth year heaping. Indeed, in the 19th century, the percentages for individuals with complete dates are closer to the percentages for individuals with incomplete dates. Overall, the prevalence of death year heaping is lower than the prevalence of birth year heaping, which suggests that when the year of death is available, it is more likely to be correct and precise. When we look at the occurrence of birth and death year heaping across different countries of birth (see Figures A.2 and A.3 in the appendix) we note similar trends, but in different magnitudes. In general, among all the considered countries, observations with complete birth dates do not seem to be affected by birth year heaping. Moreover, among those with missing birth months, the proportion of birth years ending in 0 or 5 decreases over time. There is no evident improvement in the quality of the reported death dates over time. However, observations with complete death dates exhibit fewer instances of death year heaping than those with incomplete death dates across all the considered countries.

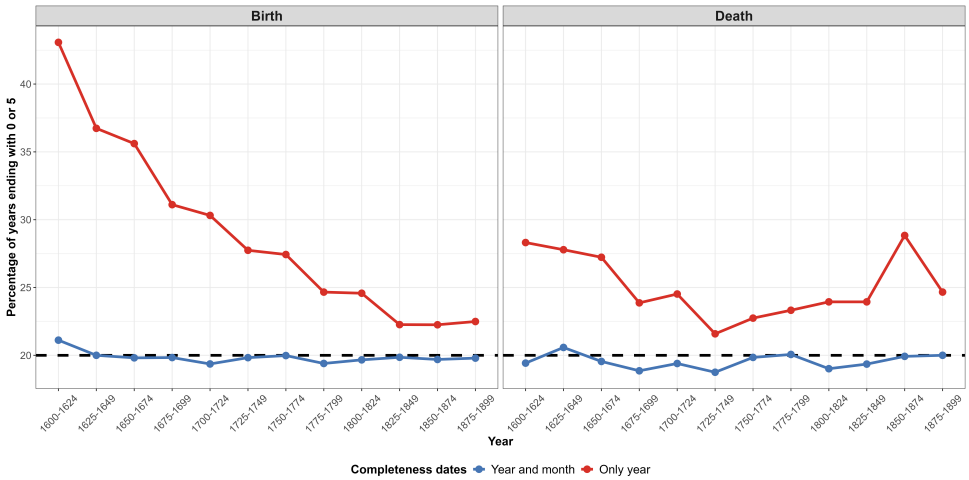


Figure 2.3: Percentages of years of birth and years of death ending with zero or five by completeness of the dates of birth and death, and by historical period.

2.4.4 Quality of Dates within Kinship Networks

We find a positive association between the quality of the birth and death dates for the focal individuals and those for their kin, net of the size of the kinship network (see Table A.4 in the appendix). This implies that possessing more accurate demographic information is associated with a higher number of relatives with demographic information of higher quality. These associations are observed across distinct types of relatives for both birth and death dates, with the former showing the strongest association. As a robustness check, we ran a logistic regression model using as the response a binary variable equal to 1 if at least one of the relatives of a given type for a focal individual has a non-missing month in the birth or death dates. As an additional sensitivity check, we tested two other modeling approaches, namely a negative binomial regression model, where the number of relatives is treated as offset, and a binomial regression model. The results, included in the appendix (Table A.6, Table A.8, Table A.10), are consistent with those of the negative binomial model presented in the main text.

Horizontal kin, especially siblings, tend to exhibit stronger associations for the variable birth month. The expected number of siblings with a non-missing month in the birth date is almost four times higher for a focal individual with a non-missing month in the birth date compared to a focal with a non-missing month. Focusing on the death month, slightly higher associations are observed for siblings, parents and children. The expected number of parents, siblings and children with non-missing month in the death date is 50% higher for a focal individual with non-missing death month compared to a focal individual with missing death month. Concerning the other relatives, the number of relatives with a non-missing month in death/birth date for a focal individual with a non-missing month in death/birth dates increases by over 20% compared to a focal with a missing month in the birth/death date.

Table A.4 in the appendix displays all the regression coefficients, including the effects of the distinct birth and death cohorts on the number of relatives without a non-missing month in the birth/death date. In general, an increase in the magnitude of these cohort effects is observed implying an improvement in the quality of the reported demographic

information. Nonetheless, if we focus on grandchildren of focal individuals from more recent birth/death cohorts, the associations are slightly lower due to the fact that FamiLinx excludes individuals that were still alive in 2015.

Figure 2.4: Exponentiated coefficients from negative binomial regression measuring the association between a focal individual and their relatives in terms of the quality of the reported demographic variables.

2.4.5 Discrepancies between the Age-sex Distribution in FamiLinx and in the Registered Population

We now compare the age-sex distribution of the digital population derived from online genealogies with that of the registered population. As an illustrative example, we concentrate on the Swedish genealogical population over the historical period of 1751-1900. Compared to other countries, Sweden stands out for its rich wealth of demographic data starting from the year 1751, including detailed population counts disaggregated by sex and age, which are available from population registers.

In Figure 2.5, we display the percentage differences in age-sex proportions between the Swedish genealogical population extracted from FamiLinx and the registered Swedish population over four calendar years: 1751, 1800, 1850, and 1900. These differences are computed for two distinct quality levels: one comprising individuals with precise birth

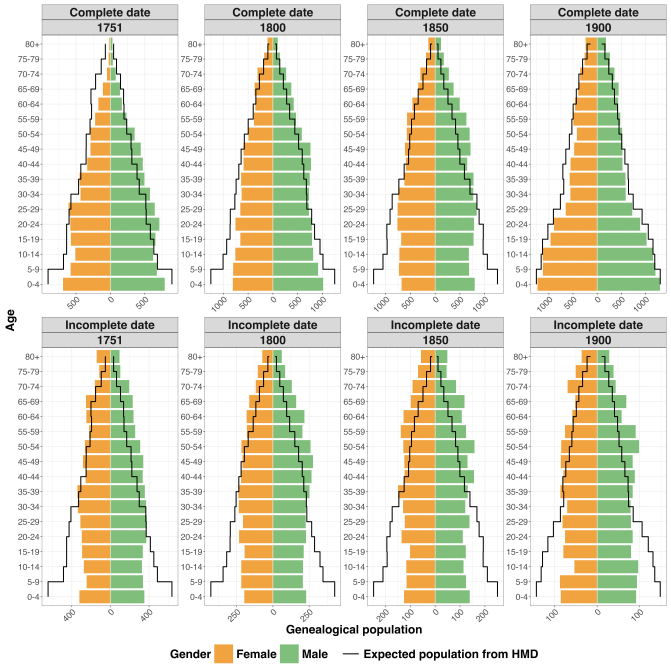


Figure 2.5: Population pyramids for the Swedish population from FamiLinx for the calendar years 1751, 1800, 1850, and 1900 by quality level.

and death dates (non-missing birth and death months), and the other comprising individuals with at least one less precise date (the birth or the death month is missing). Notably, these disparities seem to be more modest for the genealogical group with higher information quality throughout the historical period under scrutiny. If we focus on the sample of Swedish individuals with precise birth and death dates, the age-sex distribution derived from this subsample mirrors the estimates for the total Swedish population toward the end of the 19th century from the Human Mortality Database. Nonetheless, regardless of the quality of the data used, a consistent pattern is observed for the Swedish genealogical population before the end of the 19th century. Individuals at younger ages and women tend to be underrepresented, whereas more longevous male individuals are overrepresented.

Figure 2.6 shows that the underestimation of the proportions of individuals in the 0-14 age group with more accurate dates increases until the mid-19th century, but then declines rapidly toward bias levels that are close to zero. Among adult individuals (aged 15-64) with higher quality information, males exhibit an upward bias that decreases toward the end of the 19th century. Conversely, females in the same age group are underrepresented

in the second half of the 18th century (1751-1799) and of the 19th century (1851-1900), whereas they seem to be well-represented in the first part of the 19th century (1800-1850). Turning our attention to individuals aged 65 or older, we observe a consistent upward bias in the proportions for both men and women, which decreases slightly starting in the second half of the 19th century.

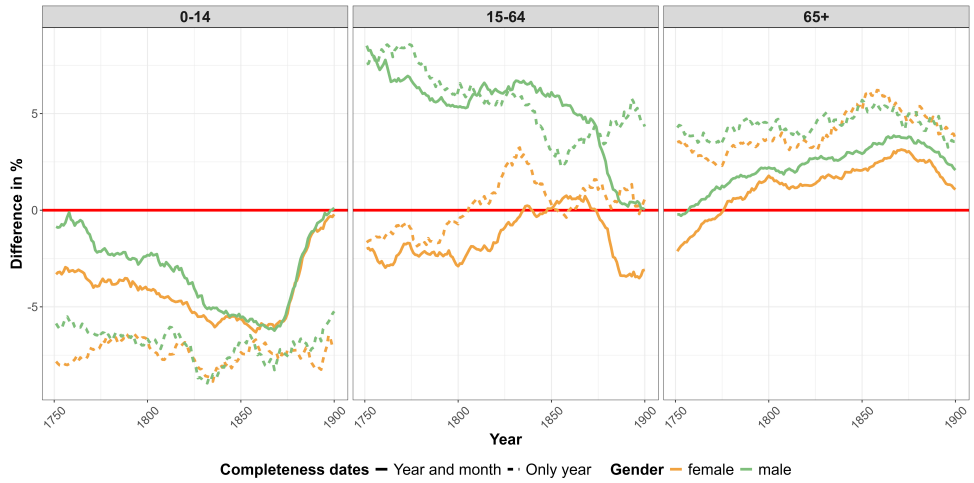


Figure 2.6: Difference between the age-sex distribution in percentage between the Swedish population from FamiLinx by quality level (precise birth and death dates against at least one non-precise date) and the registered Swedish population over the historical period 1751-1900.

2.4.6 Discrepancies between Life Expectancy in FamiLinx and in the Registered Population

We now focus on investigating life expectancy at age 30 in Sweden over the historical period of 1751-1900, specifically considering the two quality levels defined above. Our decision to evaluate life expectancy at age 30, as opposed to at birth, is motivated by the underestimation of child mortality inherent in the online genealogies (see Figure A.5 in the appendix), and the recommendation of Stelter and Alburez-Gutierrez (2022). Again, we focus on Sweden due to its long time series of national demographic estimates. However, we acknowledge that our results for Sweden may not extend to populations from other countries.

In Figure 2.7, we present the estimates of life expectancy at age 30 stratified by quality level and sex. To provide a benchmark, we incorporate life expectancy estimates from

the Human Mortality Database. It is essential to note that our analysis is limited to individuals with non-missing birth and death years who were born and died in Sweden; i.e., to a sample of highly selected individuals. The results show a pronounced survivorship bias within the genealogical Swedish male population. In line with Stelter and Alburez-Gutierrez (2022) for Germany and the Netherlands, we find that the male life expectancy at age 30 estimated from genealogical data toward the end of the 19th century seems to be slightly closer to the life expectancy derived from Swedish register data. In contrast to our analysis of male longevity, our investigation of female longevity reveals unexpected trends in life expectancy at age 30. Throughout the 18th century, this demographic indicator is consistently overestimated for the Swedish female population in FamiLinx. For the first half of the 19th century, the estimates of life expectancy at age 30 based on genealogical data align with those from the Human Mortality Database. Nonetheless, a noteworthy shift can be observed toward the end of the 19th century, as the genealogical data consistently underestimate life expectancy at age 30 for women. In general, our analysis highlights that the observed trends in life expectancy at age 30 hold true across the quality groups under comparison. Nonetheless, our results also suggest that the bias in the life expectancy at age 30 differs by gender. A possible explanation is suggested by Figure 2.6 in the appendix, in which the percentage of women in the age range 15-64 in Sweden is closer to the actual one from population registers during the period 1800-1870 in comparison to the share of men, which is more severely overestimated. On the contrary, in the last part of the 19th century, women aged 15-64 become more and more underrepresented, whereas the representation of men in the same age range improves. As a consequence, after 1870, we see a continuous increase in the underestimation of life expectancy at age 30 for women and a decrease in the overestimation for men. While this is an intriguing result, which would need further investigation, we lack sufficient tools to provide a robust explanation for the observed gender differences.

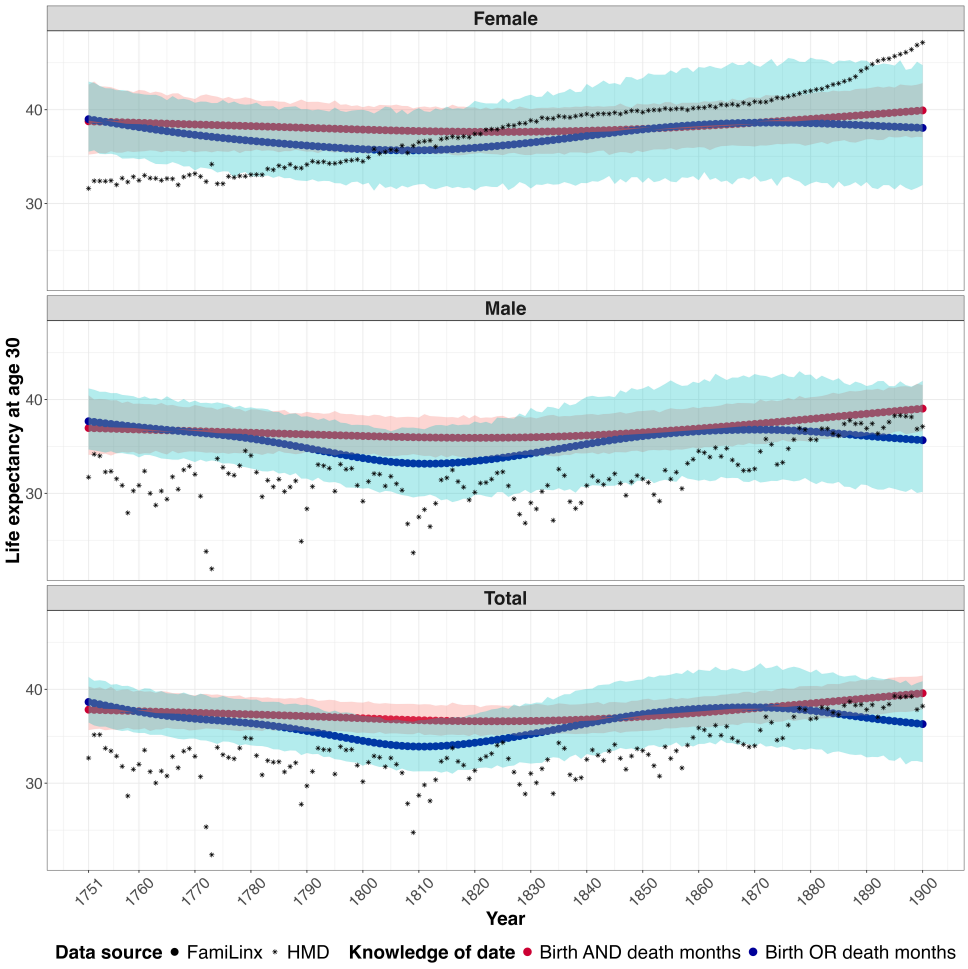


Figure 2.7: Life expectancy at age 30 in Sweden for the historical period (1751-1900) by sex and quality level (precise birth and death dates against at least one non-precise date) in FamiLinx and Swedish life expectancy at age 30 from the Human Mortality Database.

2.5 Discussion

The extensive sample size and the availability of cross-border kinship networks render FamiLinx an asset for social scientists interested in exploring past population dynamics (Hsu et al., 2021; Stelter and Alburez-Gutierrez, 2022; Cozzani et al., 2023) and the intergenerational transmission of demographic behaviors (Blanc, 2024b; Minardi et al., 2024). The availability of kinship ties and demographic information enables researchers to explore how demographic outcomes have changed within family networks. A noteworthy aspect is the extensive time period covered by the FamiLinx data, which facilitates the examination of long-term demographic processes. By drawing online digital trees, FamiLinx opens up new avenues for understanding the demographic behaviors of past populations through the lens of digital data that are less common in the field of Historical Demography compared to other non-conventional data sources (e.g. Parish Records, Obituaries, Military records, Wills). The coverage of various countries over the past four centuries provides researchers with the unique opportunity of analyzing the composition of transnational kinship networks.

In this study, we showed that when information on one demographic variable is known it is more likely that information on other demographic variables will also be known. Individuals with non-missing months in birth and death dates tend to have more precise demographic information whose quality improves over time. Furthermore, our analysis revealed that individuals with higher-quality demographic information are likely to have relatives who also have more complete and accurate demographic information available. Additionally, using Sweden as example, we observed that individuals with non-missing demographic information tend to experience higher life expectancy in comparison to the registered population throughout the considered historical period.

The majority of the previous studies portrayed FamiLinx in a positive light and underlined its potential for demographic research, leading to significant contributions, especially in the domain of Historical and Family Demography. However, we advocate for a cautious approach and provide a few recommendations to scholars who want to utilize FamiLinx for their own research.

First, as outlined in table A.1, the overrepresentation of individuals with vital events (births and deaths) in Western Countries markedly restrict the geographical scope of the possible population studies. In fact, the previous studies (Hsu et al., 2021; Stelter and Alburez-Gutierrez, 2022; Chong et al., 2022; Pojman et al., 2023; Gay et al., 2023; Cozzani et al., 2023; Blanc, 2024a,b; Minardi et al., 2024; Corti et al., 2024), which relied on this data source, predominantly concentrated on the United States of America or countries in Western Europe. Unfortunately, this is also almost inevitable in our study, as the vast majority of the individuals in the dataset lived in these countries and it is here that their family networks are extended. Furthermore, when assessing the quality of the demographic information and comparing it with the population recorded in a given historical period, it becomes necessary to limit the analysis to countries where such information is accessible.

Second, the high share of missing values in vital demographic variables, namely birth and death locations and dates, leads to a substantial reduction of the initial sample size of the data. This limitation is anticipated, as this data source was not primarily designed for population studies. Additionally, the omission of individuals who were alive in 2015, only permits the analysis of extinct birth cohorts. Hence, scholars who want to employ FamiLinx may enhance the robustness of their research by performing a careful sample selection. This selection fosters increased confidence in the completeness and the quality of the chosen kinship network, enabling researchers to conduct population studies with a more solid foundation. Nonetheless, the restriction to individuals with demographic information of higher completeness and quality results in non-negligible selectivity issues. Potential FamiLinx users approaching this dataset should be critical of the information available. Most of these individuals possess missing demographic information, and even when some of their relatives may be identified, the available information may be scarce.

Third, the age-sex distribution of online genealogical populations tends to diverge systematically from that observed in the general population. These observed divergences are a direct consequence of the under-representation of women and of individuals dying at young ages. By overlaying the age-sex distribution derived from the population register to the

genealogy-based one male individuals in older ages are overrepresented, whereas women and individuals in younger age groups are generally underrepresented. Hence, scholars who are interested in examining the evolution of demographic processes in populations originating from FamiLinx are encouraged to implement bias-correcting methods to take into account the representation issues of this data source. The implementation of such methods allows researchers to obtain more accurate measures of common demographic processes, e.g., fertility, mortality and migration. In this regard, a Bayesian modeling framework can enable researchers to calibrate genealogy-based demographic indicators with more accurate estimates originating from more traditional data such as censuses and parish records while accounting for the uncertainty of each source. For instance, Chong et al. 2022 proposed a Bayesian modeling framework to correct age-specific mortality rates by combining online genealogical data with more precise estimates from the Human Mortality Database. Future research could employ a similar modeling framework to examine other demographic processes such as fertility by integrating information from multiple data streams.

Fourth, by using Sweden as a test country, our results suggest that, regardless of the quality of demographic information, individuals from online genealogies are characterized by a persistently higher survival compared to the general population. Hence, researchers intending to harness this data source to gauge demographic outcomes should exercise caution. In general, demographic trajectories observed among individuals with non-missing birth and death years in FamiLinx are not representative of those of the broader population.

Another key consideration concerns the availability of relatives in the dataset and the completeness and quality of demographic information for the entire kinship network. FamiLinx's strength lies in its ability to provide information about relatives, facilitating the identification of kinship networks spanning across multiple generations. Notably, our regression analyses have underlined that completeness and quality are clustered at the family level. In this regard, a careful sample selection would allow researchers to conduct family-level demographic analysis. Specifically, researchers can employ the FamiLinx

database to examine how fertility and longevity spread among different types of relatives beyond parents and children. Nonetheless, while the latter analysis can provide new knowledge about the transmission of demographic behaviors over time, the results should be interpreted with caution. It should be acknowledged that genealogical populations are highly selected under a set of favorable conditions, including higher survival and higher SES. This finding aligns with the existing literature about bias and selectivity in genealogies (Hollingsworth and Hollingsworth, 1976; Zhao, 2001; Calderón Bernal et al., 2023) and in FamiLinx (Stelter and Alburez-Gutierrez, 2022; Minardi et al., 2024).

In conclusion, we encourage researchers to employ the FamiLinx data with caution. This data source displays great opportunities for demographic research, especially in the field of historical demography, due to its rich wealth of demographic information about individuals from various historical populations and its recorded kinship ties. Nonetheless, the inherent limitations of online genealogical data need to be addressed through the implementation of appropriate bias-correcting methods and through a careful sample selection. The findings and implications derived from this study are not automatically applicable to all (online) genealogical datasets. Specifically, the presented investigation is tailored to the unique attributes of the FamiLinx dataset, characterized by its availability of demographic information and linked relatives. It is essential to acknowledge that different datasets may exhibit completely distinct temporal and geographical scopes, affecting the missingness of the data and their representativeness. Nevertheless, we are confident that the methodologies and approaches employed in this study can be replicated for other genealogies, to assess the completeness of their demographic information and explore the association of these concepts within family networks.

2.6 Note on Reproducibility

To facilitate reproducibility of this research, we provide access to FamiLinx data as well as to the R codes needed to reproduce the tables and figures provided in the paper at the following Open Science Framework (OSF) repository: https://osf.io/ydzfq/.

2.7 CRediT authorship contribution statement

Riccardo Omenti: Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Conceptualization.

Andrea Colasurdo: Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Conceptualization.

Chapter 3

Bayesian Indirect Estimation of Historical Fertility in Europe and US using Online Genealogical Data

Riccardo Omenti

Monica Alexander

Nicola Barban

Bayesian Indirect Estimation of Historical Fertility in Europe and US using Online Genealogical Data

Riccardo Omenti[1,†], Monica Alexander[2,3] and 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*

[†] *Corresponding author:* riccardo.omenti2@unibo.it

Abstract

A growing number of social scientists use online genealogical data as an alternative digital census of historical populations for the examination of past demographic dynamics. However, the non-representativeness of this data source requires the development of bias-adjusting methods that would enable to obtain more accurate measures of demographic processes. We address this need by proposing a Bayesian modeling framework and an indirect estimation technique to investigate fertility trends in seven European countries and the United States of America for the historical period 1751-1910, leveraging data from the big genealogical database FamiLinx. The proposed methods estimate the period total fertility rate (TFR) using minimal data, specifically women aged 15-49 and children under age 5, while incorporating information about child mortality and data accuracy from various historical sources such as population registers and censuses. The results indicate that, with our methodological approach, online genealogical data can be fruitfully used to examine fertility patterns in regions and historical periods lacking well-functioning national vital registration systems.

Keywords: · Bayesian Methods · Digital Data · Formal Demography ·

Historical Demography · Indirect Estimation · Fertility

3.1 Introduction

The increasing availability of non-traditional data sources driven by the so-called "Data Revolution" (Cesare et al., 2018; Alburez-Gutierrez et al., 2019; Kashyap, 2021) highlights the need for the development of new statistical methods that are able to identify sources of biases and to apply appropriate corrections. Among these novel data sources, online genealogical data have gathered significant attention due to their unprecedented wealth of historical information about human societies (Alburez-Gutierrez et al., 2022). Online genealogical data are scraped from websites where users upload their own genealogical tree and insert individual-level demographic information about their ancestors, such as gender, birth and death dates and countries. Online genealogies offer an unprecedented opportunity to examine population dynamics in historical periods and countries for which we lack ground-truth population data (Stelter and Alburez-Gutierrez, 2022; Colasurdo and Omenti, 2024). Nonetheless, as these data sources are not primarily designed for population studies, they are affected by several biases that hamper their usability for demographic research (Colasurdo and Omenti, 2024). First, the bottom-up construction of the digital family trees amplifies the likelihood of omitting more distant ancestors. Second, the under-representation of various population subgroups, including women and children who died at an early age, and the over-representation of male individuals with higher socioeconomic status and better demographic conditions lead to a significant selection bias (Hollingsworth and Hollingsworth, 1976; Stelter and Alburez-Gutierrez, 2022; Calderón Bernal et al., 2023; Minardi et al., 2024). Third, an individual's inclusion in the genealogy may be affected by the so-called "selective-remembering", as the genealogist is more likely to include ancestors with a prominent role in his (her) family history (Chong et al., 2022) and to omit relatives who dishonored the family (Zhao, 2001). Fourth, the high percentage of missing values in common demographic variables, specifically birth and death dates and countries, makes only a small share of the profiles from online genealogies usable for demographic research (Minardi et al., 2024; Colasurdo and Omenti, 2024). Additionally, online genealogies are built through a bottom-up approach as users begin the construction from the bottom of their family trees and trace their lineages backwards.

As a consequence, online genealogies tend to underestimate childless individuals.

In this paper, we employ indirect estimation techniques coupled with Bayesian methods to correct fertility estimates that are derived from online genealogical populations by relying on the big genealogical database FamiLinx constructed by Kaplanis et al. (2018). Indirect estimation techniques allows for the measurement of demographic indicators using limited data inputs by borrowing information from multiple data sources. Bayesian methods, albeit more complex and computationally more intensive, not only enable to incorporate multiple data sources but also to carry out the estimation of demographic outcomes within a probabilistic framework and to quantify the uncertainty surrounding the estimates. In particular, we propose to extend the class of indirect fertility estimates introduced by Hauer and Schmertmann (2020) and the Bayesian model developed by Schmertmann and Hauer (2019) to settings, such as data from online genealogies, in which the study sample is not representative of the general population. The proposed methods yield period TFR estimates by relying on minimal input, specifically the observed number of children under age 5 and the observed number of women aged 15-49, while accounting for child mortality and sample over- and under-reporting. Importantly, our approach diverges from traditional methods by eliminating the need for information on the number of births by maternal age. In the context of populations derived from online genealogies, which are characterized by incomplete family trees, the conventional calculation of TFRs based on births disaggregated by maternal ages becomes impractical.

The employment of Bayesian methods for the measurement of demographic processes in populations, where the available data are limited or imperfect, has become increasingly popular. Bayesian methods in demography have been developed to measure migration by combining social media data with more traditional sources (Alexander et al., 2020; Rampazzo et al., 2021), to generate subnational population estimates in data-sparse contexts (Alexander and Alkema, 2022), to reconstruct past populations (Wheldon et al., 2013; Voutilainen et al., 2020), to estimate mortality in historical populations from online genealogical data (Chong et al., 2022).

In this research paper, we harness online genealogical data coupled with more traditional

data sources to estimate period TFR for seven European countries and the United States of America during the historical period 1751-1910 by means of indirect estimation and Bayesian modeling. Previous studies (Hsu et al., 2021; Cozzani et al., 2023; Gay et al., 2023; Pojman et al., 2023; Minardi et al., 2024; Corti et al., 2024; Blanc, 2024a,b), which relied on online genealogical data to study demographic outcomes in Europe and North America, operated under the assumption that these data were representative of the general population. The work by Chong et al. (2022) marked the first attempt to develop a Bayesian modeling framework to calibrate mortality rates from online genealogical data with estimates from more reliable data sources such as the Human Mortality Database. To the best of our knowledge, this research paper represents the first study to propose a method to correct the TFRs derived from online genealogical data. Overall, we believe that our proposed methods could uncover new paths for fertility estimation not only in the context of digital data from online genealogies but also in other data-sparse settings. The remainder of the paper is structured as follows. First, we provide a description of the big genealogical database *FamiLinx*. Second, we describe in detail the Bayesian modeling approach and the indirect estimation method to estimate the TFR from online genealogical data. Third, we apply the proposed methods to examine historical fertility patterns in seven European countries and the United States (US) during the historical period 1751 – 1910.

3.2 Data

3.2.1 The FamiLinx Database

This paper relies primarily on *FamiLinx*, a database derived from publicly available online genealogies on the website geni.com. The database was curated by Kaplanis et al. (2018) that scraped over 86 million profiles from the digital trees on the website geni.com. This big genealogical database contains individual-level records about 86 million individuals and information about kinship ties for approximately 43 million of these profiles. Specifically, the data incorporate micro-level records containing information about essen-

tial demographic variables, namely birth and death dates and countries. Thus, this data source generates a sizable population of individuals with life courses unfolding across multiple centuries and countries.

Despite its massive size, this data set is subject to several limitations that hamper its usability for demographic research. Besides the biases reported in the introduction, this data source significantly over-represent individuals experiencing vital events, i.e., births and deaths, in Western countries (Colasurdo and Omenti, 2024), and displays a large amount of missing values in key demographic variables, i.e., birth and death dates and locations (see table B.2 in the appendix). Furthermore, these data exhibit various reporting errors, which may include improbable ages at death or unreasonable years of birth and death. Hence, before carrying out any demographic analysis, a careful sample selection must be performed (Colasurdo and Omenti, 2024). In addition, this data set is affected by passive registration. While in active registration systems the data collection authority knows the status of the individual at all times, in passive registration systems, such as FamiLinx, only births and deaths are recorded (Colasurdo and Omenti, 2024). Consequently, this data source lacks precise information about other essential life course events, including marriages and migrations[1].

3.2.2 Sample Selection and Representativeness

Our analysis focuses on seven European (Denmark, England, Finland, France, Norway, Sweden and the Netherlands) and the United States of America[2]. We have opted to select these countries for two major reasons. First, they are among the twenty countries with the highest number of recorded births and deaths in FamiLinx. Secondly, some of these countries are characterized by a long-standing tradition of well-functioning national vital registration systems that can be leveraged to inform the bias of demographic estimates

[1]We acknowledge the possibility of estimating the timing of migration and marriages indirectly. For instance, Corti et al. (2024) used FamiLinx to examine trends in assortive mating in the US for birth cohorts from 1700 to 1910, approximating the year of marriage of a couple by the birth year of their first child.

[2]The final countries of birth and death have been established according to the rule proposed by Colasurdo and Omenti (2024). This means that text strings containing information on birth and death locations were deemed more accurate in comparison to the reported latitude and longitude that were inferred by Kaplanis et al. (2018).

calculated from online genealogical populations.

In order to create the country-specific analytical samples of online genealogical populations, we apply the following criteria:

(a) The variables sex, birth and death years and countries must not be missing.

(b) The profile must have at least one parent or one child. This ensures that each individual belongs to a family network of size strictly greater than one.

(c) The birth year must not be greater than 1910.

(d) The earliest death year cannot be less than 1751.

(e) The age at death must fall between 0 and 110.

(f) The countries of birth and death of the profile must be the same.

We do not consider individuals born later than 1910, since Kaplanis et al. (2018) omitted profiles that were still alive as of 2015 for privacy-related concerns. Hence, we can only include individuals from birth cohorts that were almost surely extinct in 2015. The earliest year of analysis is 1751 since it represents the year in which accurate demographic data started to become available for at least one country, that is Sweden, among those included in the analysis. Additionally, the age restriction enables to omit individuals with biologically implausible ages at death. We exclude migrants, defined in this study as individuals with distinct birth and death countries, due to the lack of information about the exact time of migration.

Figure 3.1 displays the age-sex distributions of the genealogy-based populations by country in selected years using population pyramids. The thick black lines indicate the true age-sex distribution when accurate population estimates are available. Overall, figure 3.1 accentuate no clear convergence in the extent to which the age-sex distributions of the genealogical populations resemble the ones obtained from more reliable data sources (see table B.1 for a detailed description). The genealogy-based age-sex distributions for the US, Sweden and France seem to become more representative towards the end of the nineteenth century. On the contrary, in England & Wales the age-sex structure of the genealogical population appears to mirror the true one more closely at the beginning of the study period and to worsen in the nineteenth century. In the other

countries we observe no clear improvement over time in terms of representativeness of online genealogical populations. Nonetheless, a general tendency of genealogical data to under-represent women and children, albeit with distinct magnitudes, is evident across the entire historical period under examination.

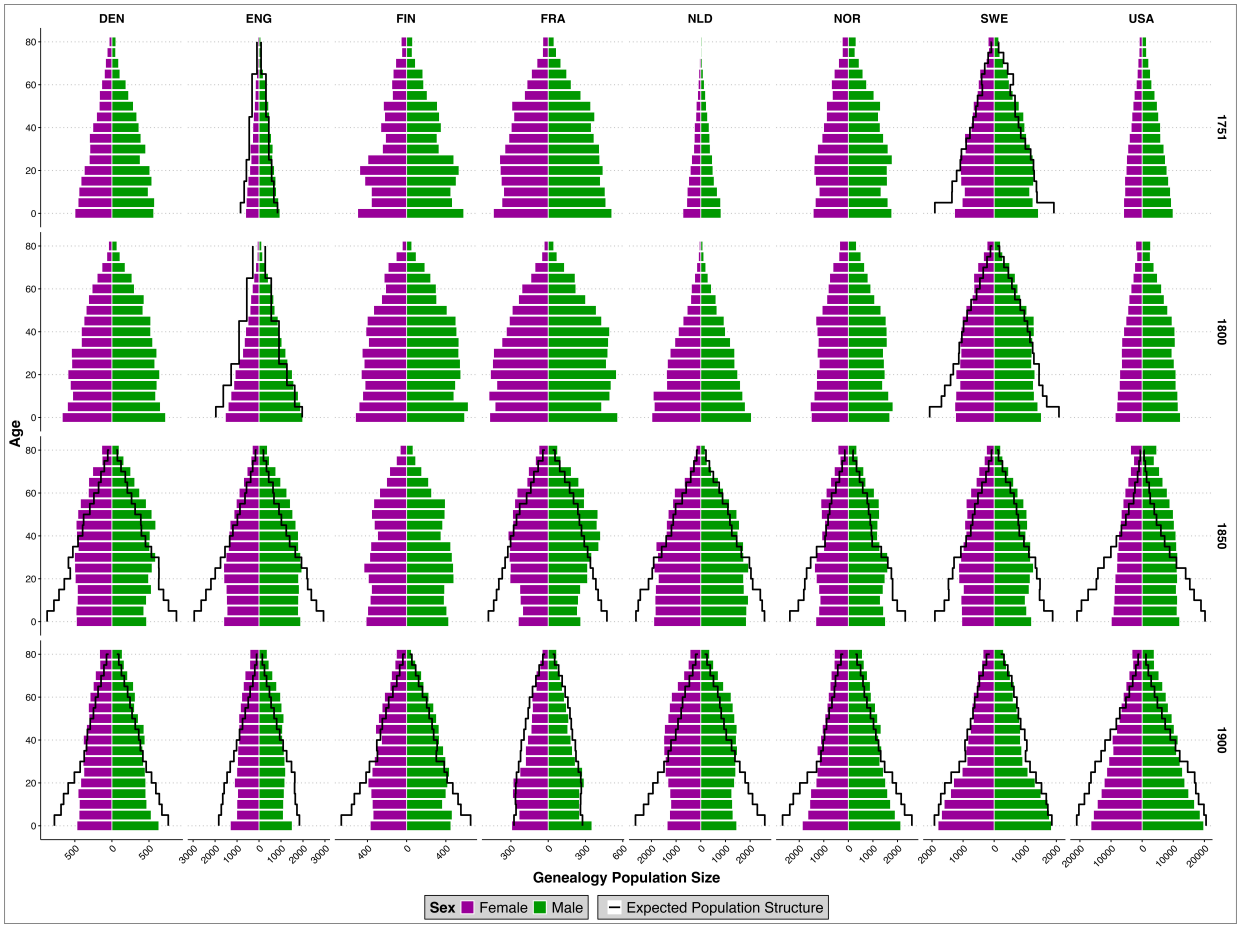


Figure 3.1: Genealogy-based and expected population counts by age and sex for selected calendar years.

3.3 Bayesian Model

The upcoming subsections provide a detailed description of the proposed Bayesian model, which builds upon the framework developed by Schmertmann and Hauer (2019). The original method by Schmertmann and Hauer (2019) is designed to estimate the TFR indirectly without needing the counts of live births classified by maternal ages. Instead, it relies on

a minimal data input: the accurate number of children under 5, the accurate number of women aged 15 – 49, and prior information on child mortality and standard age-specific fertility patterns.

In the original method, the number of children under 5 is modeled using a Poisson distribution, with the number of women aged 15 – 49 as offset and with a mean that depends on parameters with a clear demographic interpretation. Our proposed model extends the original method by modeling the number of children under 5 from online genealogical populations, which are usually not representative of the general population, and the number of children under 5 derived from more reliable data sources when available for seven European countries and the US during the historical period 1751 – 1910.

To account for the non-representativeness of the expected number of children under 5 per woman aged 15-49 in online genealogical populations, a time-varying country-specific bias-adjustment parameter is incorporated. The parameter is modeled hierarchically to allow countries with accurate data to partially inform bias-patterns in countries lacking reliable population data.

The first subsection is devoted to explaining the data model, while the subsequent subsections will provide a detailed explanation of the statistical models and prior distributions for the demographic and bias-adjustment parameters used in the mean of the data model.

3.3.1 Data Model

Drawing inspiration from the modeling framework by Schmertmann and Hauer (2019), we propose a Bayesian hierarchical model to measure the TFR using the number of children under age 5 and the number of women aged 15 – 49, while accounting for infant mortality, for the age-specific fertility schedules and for the non-representativeness of the online genealogical populations.

We model simultaneously the number of children aged 0-4 ($C_{a,t}^{(\text{gen})}$) in the genealogical population of country a in year t and the true number of children ($C_{a,t}^{(true)}$) based on the

true age-sex distribution for country a and year t.[3] We assume that both variables are Poisson-distributed:

$$C_{a,t}^{\text{gen}} | K_{x,a,t}, \tau_{a,t} \sim \mathcal{P}\left(\sum_{x=15}^{45} K_{x,a,t} \cdot W_{x,a,t}^{\text{gen}} \cdot \tau_{a,t} \right) \quad t \in \mathcal{T}_a^{gen}, a \in \mathcal{A} \tag{3.1}$$

$$C_{a,t}^{\text{true}} | K_{x,a,t} \sim \mathcal{P}\left(\sum_{x=15}^{45} K_{x,a,t} \cdot W_{x,a,t}^{\text{true}} \right) \quad t \in \mathcal{T}_a^{true}, a \in \mathcal{A} \tag{3.2}$$

where $W_{x,a,t}^{\text{gen}}$ is the number of women in age group x in country a and year t based on the genealogical sample, while $W_{x,a,t}^{\text{true}}$ denotes the true number of women in the maternal ages based on the true age-sex distribution of country a in year t. $K_{x,t,a}$ denotes the expected number of children aged $0-4$ per woman in the maternal age group x during year t at the end of five-year period. In this paper, we assume that fertility outside the age interval $[15, 50)$ is zero and that maternal ages are split into five-year age groups. The additional term $\tau_{t,a}$ allows to adjust the expected number of children aged $0-4$ per woman aged $15-49$ from the genealogical sample to take into account the non-representativeness of this data source.

$\mathcal{A}$ denotes the set of countries considered for the analysis, while $\mathcal{T}_a^{gen}$ and $\mathcal{T}_a^{true}$ indicate the set of calendar years for which genealogy-based and true population counts by age and sex are available for country a. $\mathcal{T}_a^{gen}$ spans over the temporal period $1751-1910$ for all the countries under analysis. $\mathcal{T}_a^{true}$ does not cover the entire period and varies across countries as official population estimates by age and sex for most of the countries are not available for the entire historical period.

Figure 3.2 displays a graphical summary of the proposed Bayesian modeling framework. An insightful description of the model parameters is provided in the upcoming subsections.

[3] Given the total population size ($P_{a,t}^{(\text{gen})}$) from the genealogical sample in country a during year t, we calculate the "true" number of children under 5 and of exposed women by maternal age by multiplying $P_{a,t}^{(\text{gen})}$ by the true age-sex proportions for country a and year t.

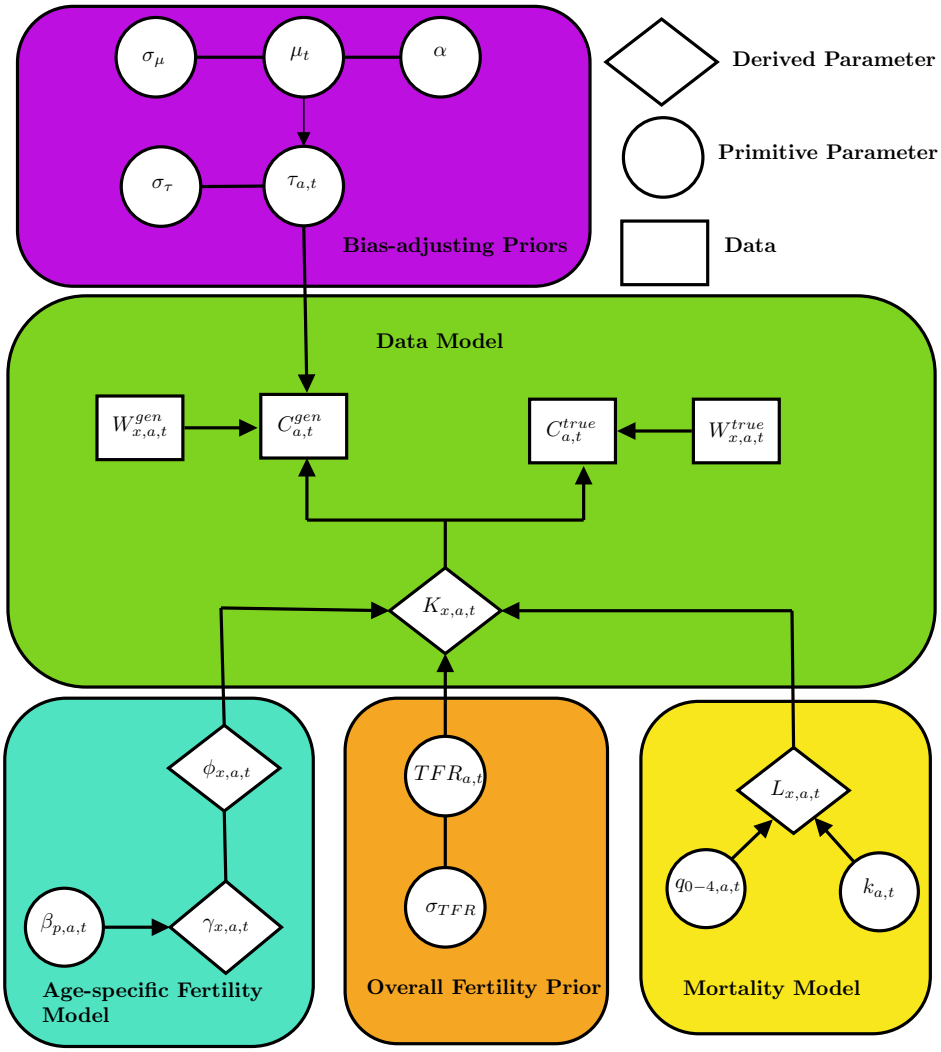


Figure 3.2: Graphical representation of the Bayesian modeling framework. Primitive parameters denote the fundamental parameters in a model that are directly assigned a prior probability distribution. Derived parameters are functions of primitive parameters and do not have prior probability distribution directly assigned to them.

3.3.2 Construction of $K_{x,a,t}$

The term $K_{x,t,a}$ follows a simple rearrangement of the Leslie matrix formulas (Wachter, 2014). However, unlike in standard cohort-component projection methods, this term defines the age of the mother as being attained at the end of the age interval rather than at the beginning.

Following the parametrization of Schmertmann and Hauer (2019), the component $K_{x,t,a}$ is defined as follows.

$$K_{x,a,t} = TFR_{a,t} \cdot \frac{L_{0,a,t}}{5} \cdot \frac{1}{2} \cdot \left[\frac{L_{x-5,a,t}}{L_{x,a,t}} \cdot \phi_{x-5,a,t} + \phi_{x,a,t} \right] \tag{3.3}$$

where $TFR_{a,t}$ denotes the TFR in country a during year t, $L_{x,a,t}$ indicates the person-years lived by women in the age group x during year t in country a, $\phi_{x,a,t}$ denotes the fraction of fertility experienced by women in age group x during year t in country a. Mathematically, $\phi_{x,a,t}$ is equal to $5\frac{F_{x,t,a}}{TFR_{a,t}}$ where $F_{x,t,a}$ is the fertility rate in age group x, year t and country a. The latter quantity is assumed to be zero outside the age interval $[15, 50)$. The implementation of a Bayesian modeling framework allows for the specification of statistical models and prior probability distributions to estimate the parameters in equation 3.3.

3.3.3 Model for Age-specific Fertility

Following Schmertmann and Hauer (2019), to incorporate knowledge about the age-specific fertility patterns, we model the ratio of the share of life time fertility in an age group x to the share of life time fertility in the earliest reproductive age group $15-19$ on the log scale as

$$\boldsymbol{\gamma}_{a,t} = \boldsymbol{m} + \boldsymbol{y}_1\beta_{1,a,t} + \boldsymbol{y}_2\beta_{2,a,t} \tag{3.4}$$

where $\gamma_{x,a,t} = \log\left(\frac{\phi_{x,a,t}}{\phi_{15,a,t}}\right)$ is an index defined as the log transformation of the ratio of the share of life time fertility in age group x to the share of life time fertility in age group $15-19$. $\boldsymbol{\gamma}_{a,t}$ is vector whose elements are $\gamma_{x,a,t}$ defined at the distinct reproductive age

groups[4]. The transformation of $\phi_{x,a,t}$ ensures that the elements of the vector $\boldsymbol{\gamma}_{a,t}$ on the left hand-side of the above equation can assume both positive and negative values[5].

$\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are components derived from a set of standard age-specific fertility curves. In particular, $\boldsymbol{m}$ is a vector containing the age-specific means of the log-transformed fertility schedules ($\gamma_{x,a,t}$), while $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are the first and second left-singular vectors which are obtained via a Singular Value Decomposition on the matrix $\boldsymbol{Y}$ whose columns are log-transformed age-specific fertility schedules. For example, in our application to online genealogical data we employ all the available national age-specific female fertility curves available from the Human Fertility Collection (Grigorieva et al., 2015) covering the historical period 1751 – 1910. The mean and the first two principal components of the log-transformed age-specific fertility schedules are displayed in figure 3.3. The mean $\boldsymbol{m}$

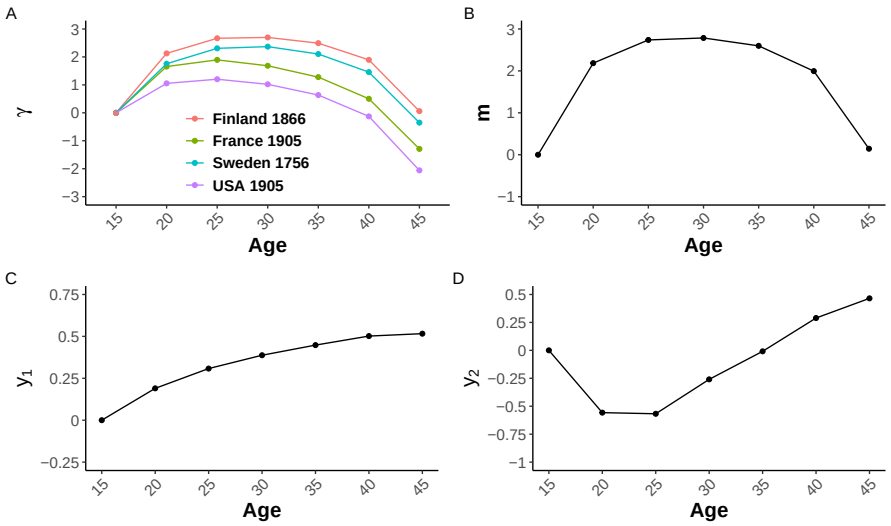


Figure 3.3: Example data (Panel A), mean (Panel B) and principal components of transformed age-specific fertility schedules (Panels C and D).

describes the overall age-specific fertility curve. As expected, age-specific fertility patterns increase up to the age class 25 – 29 and then start to taper off. The first principal component represents the tendency to postpone childbearing. The second principal component allows the modal reproductive age classes 20 – 24 and 25 – 29 to have lower fertility levels compared to the other maternal age groups.

[4] 15 – 19, 20 – 24, 25 – 29, 30 – 34, 35 – 39, 40 – 44 and 45 – 49

[5] Alternatively, a logit transformation could have been applied.

We place standard normal priors on the parameters $\beta_{1,a,t}$ and $\beta_{2,a,t}$.

$$\beta_{1,a,t}, \beta_{2,a,t} \sim \mathcal{N}(0,1) \tag{3.5}$$

From $\gamma_{x,a,t}$, we can easily derive the parameter $\phi_{x,a,t}$ that can be interpreted as the proportion of fertility experienced by women in age class x during year t in country a.

$$\phi_{x,a,t} = \frac{\exp\left(\gamma_{x,a,t}\right)}{\sum_{x=15}^{45} \exp\left(\gamma_{x,a,t}\right)} \tag{3.6}$$

3.3.4 Prior on Total Fertility Rates

We assign the Total Fertility Rate Parameters ($TFR_{a,t}$) a normal distribution centered around the corresponding historical estimate with a non-informative standard deviation parameter. Specifically, if the historical estimate ($\hat{TFR}_{a,t}$) is available for the year t, we use it as a mean. Otherwise, the mean of $TFR_{a,t}$ is set to be equal to the historical estimate closest in time to year t ($\hat{TFR}_{a,t^*}$)[6].

$$\begin{cases} TFR_{a,t} \sim \mathcal{N}(\hat{TFR}_{a,t^*}, \sigma^2_{TFR}) & t < t^* \\ TFR_{a,t} \sim \mathcal{N}(\hat{TFR}_{a,t}, \sigma^2_{TFR}) & t \geq t^* \end{cases} \tag{3.7}$$

where t^* denotes the year closest to year t for which an historical TFR estimate is available for country a. The standard deviation parameter σ_{TFR} is assigned a non-informative prior, namely a half-normal distribution with standard deviation equal to 10.

$$\sigma_{TFR} \sim \mathcal{N}^+(0, 10^2) \tag{3.8}$$

The practical implication of this choice is that the marginal posterior distribution of the TFR is almost entirely be determined by the observed data and by the prior information from the other model parameters governing the data model.

[6] An alternative approach could involve either back-projecting the TFR in the model using a reverse random walk or replacing the earliest available historical estimates with back-extrapolated values.

3.3.5 Model and Priors for Age-specific Mortality

The model for $K_{x,a,t}$ also requires the estimation of the person-years $L_{x,a,t}$ parameters. Hence, building on Schmertmann and Hauer (2019), we model child and adult mortality employing the two-dimensional mortality model by Wilmoth et al. (2012). In this model, the logarithmic transformation of the age-specific mortality risk ($\mu_{x,t,c}$) is a function of two main parameters $q_{0-4,a,t}$ and $\kappa_{a,t}$. $q_{0-4,a,t}$ indicates the probability of dying under age 5, while $\kappa_{a,t}$ is a parameter affecting the shape of the age pattern of mortality. This model can written as follows.

$$\log\left(\mu_{x,a,t}\right) = a_x + b_x \log\left(q_{0-4,a,t}\right) + c_x\Big[\log\left(q_{0-4,a,t}\right)\Big]^2 + d_x\kappa_{a,t} \tag{3.9}$$

a_x, b_x, c_x, d_x are age-specific fixed constants derived from various age-specific mortality schedules in the Human Mortality Database. $\mu_{x,t,c}$ indicates the risk of dying in the age group x at time t in country c. In order to derive the age-specific life table person-years, we employ well-known demographic relationships. The model parameters $q_{0-4,t,c}$ are assigned a Beta distribution as prior.

$$q_{0-4,t,c} \sim \mathcal{B}\Big(a\big(\hat{q}_{0-4,t,c}\big), b\big(\hat{q}_{0-4,t,c}\big)\Big) \tag{3.10}$$

where $a(\hat{q}_{0-4,t,c})$ and $b(\hat{q}_{0-4,t,c})$ are chosen so that $P(0.9\cdot\hat{q}_{0-4,t,c} \leq q_{0-4,t,c} \leq 1.1\cdot\hat{q}_{0-4,t,c}) = 0.9$. This ensures that the infant mortality in country c at time t lies fairly close to its estimated value with 90% probability. In our example, we calculate the two parameters by means of the utilities available in the R package LearnBayes (Albert, 2018).

The parameters $\kappa_{t,c}$ are assigned a standard normal prior.

$$\kappa_{t,c} \sim \mathcal{N}(0,1) \tag{3.11}$$

By relying on standard relationships of life table quantities, we are able to derive the person-years. Specifically, we know that the survival column from an abridged life table

$l_{x,a,t}$ can be written as a function of $\mu_{x,a,t}$.

$$l_{x,a,t} = \begin{cases} 1 & \text{if} \quad x = 0 \\ e^{-\mu_{0,a,t}} & \text{if} \quad x = 1 \\ l_{1,a,t} \cdot e^{-4\mu_{1,a,t}} & \text{if} \quad x = 5 \\ l_{x-5,a,t} \cdot e^{-5\mu_{x-5,a,t}} & \forall x > 5 \end{cases} \tag{3.12}$$

Similarly, the life table person-years can be determined as follow.

$$L_{x,a,t} = \begin{cases} \frac{1}{2} \cdot (l_{0,a,t} + l_{1,a,t}) + \frac{4}{2} \cdot (l_{1,a,t} + l_{5,a,t}) & x = 0 \\ \frac{5}{2} \cdot (l_{x,a,t} + l_{x+5,a,t}) & \forall x \geq 5 \end{cases} \tag{3.13}$$

A detailed discussion on how we estimate child mortality ($q_{0-4,a,t}$) in the countries under analysis can be found in the appendix in section B.6.

3.3.6 Bias-adjustment Priors

The modeling framework by Schmertmann and Hauer (2019) relies on the assumption that the number of children and women is representative of the true population of interest. Nonetheless, this assumption is violated from online genealogical populations, whose age-sex distribution generally diverge from that of the true population. For this reason, we introduce the parameter $\tau_{a,t}$ that governs the extent to which the expected number of children per woman aged $15-49$ during year t in country a is under- or over-estimated. This parameter, positive by construction, is modeled on a log-scale and is assumed to be generated from a normal distribution centered around a common time-varying parameter ν_t, which can be interpreted as a transnational mean and allows to share information across countries. In our example, in absence of ground truth population estimates for a certain country in a given year t, its patterns of non-representativeness will be informed by those of countries with accurate population data during the same year t.

$$\log(\tau_{a,t}) \sim \mathcal{N}(\nu_t, \sigma_\tau^2) \tag{3.14}$$

In addition, to ensure that $\tau_{a,t}$ varies in a relatively regular pattern over time, we impose a first-order autoregressive model (AR(1)) on ν_t[7].

$$\nu_t \sim \mathcal{N}(\rho \cdot \nu_{t-1}, \sigma_\nu^2) \tag{3.15}$$

The parameter ρ is assigned a uniform distribution on $(0, 1)$ to ensure stationarity (Gelman et al., 1995).

$$\rho \sim \mathcal{U}(0, 1) \tag{3.16}$$

The standard deviation parameters are assigned half-Normal weakly informative priors.

$$\sigma_\tau \sim \mathcal{N}^+(0, 1^2) \tag{3.17}$$

$$\sigma_\nu \sim \mathcal{N}^+(0, 1^2) \tag{3.18}$$

3.3.7 Model Implementation

The model was fitted using the R statistical package NIMBLE (de Valpine et al., 2017). The latter enables to specify the main structure of the model in R, compiles it in $C++$, and employs a Metropolis-within-Gibbs Markov Chain Monte Carlo (MCMC) algorithm to sample from the posterior distribution of the model's parameters.

Since our ultimate objective is the estimation of the period TFR, we focused on the posterior distribution of the parameter TFR and utilized its median as best estimate. In order to quantify the uncertainty around this estimate, we built 95% credible intervals by computing the 2.5% and 97.5% quantiles of the posterior distribution of the parameter TFR. For the sake of simplicity, we denote the median of the posterior distribution of the TFR parameter calculated from our proposed model by $bTFR^*$. We indicate the same estimate obtained using the Bayesian model by Schmertmann and Hauer (2019), which does not include any bias-adjustment process, with $bTFR$. Convergence was assessed

[7]We carried out sensitivity checks where $\tau_{a,t}$ was modeled via a random walk of order 1 (RW(1)) and a random walk of order 2 (RW(2)). TFR estimates were remarkably similar. The results are reported in figure B.3 in the appendix.

visually via trace plots (Gelman et al., 1995) and numerically via the potential scale reduction factor.

3.3.8 Validation of the Proposed Methods

We evaluate the performance of our proposed methods against competitors by means of the root mean squared error (RMSE) between the proposed estimates and the ground truth.

$$RMSE = \sqrt{\frac{\sum_{t \in \mathcal{T}_a^{true}} (\hat{TFR}_{a,t} - TFR_{a,t}^{true})^2}{|\mathcal{T}_a^{true}|}} \tag{3.19}$$

where $\hat{TFR}_{a,t}$ is the estimated TFR for country a during year t according to a certain method, $TFR_{a,t}^{true}$ indicates the TFR value computed from more trustworthy data sources[8] for a country a in year t, $\mathcal{T}_a^{true}$ is the set of years for which accurate historical values for the TFR are available for country a and $|\mathcal{T}_a^{true}|$ denotes the number of years in the set $\mathcal{T}_a^{true}$.

3.4 Indirect Estimation

The employment of a Bayesian modeling framework allows for the estimation of TFR from imperfect data by integrating multiple data sources while accounting for their uncertainty. The main drawback of Bayesian methods is their computational complexity. For this reason, using indirect estimation, we show how the class of TFR indicators developed by Hauer and Schmertmann (2020) can be extended to account for the biases in the counts of women and children in online genealogical populations. The idea behind the construction of these demographic indicators is similar to the Bayesian one. First, they only require the number of children under age 5 and of women aged 15-49, eliminating

[8]We acknowledge that the trustworthiness of the data sources used to produce the 'true' historical TFRs varies significantly by country. The majority of the TFR historical estimates were taken either from the Human Fertility Collection (mostly for the Scandinavian countries) or from previous studies in historical demography. Additionally, these sources do not report any measure of uncertainty around the TFR estimates, making it challenging to assess the exact accuracy of their estimates.

the need for knowledge of the number of births classified by maternal ages. Second, they allow to incorporate information about child mortality, age-specific fertility patterns and under-/over-reporting of children under age 5 and of women aged 15-49.

In contrast to the Bayesian setting, where demographic parameters are derived from probability distributions and statistical models, in indirect estimation, the values of these parameters are calculated deterministically from various data sources and are directly incorporated into the demographic indicators in the form of multipliers. Consequently, the employment of indirect estimation offers a more straightforward approach by allowing for the calculation of the TFR estimates without relying on complex statistical models. Nevertheless, it lacks the capability to incorporate the uncertainty surrounding the estimation of each demographic quantity.

3.4.1 Method description

By rearranging equation 3.3, Hauer and Schmertmann (2020) proposed to decompose the TFR as a product a three major factors.

$$TFR = \frac{1}{p_{a,t}} \cdot \frac{1}{s_{a,t}} \cdot \frac{C_{a,t}^{(gen)}}{W_{a,t}^{(gen)}} \tag{3.20}$$

The previous equation states that the TFR calculated for a country a in year t can be factorized into three major components, which includes the ratio of children aged 0-4 ($C_{a,t}^{(gen)}$) to the number of women aged 15-49 ($W_{a,t}^{(gen)}$), a multiplier for the child survival $\left(\frac{1}{s_{a,t}}\right)$, a multiplier for the age distribution of mothers at childbearing $\left(\frac{1}{p_{a,t}}\right)$. In this setting, $\left(\frac{1}{s_{a,t}}\right)$ and $\left(\frac{1}{p_{a,t}}\right)$ are treated as numerical constants derived from different data sources, while in the Bayesian modeling framework they are generated from statistical models and probabilistic distributions. Hauer and Schmertmann (2020) have set $\left(\frac{1}{s_{a,t}}\right)$ to be equal to $\left(\frac{1}{1-0.75 \cdot q_{0-4,a,t}}\right)$. By changing the approximation for $\left(\frac{1}{p_{a,t}}\right)$, Hauer and Schmertmann (2020) have identified two major classes of indicators. The first set of measures is called implied total fertility rate ($iTFR$, $iTFR^{+}$) with $\left(\frac{1}{p_{a,t}} \approx 7\right)$ and assumes that fertility levels are constant across the maternal age classes. The second measure is

named extended total fertility rate ($xTFR$, $xTFR^{+}$) allows $\left(\frac{1}{p_{a,t}}\right)$ to be different from 7 and to depend on the proportion of women aged $25-34$ ($\pi_{2534,a,t}$)in the age pyramid, specifically $\left(\frac{1}{p_{a,t}}\right)$ is approximated by $10.65-12.55\pi_{2534,a,t}$.

In general, the foundational assumption underlying equation 3.20 is that the number of children under the age of 5, as observed in an age pyramid for country a in year t, following adjustments for mortality $\left(\frac{1}{s_{a,t}}\right)$ and fertility age patterns, can serve as a reliable proxy for recent births to women aged $15-49$ within the same age pyramid. Additionally, this approach hinges on the assumption that the child-woman ratio (CWR) calculated from the age pyramid accurately reflects the ratio observed in the general population.

Nonetheless, this assumption encounters challenges when applied to populations sourced from online genealogies. Such populations are prone to various biases that compromise their representativeness. Consequently, we propose an extension to the decomposition by Hauer and Schmertmann (2020). This extension introduces a bias multiplier ($r_{a,t}$) designed to address the non-representativeness of the genealogy-based CWRs. The inclusion of this multiplier aims to refine the estimation of fertility rates when counts of children under 5 and of women aged $15-49$ from the population pyramids are biased.

Hence, by including the bias-adjustment multiplier in equation 3.20, we obtain the following expression.

$$TFR_{a,t} = r_{a,t} \cdot \frac{1}{p_{a,t}} \cdot \frac{1}{s_{a,t}} \cdot \frac{C_{a,t}^{(\text{gen})}}{W_{a,t}^{(\text{gen})}} \tag{3.21}$$

$$r_{a,t} = \begin{cases} \frac{\text{True CWR in country a}}{\text{Genealogical CWR in country a}} & \text{if} \quad a \in \mathcal{T}_a^{\text{true}} \\ \frac{\text{True CWR in country } a^*}{\text{Genealogical CWR in country } a^*} & \text{if} \quad a \notin \mathcal{T}_a^{\text{true}} \end{cases} \tag{3.22}$$

In order to mimic the borrowing of information across countries from our proposed Bayesian modeling framework, we assume that the multiplier $r_{a,t}$ is calculated from the ratio of the true CWR to the genealogical CWR when ground-truth population estimates are available for country a and year t. If this is not the case, we set the value of this multiplier to be equal to the one of another country a^* for which ground-truth population

counts are available. In our example, we borrow such information from Sweden due to its availability of high-quality demographic estimates for the entire period 1751 – 1910. By applying the previous correction factor to the decomposition proposed by Hauer and Schmertmann (2020), we proposed two new indicators for the classes of implied total fertility rates ($iTFR^*$) and extend total fertility rates ($xTFR^*$). In general, the multiplier $r_{a,t}$ is fairly similar to the parameter $\tau_{a,t}$ in that it denotes the extent to which we are under-/over-estimating the number of children per woman aged 15−49 in country a during year t. Nevertheless, $r_{a,t}$ corrects the genealogy-based CWR based on the bias patterns observed in a country with more reliable population estimates, whereas $\tau_{a,t}$ calibrate the genealogy-based CWR by borrowing information from multiple countries.

One of the key advantages of the proposed Bayesian method over indirect estimation is its ability to model the bias-adjustment multiplier more flexibly by imposing a hierarchical structure. On the contrary, the proposed indirect adjustment is simpler and straightforward but also results in a more rigid correction in the final TFR indirect estimates.

$$iTFR^* = r_{a,t} \cdot 7 \cdot \left(\frac{1}{1 - 0.75 \cdot q_{0-4,a,t}} \right) \cdot \frac{C_{a,t}^{(\text{gen})}}{W_{a,t}^{(\text{gen})}} \tag{3.23}$$

$$xTFR^* = r_{a,t} \cdot (10.65 - 12.55\pi_{2534,a,t}) \cdot \left(\frac{1}{1 - 0.75 \cdot q_{0-4,a,t}} \right) \cdot \frac{C_{a,t}^{(\text{gen})}}{W_{a,t}^{(\text{gen})}} \tag{3.24}$$

3.5 Results

We illustrate the results of the proposed model-based estimates and indirect indicators in the eight countries of interest for the historical period 1751 – 1910. Although our analysis is limited to North America and Europe, the countries under analysis exhibit considerable heterogeneity in their fertility patterns. France stands as a notable example of a country experiencing an exceptionally early fertility transition, which dates back to the beginning of the eighteenth century, roughly seventy years earlier than other nations under examination(Weir, 1984; Wrigley, 1985). United States of America and England & Wales initiated their fertility decline around 1860s and 1870s, while the other countries

did not begin their transition before 1880s (Lee, 2002).

Figure 3.4 illustrates the historical estimates of TFR by country from studies in historical demography or other reliable data sources listed in table B.1 (red asterisk), the fertility estimates based on the Bayesian model by Schmertmann and Hauer (2019), which does not account for the non-representativeness of online genealogical populations (purple squares), the estimates derived from our proposed model (green triangles). A detailed inspection of the plot underlines that our proposed Bayesian model provides TFR estimates that are closer to the historical TFR estimates. If we consider the TFR estimates obtained without any bias-adjustment, we note that they are generally downward biased as they do not account for the underestimation of children under age 5 and women aged 15 – 49 in online genealogies [9]. As illustrated in table B.1, by comparing the performance of our method against the one by Schmertmann and Hauer (2019), we observe a wide reduction in the MSE.

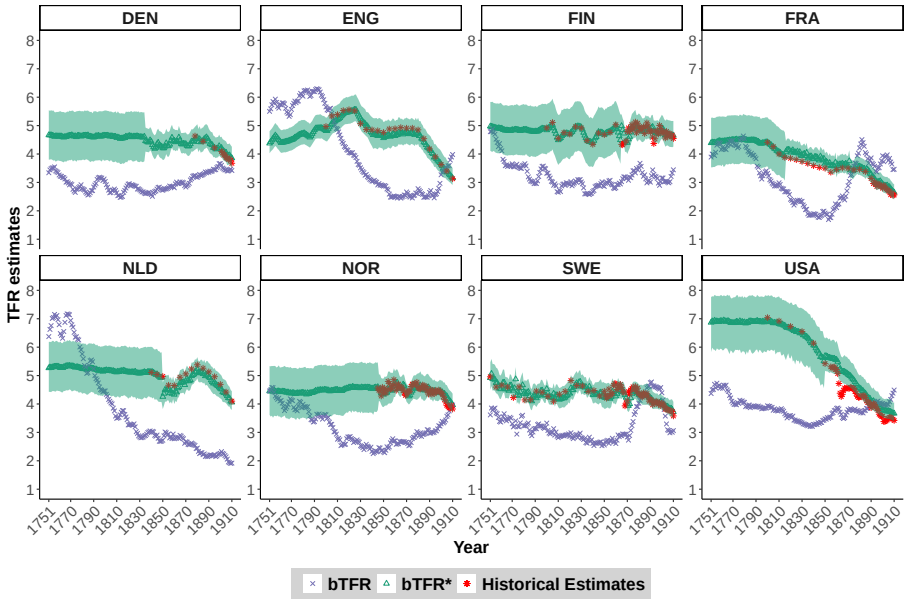


Figure 3.4: Model-based and historical TFR estimates for eight countries during the period 1751-1910. Shaded areas denote 95% credible intervals. Model-based estimates refer to the TFR medians from the corresponding posterior samples. $bTFR$ refers to median posterior estimates from the original model by Schmertmann and Hauer (2019), which does not accout for biases in population structures from FamiLinx. $bTFR^*$ indicates the median posterior estimates from our proposed model, which accounts for the non-representativeness of FamiLinx populations through the inclusion of bias-adjustment parameters.

[9] In the original model by Schmertmann and Hauer (2019), the TFR was assigned an informative uniform prior ($Unif(0, 20)$). When we fit the model, we opted to place on the TFR parameter the same prior that was used in our proposed model.

This result suggests that our proposed model effectively adjusts the biases in online genealogical populations to estimate fertility patterns more accurately. In addition, we observe a higher uncertainty around the TFR estimates in countries and historical period for which more trustworthy population data by age and sex are lacking. Nonetheless, when accurate data on children under 5 and of women aged 15−49 are available, the uncertainty around the TFR estimates decreases and the model estimates mirror the historical TFR values even more closely.

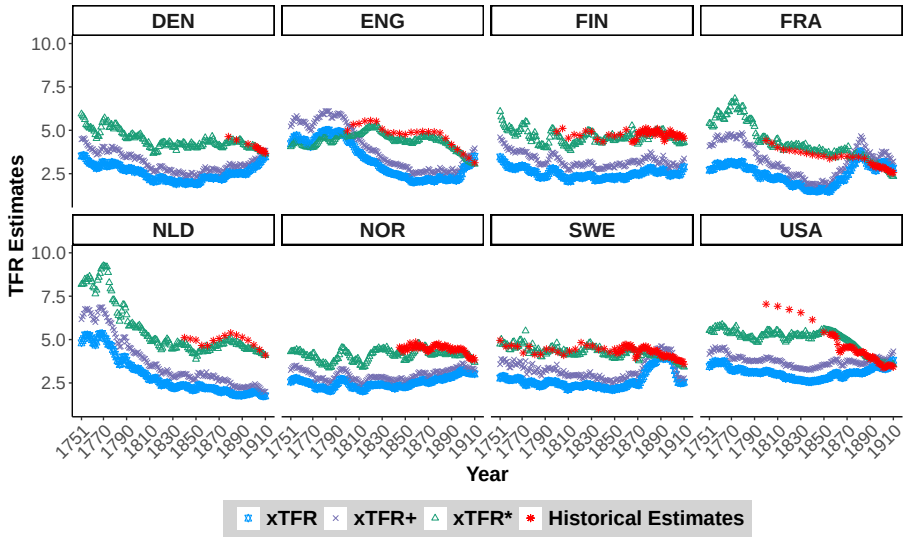


Figure 3.5: iTFR estimates and historical TFR estimates for eight countries during the period 1751-1910. $iTFR$ refers to the simplest indicator from the decomposition by Hauer and Schmertmann (2020), which does not account for child mortality and for the non-representativeness of online genealogical data. $iTFR^{+}$ is an extended version of the indicator $iTFR$ that adjusts for child mortality. $iTFR^{*}$ further refines the indicator $iTFR$ by accounting not only for both child mortality and the non-representativeness of online genealogical data.

Figure 3.5 depicts the historical estimates of TFR by country where data are available (red asterisk), the un-adjusted estimates $iTFR$ (blue stars), the mortality-adjusted estimates (purple x symbol) and our proposed estimate $iTFR^{*}$ (green triangles). In agreement with the Bayesian framework, our proposed estimate exhibits superior performance relative to both the un-adjusted and mortality-adjusted estimates. Similar results are found for $xTFR^{*}$ (see figure B.2 in the appendix). This finding underlines the necessity of including a multiplier which accounts for the bias in the CWR when applying the indirect estimation method by Hauer and Schmertmann (2020) to defective data.

Table 3.1: Performance of the different TFR estimation methods using the RMSE as a metric.

Country	bTFR*	iTFR*	xTFR*	$bTFR$	$iTFR^{+}$	$xTFR^{+}$	$iTFR$	$xTFR$
DEN	0.11	0.12	0.12	0.67	0.83	0.82	1.15	1.14
ENG	0.15	0.31	0.38	1.67	1.63	1.67	2.12	2.16
FIN	0.23	0.34	0.39	1.64	1.58	1.62	2.21	2.25
FRA	0.14	0.19	0.19	1.10	1.01	0.99	1.17	1.16
NLD	0.25	0.38	0.45	2.41	2.31	2.36	2.86	2.90
NOR	0.14	0.30	0.36	1.67	1.37	1.41	1.76	1.79
SWE	0.17	0.26	0.27	1.21	1.15	1.17	1.61	1.63
USA	0.36	0.51	0.50	1.18	1.15	1.17	1.63	1.65

Table 3.1 illustrates the performance of the proposed Bayesian model ($bTFR^{*}$) in comparison with the other TFR estimation methods using the RMSE as metric. We observe that our proposed Bayesian model outperforms all the other estimators in all countries. In general, adjusting for non-representativeness of online genealogical data improves the accuracy of the fertility estimates in both the Bayesian modeling and indirect estimation setting. In both estimation settings, the US exhibit the lowest improvement in the RMSE. While the other countries display a more substantial wealth of information on fertility and mortality levels during the historical period $1751-1910$, reliable national demographic data for the US tend to be more scant. As pointed out by Hacker and Roberts (2019), measuring fertility in the US during the previous historical period is inherently challenging as a national birth registration system was not completed until 1933. In addition, the majority of the national demographic estimates for the US are only available for the white population. In our example, we believe that this limitation is mitigated since the majority of the genealogical US population is most likely white, given the high share of individuals with ancestors from Europe in FamiLinx (Kaplanis et al., 2018; Pojman et al., 2023). The US TFR estimates ($bTFR^{*}$) from the proposed model seems to mirror the historical TFR values by Coale and Zelnik (1961) quite closely before 1850 and after 1870. Conversely, in the period between 1850 and 1870 our proposed method seems to slightly overcorrect the TFR. This finding likely arises from the unavailability of yearly US demographic estimates by age and sex, which does not allow us to properly capture the temporary TFR decline following the American Civil War during the period 1861-1865. While studying the timing of the fertility transition is outside the scope of this paper, all

countries display trends that generally agree with previous studies in historical demography. We can note that France has established itself as the front runner in fertility decline (Pison, 2012). In agreement with Jaadla et al. (2020), England & Wales experienced an increase in the TFR at the beginning of the eighteenth century and did not start their secular decline until 1880s, whereas the United States of America entered their fertility transition after the Civil War (1861–65) (Hacker and Roberts, 2019). Regarding the other countries, they do not experience any decline in fertility before 1880s with Finland notably maintaining the highest fertility levels by the end of the study period. In addition, the time series of $bTFR^{\star}$ estimates display the ability of our proposed method to capture exogenous shocks such as famines that caused temporary declines in overall fertility levels. For instance, temporary declines affected Sweden and Finland in 1868-1970 due to the last major Northern-European famine, while the Netherlands suffered from the so-called 'Potato Blight' during the historical period 1846-1847 which caused the loss of roughly 70% of the potato crop (Bergman, 1967) and an unprecedented reduction in overall fertility.

3.6 Discussion

The spread of technology has furnished population scientists with an unprecedented wealth of data sources, significantly enriching the demographic research landscape (Kashyap, 2021). This surge has led to a growing body of literature in demographic research relying on digital data. Although the majority of studies in 'digital' demography have focused on contemporary populations, we believe that historical demography stands in a unique position to benefit from these novel data streams. The digitization of traditional data sources such as parish records and censuses, alongside the development of platforms where users from different parts of the world can share their family history, offer much promise for the examination of demographic processes in the past at a more global scale. In this paper our focus is on data derived from digital genealogical trees, generated by a transnational network of genealogy enthusiasts dedicated to reconstruct their family history. As they contain demographic information about individuals

whose life courses unfolded in the last 400 years across different countries, these repositories provide an unprecedented opportunity to study population dynamics in the past. Nonetheless, as pointed out by Colasurdo and Omenti (2024), this data source presents several pitfalls, which hamper its employment in demographic research, including the under-representation of various subgroups such as women and children as well as issues related to the accuracy of the reported demographic information.

In response to these challenges, this paper proposes a methodological framework to obtain accurate TFR estimates from data that are inherently defective. Our proposed methods combine data from online genealogical populations with more traditional data sources in order to obtain TFR estimates in various historical populations. Specifically, we added a bias-adjustment process to the modeling framework developed by Schmertmann and Hauer (2019) to incorporate information about the extent to which the number of children under 5 per woman aged $15-49$ is under- or over-estimated in online genealogical data. In parallel, we extended the decomposition by Hauer and Schmertmann (2020) by incorporating a multiplier that allows for the correction of the CWR derived from the online genealogical data and for the exchange of information across countries within an indirect estimation setting.

The results suggest that the proposed adjustments, both within the Bayesian and indirect estimation frameworks, yield fairly plausible TFR estimates. While both methods enhance noticeable gains in the accuracy of the fertility estimates, the Bayesian model generally outperforms the indirect estimation approach in the majority of countries. Although the employment of indirect estimation is appealing due its straight-forward implementation, it does not allow to account for the uncertainty coming from the different data sources. Conversely, the Bayesian approach is favored as it allows to quantify the uncertainty surrounding the fertility estimates, despite being computationally more intensive.

However, this research paper is not free from limitations. First, by pooling information across countries we assume that biases observed in countries with accurate population data are similar to those observed in countries that lack such estimates. We acknowledge that this may not be the case as the extent to which population subgroups, stratified by age

and sex, in online genealogies are under- or over-represented not only varies across time but also across countries. Second, the assumption of constant child mortality in countries and historical periods lacking accurate estimates may be unrealistic. As demographic theory suggests, in historical pre-transitional populations child mortality was both high and subjected to significant fluctuations in response to exogenous factors such as pandemics and famines (Pozzi and Fariñas, 2015). Third, the lack of variables such as education and socio-economic status limits our ability to examine heterogeneity in fertility behaviors across population subgroups. Fourth, the absence of information about migration events in FamiLinx prevents us from investigating differences in fertility patterns between native and migrant women. Fourth, the anonymity of the records in *FamiLinx* does not allow for the implementation of statistical matching techniques to link these data with other more informative micro-level data sources such as censuses. Last but not least, another non-negligible limitation is the inability of the proposed method to produce fertility measures other than the TFR. The capability to estimate other fertility indicators, such as the cohort fertility rate (CFR) and the mean age at childbearing (MAB) would certainly allow for a more detailed overview of the fertility dynamics of the historical populations included in the analysis.

While our paper has employed digital data from online genealogies to examine fertility levels in past historical populations, we believe that our methodological framework holds much promise for applications in both historical and contemporary data-sparse settings such as countries lacking well-functioning civil registration systems, where non-conventional data sources could be leveraged for fertility measurement. For instance, our method could be employed to examine fertility patterns using other digital trace data such as those from Facebook's Advertising Platform that provides counts of Facebook users stratified by relevant characteristics such as age, sex and location.

To summarize, we have proposed novel extensions to existing indirect estimation techniques by Hauer and Schmertmann (2020) and to the Bayesian modeling framework by Schmertmann and Hauer (2019) to gauge fertility levels in eight countries during the historical period 1751 – 1910. Both methods allow for the combination of online genealog-

ical data with other historical data sources, whose information was included by means of multipliers in the case of indirect estimation, or through the incorporation of a statistical models and priors within the Bayesian modeling framework. To conclude, this research paper sheds new light on the potential of Bayesian and indirect estimation methods to investigate the fertility patterns in countries and historical periods with imperfect demographic data.

Chapter 4

A Bayesian Model to Estimate Male and Female Fertility Patterns at a Subnational Level

Riccardo Omenti

Monica Alexander

Nicola Barban

A Bayesian Model to Estimate Male and Female Fertility Patterns at a Subnational Level

Riccardo Omenti[1,†], Monica Alexander[2,3] and 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*

[†] *Corresponding author:* riccardo.omenti2@unibo.it

Abstract

Accurate subnational fertility estimates are crucial for shaping policy decisions across diverse sectors, including education, health care, and social welfare. However, these estimates are difficult to obtain in small populations, in which data on births classified by maternal and paternal ages may be lacking or inadequate. In this paper, we describe a Bayesian model tailored to estimate the period total fertility rates (TFR) for both men and women at a subnational level. Building on previous work by Schmertmann and Hauer (2019), the model utilizes population counts from age-sex population pyramids and models age-specific mortality and fertility patterns accounting for uncertainty and allowing for spatial and temporal dependencies. Testing the model with simulated data that mimic Australian regions, as well as with real data from US counties, demonstrates its ability to generate reasonable TFR estimates. This model shows promise for analyzing male and female fertility patterns across various subregions and time periods.

Keywords: · Bayesian Demography · Male Fertility · Small Area Estimation · Spatial Data · Indirect Estimation

4.1 Introduction

Accurate subnational fertility estimates represent an essential tool for analyzing fertility patterns within a country. Reliable subnational fertility estimates help researchers to identify compositional and contextual factors influencing fertility behaviors at a local level.

We emphasize the importance of examining fertility not only among women but also among men at a subnational level. While female fertility has been well-documented globally, male fertility tends to be neglected due to a significant lack of high quality information (Coleman, 1995). This bias mirrors the focus on female fertility in data collection efforts. Collecting data on the fertility behavior of women is comparatively easier due to a more precise definition of the childbearing age interval and the superior information quality in surveys (Greene and Biddlecom, 2000). Nonetheless, confining fertility studies exclusively to women leads researchers to neglect the distinctive aspects associated with fertility behaviors among men (Schoumaker, 2019).

In addition, estimating male fertility is crucial for matrix-oriented kinship models developed by Caswell and Song (2021). This approach requires time-varying demographic rates, namely age- and sex-specific fertility and survival rates, for the calculation of the expected number of different types of kin implied by such rates. Due to unavailability of male fertility estimates for multiple countries, the global kinship projections developed by Alburez-Gutierrez et al. (2022) assumed that the male age-specific fertility rates mirrored those of females. Incorporating accurate male fertility estimates into these models could improve the accuracy of the kinship projections by Alburez-Gutierrez et al. (2022).

Paget and Timæus (1994); Ratcliffe et al. (2000); Zhang (2010); Keilman et al. (2014); Dudel and Klüsener (2016); Schoumaker (2017); Schmertmann and Hauer (2019); Dudel et al. (2023) have documented substantial differences between male and female fertility. In general, the age-specific fertility curves of men have been found to be somewhat similar to that of women. However, unlike women, men typically have a broader reproductive age span, with lower age-specific fertility at younger ages and higher levels at older ages (Schoumaker, 2019). In addition, non-negligible disparities in sex-specific TFRs have

been found (Ratcliffe et al., 2000; Zhang, 2010; Dudel and Klüsener, 2016; Schoumaker, 2017, 2019). In low-fertility settings, fertility levels among men and women tend to be similar, with female fertility being slightly higher than male fertility (Zhang, 2010; Dudel and Klüsener, 2016)[1]. Conversely, in high-fertility settings, especially in polygynous societies[2], male fertility have been shown to be disproportionately higher than female fertility (Tragaki and Bagavos, 2014; Schoumaker, 2017). In the context of developing countries, Schoumaker (2019) found that higher differences between male and female fertility are associated with a higher prevalence of polygynous unions and with larger disparities between age at motherhood and age at fatherhood. In addition, sex ratio imbalances have been found to be major drivers of the observed differences between male and female fertility. For instance, Dudel and Klüsener (2016) found that male fertility was much lower than female fertility in eastern Germany during the 1990s, due to a high proportion of female out-migrants after the fall of the Soviet Union. Similarly, Coleman (1995) documented substantially higher male fertility in France during the 1920s due to the shortage of men following the World War I.

In general, while variations in male and female fertility have been documented at a national level, less attention has been devoted to the study of male and female fertility at a subnational level. One of the major challenges in producing subnational fertility estimates can be attributed to small populations in which variations in birth counts tend to be fairly high. Furthermore, detailed information on births classified by parental ages, especially fathers, is often lacking for subnational populations due to various reasons: data confidentiality concerns, unavailability of paternal information in birth records, low coverage of certain areas within a country, lack of well-functioning vital registration systems, lack of nationally-representative surveys.

Building on the methodological framework by Schmertmann and Hauer (2019), this research paper proposes a Bayesian model for estimating the period TFR across multiple

[1] Dudel and Klüsener (2021) explained the slightly higher fertility levels observed among women in high-income countries with the so-called "birth squeeze" hypothesis. In many couples from high-income countries, the male partner tends to be older than the female partner, who comes from smaller birth cohorts. Hence, this move from bigger to smaller cohorts puts males at a disadvantage.

[2] Polygyny refers to a form of polygamy entailing the marriage of a man to several women.

geographical areas without the knowledge of the number of births classified by parental ages. The proposed model allows for the estimation of the period TFR using minimal input data, specifically the number of children aged $0-4$, the number of women in the reproductive age interval $15-49$ and the number of men aged $15-59$. Unlike the original model by Schmertmann and Hauer (2019), which focuses solely on the estimation of female TFRs, our objective is to expand this model in two ways: first, by enabling the estimation of male fertility, and second, by accounting for temporal and spatial dependencies among adjacent areas and consecutive years.

To the best of our knowledge, this research paper represents one of the first attempts to develop a Bayesian model for estimating male fertility at a subnational level. Our proposed model relies on a Bayesian hierarchical structure, which allows for the incorporation of prior information about subnational mortality schedules and standard age-specific fertility patterns. In addition, it allows to construct credible intervals for the TFR estimates, helping to understand fertility patterns in smaller geographical areas where uncertainty is higher.

The following section will briefly describe existing methods for the estimation of male fertility from traditional data sources, such as surveys and birth registers, and provide a brief overview on the use of Bayesian methods for demographic estimation. We then provide a detailed description of the proposed model followed by a model-based simulation using Australian subnational fertility data. Finally, following a brief outline of the data sources, we present an application of the proposed model to population data from US counties over the period 1982-2019.

4.2 Background

4.2.1 Existing Methods for Male Fertility Estimation

Previous research on the estimation of male fertility indicators has relied mostly on either nationally-representative surveys or birth registers. For instance, Schoumaker (2017) proposed a revised version of the own-children method to compute male fertility indicators

in developing countries from the Demographic and Health Surveys (DHS). Focusing on countries with well-functioning registration systems, Dudel and Klüsener (2016) employed standard demographic techniques to estimate male fertility from birth register data. To account for births with missing ages at fatherhood, both methods used imputation techniques.

These methods have shown to measure male fertility accurately when high quality birth register data and nationally representative surveys are available. However, these data sources are not always accessible. Many low-income countries lack well-functioning vital registration systems, which would provide the necessary data to compute male fertility according to the method by (Dudel and Klüsener, 2016). While nationally-representative surveys have been conducted in many developing countries, they are typically available only for selected years, being both time-consuming and expensive to implement. Furthermore, both of these methods do not account for potential measurement errors in the data or random variation in demographic processes, which is particularly crucial for subnational populations.

Our proposed Bayesian model is particularly useful in a subnational setting. First, it allows to account for various types of errors and to effectively incorporate spatial and temporal dependencies. Second, our proposed model does not require information on births classified by parental ages. Instead, it requires accurate subnational age-sex population counts, along with standard age-specific fertility schedules and subnational mortality estimates. Generally, subnational population estimates by age and sex are more readily available in comparison to subnational birth data classified by parental ages. In developed countries, accurate subnational population estimates by age and sex are generally provided by statistical offices without restrictions. In low-income countries, obtaining these estimates can be significantly more challenging (Alexander and Alkema, 2022). Nonetheless, several efforts have been made to produce subnational population estimates by age and sex in developing countries. For instance, subnational population estimates and projections by age and sex are provided by the US Census Bureau for several developing

nations for the period 2000 – 2030 using data from the most recent censuses (U.S. Census Bureau, 2017).

4.2.2 Bayesian Methods in Demography

Bayesian methods[3] have become increasingly common in demography due to their ability to combine multiple data sources in the same model and to account for various types of errors (Bijak and Bryant, 2016; Bryant and Zhang, 2018). The employment of Bayesian methods has allowed unprecedented advancements in the estimation and forecast of national populations (Raftery et al., 2012, 2014), migration (Bijak, 2008; Bijak and Wiśniowski, 2010; Abel et al., 2013), fertility (Alkema et al., 2011; Schmertmann et al., 2014; Ellison et al., 2020; Batyra et al., 2023; Ellison et al., 2024) and mortality (Raftery et al., 2013; Alkema and New, 2014; Alexander and Alkema, 2018).

In the context of subnational estimation, Bayesian methods have also been widely used. For subnational mortality, Alexander et al. (2017) and Schmertmann and Gonzaga (2018) developed Bayesian hierarchical models to estimate subnational mortality in contexts where data availability is limited. Regarding subnational populations, Bayesian methods have again played a crucial role. Bryant and Graham (2013) proposed a formal Bayesian framework to obtain subnational population estimates for six regions in New Zealand relying on multiple data sources. More recently, Alexander and Alkema (2022) developed a Bayesian method to estimate and forecast subnational population estimates by age and sex in contexts with limited data. In addition, the "digital revolution" has led researchers to produce subnational population estimates from non-conventional data sources within a Bayesian framework. For instance, Leasure et al. (2020) developed a Bayesian hierarchical model to obtain subnational population estimates using geo-located data.

Concerning subnational fertility estimation, Ševčíková et al. (2018) proposed a Bayesian model to estimate and forecast subnational TFRs that are consistent with the national

[3]The literature review of Bayesian methods in demography presented in this chapter is an expanded version of the concise overview provided in Section 3.1 of Chapter 3. While Section 3.1 of Chapter 3 offers a brief introduction to Bayesian approaches for demographic research, Section 4.2.2 provides a more comprehensive examination of this research field, exploring recent advancements and applications in greater detail.

estimates produced by the United Nations. Schmertmann et al. (2013) employed empirical Bayesian methods to smooth volatile regional fertility data and then applied a new variant of the Brass relational model.

Despite the unprecedented developments in demographic estimation with Bayesian methods, male fertility estimation has been largely neglected, especially at a subnational level. Building on the methodological framework by Schmertmann and Hauer (2019), we present a Bayesian model to examine the evolution of subnational male and female TFR estimates over time and space. By combining multiple data sources, this model allows uncertainty to be incorporated and estimates to be driven by the available data.

While the proposed model requires the availability of accurate population estimates by age and sex, it does not require any knowledge of subnational age-specific fertility patterns. Instead, we utilize a statistical model based on principal components derived from national age-specific fertility patterns. To incorporate the mortality experienced by children and their parents, our model demands the availability of subnational mortality estimates. Although this is a strong requirement, we acknowledge that this limitation could be overcome by using more complex mortality model to address situations where accurate subnational mortality estimates are unavailable.

4.3 Method

4.3.1 Model Setup

Following the approach by Schmertmann and Hauer (2019)[4], we let $C_{a,t}$ be the observed number of children in area a and year t. We assume that it can be modeled as a Poisson distribution:

$$C_{a,t}|K^s_{x,a,t} \sim \text{Pois}\left(\sum_{x=15}^{\omega^s} K^s_{x,a,t} E^s_{x,a,t}\right) \tag{4.1}$$

[4]The modeling framework outlined in Section 4.3 of this chapter resembles quite closely the approach presented in Section 3.3 of Chapter 3, leading to the unavoidable repetition of some formulae. However, the modeling framework from Chapter 3 aims to produce female TFR estimates by combining imperfect population data with more accurate population data, while the modeling approach of this chapter leverages accurate small-area population data classified by age and sex to obtain subnational male and female TFR estimates.

where $K^s_{x,a,t}$ is the expected number of children per individual of sex s in age group x, area a and year t at the end of the five-year period[5], $E^s_{x,a,t}$ indicates the observed number of individuals of sex s in age group x, area a and year t, ω^s is the last reproductive age group for individuals of sex s. The value of ω^s is assumed to be $45-49$ for women and $55-59$ for men. In addition, throughout this paper, we will consider demographic quantities calculated for five-year age groups.

By harnessing standard approximations from cohort-component projection methods (see Keyfitz et al. (2005) for a review), we can define $K^s_{x,a,t}$ as follows.

$$K^s_{x,a,t} = \left[\frac{L^s_{x-5,a,t}}{L^s_{x,a,t}} \cdot F^s_{x-5,a,t} + F^s_{x,a,t}\right] \cdot \frac{L_{0,a,t}}{2} \tag{4.2}$$

where $L^s_{x,a,t}$ denotes the expected person-years lived by individuals of sex s in age group x in area a and year t. $L_{0,a,t}$ denotes the person-years lived by individuals in the age group $0-4$ in area a and year t. $F^s_{x,a,t}$ denotes the expected fertility experienced by individuals of sex s in the age interval $[x, x+5)$ in area a and year t. We set $F^s_{x,a,t}$ to be zero outside the interval $[15, 60)$ for men and outside the interval $[15, 50)$ for women.

Following Hauer and Schmertmann (2020), we can rearrange equation 4.2, which can be rewritten as

$$K^s_{x,t,c} = TFR^s_{a,t} \cdot \frac{L_{0,a,t}}{5} \cdot \frac{1}{2} \cdot \left[\frac{L^s_{x-5,a,t}}{L^s_{x,a,t}} \cdot \phi^s_{x-5,a,t} + \phi^s_{x,a,t}\right] \tag{4.3}$$

where $\phi^s_{x,a,t} = \frac{5 \cdot F^s_{x,a,t}}{TFR^s_{a,t}}$ is the fraction of life time fertility occurring to individuals of sex s in the age group x if they are subject to the age-specific fertility rates observed in area a and year t throughout their reproductive ages. $\frac{L_{0,a,t}}{5}$ denotes the expected fraction of still alive among children aged $0-4$ in area a and year t. $TFR^s_{a,t}$ is the period total fertility rate experienced by either men or women in area a and year t. This demographic measure can be interpreted as the expected number of children per man (woman) in area a and year t if he (she) is subject to the current period age-specific fertility rates throughout his (her) parental ages.

[5]The term $K_{x,t,a}$ follows a simple rearrangement of the Leslie matrix formulas (Wachter, 2014). However, unlike in standard cohort-component projection methods, this term defines the age of the mother as being attained at the end of the age interval rather than at the beginning.

Our proposed Bayesian model 4.3 incorporates demographic knowledge and uncertainty about demographic quantities by placing statistical models and priors on mortality and fertility parameters in equation 4.2.

4.3.2 Model for Age-specific Fertility Schedules

To incorporate prior knowledge about age-specific fertility schedules, we model the log transformation of the ratio of the share of life time fertility experienced by individuals of sex s in the age group x ($\phi^s_{x,a,t}$) to the share of life time fertility experienced by individuals of sex s in the age group $15-19$ ($\phi^s_{15,a,t}$). By applying this transformation, we make sure that the age-specific fertility schedules can assume both positive and negative values.

$$\gamma^s_{x,a,t} = \log\left(\frac{\phi^s_{x,a,t}}{\phi^s_{15,a,t}}\right) \tag{4.4}$$

Then, we model $\gamma^s_{x,a,t}$ as follows

$$\gamma^s_{x,a,t} = m^s_x + y^s_{1,x}\beta^s_{1,a,t} + y^s_{2,x}\beta^s_{2,a,t} + \nu^s_a + \delta^s_t + \epsilon^s_{a,t} \tag{4.5}$$

where $\boldsymbol{m}^s$, $\boldsymbol{y}^s_1$ and $\boldsymbol{y}^s_2$ are components derived from a set of standard age-specific fertility curves. In particular, $\boldsymbol{m}^s$ is a vector containing the age-specific means of the log-transformed age-specific fertility schedules ($\gamma^s_{x,a,t}$), while $\boldsymbol{y}^s_1$ and $\boldsymbol{y}^s_2$ are the first and second left-singular vectors which are obtained via a singular value decomposition (SVD) on the matrix $\boldsymbol{Y}^s$ whose columns are log-transformed male (female) age-specific fertility schedules ($\gamma^s_{a,t}$).

For example, in our application to US counties we employ the US national age-specific female fertility curves for the period 1982–2021 retrieved from the Human Fertility Database (Jasilioniene et al., 2015). US national age-specific male fertility curves are derived from the Human Fertility Collection (Grigorieva et al., 2015) and cover the period $1982-2015$.

The mean $\boldsymbol{m}^s$ describes the overall age-specific fertility curve. As expected, both male and female age-specific fertility patterns increase up to the age $30-34$ and then start to taper off. The decrease is substantially faster for women compared to men due the nar-

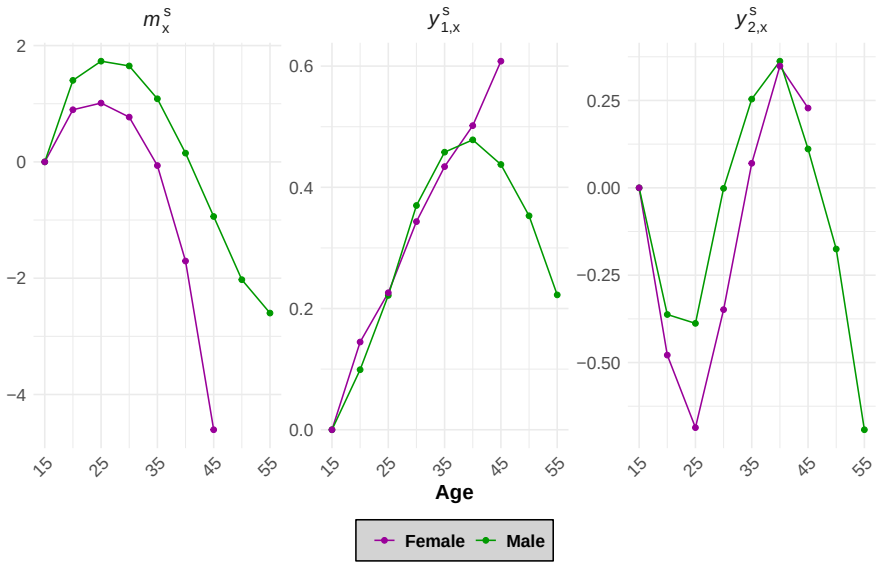


Figure 4.1: The figure provides an example of the SVD applied to logged US age- and sex-specific fertility proportions ($\gamma^s_{x,a,t}$). The first plot illustrates the average logged fertility rates proportions across the distinct reproductive age classes by sex ($\mathbf{m}^s$). The second and third plots display the values of the first ($\boldsymbol{y}^s_1$) and second ($\boldsymbol{y}^s_2$) left-singular vectors of the matrix ($\boldsymbol{Y}^s$) separately for men and women.

rower female reproductive age span. For both men and women, $\boldsymbol{y}^s_1$ seems to allow for the postponement of the mean age at parenthood, strictly increasing for women throughout the reproductive age period and for men up to the age class $40-44$. This is coherent with the "postponement transition" suggested by Kohler et al. (2002). $\boldsymbol{y}^s_2$ allows for higher fertility levels in the reproductive age interval $35-44$. For instance, in regions with a large share of highly educated men and women in reproductive ages, fertility in the age interval $35-44$ may be higher than the mean age fertility pattern.

$\beta^s_{1,a,t}$ and $\beta^s_{2,a,t}$ are defined as shape parameters drawn independently from a standard normal distribution for each combination of area a and time t.

$$\beta^s_{p,a,t} \sim \mathcal{N}(0,1) \tag{4.6}$$

The parameters ν^s_a controls the spatial autocorrelation among adjacent areas. These spatial effects are modeled via an intrinsic conditionally autoregressive model (ICAR)

(Besag et al., 1991; Besag and Kooperberg, 1995).

$$\nu_a^s|\nu_b^s \sim \mathcal{N}\left(\frac{1}{n_{\Delta_a}} \cdot \sum_{b\epsilon\Delta_a} w_{a,b}\nu_a^s, n_{\Delta_a}\lambda_\nu^s\right) \tag{4.7}$$

where $w_{a,b}$ is a weight being equal to 1 if area b is a neighbor of area a or 0 otherwise. Δ_a indicates the set of neighbors for area a. n_{Δ_a} denotes the number of neighbors of area a. λ_ν^s is the precision of the spatial effect ν_a^s. Following the suggestions by Knorr-Held (2000), we assign a non-informative prior to λ_ν^s. Namely,

$$\lambda_\nu^s \sim \text{gamma}(1, 0.01) \tag{4.8}$$

The parameter δ_t^s accounts for the temporal dependence of age-specific fertility patterns between consecutive years. This temporal effect is modeled via a first-order random walk.

$$\delta_t^s \sim \mathcal{N}(\delta_{t-1}^s, \sigma_\delta^2) \tag{4.9}$$

The parameter $\epsilon_{a,t}^s$ captures the residual spatio-temporal autocorrelation that is not accounted by the other independent spatial and temporal effects. It is assigned a normal distribution with zero mean.

$$\epsilon_{a,t}^s \sim \mathcal{N}(0, \sigma_\epsilon^2) \tag{4.10}$$

We place weakly informative priors on the standard deviation parameters σ_δ and σ_ϵ[6].

$$\sigma_\delta \sim \mathcal{N}^+(0, 1) \tag{4.11}$$

$$\sigma_\epsilon \sim \mathcal{N}^+(0, 1) \tag{4.12}$$

[6]A simulation exercise for understanding the ability of model 4.5 to capture age-specific fertility patterns is included in the appendix (see figure C.1).

Priors on Total Fertility Rates

The total fertility rate $TFR^s_{t,a}$ from equation 4.3 is assigned a normal distribution, whose mean is assumed to be equal to the regional-level TFR observed in year t ($TFR^{\text{region},s}_t$)[7] for either men or women[8].

$$TFR^s_{t,a} \sim \mathcal{N}(TFR^{\text{region},s}_t, \sigma^2_{TFR^s_{t,a}}) \tag{4.13}$$

The standard deviation parameter $\sigma_{TFR^s_{t,a}}$ is assigned a weakly informative prior.

$$\sigma_{TFR^s_{t,a}} \sim \mathcal{N}^+(0,1) \tag{4.14}$$

The practical implication of centering the prior distribution of $TFR^s_{t,a}$ to the state value $TFR^{\text{region},s}_t$ is to shrink counties with a small population towards the state average. In this manner, fertility levels in small counties are partially informed by the state fertility level. On the contrary, fertility levels in larger counties are primarily determined by the composition of their population in terms of age and sex. In addition, if state-level TFR values are unavailable, we center the distribution of the TFR parameter around the national value. In our example, US male TFR values are available until 2004 at the state level. Therefore, starting from 2005, we let the mean of the male TFR parameters be equal to the corresponding national values[9].

[7] In the United States, "region" can refer to various geographic units when discussing the TFR, such as individual states or broader areas. When state-specific TFR data is unavailable, national TFR figures are often used.

[8] We acknowledge that the US is a country with a well-functioning civil registration system. National or state-level male TFR estimates may be unavailable in other contexts, including several developing countries. As an alternative, one could place an uninformative prior on the TFR parameter (e.g. a uniform distribution on the interval $[0, 12]$).

[9] National male TFR values are available from Human Fertility Collection until 2015. For the years 2016-2019, we assume that the mean of the TFR parameters to be equal to the national TFR value recorded in 2015.

4.3.3 Priors on Mortality Parameters

Schmertmann and Hauer (2019) modeled child and adult mortality with the log-quadratic mortality model by Wilmoth et al. (2012). Without delving into technical details, they established prior distributions for the two parameters of the log-quadratic model by Wilmoth et al. (2012) and recovered the age-specific person-years using standard life table relationships.

Our proposed model incorporates information about child and adult mortality by placing a prior probability distribution directly on the person-years parameters (e.g. $L_{0,a,t}$ and $L^s_{x,a,t}$) of equation 4.3. In our data example, we used county-specific life tables for the historical period 1982−2019 from the US Mortality Database. However, we are aware that the direct modeling of the person-years is feasible provided that subnational life tables are available for the country of interest. In case we lack detailed subnational mortality data, the person-years parameters could estimated via mortality models or using national life tables from the United Nations World Population Prospects (UNDESA, 2024) or the World Health Organization (WHO, 2022).

Operationally, in order to include the uncertainty associated to the subnational person-years estimates in 4.3, we assume that the person-years for individuals of sex s in an age group x in area a at time t ($L^s_{x,a,t}$) are normally distributed with a mean equal to the corresponding person-years estimate for the age group x from the subnational life table data referring to sex s, area a and, time t. The variance is calculated through empirical simulations[10].

$$L_{0,a,t} \sim \mathcal{N}\left(\hat{L}_{0,a,t}, \hat{\sigma}^2_{\hat{L}_{0,a,t}}\right) \tag{4.15}$$

$$L^s_{x,a,t} \sim \mathcal{N}\left(\hat{L}^s_{x,a,t}, \hat{\sigma}^2_{\hat{L}^s_{x,a,t}}\right) \tag{4.16}$$

where $\hat{L}_{0,a,t}$ indicates the estimated person-years for the age-group $0-4$ from a period life table constructed using mortality rates observed in area a at time t. $\hat{L}_{0,a,t}$ is included to incorporate information about mortality among children under age 5.

[10] More details on the calculation of the variances of the person-years parameters are provided in the appendix in section C.1.

Similarly, $\hat{L}^s_{x,a,t}$ denotes the estimated sex-specific person-years for the age classes $x = 10-14, 15-19, \ldots, 45-49$ for women and $x = 10-14, 15-19, \ldots, 55-59$ for men. These estimates are obtained from a sex-specific subnational life tables referring to area a and year t. The inclusion of $\hat{L}^s_{x,a,t}$ is to account for mortality experienced by men and women during their reproductive ages. $\hat{\sigma}^2_{\hat{L}^s_{x,a,t}}$ and $\hat{\sigma}^2_{\hat{L}_{0,a,t}}$ indicate the variances of the person-years parameters calculated by means of simulations[11].

4.3.4 Model Summary

The model is fitted separately by sex and state. The figure 4.2 provides a graphical representation of the proposed model. Broadly speaking, the number of children aged $0-4$ is

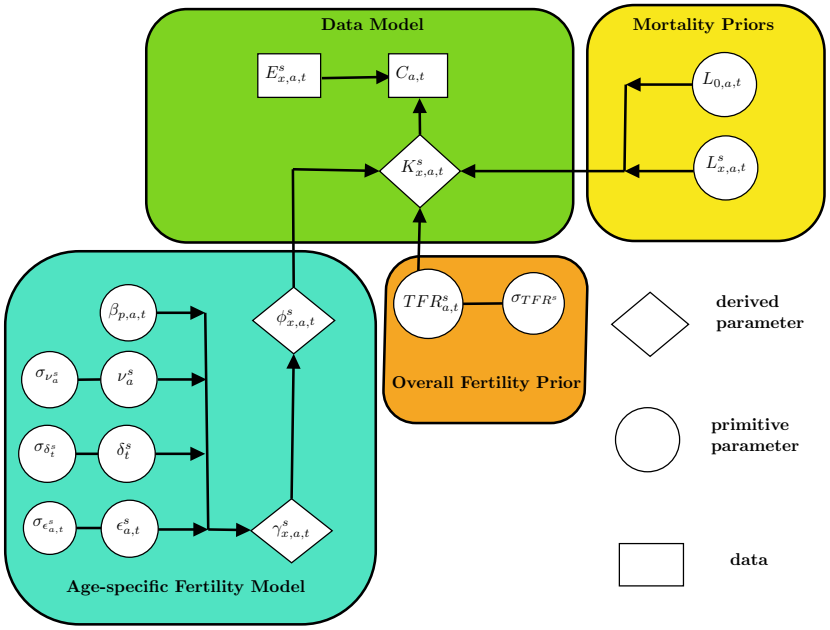


Figure 4.2: Graphical representation of the Bayesian modeling framework. Primitive parameters denote the fundamental parameters in a model that are directly assigned a prior probability distribution. Derived parameters are functions of primitive parameters and do not have prior probability distribution directly assigned to them.

assumed to be Poisson-distributed. The fertility and mortality parameters, which govern the mean of the Poisson distribution, are specified by imposing a hierarchical structure. Preliminary information about the overall fertility level is included by placing a probability distribution on the TFR parameter and on its standard deviation. Prior knowledge about the age-specific fertility patterns is incorporated by specifying a statistical model

[11] Details on the estimation of the variances are reported in the appendix.

on national age-specific fertility patterns, in which we also account for spatial and temporal dependencies, and probability distributions on its parameters. Finally, we include prior knowledge about child and adult mortality by placing prior probability distributions directly on the person-years parameters.

4.3.5 Model Implementation

Operationally, posterior samples for the TFR parameters were obtained using the R package *nimble* (de Valpine et al., 2017). This package allows to specify the main structure of our model in R, compiles the model in C++ and implement an adaptive Metropolis-within-Gibbs MCMC algorithm to simulate from the posterior distribution of the TFR parameters.

Best estimates of the TFR parameters were taken to be the medians of the corresponding posterior samples. 95% credible intervals for the TFR parameters were constructed by computing the 2.5% and 97.5% quantiles from the relevant posterior samples.

4.4 Model Simulations

4.4.1 Simulations on Data from Australia

To evaluate the performance of our model, we created a simulated data set of children generated using existing subnational fertility and mortality patterns for eight Australian territories. Given the availability of subnational fertility and mortality estimates by age and sex for the period 2001 – 2020[12], we computed the true number of expected children ($K^s_{x,a,t}$) at each parental age group using equation 4.2. Next, we simulated the overall number of children ($C_{a,t}$), according to the relationship shown in equation 4.1[13], and then applied our proposed model to the simulated data[14]. The number of exposed men and women were treated as fixed, and the dependency structure among regions was specified

[12]Details on the data used for the simulation are reported in table C.1 in the appendix.

[13]The 'true' subnational age-specific fertility rates and person-years were incorporated in equation 4.2.

[14]Standard errors for subnational mortality data are not available. For this reason, we assumed that the subnational values for the person-years were accurate and that the standard deviations of the probability distributions of the person-years parameters were equal to 0.001.

using an adjacency matrix matching the neighboring structure of Australian regions. Fig-

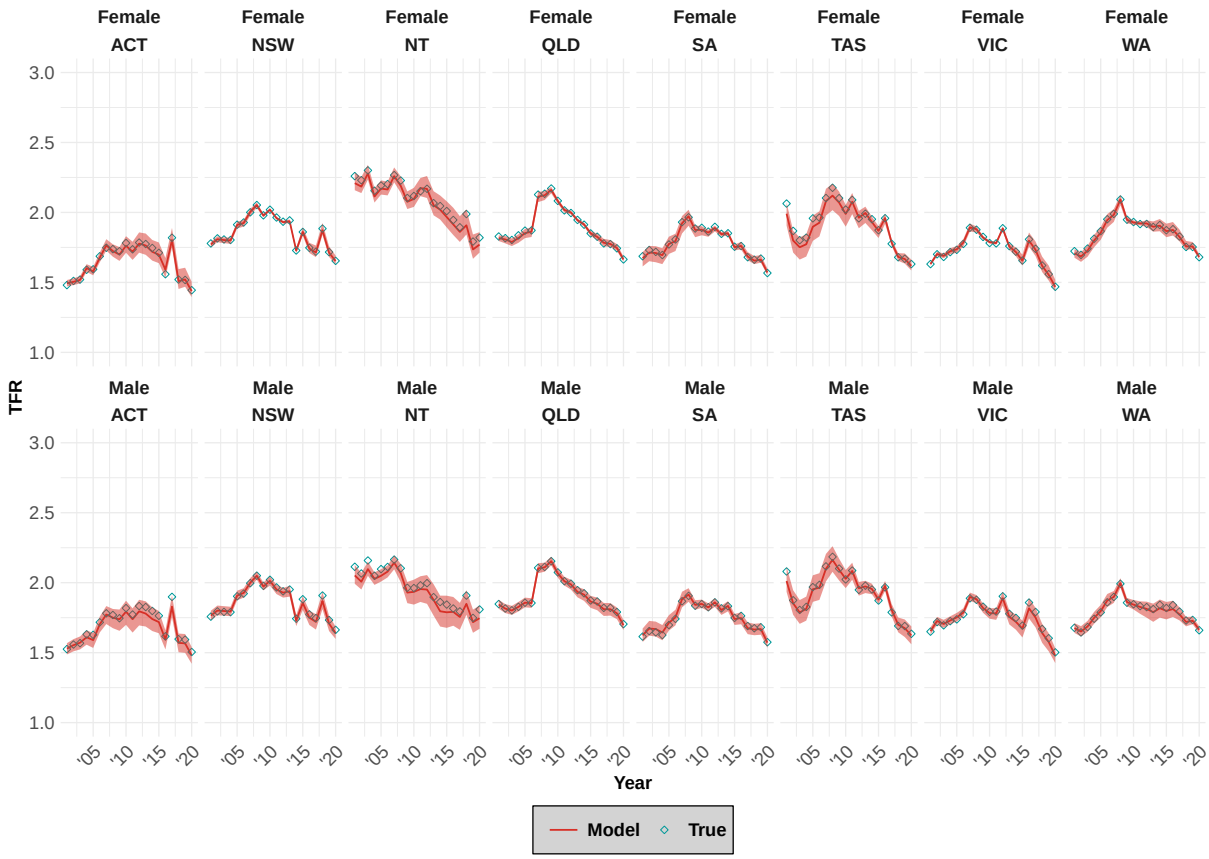


Figure 4.3: True and model-based total fertility rates for eight Australian territories during the period 2001-2020. 95% credible intervals are also provided via shaded areas. TFR estimates are the medians from the corresponding posterior samples.

ure 4.3 displays the true and estimated TFRs for the eight Australian regions. The dots indicate the true TFRs that are computed from national birth register data using standard demographic methods. The solid lines denote the TFR estimates that were obtained by fitting the proposed model to the simulated data. The corresponding 95% credible intervals are displayed via shaded regions. Male and female TFR estimates seem to mirror the true TFR values quite closely. In addition, regions with larger absolute numbers of births show lower uncertainty around their estimates. Regions such as Tasmania (TS) and Northern Territory (NT) with smaller populations exhibit wider credible intervals.

To evaluate the performance of the model, we compare it with the implied total fertility rate (iTFR) that is obtained via an indirect estimation method developed by Hauer and

Schmertmann (2020). $iTFR$ is computed as follows.

$$iTFR = \frac{1}{1 - 0.75 \cdot q_{0-4}} \cdot 7 \cdot \frac{C_{0-4}}{W_{15-49}} \tag{4.17}$$

where q_{0-4} is the probability of death under 5 to account for child survival, C_{0-4} is number of children under 4 and W_{15-49} is the number of women aged $15-49$. The multiplicative constant 7 allows age-specific fertility levels to be equal across the distinct childbearing age groups, namely $15-19, \ldots, 45-49$[15]. By replacing W_{15-49} with the number of men aged $15-59$ and the multiplicative constant with 9, we can produce an indicator equivalent to the iTFR by Hauer and Schmertmann (2020) for the measurement of male fertility.

For each region, we compare these methods of estimation using the root mean squared error (RMSE), defined as

$$RMSE = \sqrt{\frac{1}{T}\left(\sum_{t=1}^{T} \hat{TFR}_{a,t} - TFR^{*}_{a,t}\right)^2} \tag{4.18}$$

where T is the total number of years, $\hat{TFR}_{a,t}$ is the TFR estimated from the model for region a and year t, and $TFR^{*}_{a,t}$ is the true TFR of region a in year t. Table 4.1 compares the performance of the proposed Bayesian model with the indirect estimation method by Hauer and Schmertmann (2020), using the root mean squared error (RMSE) as the evaluation metric.

Table 4.1: Performance of the proposed model and of the indirect method by Hauer and Schmertmann (2020) using the RMSE as metric.

	Male		Female	
State	**Model**	**Indirect**	**Model**	**Indirect**
ACT	0.0323	0.0954	0.0159	0.0720
NSW	0.0167	0.0709	0.0046	0.0534
NT	0.0445	0.1850	0.0380	0.1350
SA	0.0121	0.0437	0.0086	0.0336
QLD	0.0071	0.0437	0.0093	0.0230
TAS	0.0202	0.0997	0.0361	0.0742
VIC	0.0203	0.0826	0.0049	0.0600
WA	0.0131	0.0840	0.0075	0.0590

[15] A similar indicator with a different multiplicative constant for the age, called extended total fertility rate is presented in the paper by Hauer and Schmertmann (2020).

Table 4.1 demonstrates the superiority of our Bayesian method over the indirect estimation approach proposed by Hauer and Schmertmann (2020), as indicated by consistently lower RMSE values across all Australian regions.

While the indirect estimation approach by Hauer and Schmertmann (2020) assumes assumption that fertility remains constant across the different age groups, the proposed model integrates information on age-specific fertility patterns through a principal component regression on national age-specific fertility curves. Additionally, our proposed modeling framework accounts for the survival of both children under 5 and their parents by incorporating person-years estimates from subnational life tables, whereas the indirect estimation method by Hauer and Schmertmann (2020) only accounts for child survival. These incorporations allow our proposed model to generate more accurate TFR estimates in comparison to the indirect estimation approach by Hauer and Schmertmann (2020).

Furthermore, in regions with smaller populations, such as Northern territories, our proposed model exhibits a significantly lower RMSE compared to the indirect estimation proposed by Hauer and Schmertmann (2020). This result highlights the ability of our model to account for the higher stochastic variation in the number of children under 5 observed in smaller populations, where fertility indicators are more challenging to estimate with precision.

Regarding differences in model performance by sex, the proposed model performs slightly better for female fertility compared male fertility. A potential explanation for this lies in the higher accuracy of the age-specific female fertility rates used in the simulation. While the maternal age is available for all births, the paternal age is missing in a non-negligible number of cases. To address this issue, the births with missing paternal ages have been redistributed according to the observed distribution of the births by paternal ages[16]. The imputation of paternal ages for births lacking this information has inevitably introduced greater imprecision into the male age-specific fertility estimates used to simulate the number of children under 5. As a result, the performance of the proposed model for men is slightly worse than for women.

[16]This imputation technique was developed by Dudel and Klüsener (2016)

4.5 Data

4.5.1 Mortality

To account for the mortality of parents and children, we model directly the age-specific person-years. Our data example refers to the US context that provides a rich set of subnational life tables. Information on county-level mortality are retrieved from the US Mortality Database. This data source provides life tables by sex for both US states (1969–2020) and counties (1982–2019). Additionally, to account for the uncertainty in the county-level mortality indicators, standard errors for the age-specific death probabilities are provided. Building on the standard errors of the age-specific death probabilities, we can easily exploit simulations and life table relationships to obtain uncertainty measures for the age-specific person-years.

4.5.2 Fertility

Neither male nor female age-specific fertility rates are available for all the US counties. Hence, we rely on national age-specific fertility rates from the Human Fertility Database for women (Jasilioniene et al., 2016) and from the Human Fertility Collection for men (Grigorieva et al., 2015).

TFR estimates at the state level are available for women for the entire study (1982-2019)[17] and for men for the period (1982-2004). The US National Bureau of Economic Research provides birth records with information on the parental ages and on the state where the birth occurred for the period 1982-2004 [18]. Using the method proposed by Dudel and Klüsener (2016), we can easily obtain male and female TFR estimates for the period 1982 – 2004. Concerning the period 2005 – 2019, the National Center for Health Statistics provides TFR values at the state level for only women. Hence, for the years after 2004, we used as prior information the national male TFR values available from the Human Fertility Collection.

[17] From 2005 to 2019, state-level female TFR values are taken from the National Vital Statistics.
[18] More information at the website https://data.nber.org/natality/

4.5.3 Population Counts by Age and Sex

Population counts by age and sex are taken from National Cancer Institute, which produces annual population estimates from 1969 on in collaboration with US Census Bureau and the National Center for Health Statistics.

Using these data, we retrieved the number of children under age 5, the number of women in the age group 15 – 49 and the number of men aged 15 – 59 for the period 1982 – 2019 and for all the US counties.

4.6 Results

4.6.1 Application to US Counties

We applied our proposed model to produce period male and female TFR estimates in US counties for the historical period 1982 – 2019. As of 2019, there were 3,244 counties in the US, including the states of Alaska and Hawaii. The population sizes of US counties display a significant heterogeneity ranging from counties with over 10 million residents such as Los Angeles to others with less than 100 inhabitants. As a data example, we illustrate the model results for the counties of two US states: California and Utah[19]. The former is the state with largest population size in the country, while the latter was the state with the highest recorded fertility levels until the middle of the 2010s[20]. The proposed model was fitted separately by sex and state. Figure 4.4 displays the evolution of TFR estimates from 1982 to 2019 of six counties, three in California and three in Utah, disaggragated by sex. TFR estimates are the medians of the corresponding posterior samples.

The three counties in California, which present higher absolute numbers of children under age 5, display TFR estimates with lower uncertainty in comparison to the counties in Utah. When examining fertility differentials by sex, we observe no staggering discrepancies between men and women in the counties of both states. This result is coherent with previous research on male fertility in low-fertility countries (Dudel and Klüsener, 2021).

[19] Results for all the counties in continental US are presented in the appendix.

[20] South Dakota has been the US state with the highest TFR since 2016.

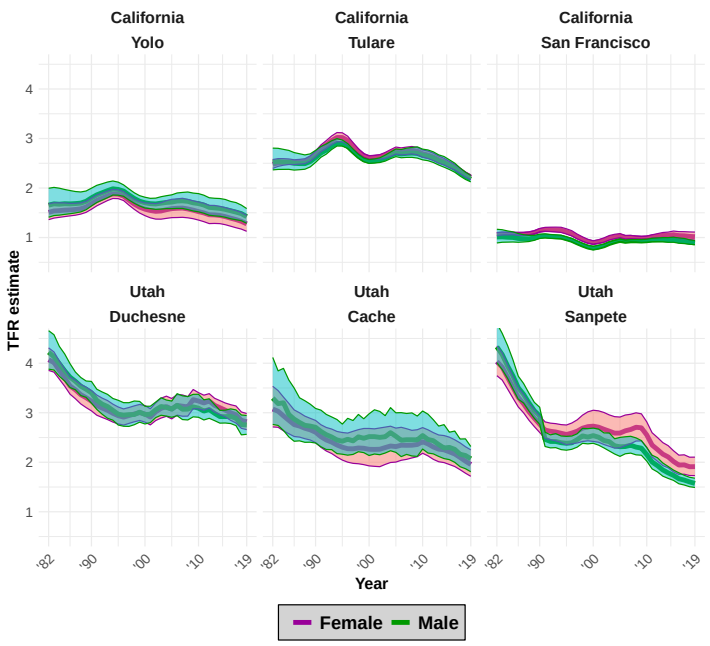


Figure 4.4: Male and female period TFR estimates for six US counties across the period 1982 – 2019. 95% credible intervals are also provided via shaded areas. TFR estimates are the medians of the corresponding posterior samples.

Concerning the temporal evolution of the TFR estimates by sex, we find distinct trajectories in the counties of the two states. In California we observe an increase in the TFR estimates for both sexes during the 1990s followed by a decline starting from the 2000s. Previous research (Hill and Johnson, 2002; Johnson and Li, 2007) attributed this finding to a growing population of young adult migrants from Latin American, who tended to display higher fertility levels in comparison to the US-born population. In California, Tulare consistently exhibited similar male and female TFR estimates above the replacement level of 2.1 throughout the entire study period. In Yolo, men displayed slightly higher TFR estimates in comparison to women. Nonethless, both men and women in Yolo were found to be consistently below the replacement level for most of the period. San Francisco showed extremely low fertility TFR estimates, with women experiencing slightly higher fertility than men. This trend may be linked to the high living costs and to a relatively high number of working-age individuals, who might temporarily reside in this county for job-related reasons.

Counties in Utah experienced a continuous decline in male and female TFRs from the 1980s until the early 2000s. Men and women in these counties used to bear on average

more than three children as early as 1982. Nonetheless, in the year 2000, their male and female TFR estimates declined by at least one unit. Fertility remained roughly constant throughout the period 2000 – 2009 with TFR estimates slightly above the replacement level of 2.1. After year 2009, TFR estimates began to decline again in the Utah counties, possibly a response to the global financial crisis (Comolli, 2017). In this regard, both Sanpete and Cache counties in Utah reached levels below the replacement levels of 2.1 during the period 2010-2019. Hence, net of migration, the populations of these counties would start to decrease.

Figure 4.5 shows the spatial distribution of male and female median TFR estimates for Utah and California in 2019. TFR estimates are generally higher in Utah compared to California. Most of the Utah counties display a TFR estimate higher than 2.1. Conversely, in California, a high proportion of counties display male and female TFR estimates below the replacement level.

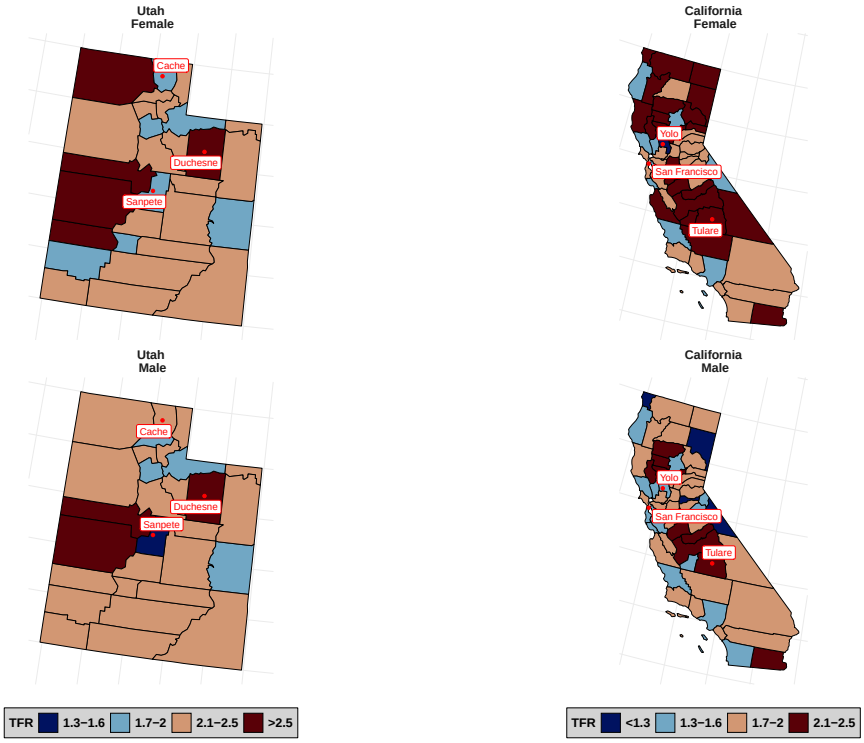


Figure 4.5: Spatial distributions of male and female TFR estimates for Utah and California in 2019. TFR estimates are the medians of the corresponding posterior samples.

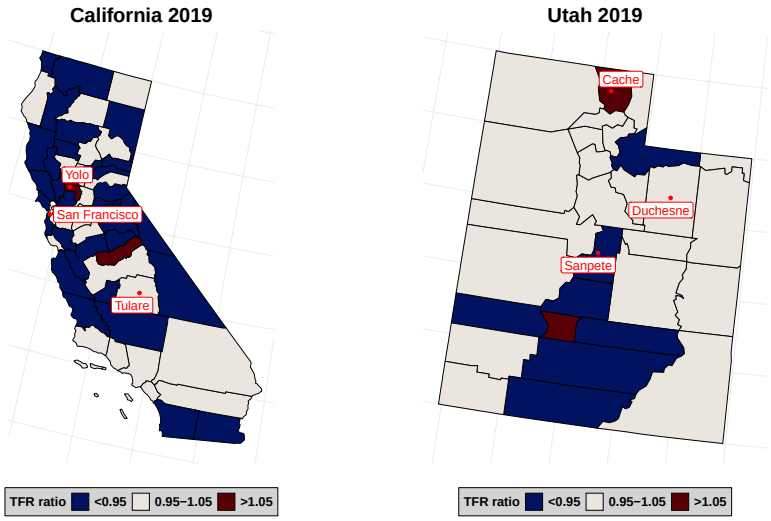


Figure 4.6: Spatial distributions of the male to female TFR ratios for Utah and California in 2019. Male and female TFR estimates are the medians of the corresponding posterior samples.

Figure 4.6 displays the ratio of the male TFR estimates to the female TFR estimates in 2019 in Utah and California. In general, the majority of the counties in both states exhibit ratios between 0.95 and 1.05, implying similar levels of fertility between men and women. This is consistent with what has been found at the national level by Dudel and Klüsener (2021) and at the state level by Schubert and Dudel (2024). However, a non-negligible number of counties display TFR ratios lower than 0.95, with women presenting higher fertility levels in comparison to men. On the contrary, only few counties display a TFR ratio grater than 1.05 both in California and Utah.

4.7 Discussion

In this article, we propose a Bayesian modeling framework to estimate subnational TFRs for both men and women in situations where the amount of available data is limited. The proposed method extends the modeling framework developed by Schmertmann and Hauer (2019) in two ways. First, it allows for the estimation of both male and female fertility in contexts with limited birth data. Second, it accounts for possible spatial and temporal dependencies in fertility patterns.

When tested with simulated data from Australian regions, the model outperforms the indirect estimation method by Hauer and Schmertmann (2020). Additionally, in compar-

ison to non-probabilistic indirect estimation methods, our Bayesian modeling framework allows for the quantification of uncertainty around the TFR estimates.

After a validation with simulated data, the model was fitted to estimate male and female TFR estimates at the county level in the United States from 1982 to 2019. The results suggest that men and women in the US experience similar fertility patterns across this period, with women exhibiting slightly higher fertility levels than men. The degree of uncertainty around the TFR estimates is linked to the absolute number of children under age 5, with higher uncertainty in counties with a smaller absolute number of children under age 5. A comparison between Utah and California showed how states within the same country may experience distinct fertility patterns. The majority of counties in Utah displayed TFR estimates above the replacement level of 2.1 in 2019, whereas only a small portion of counties in California did so.

One of the main advantages of the model is that it does not require any knowledge on the births classified by parental ages. In developed countries, subnational information on births by parental ages is often inaccessible to researchers due to privacy concerns. In countries lacking well-functioning vital registration systems, births classified by parental ages can only be computed from nationally-representative surveys that are not carried out timely due to the high associated costs.

In addition, we argue that our modeling framework can be easily applied to other countries. In the majority of the high-income countries, subnational mortality estimates, e.g. through the Human LifeTable Database (Shkolnikov et al., 2017), and subnational population counts by age and sex, e.g. through national statistical offices, are made available to researchers without particular restrictions. Regarding developing countries, subnational population estimates by age and sex from the Census Bureau (U.S. Census Bureau, 2017) could be leveraged to examine male and female fertility at a subregional level[21]. Since the majority of the developing countries lacks subnational mortality estimates, national mortality estimates from the United Nations World Population Prospects (UNDESA, 2024)

[21]The Census Bureau provides population estimates by age and sex for a limited amount of developing countries. Hence, fitting the proposed model to these population data would allow to investigate subnational male and female fertility for a subset of developing countries.

or from the World Health Organization Mortality Database (WHO, 2022) could be employed. The uncertainty coming from the absence of subnational mortality data could be incorporated by placing higher variances on the mortality parameters or by modeling the subnational mortality rates as a linear combination of the principal components computed from national mortality data as previously done by Alexander and Alkema (2022).

However, we acknowledge that our proposed method is not free from caveats. First, it relies on the knowledge of national age-specific male fertility patterns, which are only available for only a limited amount of countries. In many countries, especially those without well-functioning civil registration systems, age-specific male fertility rates can only be estimated from nationally-representative surveys, such as DHS data.

Second, our chosen age range for male reproduction may be overly restrictive for some countries. For example, Schoumaker (2017) showed that in world regions such as Sub-Saharan Africa, male age-specific fertility rates can remain above zero until age 75.

Third, our model does not properly account for internal migration. For instance, counties with a higher share of college students in comparison to the total population tend to experience temporary population surges in the reproductive age classes $15-19$ and $20-24$, which can lead to an underestimation of the actual TFR. Fourth, the proposed model only allows to measure fertility using the TFR. Even though the TFR is most popular fertility indicator, the proper understanding of the fertility behavior of an area demands the calculation of other measures such as the mean age at childbearing (MAB) or the cohort fertility rate (CFR).

In conclusion, building on the methodological framework by Schmertmann and Hauer (2019), the proposed method allows to obtain reliable male and female TFR estimates at a subnational level. The Bayesian modeling framework integrates multiple data sources while incorporating their corresponding uncertainty. In addition, in comparison to traditional indirect deterministic methods, our approach offers increased flexibility in modeling fertility processes by accounting for spatial and temporal dependencies in fertility patterns.

Chapter 5

Conclusion

5.1 Contributions

In this section, I present the main findings of the three studies included in this thesis and briefly discuss their key contributions.

5.1.1 Using online genealogies for demographic research

Chapter 2 addressed the promises and pitfalls of online genealogical data for demographic research. First, novel indicators to assess the quality and completeness of demographic information in the big genealogical database FamiLinx were introduced. Second, regression analyses were conducted to investigate the behavior of the completeness and quality of demographic variables across family networks. Third, the age-sex stracture and mortality of the Swedish population from FamiLinx was compared with those derived from more reliable data sources.

The first part of the study evaluated the concept of completeness of demographic variables in FamiLinx by computing the percentage of non-missing values in common demographic variables, namely birth and death years and countries. To assess the quality of the available demographic variables, the percentage of individuals with non-missing months in their birth and death dates was computed. The latter was considered as an indicator of more reliable demographic information, as individuals with available month information were deemed to exhibit more accurate demographic information in comparison to those with only the year available.

The second part of the study investigated the completeness and quality of demographic information across family networks using negative binomial regression binomial regression models. These models compared the expected number of relatives of a certain type with a non-missing value in a demographic variable, based on whether a focal individual had a missing or non-missing value in the same demographic variable. The results of the regression models revealed that a focal individual with a non-missing demographic variable was expected to be embedded in a family network of members with more accurate and more complete demographic variables. This finding underscores the tremendous potential of FamiLinx as an innovative data source for the investigation of the intergenerational

transmission of demographic behaviors.

The third part of the study revealed two non-negligible findings by comparing mortality indicators and the age-sex proportions of the Swedish population from FamiLinx to those from more reliable data sources. First, in agreement with the previous literature on mortality in genealogies (Zhao, 2001; Stelter and Alburez-Gutierrez, 2022; Chong et al., 2022), individuals in genealogies tended to experience lower mortality levels in comparison to the general population. Second, the age-sex distribution of populations from online genealogies did not align with the expected one from more reliable data sources. Specifically, the number of children and women was usually underestimated, whereas more longevous individuals and men were generally overrepresented. This finding sheds new light on the necessity of developing appropriate statistical methodologies to enable a more accurate measurement of demographic processes at a population level from non-representative data sources, including online genealogies.

To summarize, this chapter provided three significant contributions:

(a) Though a detailed analysis of the demographic information in FamiLinx, it was revealed that the non-missingness of demographic variables is highly selective.

(b) One of the key relevant features of FamiLinx is the possibility to establish family ties among its individuals. Standard regression models revealed the presence of two distinct groups: one consisting of family networks with high quality and complete demographic information, and another with a high share of missing demographic information.

(c) The non-representativeness of FamiLinx prevents researchers to solely rely on this data source for the estimation of demographic indicators when examining population-level demographic trends in historical populations.

5.1.2 Correcting fertility in online genealogical data

Chapter 3 aimed to develop a Bayesian modeling framework for producing TFR estimates by integrating online genealogical data from FamiLinx with more reliable data sources, such as censuses and population registers. The proposed model allowed to reconstruct

historical TFR time series for seven European countries and the United States over the period 1751 – 1910.

This approach extended the Bayesian modeling framework developed by Schmertmann and Hauer (2019), which assumed that the observed age-sex structure in the data reflects the structure of the true population. However, as demonstrated in Chapter 2, the age-sex composition of online genealogical populations does not mirror that of the true population. To address this discrepancy, the proposed method modeled simultaneously the number of children aged 0–4 from FamiLinx and the true number of children aged 0–4 based on more reliable data sources. A bias-adjustment parameter was introduced to correct for potential biases in the expected number of children under 5 per woman aged 15 – 49 in FamiLinx. This parameter was assumed to be governed by a global time-varying mean parameter, which captures the overall bias across all countries over time. By incorporating this parameter, bias patterns in countries with limited accurate data were partially informed by those observed in countries with reliable population counts by age and sex.

In parallel, the indirect estimation method developed by Hauer and Schmertmann (2020) was extended to account for the non-representativeness of children under 5 and women aged 15 – 49 in online genealogical populations. Specifically, an additional multiplier was introduced in the TFR decomposition by Hauer and Schmertmann (2020) to adjust for the biases in the estimation of the child-woman ratio (CWR). In order to partially mimic the exchange of information across countries proposed in the Bayesian, in absence of accurate population data the multiplier was estimated based on the patterns observed in countries with accurate population data available.

The results indicated that the proposed adjustments enabled to produce more accurate TFR estimates in comparison to the original Bayesian model by Schmertmann and Hauer (2019) and to the indirect estimation method by Hauer and Schmertmann (2020). This finding underlined that using online genealogical data alone to estimate demographic indicators in historical population could lead to severely biased TFR estimates. This study also demonstrated that the reconstruction of fertility in historical populations from online genealogies required the incorporation of information from more accurate data

sources. Furthermore, the TFR estimates of the proposed Bayesian model closely aligned with historical TFR values reported in historical demography studies or from reliable sources such as Human Fertility Collection.

To summarize, this chapter provided four significant contributions:

(a) The Bayesian modeling framework by Schmertmann and Hauer (2019) was extended to handle TFR estimation with imperfect data.

(b) The indirect estimation method Hauer and Schmertmann (2020) was expanded to obtain TFR estimates with biased CWRs.

(c) New TFR estimates with their corresponding credible intervals were obtained from multiple countries with partial availability of accurate population data for the historical period 1751 – 1910.

5.1.3 Measuring male and female fertility at a subnational level

In Chapter 4, the Bayesian modeling framework by Schmertmann and Hauer (2019) was expanded to tackle male and female fertility estimation at a subnational level. To estimate the male TFR, the number of men aged 15-59, classified in 5-year age groups, was treated as offset, considering them as exposed to the risk of fathering a child. A principal component regression based on national male age-specific fertility curves was incorporated to capture age-specific fertility patterns. As an extentision of the original modeling framework by Schmertmann and Hauer (2019), temporal and spatial parameters were introduced into this regression model. The inclusion of temporal parameters allowed TFR estimates to vary more smoothly over time, whereas spatial parameters accounted for spatial dependencies in fertility behaviors among neighboring areas. Additionally, unlike the original model by Schmertmann and Hauer (2019), person-years parameters were directly informed by estimates from subnational life tables, avoiding reliance on the log-quadratic mortality model by Wilmoth et al. (2012). TFR parameters were partially informed either national or regional TFR values.

When tested with simulated data from Australian regions, the proposed model demon-

strated superior performance compared to the deterministic indirect estimation method developed by Hauer and Schmertmann (2020). The superiority of the proposed model was particularly evident for regions with smaller populations.

An empirical application to population counts classified by age and sex from US counties provided new subnational TFR estimates for both men and women during the period 1982-2019. These estimates offer novel insights into how male and female fertility behaviors have changed over time and space, particularly in the context of the second demographic transition (Lesthaeghe, 2014). Focusing on Utah and California as case studies, male TFR estimates were found to be fairly similar to female TFR estimates. However, the observed TFR estimates were found to be spatially heterogeneous. In Utah, most counties in 2019 exhibited TFR estimate above the replacement level of 2.1, while in California, an increasing number of counties was approaching TFR levels as low as 1.5. Broadly speaking, this study made three notable contributions:

(a) The Bayesian modeling framework by Schmertmann and Hauer (2019) was extented to obtain male TFR estimates at a subnational level.

(b) Temporal and spatial parameters were incorporated, accounting for dependencies in fertility behaviors over time and between adjacent areas.

(c) The proposed model produced new subnational TFR estimates by sex for all US counties over the historical period 1982 – 2019.

5.2 Limitations and Extensions

In this section, I reflect on the limitations of the studies presented in the thesis. Furthermore, I discuss how these studies can be extended and improved.

5.2.1 Online genealogies for demographic research

The first chapter of this thesis provided a detailed assessment about the potential of online genealogical data in demographic research. By allowing users worldwide to recon-

struct their family trees and insert information, including demographic data, about their ancestors, online genealogies represent a rich data source, especially for historical demographers. The focus of the first chapter was on the big genealogical database FamiLinx, which was constructed by Kaplanis et al. (2018) by merging individuals from digital family trees available on the website geni.com.

One of the main limitations of this study is its sole focus on FamiLinx, whose features may not always be generalized to other online genealogical data. A potential extension would be to explore population data from other genealogical websites such as ancestry.com, and investigate whether the biases observed in FamiLinx are generalizable to data from other online genealogies. Additionally, as Kaplanis et al. (2018) only selected individuals, who died by the end of 2015, the accuracy of demographic indicators from FamiLinx in more contemporary periods cannot be properly assessed. A valuable extension in this direction would be to scrape more data from geni.com, including data about individuals who are still alive at the time of data collection. Integrating the updated data with modern population censuses would lead to a deeper understanding of the extent to which we can online genealogies for population-generalizable measurements in a contemporary setting.

In addition, while Sweden was used as a case study to investigate the degree to which mortality indicators in FamiLinx are biased in comparison to those derived from more reliable data sources, future studies should perform similar comparisons with other countries.

Another limitation is that most of the users in geni.com are based in western countries, as are the majority of their ancestors. Consequently, FamiLinx severely under-represents individuals from the Global South, hindering its use for the estimation of demographic indicators in countries lacking well-functioning civil registration systems. A valuable step moving forward would be to conduct similar analyses on the completeness and quality of data stemming from genealogical websites, which may be commonly used in some developing countries, and potentially combine these data with nationally representative population surveys such as those from the Demographic and Health Surveys (DHS).

5.2.2 Fertility estimation with imperfect data

Chapter 3 extends the Bayesian modeling framework by Schmertmann and Hauer (2019) to enable the indirect estimation of the TFR in presence of imperfect population data.

One of the main limitations of this study is the inability to analyze fertility patterns within population subgroups known to have distinct reproductive behaviors. The lack of micro-level data on socio-economic status in FamiLinx prevents an investigation into whether women from wealthier families were forerunners in the fertility decline observed within the countries under analyses during the historical period 1751 – 1910. Similarly, the absence of data on the exact time of migration of an individual in FamiLinx makes it challenging to examine differences in fertility levels between native-born women and migrants.

A potential solution to address this limitation could be to link FamiLinx data with micro-level censuses from the Integrated Public Use Microdata Series (IPUMS) (Ruggles et al., 2011). However, since micro-level data in FamiLinx are anonymized, performing data linkage would become impractical. A promising advancement in this direction could involve re-scraping data from `geni.com` while also collecting the first names and last names of deceased individuals in the digital family trees. These names could be then linked to micro-level census records using ad-hoc matching algorithms, such as the one proposed by Abramitzky et al. (2020).

Concerning the proposed Bayesian model, a next step would be to model the bias-adjustment parameters more flexibly. One interesting approach could involve modeling these parameters through basis-splines, which would provide greater flexibility in capturing variations in bias-patterns both over years and across countries.

Similarly, the calculation of the bias-adjustment multiplier in the indirect estimation method could be made significantly more flexible. Specifically, instead of relying solely on information from Sweden when a country lacks accurate data for a given year, a more refined approach would involve modeling the observed adjustment multipliers from both the country in question and Sweden for the years in which both have accurate population. This could be done using generalized additive models (GAM), which would allow to flex-

ibly model the variation in bias-adjustment multiplier over time. This could be achieved by modeling the bias-adjustment multiplier as a function of smooth time trends, such as P-splines (Eilers et al., 2006), along with country-specific effects. This model could be then used to predict the multiplier values for the years in which the country of interest lacks accurate population data.

Furthermore, a noteworthy extension could involve the application of a similar modeling procedures to other online genealogical data as well as to other types of historical data, including parish records. A valuable contribution in historical demography would be to use data from Italian parish records in combination with more trustworthy data sources, such as the Human Fertility Collection and censuses, to obtain time series of TFR estimates with the corresponding credible intervals across various Italian subregions.

5.2.3 Male and female fertility estimation with subnational population data

In Chapter 4, a Bayesian modeling framework was developed to allow for the indirect estimation of the subnational male and female TFR.

One of the main limitations of this approach is its inability to produce fertility indicators beyond the TFR. While the TFR is widely used to assess the overall fertility of a country, it provides only a partial view of the reproductive behavior within the country. As a quantum measure, the TFR denotes the average number of children a woman or a man is expected to bear throughout her (his) reproductive ages.

Conditional on the availability of subnational birth counts classified by parental ages, a promising extension would be the development of a Bayesian framework that directly models the number of births across different maternal and paternal ages at a subnational level. Modeling the subnational number of births classified by maternal and paternal ages would allow to obtain subnational age- and sex-specific fertility estimates that could be leveraged to compute a wide range of fertility measures at a subnational level. Furthermore, these age-specific fertility rates could be employed as input to perform subnational projections of kinship networks using the matrix-oriented formal demographic method by

Caswell and Song (2021).

Another limitation concerns the direct modeling of subnational person-years. While many high-income countries provide subnational life tables, such data are usually unavailable in middle- and low-income countries. Researchers willing to apply the proposed model to these contexts may need alternative approaches to estimate person-years, such as modeling subnational mortality rates through a principal component regression on national mortality rates, as previously done by Alexander and Alkema (2022), and deriving the corresponding person-year estimates using standard life table relationships.

Additionally, the proposed model does not allow for the TFR estimation for population subgroups defined by socio-economic characteristics such as education level, household income, or religious affiliation. Future work could incorporate these dimensions, offering the opportunity to investigate how male fertility behaviors may change across various socio-economic subgroups.

Furthermore, the modeling framework presented in Chapter 4 offers novel opportunities to examine sex-specific differences in TFR estimates, particularly in regions where male and female fertility patterns diverge. For example, in certain Sub-Saharan African countries, where polygynous unions are common, men tend to have more children than women over their reproductive lifetimes (Schoumaker, 2017). Conversely, in countries such as India and China, a strong preference for sons has led to sex ratios at births above the biological value of 1.05 (Alkema et al., 2014; Kashyap et al., 2015), resulting in notable differences in male and female fertility in the opposite direction.

Another valuable application would be to investigate how subnational variations in male and female TFR may correlate with macro-level socio-economic indicators. For instance, an interesting application could involve investigating whether regions with male-to-female TFR ratios significantly above or below 1 exhibit greater gender inequality compared to regions where these ratios are close to 1.

Bibliography

Abbasi-Shavazi, M. J. and McDonald, P. (2000). Fertility and multiculturalism: Immigrant fertility in Australia, 1977–1991. *International Migration Review*, 34(1):215–242.

Abel, G., Bijak, J., Findlay, A., McCollum, D., and Wiśniowski, A. (2013). Forecasting environmental migration to the United Kingdom: an exploration using Bayesian models. *Population and Environment*, 35(2):183–203.

Abramitzky, R., Boustan, L., Eriksson, K., Feigenbaum, J., and Pérez, S. (2021). Automated linking of historical data. *Journal of Economic Literature*, 59(3):865–918.

Abramitzky, R., Mill, R., and Pérez, S. (2020). Linking individuals across historical sources: A fully automated approach. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 53(2):94–111.

Albert, J. (2018). *Package LearnBayes*. R package version 2.7.3.

Alburez-Gutierrez, D., Aref, S., and Gil-Clavel, S. (2019). Demography in the digital era: New data sources for population research. In Arbia, G., Peluso, S., Pinna, A., et al., editors, *Book of Short Papers SIS 2019*, pages 22–33. Pearson.

Alburez-Gutierrez, D., Barban, N., Caswell, H., Kolk, M., Margolis, R., Smith-Greenaway, E., Song, X., Verdery, A. M., and Zagheni, E. (2022). Kinship, demography, and inequality: Review and key areas for future development. *Working Paper in SocArXiv*.

Alexander, M. and Alkema, L. (2018). Global estimation of neonatal mortality using a Bayesian hierarchical splines regression model. *Demographic Research*, 38:335–372.

Alexander, M. and Alkema, L. (2022). A Bayesian cohort component projection model to estimate women of reproductive age at the subnational level in data-sparse settings. *Demography*, 59(5):1713–1737.

Alexander, M., Polimis, K., and Zagheni, E. (2020). Combining social media and survey data to nowcast migrant stocks in the United States. *Population Research and Policy Review*, 41:1–28.

Alexander, M. and Root, L. (2022). Competing effects on the average age of infant death. *Demography*, 59(2):587–605.

Alexander, M., Zagheni, E., and Barbieri, M. (2017). A flexible Bayesian model for estimating subnational mortality. *Demography*, 54(6):2025–2041.

Alkema, L., Chao, F., You, D., Pedersen, J., and Sawyer, C. C. (2014). National, regional, and global sex ratios of infant, child, and under-5 mortality and identification of countries with outlying ratios: a systematic assessment. *The Lancet Global Health*, 2(9):521–530.

Alkema, L. and New, J. R. (2014). Global estimation of child mortality using a Bayesian B-spline bias-reduction model. *The Annals of Applied Statistics*, 8(9):2122–2149.

Alkema, L., Raftery, A. E., Gerland, P., Clark, S. J., Pelletier, F., Buettner, T., and Heilig, G. K. (2011). Probabilistic projections of the total fertility rate for all countries. *Demography*, 48(3):815–839.

Aparicio Castro, A., Wiśniowski, A., and Rowe, F. (2024). A Bayesian approach to estimate annual bilateral migration flows for South America using census data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 187(2):410–435.

Arpino, B., Le Moglie, M., and Mencarini, L. (2022). What tears couples apart: An analysis of union dissolution in Germany with machine learning. *Demography*, 59(1):161–186.

Balbo, N., Billari, F. C., and Mills, M. (2013). Fertility in advanced societies: A review of research. *European Journal of Population*, 29(1):1–38.

Batyra, E., Leone, T., and Myrskylä, M. (2023). Forecasting of cohort fertility by educational level in countries with limited data availability: The case of Brazil. *Population Studies*, 77(2):179–195.

Becker, R. A., Wilks, A. R., and Brownrigg, R. (2022). Mapdata: Extra Map Databases. R package version 2.3.1.

Bergman, M. (1967). The potato blight in the Netherlands and its social consequences (1845–1847). *International Review of Social History*, 12(3):390–431.

Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, 82(4):733–746.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20.

Bijak, J. (2008). Bayesian methods in international migration forecasting. In Raymer, J. and Willekens, F., editors, *International Migration in Europe: Data, Models and Estimates*, pages 255–282. John Wiley & Sons, Chichester, UK.

Bijak, J. (2022). *Towards Bayesian model-based demography: Agency, complexity and uncertainty in migration studies*. Springer Nature.

Bijak, J. and Bryant, J. (2016). Bayesian demography 250 years after Bayes. *Population Studies*, 70(1):1–19.

Bijak, J. and Wiśniowski, A. (2010). Bayesian forecasting of immigration to selected european countries by using expert knowledge. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 173(4):775–796.

Billari, F. C. and Zagheni, E. (2017). Big data and population processes: A revolution? In Petrucci, A. and Verde, R., editors, *SIS 2017. Statistics and Data science: New challenges, new generations. 28–30 June 2017 Florence (Italy). Proceedings of the Conference of the Italian Statistical Society*, pages 167–178, Firenze, Italy. Firenze University Press.

Blanc, G. (2024a). The cultural origins of the demographic transition in France. Working paper, Manchester, UK: University of Manchester.

Blanc, G. (2024b). Demographic transitions, rural flight, and intergenerational persistence: Evidence from crowdsourced genealogies. Working paper, Manchester, UK: University of Manchester.

Blayo, Y. (1975). La mortalité en France de 1740 à 1829. *Population (French Edition)*, 30:123–142.

Booth, H. (2006). Demographic forecasting: 1980 to 2005 in review. *International journal of forecasting*, 22(3):547–581.

Breschi, M., Kurosu, S., and Michel, O., editors (2003). *The Own-Children Method of Fertility Estimation: Applications in Historical Demography*. Editrice Universitaria Udine, Udine.

Bryant, J. and Zhang, J. L. (2018). *Bayesian demographic estimation and forecasting*. Chapman and Hall/CRC.

Bryant, J. R. and Graham, P. J. (2013). Bayesian demographic accounts: Subnational population estimation using multiple data sources. *Bayesian Analysis*, 8(3):591–622.

Calderón Bernal, L., Alburez-Gutierrez, D., and Zagheni, E. (2023). Analyzing biases in genealogies using demographic microsimulation. MPIDR Working Paper Series WP-2023-034, Max Planck Institute for Demographic Research, Rostock.

Camarda, C. G. (2012). Mortalitysmooth: An R package for smoothing Poisson counts with P-splines. *Journal of Statistical Software*, 50(1):1–24.

Caswell, H. and Song, X. (2021). The formal demography of kinship iii. *Demographic Research*, 45(16):517–546.

Cesare, N., Lee, H., McCormick, T., Spiro, E., and Zagheni, E. (2018). Promises and pitfalls of using digital traces for demographic research. *Demography*, 55(5):1979–1999.

Chesnais, J.-C. (1992). *The demographic transition: stages, patterns, and economic implications: a longitudinal study of sixty-seven countries covering the period 1720-1984*. Oxford University Press.

Chong, M., Alburez-Gutierrez, D., Alexander, M., and Zagheni, E. (2022). Identifying and correcting bias in big crowd-sourced online genealogies. MPIDR Working Paper Series WP-2022-005, Max Planck Institute for Demographic Research, Rostock.

Chong, M. Y. and Alexander, M. (2024). Estimating the timing of stillbirths in countries worldwide using a bayesian hierarchical penalized splines regression model. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 73(4):902–920.

Coale, A. J. and Zelnik, M. (1961). *New estimates of fertility and population in the United States*, volume 2198. Princeton University Press.

Colasurdo, A. and Omenti, R. (2024). Using online genealogical data for demographic research: An empirical examination of the FamiLinx database. *Demographic Research*, 51(41):1299–1350.

Coleman, D. A. (1995). *Male fertility trends in industrial countries: Theories in search of some evidence*. International Union for the Scientific Study of Population.

Comolli, C. L. (2017). The fertility response to the great recession in Europe and the United States: Structural economic conditions and perceived economic uncertainty. *Demographic Research*, 36(51):1549–1600.

Corti, G., Minardi, S., and Barban, N. (2024). Trends in assortative mating in the United States, 1700–1910. Evidence from FamiLinx data. *The History of the Family*, pages 1–21.

Cozzani, M., Minardi, S., Corti, G., and Barban, N. (2023). Birth month and adult lifespan: A within-family, cohort, and spatial examination using FamiLinx data in the United States (1700–1899). *Demographic Research*, 49(9):201–218.

Cummins, N. (2017). Lifespans of the European elite, 800–1800. *The Journal of Economic History*, 77(2):406–439.

Daponte, B. O., Kadane, J. B., and Wolfson, L. J. (1997). Bayesian demography: projecting the Iraqi Kurdish population, 1977–1990. *Journal of the American Statistical Association*, 92(440):1256–1267.

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., and Bodik, R. (2017). Programming with models: Writing statistical algorithms for general model structures with NIMBLE. *Journal of Computational and Graphical Statistics*, 26(2):403–413.

Disney, G., Wiśniowski, A., Forster, J. J., Smith, P. W., and Bijak, J. (2015). Evaluation of existing migration forecasting methods and models. Report for the migration advisory committee: Commissioned research, ESRC Centre for Population Change, University of Southampton.

Dubuc, S. (2009). Application of the own-children method for estimating fertility by ethnic and religious groups in the UK. *Journal of Population Research*, 26:207–225.

Dudel, C., Cheng, Y.-h. A., and Klüsener, S. (2023). Shifting parental age differences in high-income countries: Insights and implications. *Population and Development Review*, 49(4):879–908.

Dudel, C. and Klüsener, S. (2016). Estimating male fertility in Eastern and Western Germany since 1991: A new lowest low? *Demographic Research*, 35(53):1549–1560.

Dudel, C. and Klüsener, S. (2021). Male–female fertility differentials across 17 high-income countries: Insights from a new data resource. *European Journal of Population*, 37(2):417–441.

Eilers, P. H., Currie, I. D., and Durbán, M. (2006). Fast and compact smoothing on large multidimensional grids. *Computational Statistics & Data Analysis*, 50(1):61–76.

Ellison, J., Berrington, A., Dodd, E., and Forster, J. J. (2024). Combining individual-and population-level data to develop a bayesian parity-specific fertility projection model. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 73(2):275–297.

Ellison, J., Dodd, E., and Forster, J. J. (2020). Forecasting of cohort fertility under a hierarchical bayesian approach. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 183(3):829–856.

Garrett, E., Reid, A., and Szreter, S. (2010). Fertility and child mortality in their household setting: a variety of perspectives from UK censuses, 1861-1911. *Popolazione e Storia*, 11(2):59–82.

Gavrilova, N. S. and Gavrilov, L. A. (2007). Search for predictors of exceptional human longevity: Using computerized genealogies and internet resources for human longevity studies. *North American Actuarial Journal*, 11(1):49–67.

Gay, V., Gobbi, P., and Goñi, M. (2023). Revolutionary transition: Inheritance change and fertility decline. HAL Working Paper Series hal-04285818, HAL.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

Girosi, F. and King, G. (2008). *Demographic forecasting*. Princeton University Press.

Grabill, W. R. and Cho, L. J. (1965). Methodology for the measurement of current fertility from population data on young children. *Demography*, 2:50–73.

Graunt, J. (1662). *Natural and political observations made upon the bills of mortality*. Martyn, London.

Greene, M. E. and Biddlecom, A. E. (2000). Absent and problematic men: Demographic accounts of male reproductive roles. *Population and development review*, 26(1):81–115.

Grigorieva, O., Jasilioniene, A., Jdanov, D., Grigoriev, P., Sobotka, T., Zeman, K., and Shkolnikov, V. (2015). Methods protocol for the human fertility collection.

Hacker, J. D. (2010). Decennial life tables for the white population of the United States, 1790–1900. *Historical methods*, 43(2):45–79.

Hacker, J. D. and Roberts, E. (2019). Fertility decline in the United States, 1850-1930: New evidence from complete-count datasets. *Annales de démographie historique*, 138(2):143–177.

Hauer, M., Baker, J., and Brown, W. (2013). Indirect estimates of total fertility rate using child woman/ratio: A comparison with the bogue-palmore method. *plos one*, 8(6):e67226.

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

Henry, L. (1968). Historical demography. *Daedalus*, 97(2):385–396.

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.

Hilbert, M. and López, P. (2011). The world's technological capacity to store, communicate, and compute information. *Science*, 332(6025):60–65.

Hill, L. E. and Johnson, H. P. (2002). *Understanding the future of Californians' fertility: the role of immigrants*. Public Policy Instit. of CA.

Hollingsworth, T. H. and Hollingsworth, T.-T. (1976). Genealogy and historical demography. *Annales de demographie historique*, pages 167–170.

Hsu, C.-H., Posegga, O., Fischbach, K., and Engelhardt, H. (2021). Examining the trade-offs between human fertility and longevity over three centuries using crowdsourced genealogy data. *PloS one*, 16(8):e0255528.

Jaadla, H., Reid, A., Garrett, E., Schürer, K., and Day, J. (2020). Revisiting the fertility transition in England and Wales: The role of social class and migration. *Demography*, 57(4):1543–1569.

Jasilioniene, A., Jdanov, D. A., Sobotka, T., Andreev, E. M., Zeman, K., Shkolnikov, V. M., Goldstein, J. R., Philipov, D., and Rodriguez, G. (2015). Methods protocol for the human fertility database. *Rostock: Max Planck Institute for Demographic Research*.

Jasilioniene, A., Sobotka, T., Jdanov, D. A., Zeman, K., Kostova, D., Andreev, E. M., Grigoriev, P., and Shkolnikov, V. M. (2016). Data resource profile: the human fertility database. *International Journal of Epidemiology*, 45(4):1077–1078.

Johnson, H. P. and Li, Q. (2007). *Birth rates in California*, volume 9. Public Policy Institute of California San Francisco, CA.

Kaplanis, J., Gordon, A., Shor, T., Weissbrod, O., Geiger, D., Wahl, M., Gershovits, M., Markus, B., Sheikh, M., Gymrek, M., et al. (2018). Quantitative analysis of population-scale family trees with millions of relatives. *Science*, 360(6385):171–175.

Kasakoff, A. B. and Adams, J. W. (1995). The effect of migration on ages at vital events: a critique of family reconstitution in historical demography. *European Journal of Population/Revue Européenne de Démographie*, pages 199–242.

Kashyap, R. (2021). Has demography witnessed a data revolution? Promises and pitfalls of a changing data ecosystem. *Population Studies*, 75(sup1):47–75.

Kashyap, R., Esteve, A., and García-Román, J. (2015). Potential (mis) match? Marriage markets amidst sociodemographic change in India, 2005–2050. *Demography*, 52(1):183–208.

Kashyap, R. and Zagheni, E. (2023). Leveraging digital and computational demography for policy insights. In Springer, editor, *Handbook of Computational Social Science for Policy*, pages 327–344. Springer.

Keilman, N., Tymicki, K., and Skirbekk, V. (2014). Measures for human reproduction should be linked to both men and women. *International Journal of Population Research*, 2014(1):908385.

Keyfitz, N., Caswell, H., et al. (2005). *Applied mathematical demography*, volume 47. Springer.

Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in medicine*, 19(17-18):2555–2567.

Kohler, H.-P., Billari, F. C., and Ortega, J. A. (2002). The emergence of lowest-low fertility in Europe during the 1990s. *Population and development review*, 28(4):641–680.

Leasure, D. R., Jochem, W. C., Weber, E. M., Seaman, V., and Tatem, A. J. (2020). National population mapping from sparse survey data: A hierarchical Bayesian modeling framework to account for uncertainty. *Proceedings of the National Academy of Sciences*, 117(39):24173–24179.

Lee, R. (2002). The demographic transition: three centuries of fundamental change. *Journal of economic perspectives*, 17(4):167–190.

Lesthaeghe, R. (2014). The second demographic transition: A concise overview of its development. *Proceedings of the National Academy of Sciences*, 111(51):18112–18115.

Mare, R. D. (2011). A multigenerational view of inequality. *Demography*, 48(1):1–23.

Minardi, S., Corti, G., and Barban, N. (2024). Historical patterns in the intergenerational transmission of lifespan and longevity: A research note on US cohorts born between 1700 and 1900. *Demography*, 61(4):979–994.

Otterstrom, S. M. and Bunker, B. E. (2013). Genealogy, migration, and the intertwined geographies of personal pasts. *Annals of the Association of American Geographers*, 103(3):544–569.

Paget, W. J. and Timæus, I. M. (1994). A relational gompertz model of male fertility: Development and assessment. *Population Studies*, 48(2):333–340.

Pison, G. (2012). France and Germany: a history of criss-crossing demographic curves. *Population Societies*, 487(3):1–4.

Pojman, E., Mwedzi, D. E., Bucaro, O. O., Zhang, S., Chong, M., Alexander, M., and Alburez-Gutierrez, D. (2023). Leaving for life: Using online crowd-sourced genealogies to estimate the migrant mortality advantage for the United Kingdom and Ireland during the 18th and 19th centuries. MPIDR Working Paper Series WP-2023-050, Max Planck Institute for Demographic Research, Rostock.

Poole, D. and Raftery, A. E. (2000). Inference for deterministic simulation models: the Bayesian melding approach. *Journal of the American Statistical Association*, 95(452):1244–1255.

Post, W., Van Poppel, F., Van Imhoff, E., and Kruse, E. (1997). Reconstructing the extended kin-network in the Netherlands with genealogical data: Methods, problems, and results. *Population Studies*, 51(3):263–278.

Pozzi, L. and Fariñas, D. R. (2015). Infant and child mortality in the past. *Annales de démographie historique*, 129(1):55–75.

Raftery, A. E., Alkema, L., and Gerland, P. (2014). Bayesian population projections for the United Nations. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 29(1):58–68.

Raftery, A. E., Chunn, J. L., Gerland, P., and Ševčíková, H. (2013). Bayesian probabilistic projections of life expectancy for all countries. *Demography*, 50(3):777–801.

Raftery, A. E., Li, N., Ševčíková, H., Gerland, P., and Heilig, G. K. (2012). Bayesian probabilistic population projections for all countries. *Proceedings of the National Academy of Sciences*, 109(35):13915–13921.

Rampazzo, F., Bijak, J., Vitali, A., Weber, I., and Zagheni, E. (2021). A framework for estimating migrant stocks using digital traces and survey data: An application in the United Kingdom. *Demography*, 58(6):2193–2218.

Ratcliffe, A. A., Hill, A. G., and Walraven, G. (2000). Separate lives, different interests: male and female reproduction in the Gambia. *Bulletin of the World Health Organization*, 78(5):570–579.

Reid, A., Jaadla, H., Garrett, E., and Schürer, K. (2020). Adapting the own children method to allow comparison of fertility between populations with different marriage regimes. *Population Studies*, 74(2):197–218.

Ruggles, S., Roberts, E., Sarkar, S., and Sobek, M. (2011). The North Atlantic population project: Progress and prospects. *Historical methods*, 44(1):1–6.

Scalone, F. and Dribe, M. (2017). Testing child-woman ratios and the own-children method on the 1900 Sweden census: Examples of indirect fertility estimates by socioeconomic status in a historical population. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 50(1):16–29.

Schmertmann, C., Zagheni, E., Goldstein, J. R., and Myrskylä, M. (2014). Bayesian forecasting of cohort fertility. *Journal of the American Statistical Association*, 109(506):500–513.

Schmertmann, C. P., Cavenaghi, S. M., Assunção, R. M., and Potter, J. E. (2013). Bayes plus Brass: estimating total fertility for many small areas from sparse census data. *Population Studies*, 67(3):255–273.

Schmertmann, C. P. and Gonzaga, M. R. (2018). Bayesian estimation of age-specific mortality and life expectancy for small areas with defective vital records. *Demography*, 55(4):1363–1388.

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

Schoumaker, B. (2017). Measuring male fertility rates in developing countries with demographic and health surveys: An assessment of three methods. *Demographic Research*, 36(1):803–850.

Schoumaker, B. (2019). Male fertility around the world and over time: How different is it from female fertility? *Population and Development Review*, 45(3):459–487.

Schubert, H.-A. and Dudel, C. (2024). Same but different? Male-female fertility differences at the subnational level over time and across countries. In *Paper presented at European Population Conference 2024 in Edinburgh, UK*.

Ševčíková, H., Raftery, A. E., and Gerland, P. (2018). Probabilistic projection of subnational total fertility rates. *Demographic research*, 38:1843–1884.

Shkolnikov, V. M., Boe, J. R. W., and Gellers-Barkmann, S. (2017). Methods protocol for the human life-table database.

Song, X. and Campbell, C. D. (2017). Genealogical microdata and their significance for social science. *Annual Review of Sociology*, 43(1):75–99.

Spoorenberg, T. and Dutreuilh, C. (2007). Quality of age reporting: extension and application of the modified whipple's index. *Population*, 62(4):729–741.

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

Stockwell, E. G. and Wicks, J. W. (1974). Age heaping in recent national censuses. *Social Biology*, 21(2):163–167.

Tragaki, A. and Bagavos, C. (2014). Male fertility in Greece: Trends and differentials by education level and employment status. *Demographic Research*, 31(6):137–160.

UNDESA (1983). *Indirect techniques for demographic estimation*, volume 10. New York: United Nations.

UNDESA (2024). World population prospects 2024: Data sources. Technical report, United Nations, Department of Economic and Social Affairs, Population Division.

United Nations (2016). A review of key concepts: Coverage completeness (UN expert group meeting, 3–4 november 2016). Technical report, New York: United Nations.

U.S. Census Bureau (2017). Subnational population by sex, age, and geographic area. Data set. Available from https://www.census.gov/geographies/mapping-files/time-series/demo/international-programs/subnationalpopulation.html.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Voutilainen, M., Helske, J., and Högmander, H. (2020). A Bayesian reconstruction of a historical population in Finland, 1647–1850. *Demography*, 57(3):1171–1192.

Wachter, K. W. (2014). *Essential demographic methods*. Harvard University Press.

Weir, D. R. (1984). Fertility transition in rural France, 1740–1829. *The Journal of Economic History*, 44(2):612–614.

Wheldon, M. C., Raftery, A. E., Clark, S. J., and Gerland, P. (2013). Reconstructing past populations with uncertainty from fragmentary data. *Journal of the American Statistical Association*, 108(501):96–110.

WHO (2022). World health organization mortality database.

Wilmoth, J., Zureick, S., Canudas-Romo, V., Inoue, M., and Sawyer, C. (2012). A flexible two-dimensional mortality model for use in indirect estimation. *Population Studies*, 66(1):1–28.

Wilmoth, J. R., Andreev, K., Jdanov, D., Glei, D. A., Boe, C., Bubenheim, M., Philipov, D., Shkolnikov, V., and Vachon, P. (2007). Methods protocol for the human mortality database. *University of California, Berkeley, and Max Planck Institute for Demographic Research, Rostock. URL: http://mortality. org [version 31/05/2007]*, 9:10–11.

Woods, R. (2000). *The demography of Victorian England and Wales*, volume 35. Cambridge University Press.

Wrigley, E. A. (1981). The prospects for population history. *The Journal of Interdisciplinary History*, 12(2):207–226.

Wrigley, E. A. (1985). The fall of marital fertility in nineteenth-century France: Exemplar or exception? *European Journal of Population*, 1(1):31–60.

Wrigley, E. A., Davies, R. S., Oeppen, J. E., and Schofield, R. S. (1997). *English population history from family reconstitution 1580-1837*. Cambridge University Press.

Yu, C. C., Ševčíková, H., Raftery, A. E., and Curran, S. R. (2023). Probabilistic county-level population projections. *Demography*, 60(3):915–937.

Zhang, L. (2010). *Male fertility patterns and determinants*. Springer, Dordrecht.

Zhao, Z. (2001). Chinese genealogies as a source for demographic research: A further assessment of their reliability and biases. *Population Studies*, 55(2):181–193.

Appendix A

Appendix A: Supplemental Information

A.1 Details on Demographic Variables in FamiLinx

Table A.1: Number of births and deaths by country

Country	Number of Births	Number of Deaths
USA	2,479,761	2,122,063
UK	936,188	324,630
NORWAY	468,391	281,471
SWEDEN	359,999	222,005
NETHERLANDS	301,079	184,430
GERMANY	298,271	137,164
ESTONIA	267,137	121,194
CANADA	248,248	185,322
DENMARK	180,569	97,780
FRANCE	177,715	112,167
POLAND	112,382	58,575
FINLAND	111,272	73,401
AUSTRALIA	94,687	90,788
SPAIN	81,812	24,752
IRELAND	69,739	17,991
BELGIUM	67,638	46,338
INDIA	67,132	52,773
SWITZERLAND	55,116	17,388
SOUTH AFRICA	50,815	44,364
RUSSIA	49,605	24,145
ITALY	36,962	16,487
CZECH REPUBLIC	24,237	13,971
NEW ZEALAND	20,368	23,738
ISRAEL	10,890	35,030

Table A.2: Absolute frequencies and percentage of missing and non-missing values in relevant demographic variables in the complete sample and in the analytical sample.

Variable	Complete Sample	Analytical Sample
Sample Size	86,124,644	7,618,651
Gender		
Missing	14,071,200 (16.34%)	4,708 (0.06%)
Male	37,998,030 (44.12%)	4,108,522 (53.93%)
Female	34,055,414 (39.54%)	3,505,421 (46.01%)
Birth Date Information		
Missing	52,405,914 (60.85%)	677,215 (8.89%)
Only year	13,692,092 (15.90%)	2,389,882 (31.37%)
Year and Month	849,377 (0.99%)	195,874 (2.57%)
Complete Date	19,177,261 (22.27%)	4,355,680 (57.17%)
Death Date Information		
Missing	64,383,957 (74.77%)	2,656,270 (34.87%)
Only year	6,736,492 (7.82%)	1,143,731 (15.01%)
Year and Month	853,888 (0.99%)	217,310 (2.85%)
Complete Date	14,150,307 (16.43%)	3,601,340 (47.27%)
Birth Location Information		
Missing	70,464,808 (81.82%)	1,048,638 (13.76%)
Reported	15,659,836 (18.18%)	6,570,013 (86.24%)
Death Location Information		
Missing	74,861,173 (86.92%)	3,290,684 (43.19%)
Reported	11,263,471 (13.08%)	4,327,967 (56.81%)
Parent/Child Linkage		
Missing	47,172,309 (54.77%)	
At least one link	38,952,335 (45.23%)	7,618,651 (100.00%)

A.2 Details on the Estimation of the Smoothed Mortality Rates

To estimate life expectancy at age 30 from online genealogies, we rely on the R package MortalitySmooth developed by Camarda (2012) that allows to smooth mortality rates over years and ages.

We consider mortality experienced by individuals, who were born and died in Sweden, during the historical period 1751-1900. To obtain smoothed estimates of mortality rates by year and age, we model the number of deaths in a year t at an age x Y_{xt} in Sweden as

a Poisson distribution.

$$Y_{xt} \sim \mathcal{P}(E_{xt} \cdot \mu_{xt})$$

$$x = 30, \ldots, 80$$

$$t = 1751, \ldots, 1900$$

E_{xt} indicates the number of exposed Swedish individuals in year t and age x and μ_{xt} denotes the risk of death for Swedish individuals aged x in year t.

For the performance of the mortality analysis, death counts, exposure and mortality risks by year and age are arranged in rectangular matrices, called Y, M and E, in which rows represent ages and columns refer to years. The smoothness is achieved by incorporating two-dimensional P-splines. Specifically, we model the mean of the Poisson distribution of the number of deaths as follows.

$$\log\left(E(\mathbf{Y})\right) = \log\left(\mathbf{E}\right) + \log\left(\mathbf{M}\right) = \log\left(\mathbf{E}\right) + \mathbf{B}_y\mathbf{A}\mathbf{B}_a^{'}$$

In the model, the B-splines spaced over the ages are stored in the regression matrix $\mathbf{B}_a$ of dimension $k_a \times k_a$. The B-splines spaced over the years are stored in the regression $\mathbf{B}_y$ of dimension $k_y \times k_y$. Both $\mathbf{B}_a$ and $\mathbf{B}_y$ have an associated set of regression coefficients. Note that the numbers k_a and k_y indicate the number of B-splines chosen over the ages (k_a) and years (k_y). Following the guidelines by Camarda (2012), we chose B-splines that are equally spaced over the years and the ages. The rows of the matrix $\mathbf{A}$ of dimension $k_a \times k_y$ denote the regression coefficients for $\mathbf{B}_a$, whereas its columns indicate the regression coefficients for $\mathbf{B}_y$. The estimation of the regression parameters is performed via Iterative Regression Weighted Least Squares (IRWLS). We set the diagonal matrix of weights required for this estimation procedure to be equal to be the identity.

To reduce the number of parameters in the model, we can choose the number of B-splines with an additional two-dimensional penalty P on the regression coefficients.

$$\mathbf{P} = \lambda_a\left(I_{k_y} \otimes \mathbf{D}_a^{'}\mathbf{D}_a\right) + \lambda_y\left(I_{k_a} \otimes \mathbf{D}_y^{'}\mathbf{D}_y\right)$$

Where $\lambda_a$and λ_y are the smoothing parameters used for the ages and the years. $\mathbf{D}_a$ and $\mathbf{D}_y$ are the difference matrices. I_{k_y} and I_{k_a} are identity matrices of dimension k_y and k_a respectively. The symbol $\otimes$ stands for the Kronecker product. The optimal values for $\lambda_a$and λ_y are chosen so that either Bayesian Information Criterion (BIC) or Akaike Information Criterion are minimized. To smooth mortality rates in R, we employed the function mort2Dsmooth(x,y,Z,offset) from the R package MortalitySmooth by Camarda (2012). The function requires the following arguments:

(a) A vector of ages named x (in our application x= 30,. . . ,80)

(b) A vector of years named y (in our application y=1751,. . . ,1900)

(c) A matrix of death counts over ages (rows) and years (columns) named Z (matrix $\mathbf{Y}$ in the model notation)

(d) A matrix of logged population counts over ages (rows) and years (columns) named offset (matrix $\mathbf{E}$ with log-transformed entries in the model notations)

Concerning the remaining arguments, we opted for the default options. Optional arguments include:

(a) The degree of the polynomials for the construction of B-splines (q), whose default option is set to be q=3 (necessary for the construction of matrix $\mathbf{B}$)

(b) The order of the differences for the penalty matrix (d), whose default option is set to be d=2 (necessary for the specification of penalty matrices $\mathbf{D}_a$ and $\mathbf{D}_y$)

(c) A matrix of weights over the ages and years (W) which is set by default to be equal to the identity matrix (necessary for the specification of the diagonal matrix of weights to be used in the estimation of the regression coefficients)

(d) The selection of the smoothing parameters ($\lambda_a$and λ_y in the model notation) is carried out by default using Bayesian Information Criterion (BIC). Alternative selection criteria can be specified via the option method

As part of the output, the function *mort2Dsmooth* provides a matrix of smoothed mortality rates over the ages and the years. Exploiting standard life table relationships, these smoothed mortality rates by ages and years are then used to obtain smoothed estimates of life expectancy at birth and at age 30.

A.3 Additional Plots on Quality and Completeness of Demographic Information and Regression Tables

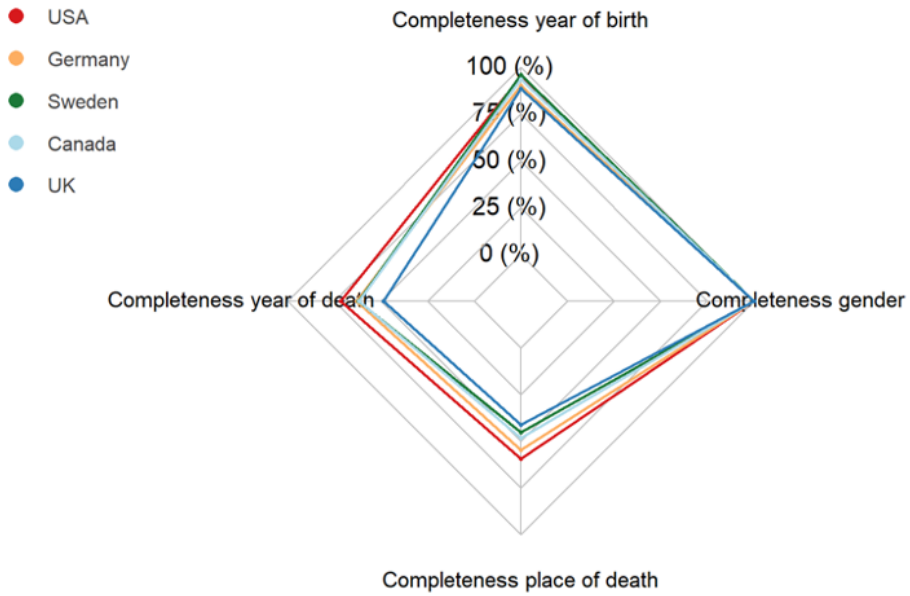


Figure A.1: Percentage of non-missing values for 4 relevant demographic variables in the dataset, by country of birth of the focal individual.

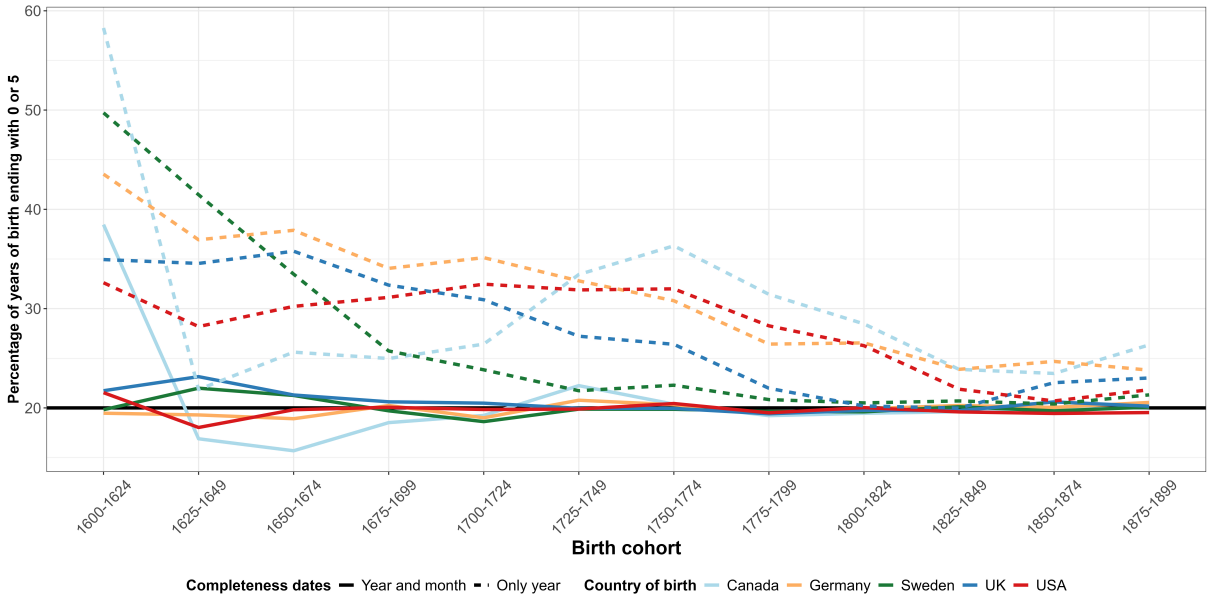


Figure A.2: Percentage of years of birth ending with 0 or 5, by country of birth and birth cohort.

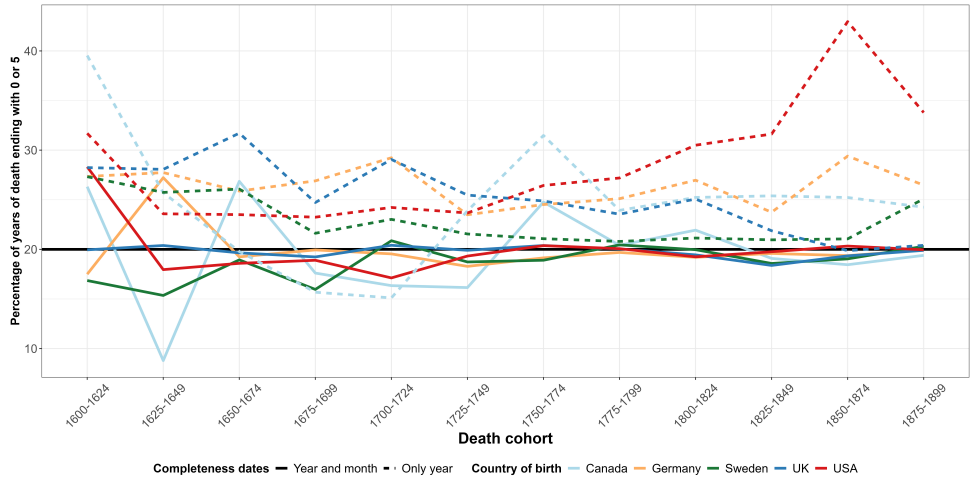


Figure A.3: Percentage of years of death ending with 0 or 5, by country of birth and death cohort.

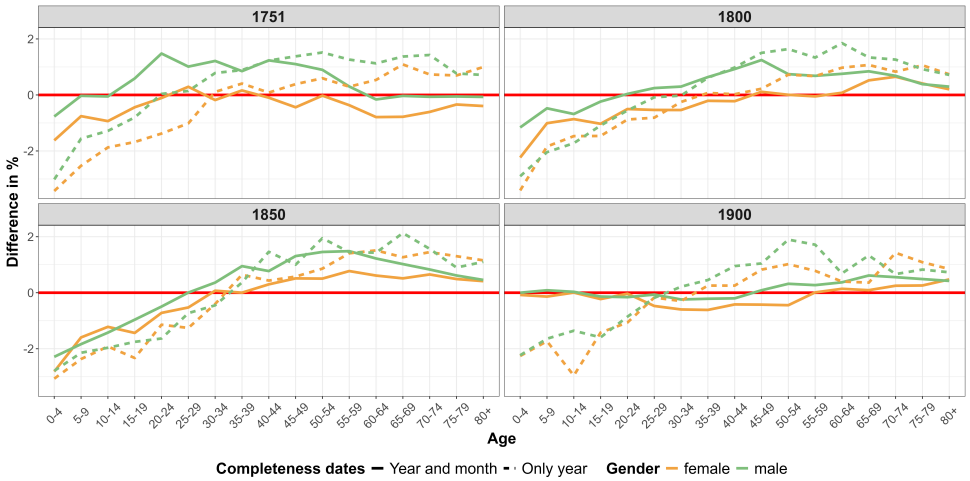


Figure A.4: Difference between the age-sex distribution in percentage between the Swedish population from FamiLinx by quality level (precise birth and death dates against at least one non-precise date) and the registered Swedish population over the years 1751, 1800, 1850 and 1900.

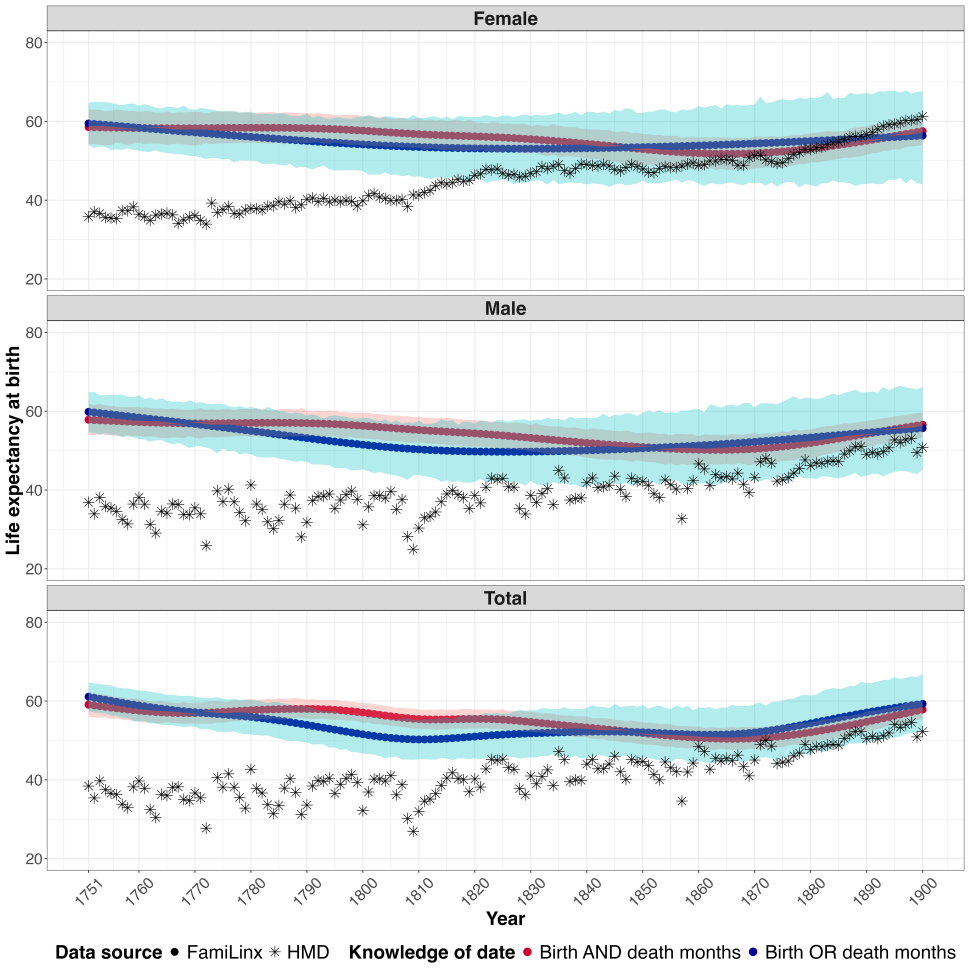


Figure A.5: Life expectancy at birth in Sweden for the historical period (1751-1900) by sex and quality level (precise birth and death dates against at least one non-precise date) in FamiLinx and Swedish life expectancy at birth from the HMD.

Table A.3: Coefficients of the negative binomial regression models to test the association in terms of completeness, by type of relative and demographic variable.

Demographic Variable	Effect	Children	Parents	Grandparents	Siblings	Aunts & Uncles	Cousins	Grandchildren
Birth Year	Intercept	-0.9420 (0.0015)	-2.1268 (0.0021)	-0.6295 (0.0011)	-0.3781 (0.0013)	0.4618 (0.0011)	0.7929 (0.0016)	-0.3481 (0.0020)
	Yes	0.7193 (0.0015)	0.7448 (0.0009)	0.4147 (0.0010)	1.0749 (0.0013)	0.4516 (0.0011)	0.6863 (0.0016)	0.4259 (0.0020)
	Nr of Relatives	0.2361 (0.0001)	0.9680 (0.0010)	0.3507 (0.0002)	0.1412 (0.0000)	0.1139 (0.0000)	0.0510 (0.0000)	0.1334 (0.0000)
Death Year	Intercept	-1.1837 (0.0012)	-2.0984 (0.0031)	-0.5676 (0.0008)	-0.5448 (0.0008)	0.0237 (0.0008)	0.5744 (0.0010)	-0.4926 (0.0015)
	Yes	0.6840 (0.0012)	0.4690 (0.0006)	0.1647 (0.0005)	0.6939 (0.0007)	0.3192 (0.0008)	0.4019 (0.0010)	0.2377 (0.0016)
	Nr of Relatives	0.2147 (0.0001)	1.0318 (0.0016)	0.0909 (0.0002)	0.1683 (0.0001)	0.1373 (0.0001)	0.0536 (0.0000)	0.1185 (0.0000)
Birth Country	Intercept	-1.1167 (0.0015)	-2.2153 (0.0036)	-0.9520 (0.0016)	-0.8279 (0.0010)	-0.3143 (0.0013)	0.3158 (0.0014)	-0.6374 (0.0016)
	Yes	0.4913 (0.0015)	0.5097 (0.0012)	0.3544 (0.0013)	0.4845 (0.0010)	0.6805 (0.0011)	0.7612 (0.0014)	0.3553 (0.0017)
	Nr of Relatives	0.2311 (0.0001)	0.9417 (0.0017)	0.3389 (0.0003)	0.0244 (0.0001)	0.1387 (0.0001)	0.0521 (0.0000)	0.1237 (0.0000)
Death Country	Intercept	-1.4895 (0.0014)	-2.2700 (0.0043)	-0.8599 (0.0014)	-0.9690 (0.0011)	-0.5005 (0.0012)	0.0795 (0.0012)	-0.8895 (0.0015)
	Yes	0.6586 (0.0014)	0.4457 (0.0008)	0.2254 (0.0009)	0.6926 (0.0010)	0.3858 (0.0011)	0.4506 (0.0013)	0.2517 (0.0017)
	Nr of Relatives	0.1982 (0.0002)	0.9804 (0.0022)	0.3275 (0.0004)	0.1747 (0.0001)	0.1437 (0.0001)	0.0536 (0.0000)	0.1067 (0.0001)
Overall Nr. of Relatives		14,589,754	10,633,969	11,104,591	25,042,881	21,380,793	39,633,282	16,907,137
Overall Nr. of Focal Individuals		4,323,112	5,549,757	4,173,650	4,295,590	3,107,106	2,334,853	2,932,190

Notes: All p-values are smaller than 0.001 and standard errors are shown in parentheses. Coefficients of the negative binomial regression models to test association in terms of completeness, by type of relative and demographic variable. The models are fitted considering the individuals of the analytical sample and their kinship network. For each type of relative, if the focal individual does not have that type of relative, they are omitted from the regression model. Sample size of focal individuals employed for each model and the size of their kinship network are also included.

Table A.4: Coefficients of the negative binomial regression models to test the association in terms of quality, by type of relative and demographic variable.

Demographic Variable	Effect	Children	Parents	Grandparents	Siblings	Aunts & Uncles	Cousins	Grandchildren
Birth Date	Intercept	-1.0791 (0.0041)	-2.6315 (0.0085)	-2.6170 (0.0113)	-0.8550 (0.0045)	-0.7846 (0.0082)	0.1840 (0.0098)	-0.2631 (0.0038)
	Non-missing birth month	0.6772 (0.0010)	0.7982 (0.0011)	0.6793 (0.0013)	1.2354 (0.0009)	0.7056 (0.0011)	0.7204 (0.0012)	0.3632 (0.0013)
	Nr of relatives	0.2129 (0.0001)	0.6265 (0.0026)	0.3616 (0.0005)	0.1468 (0.0001)	0.1471 (0.0001)	0.0538 (0.0000)	0.1150 (0.0001)
	Birth Period (Ref. 1800-1824)							
	1625-1649	0.1674 (0.0051)	0.1987 (0.0084)	0.3613 (0.0136)	0.0559 (0.0056)	0.1251 (0.0103)	0.1090 (0.0123)	0.0671 (0.0050)
	1650-1674	0.2303 (0.0047)	0.3442 (0.0077)	0.5792 (0.0123)	0.1015 (0.0050)	0.1822 (0.0093)	0.2555 (0.0110)	0.1307 (0.0046)
	1675-1699	0.2768 (0.0045)	0.5427 (0.0073)	0.7764 (0.0118)	0.1509 (0.0048)	0.3203 (0.0088)	0.3463 (0.0104)	0.1650 (0.0045)
	1700-1724	0.2841 (0.0044)	0.6479 (0.0071)	0.9946 (0.0115)	0.1791 (0.0047)	0.3709 (0.0085)	0.3629 (0.0102)	0.1991 (0.0044)
	1725-1749	0.2984 (0.0044)	0.7310 (0.0070)	1.1281 (0.0114)	0.1947 (0.0046)	0.4343 (0.0084)	0.3556 (0.0101)	0.2859 (0.0043)
	1750-1774	0.3455 (0.0043)	0.7460 (0.0070)	1.1999 (0.0114)	0.1885 (0.0046)	0.4569 (0.0084)	0.3605 (0.0100)	0.3661 (0.0042)
	1775-1799	0.4151 (0.0042)	0.7807 (0.0070)	1.2028 (0.0114)	0.2064 (0.0045)	0.4426 (0.0083)	0.3902 (0.0100)	0.3896 (0.0042)
	1800-1824	0.3879 (0.0042)	0.8312 (0.0069)	1.1996 (0.0113)	0.2306 (0.0045)	0.4398 (0.0083)	0.4193 (0.0099)	0.4447 (0.0042)
	1825-1849	0.4260 (0.0042)	0.9354 (0.0069)	1.2760 (0.0113)	0.2224 (0.0045)	0.5136 (0.0083)	0.4255 (0.0099)	0.5089 (0.0042)
	1850-1874	0.5380 (0.0042)	1.0499 (0.0069)	1.3887 (0.0113)	0.2335 (0.0045)	0.5537 (0.0082)	0.4420 (0.0099)	0.4438 (0.0043)
	1875-1900	0.5432 (0.0042)	1.1066 (0.0069)	1.5278 (0.0113)	0.2715 (0.0045)	0.5739 (0.0082)	0.4877 (0.0099)	0.2283 (0.0048)
Overall Nr. of Relatives		9,161,737	6,303,303	7,607,857	19,195,687	15,662,141	29,114,532	9,732,959
Overall Nr. of Focal Individuals		2,692,055	3,238,700	2,985,991	3,232,826	2,330,888	6,918,263	1,713,717
Death Date	Intercept	-0.8865 (0.0060)	-1.7394 (0.0081)	-1.2475 (0.0085)	-0.3347 (0.0068)	-0.2626 (0.0106)	0.3442 (0.0136)	-0.3524 (0.0066)
	Non-missing death month	0.5103 (0.0013)	0.5223 (0.0013)	0.4158 (0.0015)	0.5491 (0.0012)	0.3973 (0.0015)	0.3857 (0.0018)	0.3287 (0.0015)
	Nr of relatives	0.2164 (0.0001)	0.6759 (0.0026)	0.3893 (0.0006)	0.1616 (0.0001)	0.1534 (0.0001)	0.0719 (0.0001)	0.1338 (0.0001)
	Death Period (Ref. 1800-1824)							
	1625-1649	0.0752 (0.0076)	0.0221 (0.0085)	0.0497 (0.0110)	0.0205 (0.0088)	0.0109 (0.0138)	0.0514 (0.0177)	0.0698 (0.0084)
	1650-1674	0.0831 (0.0070)	0.0281 (0.0079)	0.0713 (0.0102)	0.0456 (0.0080)	0.0496 (0.0127)	0.0918 (0.0162)	0.0654 (0.0077)
	1675-1699	0.1300 (0.0065)	0.0994 (0.0072)	0.1252 (0.0094)	0.0903 (0.0074)	0.1206 (0.0116)	0.1773 (0.0148)	0.0675 (0.0072)
	1700-1724	0.1470 (0.0064)	0.1654 (0.0070)	0.1964 (0.0090)	0.1194 (0.0071)	0.1491 (0.0112)	0.2028 (0.0143)	0.1038 (0.0071)
	1725-1749	0.1690 (0.0063)	0.1938 (0.0068)	0.2516 (0.0087)	0.1193 (0.0070)	0.1684 (0.0109)	0.2177 (0.0140)	0.1666 (0.0069)
	1750-1774	0.2124 (0.0062)	0.2047 (0.0067)	0.2877 (0.0086)	0.1383 (0.0069)	0.1852 (0.0107)	0.2357 (0.0139)	0.2310 (0.0069)
	1775-1799	0.2637 (0.0061)	0.1993 (0.0067)	0.2893 (0.0086)	0.1509 (0.0069)	0.1926 (0.0107)	0.2494 (0.0138)	0.2857 (0.0068)
	1800-1824	0.3176 (0.0061)	0.2000 (0.0066)	0.2848 (0.0085)	0.1792 (0.0068)	0.2064 (0.0107)	0.2811 (0.0137)	0.3521 (0.0068)
	1825-1849	0.3530 (0.0061)	0.2144 (0.0066)	0.2745 (0.0085)	0.2040 (0.0068)	0.2377 (0.0106)	0.3103 (0.0137)	0.3930 (0.0068)
	1850-1874	0.3905 (0.0060)	0.2607 (0.0066)	0.2980 (0.0085)	0.2345 (0.0068)	0.2791 (0.0106)	0.3326 (0.0137)	0.4222 (0.0067)
	1875-1900	0.4089 (0.0060)	0.2947 (0.0066)	0.3477 (0.0085)	0.2530 (0.0068)	0.3074 (0.0106)	0.3414 (0.0136)	0.4433 (0.0067)
Overall Nr. of Relatives		3,420,063	7,531,416	3,008,568	5,514,632	4,085,603	6,918,263	4,440,454
Overall Nr. of Focal Individuals		1,379,573	3,877,896	1,567,008	1,173,828	816,322	615,066	1,051,907

Notes: Notes: all p-values are smaller than 0.001 and standard errors are shown in parentheses. Here, we used a binary response indicating whether a focal individual has at least one relative with a non-missing value in a demographic variable. Coefficients of the logistic regression models to test association in terms of completeness, by type of relative and demographic variable. The models are fitted considering the individuals of the analytical and their kinship network. For each type of relative, if the focal individual does not have that type of relative, he is omitted from the regression model. Sample size of focal individuals employed for each model and the size of their kinship network are also included.

Table A.5: Coefficients of the logistic regression models to test the association in terms of completeness, by type of relative and demographic variable.

Demographic Variable	Effect	Children	Parents	Grandparents	Siblings	Aunts & Uncles	Cousins	Grandchildren
Birth Year	Intercept	-0.6938 (0.0033)	-3.3657 (0.0076)	-1.0098 (0.0061)	-1.7600 (0.0057)	0.7014 (0.0071)	0.6504 (0.0076)	-0.3283 (0.0039)
	Yes	1.5491 (0.0033)	2.4178 (0.0036)	1.7874 (0.0045)	3.1204 (0.0051)	2.4317 (0.0066)	2.3151 (0.0073)	0.8848 (0.0040)
	Nr of Relatives	0.1427 (0.0004)	1.7515 (0.0036)	0.6921 (0.0021)	0.4011 (0.0010)	0.3556 (0.0012)	0.1751 (0.0007)	0.0666 (0.0002)
Death Year	Intercept	-0.7980 (0.0023)	-2.4095 (0.0065)	-0.4066 (0.0038)	-1.2217 (0.0030)	-0.5456 (0.0036)	-0.4848 (0.0038)	-0.2924 (0.0023)
	Yes	1.0737 (0.0023)	1.3920 (0.0021)	0.8335 (0.0026)	1.6112 (0.0027)	1.0649 (0.0034)	1.1940 (0.0041)	0.2956 (0.0029)
	Nr of Relatives	0.1621 (0.0004)	1.4078 (0.0033)	0.5480 (0.0014)	0.3391 (0.0005)	0.3035 (0.0006)	0.1576 (0.0004)	0.0855 (0.0003)
Birth Country	Intercept	-0.7375 (0.0027)	-2.3430 (0.0066)	-0.8112 (0.0043)	-1.0073 (0.0038)	-0.7838 (0.0037)	-0.5777 (0.0041)	-0.6096 (0.0030)
	Yes	0.7711 (0.0027)	1.0423 (0.0027)	0.7506 (0.0033)	1.5841 (0.0033)	1.4575 (0.0032)	1.3856 (0.0040)	0.5212 (0.0027)
	Nr of Relatives	0.1219 (0.0004)	1.1878 (0.0031)	0.4415 (0.0012)	0.1857 (0.0004)	0.1722 (0.0004)	0.0934 (0.0002)	0.0744 (0.0002)
Death Country	Intercept	-1.2219 (0.0022)	-2.1237 (0.0064)	-0.6300 (0.0031)	-1.3860 (0.0023)	-0.8676 (0.0028)	-0.7655 (0.0030)	-0.8117 (0.0023)
	Yes	0.8803 (0.0022)	0.8904 (0.0018)	0.4692 (0.0021)	1.1758 (0.0023)	0.7714 (0.0027)	0.8570 (0.0034)	0.2280 (0.0025)
	Nr of Relatives	0.1540 (0.0003)	1.0258 (0.0033)	0.3681 (0.0011)	0.2622 (0.0004)	0.2153 (0.0004)	0.1147 (0.0002)	0.0905 (0.0002)
Overall Nr. of Relatives		14,589,754	10,633,969	11,104,591	25,042,881	21,380,793	39,633,282	16,907,137
Overall Nr. of Focal Individuals		4,323,112	5,549,757	4,173,650	4,295,590	3,107,106	2,334,853	2,932,190

Notes: all p-values are smaller than 0.001 and standard errors are shown in parentheses. Here, we used a binary response indicating whether a focal individual has at least one relative with a non-missing value in a demographic variable. Coefficients of the logistic regression models to test association in terms of completeness, by type of relative and demographic variable. The models are fitted considering the individuals of the analytical and their kinship network. For each type of relative, if the focal individual does not have that type of relative, he is omitted from the regression model. Sample size of focal individuals employed for each model and the size of their kinship network are also included.

Table A.6: Coefficients of the negative binomial regression models to test the association in terms of quality, by type of relative and demographic variable.

Demographic Variable	Effect	Children	Parents	Grandparents	Siblings	Aunts & Uncles	Cousins	Grandchildren
Birth Date	Intercept	-1.1422 (0.0086)	-2.7385 (0.0166)	-3.0844 (0.0171)	-1.9024 (0.0160)	-2.1304 (0.0215)	-1.2527 (0.0284)	-0.4264 (0.0090)
	Non-missing birth month	1.8989 (0.0033)	1.7805 (0.0027)	1.2973 (0.0027)	3.1068 (0.0039)	1.9150 (0.0038)	2.1687 (0.0054)	1.2711 (0.0045)
	Nr of relatives	0.1771 (0.0006)	0.5114 (0.0057)	0.3874 (0.0014)	0.2340 (0.0060)	0.2269 (0.0006)	0.1136 (0.0004)	0.3188 (0.0016)
	Birth Period (Ref. 1800-1824)							
	1625-1649	0.2587 (0.0113)	0.3021 (0.0160)	0.4816 (0.0206)	0.1350 (0.0208)	0.2380 (0.0273)	0.1241 (0.0365)	0.1241 (0.0365)
	1650-1674	0.3667 (0.0107)	0.4504 (0.0148)	0.7692 (0.0189)	0.1821 (0.0193)	0.3559 (0.0250)	0.2381 (0.0336)	0.2381 (0.0336)
	1675-1699	0.4610 (0.0104)	0.7343 (0.0141)	1.0403 (0.0181)	0.3437 (0.0184)	0.5421 (0.0249)	0.4698 (0.0320)	0.4698 (0.0320)
	1700-1724	0.4945 (0.0101)	0.9118 (0.0138)	1.2970 (0.0176)	0.3769 (0.0179)	0.6104 (0.0229)	0.5578 (0.0308)	0.5578 (0.0308)
	1725-1749	0.6056 (0.0010)	1.0753 (0.0136)	1.4905 (0.0174)	0.4457 (0.0175)	0.7636 (0.0225)	0.6310 (0.0302)	0.6307 (0.0302)
	1750-1774	0.7912 (0.0099)	1.1161 (0.0135)	1.6396 (0.0173)	0.5445 (0.0172)	0.8384 (0.0221)	0.6795 (0.0298)	0.6795 (0.0298)
	1775-1799	1.0089 (0.0099)	1.1923 (0.0133)	1.6767 (0.0172)	0.6529 (0.0170)	0.8919 (0.0220)	0.7628 (0.0294)	0.7628 (0.0295)
	1800-1824	0.8771 (0.0098)	1.2782 (0.0133)	1.6752 (0.0171)	0.8224 (0.0169)	0.9105 (0.0219)	0.8877 (0.0294)	0.8877 (0.0294)
	1825-1849	0.8083 (0.0098)	1.5914 (0.0132)	1.8173 (0.0170)	0.6775 (0.0166)	1.1986 (0.0218)	0.8772 (0.0288)	0.8772 (0.0291)
	1850-1874	1.0600 (0.0099)	2.0107 (0.0134)	2.0420 (0.0170)	0.6391 (0.0165)	1.2891 (0.0217)	0.7806 (0.0290)	0.7810 (0.0230)
	1875-1900	1.3812 (0.0110)	2.2704 (0.0134)	2.4092 (0.0170)	0.7833 (0.0166)	1.2924 (0.0216)	0.8716 (0.0289)	0.8712 (0.0289)
Overall Nr. of Relatives		9,161,737	6,303,303	7,607,857	19,195,687	15,662,141	29,114,532	9,732,959
Overall Nr. of Focal Individuals		2,692,055	3,238,700	2,985,991	3,232,826	2,330,888	6,918,263	1,713,717
Death Date	Intercept	-0.8865 (0.0060)	-1.7394 (0.0081)	-1.2475 (0.0085)	-0.3347 (0.0068)	-0.2626 (0.0106)	0.3442 (0.0136)	-0.3524 (0.0066)
	Non-missing death month	0.5103 (0.0013)	0.5223 (0.0013)	0.4158 (0.0015)	0.5491 (0.0012)	0.3973 (0.0015)	0.3857 (0.0018)	0.3287 (0.0015)
	Nr of relatives	0.2164 (0.0001)	0.6759 (0.0026)	0.3893 (0.0006)	0.1616 (0.0001)	0.1534 (0.0001)	0.0881 (0.0001)	0.2353 (0.0003)
	Death Period (Ref. 1800-1824)							
	1625-1649	0.2823 (0.0082)	0.1913 (0.0091)	0.3275 (0.0127)	0.1337 (0.0124)	0.1628 (0.0198)	0.0326 (0.0203)	0.1707 (0.0198)
	1650-1674	0.4101 (0.0082)	0.2435 (0.0104)	0.4754 (0.0119)	0.2317 (0.0140)	0.2296 (0.0193)	0.1348 (0.0202)	0.2148 (0.0198)
	1675-1699	0.5455 (0.0081)	0.4136 (0.0103)	0.7594 (0.0118)	0.3596 (0.0138)	0.3742 (0.0192)	0.2601 (0.0201)	0.2567 (0.0201)
	1700-1724	0.6414 (0.0080)	0.5414 (0.0101)	0.9491 (0.0116)	0.4136 (0.0135)	0.4939 (0.0189)	0.4249 (0.0199)	0.4141 (0.0199)
	1725-1749	0.7745 (0.0079)	0.6816 (0.0100)	1.1014 (0.0114)	0.5071 (0.0133)	0.6114 (0.0188)	0.5528 (0.0197)	0.5427 (0.0197)
	1750-1774	1.0485 (0.0078)	0.7372 (0.0100)	1.2456 (0.0112)	0.6459 (0.0132)	0.6555 (0.0187)	0.6268 (0.0196)	0.6162 (0.0196)
	1775-1799	1.1999 (0.0078)	0.7908 (0.0099)	1.2829 (0.0111)	0.7622 (0.0131)	0.7097 (0.0185)	0.7253 (0.0195)	0.7253 (0.0195)
	1800-1824	0.8293 (0.0077)	0.8868 (0.0098)	1.2817 (0.0110)	0.8924 (0.0131)	0.7723 (0.0183)	0.7716 (0.0193)	0.7714 (0.0195)
	1825-1849	0.6554 (0.0077)	1.2043 (0.0098)	1.3985 (0.0110)	0.7723 (0.0129)	1.0031 (0.0182)	0.7681 (0.0193)	0.7681 (0.0193)
	1850-1874	0.7993 (0.0078)	1.5275 (0.0099)	1.5711 (0.0111)	0.8496 (0.0129)	1.0248 (0.0183)	0.6789 (0.0194)	0.6689 (0.0194)
	1875-1900	0.9965 (0.0087)	1.8017 (0.0100)	1.8971 (0.0111)	0.9443 (0.0129)	1.0394 (0.0183)	0.6907 (0.0194)	0.6907 (0.0195)

Notes: all p-values are smaller than 0.01, standard errors are shown in parentheses and Y denotes the inclusion of controls in the logistic regression models. The models are fitted considering the individuals of the analytical sample, who were born and/or died in the historical period 1600-1900, and their kinship network. For each type of relative, if the focal individual does not have that type of relative, he is omitted from the regression model. We include the birth and death years as control which are grouped in 25-year classes. These classes are entered in the regression as a series of dummies. Number of relatives and number of focal individuals with non-missing birth (death) years for each regression model are also included in the table.

Table A.7: Coefficients of the negative binomial regression models, in which the number of relatives is treated as offset, to test the association in terms of completeness, by type of relative and demographic variable.

Demographic Variable	Effect	Children	Parents	Grandparents	Siblings	Aunts & Uncles	Cousins	Grandchildren
Birth Year	Intercept	1.0412 (0.0014)	-0.9092 (0.0009)	-0.5998 (0.0009)	-1.0632 (0.0010)	-0.4737 (0.0013)	-0.7321 (0.0007)	-0.8762 (0.0022)
	Yes	0.7078 (0.0014)	0.7606 (0.0009)	0.4027 (0.0010)	0.9638 (0.0010)	0.3885 (0.0013)	0.5841 (0.0013)	0.3765 (0.0023)
Death Year	Intercept	-1.4047 (0.0011)	-0.5721 (0.0005)	-1.2772 (0.0005)	-1.1397 (0.0010)	-0.7770 (0.0006)	-0.9973 (0.0007)	-1.1491 (0.0016)
	Yes	0.6781 (0.0012)	0.2376 (0.0006)	0.1279 (0.0006)	0.6973 (0.0010)	0.3226 (0.0007)	0.4017 (0.0009)	0.2166 (0.0018)
Birth Country	Intercept	-2.0997 (0.0020)	-1.0360 (0.0012)	-0.9683 (0.0013)	-2.4134 (0.0017)	-1.1055 (0.0011)	-1.2938 (0.0012)	-1.2880 (0.0016)
	Yes	0.3901 (0.0022)	0.5110 (0.0012)	0.3536 (0.0013)	0.4499 (0.0018)	0.6743 (0.0012)	0.7649 (0.0014)	0.3754 (0.0019)
Death Country	Intercept	-1.7927 (0.0013)	-1.0123 (0.0007)	-0.9112 (0.0007)	-1.5288 (0.0010)	-1.2557 (0.0008)	-1.4976 (0.0009)	-1.6462 (0.0015)
	Yes	0.6625 (0.0015)	0.4434 (0.0008)	0.2251 (0.0009)	0.7002 (0.0010)	0.3893 (0.0011)	0.4567 (0.0012)	0.2519 (0.0018)
Overall Nr. of Relatives		14,589,754	10,633,969	11,104,591	25,042,881	21,380,793	39,633,282	16,907,137
Overall Nr. of Focal Individuals		4,323,112	5,549,757	4,173,650	4,295,590	3,107,106	2,334,853	2,932,190

Notes: all p-values are smaller than 0.001 and standard errors are shown in parentheses. Coefficients of the negative binomial regression models with the number of relatives as offset to test association in terms of completeness, by type of relative and demographic variable. The models are fitted considering the individuals of the analytical and their kinship network. For each type of relative, if the focal individual does not have that type of relative, he is omitted from the regression model. Sample size of focal individuals employed for each model and the size of their kinship network are also included.

Table A.8: Coefficients of the negative binomial regression models, in which the number of relatives is treated as offset, to test the association in terms of quality, by type of relative and demographic variable.

Demographic Variable	Effect	Children	Parents	Grandparents	Siblings	Aunts & Uncles	Cousins	Grandchildren
Birth Date	Intercept	-1.1446 (0.0037)	-2.0640 (0.0069)	-2.5501 (0.0124)	-1.5306 (0.0039)	-1.3671 (0.0081)	-1.3671 (0.0090)	-0.7800 (0.0037)
	Non-missing month	0.6773 (0.0010)	0.7979 (0.0011)	0.6790 (0.0014)	1.2034 (0.0086)	0.6933 (0.0011)	0.6945 (0.0011)	0.3687 (0.0012)
	Birth Cohort (Ref. 1600-1624)							
	1625-1649	0.1493 (0.0046)	0.1975 (0.0083)	0.3559 (0.0149)	0.0533 (0.0048)	0.1030 (0.0101)	0.0680 (0.0112)	0.0585 (0.0047)
	1650-1674	0.1872 (0.0043)	0.3425 (0.0077)	0.5730 (0.0135)	0.1084 (0.0043)	0.1614 (0.0092)	0.1798 (0.0100)	0.0719 (0.0044)
	1675-1699	0.2093 (0.0041)	0.5401 (0.0073)	0.7674 (0.0130)	0.1451 (0.0041)	0.2978 (0.0086)	0.2530 (0.0095)	0.0710 (0.0043)
	1700-1724	0.1986 (0.0040)	0.6443 (0.0071)	0.9832 (0.0127)	0.1508 (0.0040)	0.3500 (0.0086)	0.2707 (0.0093)	0.0793 (0.0042)
	1725-1749	0.2013 (0.0040)	0.7264 (0.0071)	1.1139 (0.0126)	0.1488 (0.0040)	0.4004 (0.0084)	0.2706 (0.0092)	0.1358 (0.0041)
	1750-1774	0.2300 (0.0039)	0.7409 (0.0071)	1.1847 (0.0125)	0.1362 (0.0040)	0.3999 (0.0082)	0.2577 (0.0092)	0.1684 (0.0040)
	1775-1799	0.2671 (0.0039)	0.7751 (0.0070)	1.1863 (0.0125)	0.1429 (0.0039)	0.3722 (0.0082)	0.2645 (0.0092)	0.1451 (0.0040)
	1800-1824	0.2035 (0.0039)	0.8253 (0.0069)	1.1819 (0.0124)	0.1583 (0.0039)	0.3578 (0.0082)	0.2754 (0.0091)	0.1690 (0.0039)
	1825-1849	0.2152 (0.0038)	0.9292 (0.0069)	1.2573 (0.0124)	0.1451 (0.0039)	0.4188 (0.0082)	0.2657 (0.0091)	0.2296 (0.0040)
	1850-1874	0.3044 (0.0038)	1.0436 (0.0069)	1.3688 (0.0124)	0.1584 (0.0039)	0.4510 (0.0081)	0.2748 (0.0091)	0.2572 (0.0041)
	1875-1900	0.3374 (0.0038)	1.1002 (0.0069)	1.5069 (0.0124)	0.2067 (0.0039)	0.4674 (0.0081)	0.3264 (0.0091)	0.2760 (0.0047)
Overall Nr. of Relatives		9,161,737	6,303,303	7,607,857	19,195,687	15,662,141	29,114,532	9,732,959
Overall Nr. of Focal Individuals		2,692,055	3,238,700	2,985,991	3,232,826	2,330,888	6,918,263	1,713,717
Death Date	Intercept	-0.9188 (0.0051)	-1.0783 (0.0065)	-1.1089 (0.0110)	-0.8673 (0.0056)	-0.7759 (0.0110)	-0.7041 (0.0107)	-0.7403 (0.0067)
	Non-missing month	0.4893 (0.0011)	0.5222 (0.0013)	0.4164 (0.0019)	0.5238 (0.0001)	0.3744 (0.0015)	0.3336 (0.0013)	0.3006 (0.0014)
	Death Cohort (Ref. 1600-1624)							
	1625-1649	0.0600 (0.0065)	0.0219 (0.0085)	0.0497 (0.0143)	0.0186 (0.0072)	0.0062 (0.0143)	0.0256 (0.0140)	0.0664 (0.0084)
	1650-1674	0.0760 (0.0060)	0.0278 (0.0079)	0.0703 (0.0133)	0.0591 (0.0065)	0.0344 (0.0131)	0.0280 (0.0126)	0.0607 (0.0077)
	1675-1699	0.1060 (0.0056)	0.0989 (0.0072)	0.1233 (0.01217)	0.0967 (0.0060)	0.0878 (0.0120)	0.0730 (0.0110)	0.0715 (0.0075)
	1700-1724	0.1094 (0.0054)	0.1685 (0.0070)	0.1891 (0.0117)	0.1000 (0.0058)	0.1020 (0.0110)	0.0863 (0.0112)	0.0841 (0.0070)
	1725-1749	0.1102 (0.0053)	0.1993 (0.0067)	0.2499 (0.0114)	0.0959 (0.0059)	0.1201 (0.0112)	0.0906 (0.0108)	0.1104 (0.0070)
	1750-1774	0.1321 (0.0053)	0.2063 (0.0066)	0.2873 (0.0112)	0.0937 (0.0057)	0.1218 (0.0111)	0.0953 (0.0109)	0.1517 (0.0069)
	1775-1799	0.1693 (0.0053)	0.1978 (0.0067)	0.2853 (0.0111)	0.0960 (0.0057)	0.1189 (0.0111)	0.1032 (0.0108)	0.1864 (0.0069)
	1800-1824	0.2042 (0.0052)	0.1987 (0.0066)	0.2791 (0.0111)	0.1157 (0.0056)	0.1137 (0.0110)	0.1231 (0.0108)	0.2244 (0.0068)
	1825-1849	0.2221 (0.0052)	0.2113 (0.0066)	0.2667 (0.0110)	0.1355 (0.0056)	0.1260 (0.0106)	0.1389 (0.0108)	0.2298 (0.0068)
	1850-1874	0.2344 (0.0052)	0.2592 (0.0066)	0.2881 (0.0110)	0.1564 (0.0056)	0.1560 (0.0110)	0.1584 (0.0107)	0.2305 (0.0068)
	1875-1900	0.2350 (0.0052)	0.2947 (0.0065)	0.3372 (0.0110)	0.1715 (0.0056)	0.1804 (0.0110)	0.1681 (0.0107)	0.2361 (0.0068)
Overall Nr. of Relatives		3,420,063	7,531,416	3,008,568	5,514,632	4,085,603	6,918,263	4,440,454
Overall Nr. of Focal Individuals		1,379,573	3,877,896	1,567,008	1,173,828	816,322	615,066	1,051,907

Notes: all p-values are smaller than 0.001 and standard errors are shown in parentheses. Coefficients of the negative binomial regression models with the number of relatives as offset to test association in terms of completeness, by type of relative and demographic variable. The models are fitted considering the individuals of the analytical and their kinship network. For each type of relative, if the focal individual does not have that type of relative, he is omitted from the regression model. Sample size of focal individuals employed for each model and the size of their kinship network are also included.

Table A.9: Coefficients of the binomial regression models to test the association in terms of completeness, by type of relative and demographic variable.

Demographic Variable	Effect	Children	Parents	Grandparents	Siblings	Aunts & Uncles	Cousins	Grandchildren
Birth Year	Intercept	-1.0350 (0.0014)	-0.9092 (0.0022)	-0.5997 (0.0020)	-1.0632 (0.0018)	-0.4737 (0.0016)	-0.7065 (0.0013)	-0.8168 (0.0017)
	Yes	0.7105 (0.0019)	0.7606 (0.0009)	0.4027 (0.0020)	0.9637 (0.0018)	0.3885 (0.0016)	0.5634 (0.0013)	0.3790 (0.0018)
Death Year	Intercept	-1.4161 (0.0013)	-0.7407 (0.0008)	-0.5721 (0.0008)	-1.1293 (0.0008)	-0.7641 (0.0006)	-0.9384 (0.0005)	-1.1632 (0.0012)
	Yes	0.6939 (0.0013)	0.4693 (0.0010)	0.2377 (0.0006)	0.6946 (0.0008)	0.3155 (0.0008)	0.3636 (0.0006)	0.2539 (0.0013)
Birth Country	Intercept	-2.0997 (0.0020)	-1.0360 (0.0018)	-0.9683 (0.0017)	-2.4134 (0.0021)	-1.0932 (0.0009)	-1.2933 (0.0007)	-1.2887 (0.0010)
	Yes	0.3901 (0.0022)	0.5111 (0.0018)	0.3536 (0.0018)	0.4499 (0.0022)	0.6657 (0.0010)	0.7693 (0.0008)	0.3814 (0.0011)
Death Country	Intercept	-1.2826 (0.0014)	-1.0123 (0.0008)	-0.9112 (0.0008)	-1.5067 (0.0007)	-1.2328 (0.0006)	-1.4424 (0.0005)	-1.7035 (0.0011)
	Yes	0.5144 (0.0015)	0.4434 (0.0011)	0.2251 (0.0011)	0.6910 (0.0008)	0.3752 (0.0009)	0.4231 (0.0007)	0.2966 (0.0013)
Overall Nr. of Relatives		14,589,754	10,633,969	11,104,591	25,042,881	21,380,793	39,633,282	16,907,137
Overall Nr. of Focal Individuals		4,323,112	5,549,757	4,173,650	4,295,590	3,107,106	2,334,853	2,932,190

Notes: all p-values are smaller than 0.001 and standard errors are shown in parentheses. Coefficients of the binomial regression models to test association in terms of completeness, by type of relative and demographic variable. The models are fitted considering the individuals of the analytical and their kinship network. For each type of relative, if the focal individual does not have that type of relative, he is omitted from the regression model. Sample size of focal individuals employed for each model and the size of their kinship network are also included.

Table A.10: Coefficients of the binomial regression models to test the association in terms of quality, by type of relative and demographic variable.

Demographic Variable	Effect	Children	Parents	Grandparents	Siblings	Aunts & Uncles	Cousins	Grandchildren
Birth Date	Intercept	-1.1434 (0.0051)	-2.0721 (0.0010)	-2.5528 (0.0131)	-1.5421 (0.0049)	-1.3671 (0.0078)	-1.2277 (0.0075)	-0.7414 (0.0036)
	Non-missing month	0.6782 (0.0012)	0.8019 (0.0017)	0.6799 (0.0016)	1.2033 (0.0011)	0.6744 (0.0010)	0.6391 (0.00074)	0.3637 (0.0011)
	Birth Cohort (Ref. 1600-1624)							
	1650-1674	0.1891 (0.0063)	0.3393 (0.0112)	0.5728 (0.0144)	0.1145 (0.0055)	0.1478 (0.0088)	0.1708 (0.0081)	0.0570 (0.0043)
	1675-1699	0.2103 (0.0057)	0.5373 (0.0110)	0.7675 (0.0138)	0.1546 (0.0053)	0.2912 (0.0082)	0.2333 (0.0077)	0.0447 (0.0042)
	1700-1724	0.1982 (0.0055)	0.6422 (0.0104)	0.9805 (0.0135)	0.1580 (0.0052)	0.34134 (0.0081)	0.2536 (0.0076)	0.0495 (0.0040)
	1725-1749	0.2009 (0.0054)	0.7252 (0.0103)	1.1117 (0.0133)	0.1551 (0.0051)	0.3865 (0.0080)	0.2479 (0.0076)	0.1006 (0.0040)
	1750-1774	0.2315 (0.0054)	0.7415 (0.0102)	1.1851 (0.0133)	0.1468 (0.0050)	0.3812 (0.0079)	0.2262 (0.0075)	0.1216 (0.0039)
	1775-1799	0.2703 (0.0054)	0.7775 (0.0101)	1.1883 (0.0132)	0.1563 (0.0050)	0.3479 (0.0079)	0.2225 (0.0075)	0.0982 (0.0038)
	1800-1824	0.1992 (0.0053)	0.8306 (0.0101)	1.1855 (0.0132)	0.1777 (0.0050)	0.3314 (0.0079)	0.2242 (0.0075)	0.1328 (0.0038)
	1825-1849	0.2079 (0.0052)	0.9374 (0.0101)	1.2624 (0.0132)	0.1584 (0.0049)	0.3914 (0.0078)	0.2120 (0.0075)	0.1944 (0.0039)
	1850-1874	0.2999 (0.0052)	1.0533 (0.0101)	1.3724 (0.0131)	0.1687 (0.0049)	0.4224 (0.0078)	0.2321 (0.0075)	0.2196 (0.0041)
	1875-1900	0.3375 (0.0054)	1.1071 (0.0100)	1.5080 (0.0131)	0.2169 (0.0049)	0.4395 (0.0078)	0.2867 (0.0075)	0.2373 (0.0049)
Overall Nr. of Relatives		9,161,737	6,303,303	7,607,857	19,195,687	15,662,141	29,114,532	9,732,959
Overall Nr. of Focal Individuals		2,692,055	3,238,700	2,985,991	3,232,826	2,330,888	6,918,263	1,713,717
Death Date	Intercept	-0.9188 (0.0051)	-1.0783 (0.0065)	-1.1089 (0.0110)	-0.8673 (0.0056)	-0.7759 (0.0110)	-0.7041 (0.0107)	-0.7403 (0.0067)
	Non-missing month	0.4893 (0.0011)	0.5222 (0.0013)	0.4164 (0.0019)	0.5238 (0.0001)	0.3744 (0.0015)	0.3336 (0.0013)	0.3006 (0.0014)
	Death Cohort (Ref. 1600-1624)							
	1625-1649	0.0600 (0.0065)	0.0219 (0.0085)	0.0497 (0.0143)	0.0186 (0.0072)	0.0062 (0.0143)	0.0256 (0.0140)	0.0664 (0.0084)
	1650-1674	0.0760 (0.0060)	0.0278 (0.0079)	0.0703 (0.0133)	0.0591 (0.0065)	0.0344 (0.0131)	0.0280 (0.0126)	0.0607 (0.0077)
	1675-1699	0.1060 (0.0056)	0.0989 (0.0072)	0.1233 (0.01217)	0.0967 (0.0060)	0.0878 (0.0120)	0.0730 (0.0110)	0.0715 (0.0075)
	1700-1724	0.1094 (0.0054)	0.1685 (0.0070)	0.1891 (0.0117)	0.1000 (0.0058)	0.1020 (0.0110)	0.0863 (0.0112)	0.0841 (0.0070)
	1725-1749	0.1102 (0.0053)	0.1993 (0.0067)	0.2499 (0.0114)	0.0959 (0.0059)	0.1201 (0.0112)	0.0906 (0.0108)	0.1104 (0.0070)
	1750-1774	0.1321 (0.0053)	0.2063 (0.0066)	0.2873 (0.0112)	0.0937 (0.0057)	0.1218 (0.0111)	0.0953 (0.0109)	0.1517 (0.0069)
	1775-1799	0.1693 (0.0053)	0.1978 (0.0067)	0.2853 (0.0111)	0.0960 (0.0057)	0.1189 (0.0111)	0.1032 (0.0108)	0.1864 (0.0069)
	1800-1824	0.2042 (0.0052)	0.1987 (0.0066)	0.2791 (0.0111)	0.1157 (0.0056)	0.1137 (0.0110)	0.1231 (0.0108)	0.2244 (0.0068)
	1825-1849	0.2221 (0.0052)	0.2113 (0.0066)	0.2667 (0.0110)	0.1355 (0.0056)	0.1260 (0.0106)	0.1389 (0.0108)	0.2298 (0.0068)
	1850-1874	0.2344 (0.0052)	0.2592 (0.0066)	0.2881 (0.0110)	0.1564 (0.0056)	0.1560 (0.0110)	0.1584 (0.0107)	0.2305 (0.0068)
	1875-1900	0.2350 (0.0052)	0.2947 (0.0065)	0.3372 (0.0110)	0.1715 (0.0056)	0.1804 (0.0110)	0.1681 (0.0107)	0.2361 (0.0068)
Overall Nr. of Relatives		3,420,063	7,531,416	3,008,568	5,514,632	4,085,603	6,918,263	4,440,454
Overall Nr. of Focal Individuals		1,379,573	3,877,896	1,567,008	1,173,828	816,322	615,066	1,051,907

Notes: all p-values are smaller than 0.001 and standard errors are shown in parentheses. Coefficients of the binomial regression models to test association in terms of completeness, by type of relative and demographic variable. The models are fitted considering the individuals of the analytical and their kinship network. For each type of relative, if the focal individual does not have that type of relative, he is omitted from the regression model. Sample size of focal individuals employed for each model and the size of their kinship network are also included.

Appendix B

Appendix B: Supplemental Information

B.1 Additional Descriptive Tables and Figures

Table B.1: Data sources for mortality, fertility and population estimates by country

Country	Child Mortality (q_{0-4})	Fertility (TFR)	Age-sex proportions
USA	Hacker (2010) (1790-1899) †,‡ Human Life-Table Database (1900-1910)	Coale and Zelnik (1961) (1800-1910) ‡	IPUMS censuses (1850-1910)†
Norway	Human Mortality Database (1846-1910)	Human Fertility Collection (1845-1910)	Human Mortality Database (1846-1910)
Netherlands	Human Mortality Database (1850-1910)	Chesnais (1992) (1830-1900) ‡ Human Fertility Collection (1906-1910)	Human Mortality Database (1856-1910)
Sweden	Human Mortality Database (1751-1910)	Human Fertility Collection (1751-1910)	Human Mortality Database (1751-1910)
England & Wales	Wrigley et al. (1997) (1751-1836) ‡,‡ Human Mortality Database (1841-1910) ‡,*	Woods (2000) (1800-1910) ‡	Wrigley et al. (1997) (1751-1836)‡ Human Mortality Database (1841-1910)
Denmark	Human Mortality Database (1835-1910)	Human Fertility Collection (1878-1910)	Human Mortality Database (1835-1910)
Finland	Human Life-Table Database (1751-1877) † Human Mortality Database (1878-1910)	Human Fertility Collection (1866-1910)	Human Mortality Database (1878-1910)
France	Blayo (1975)(1751-1815) † Human Mortality Database (1816-1910)	Pison (2012) (1800-1891) ‡ Human Fertility Collection (1892-1910)	Human Mortality Database (1816-1910)

† Only ten-year estimates are available.
‡ Only five-year population estimates are available.
* Population counts by age and sex from England and Wales are provided by the HMD. However, decennial censuses from the years 1851, 1861, 1871, 1881, 1891 and 1901 are also available as an alternative data source from IPUMS (Ruggles et al., 2011).

Table B.1 illustrates the different data sources with accurate demographic rates, namely the probability of death under age 5 (q_{0-4}), total fertility rate (TFR) and proportion of individuals by age and sex. The availability of accurate demographic information varies by country and over time. Additionally, demographic estimates are not always available on a yearly basis. In most countries, we only observe quinquennial or decennial estimates. To overcome this issue, we employ linear interpolation.

Table B.2 shows the distribution of missing values in the main demographic variables in FamiLinx. As the data is generated from users, the data are affected by a high of missing values. A detailed discussion on the quality of FamiLinx can be found in Chapter 2.

Table B.2: Distribution of the missing values in the main demographic variables in FamiLinx.

Variable	Absolute size	%
Sample Size	86,124,644	
Sex		
Missing	14,071,200	16.34%
Male	37,998,030	44.12%
Female	34,055,414	39.54%
Birth Date		
Missing	52,405,914	60.85%
Only Year	13,692,092	15.90%
Year and Month	849,377	0.99%
Complete Date	14,150,307	22.27%
Death Date		
Missing	64,383,957	74.77%
Only Year	6,736,492	7.82%
Year and Month	853,888	0.99%
Complete Date	14,150,307	16.43%
Birth Location		
Missing	74,464,173	81.82%
Reported	15,659,836	18.18%
Death Location		
Missing	74,861,173	86.92%
Reported	11,263,471	13.08%

Table B.3: Sample sizes and person-years by country in the selected analytical sample.

Country	Sample Size (%)	person-years
USA	953,236 (57.20)	63,102,666
Norway	145,515 (8.73)	9,229,826
Netherlands	126,307 (7.57)	7,181,535
Sweden	125,496 (7.52)	7,414,288
England & Wales	117,843 (7.07)	6,797,670
Denmark	49,509 (2.97)	2,961,458
Finland	43,815 (2.63)	2,475,602
France	38,388 (2.30)	2,150,470

Table B.3 displays the country-specific sample sizes, percentages and person-years in the analytical sample. The US is the country with the largest number of individuals. The selected countries are among the top 20 countries with largest number of recorded vital events (either births or deaths).

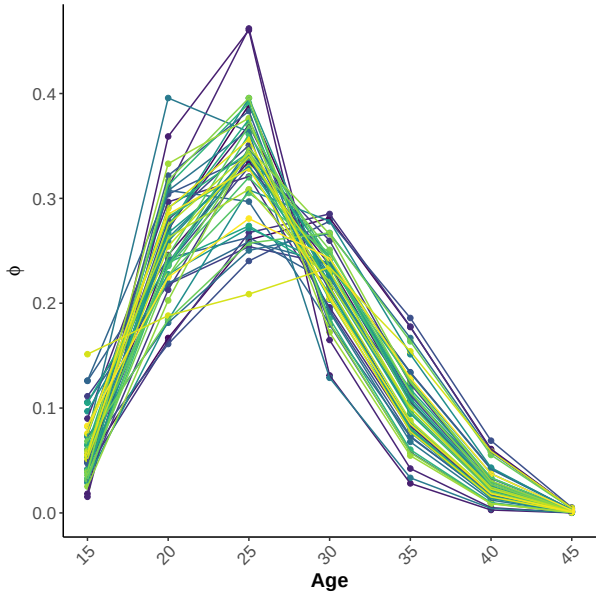


Figure B.1: Simulated patterns for the age-specific fertility proportions ($\phi_{x,a,t}$).

Figure B.1 displays 200 draws for $\phi_{x,t,c}$ based on Equation 4.5. Based on the simulations, it seems that we are able to capture various age-specific fertility curves.

Figure B.2 displays the indirect extended TFR estimates based on the class of indicators introduced by Hauer and Schmertmann (2020). $xTFR^{*}$ indicates our proposed indirect indicator that accounts for the non-representativeness of the online genealogical populations, while $xTFR$ and $xTFR^{+}$ denote the unadjusted and mortality-adjusted $xTFR$ estimates. Consistently with the results for $iTFR$, by accounting for the mortality of children and the non-representativeness of children under 5 and of women aged 15-49, we obtain more accurate TFR estimates.

In Figure B.3, we show the bivariate relationships between the median TFR estimates that we can obtain by specifying different time series model on the parameter ν_t. Based upon the reported results, employing a RW(1) or a RW(2) for the parameter ν_t would have led to quite similar results as the estimates are highly correlated with a correlation coefficient close to 1.

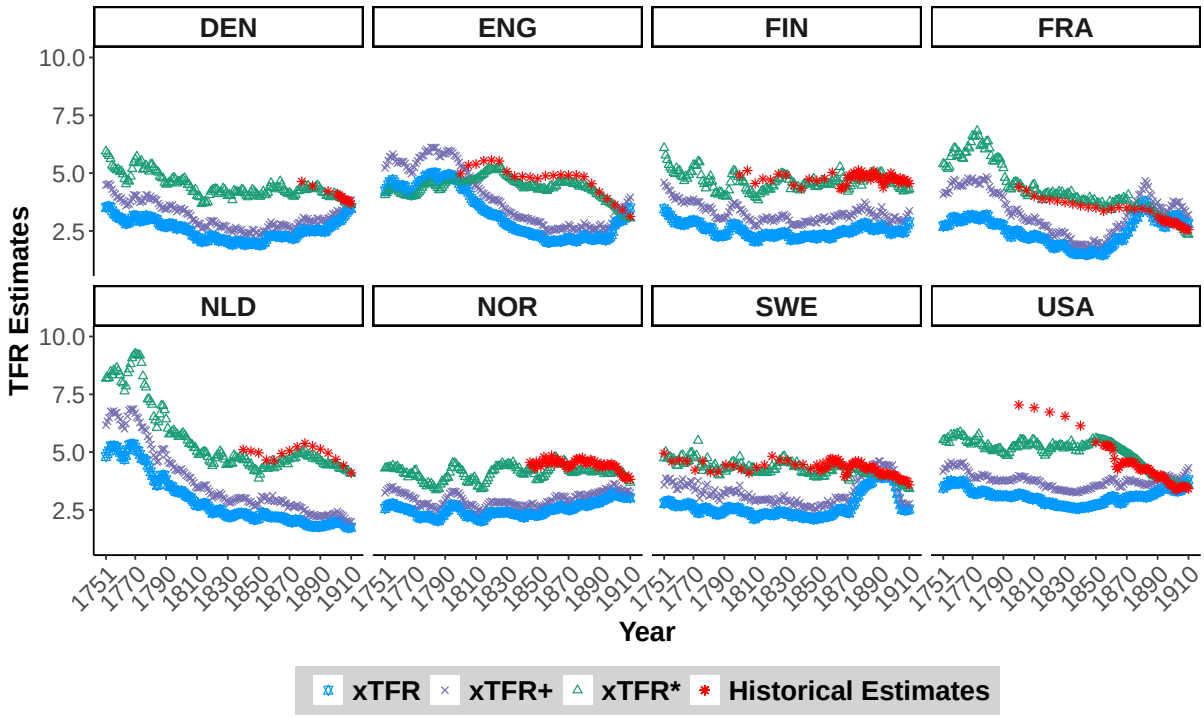


Figure B.2: Time series of xTFR estimates for the historical period 1751-1910 by country. $xTFR$ refers to the simplest indicator from the decomposition by Hauer and Schmertmann (2020), which does not account for child mortality and for the non-representativeness of online genealogical data. $xTFR^{+}$ is an extended version of the indicator $xTFR$ that adjusts for child mortality. $xTFR^{*}$ further refines the indicator $xTFR$ by accounting not only for both child mortality and the non-representativeness of online genealogical data.

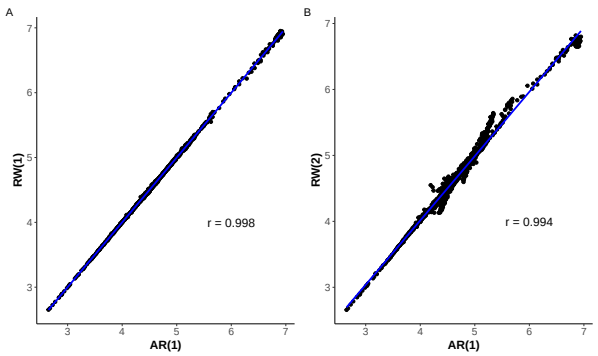


Figure B.3: Posterior median TFR estimates according a AR(1) specification for the parameter ν_t against those obtained either with a RW(1) (Panel A) or with a RW(2) (Panel B). The value of the correlation coefficient (r) between the TFR estimates is reported.

B.2 Posterior Estimates for the Bias-adjustment Parameters

Figure B.4 shows the median posterior estimates of the parameters ν_t coupled with the 2.5% and 97.5% quantiles for the period 1751 – 1910. Values values of ν_t greater than 0 imply an over-representation of the expected number of children under age 5 per woman aged 15 – 49 in the genealogical populations compared to the real one. In the opposite case the same quantity in the genealogical sample is underrepresented. We note that the

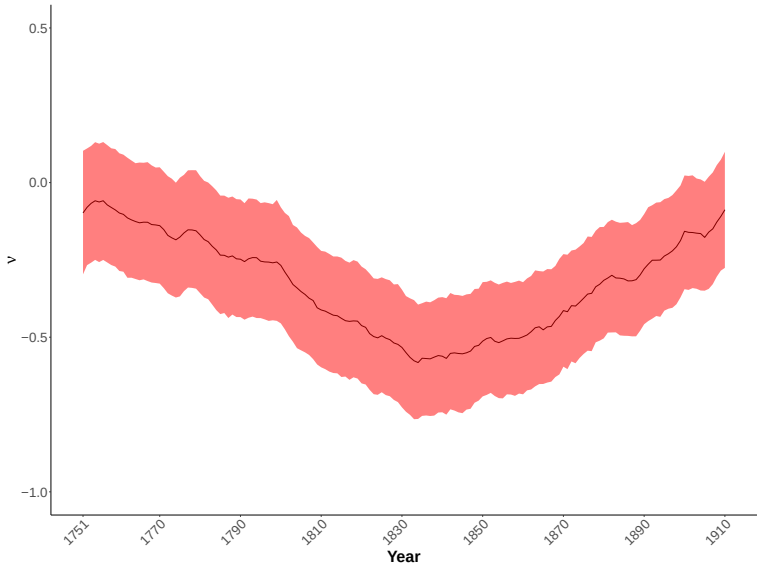


Figure B.4: Median posteriors for the parameters ν_t with 95% credible intervals represented through shaded areas.

estimates of the parameter in the first phase of the historical period are mostly informed by Sweden and England, the only countries with accurate population percentages by age and sex for the entire period under analysis. Starting from the beginning of the 19th century an increasing number of countries presents representative population estimates by age and sex. The median posterior estimates for the parameters ν_t decreases up to the middle of the 19th century. This result could be driven by two potential mechanisms. First, by 1850 an increasing number of countries presents representative population estimates by age and sex and their contribution results in a decrease of the estimates for ν. Second, this decrease could be explained by the impact of transatlantic migrants. The exclusion of individuals with distinct birth and death years from our analytical sample may lead

to an increase in the under-representation of the individuals that were born in Europe during the period 1800-1850 and migrated to the US during their life course. Starting from 1850, the parameter increases and approaches the value of 0. This result is consistent with the previous literature (Stelter and Alburez-Gutierrez, 2022; Colasurdo and Omenti, 2024; Minardi et al., 2024), which underlines that genealogical populations from FamiLinx become more representative towards the end of the 19th century.

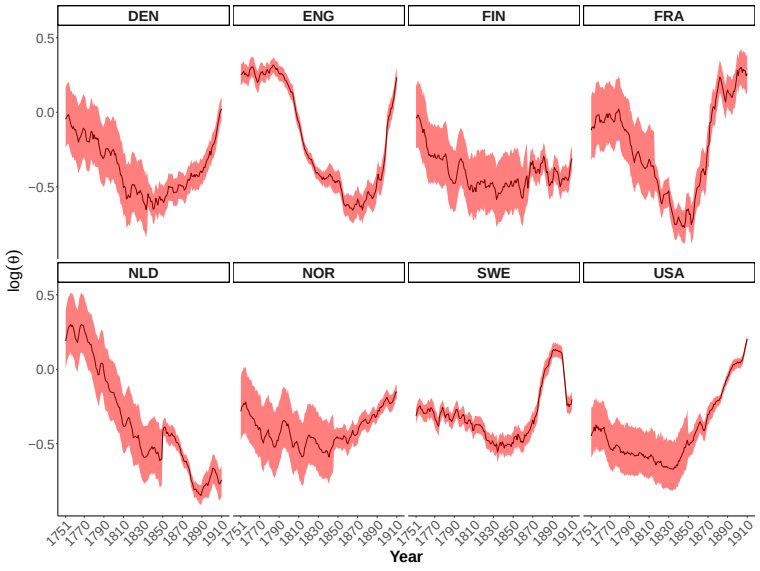


Figure B.5: Median posteriors for the (logged) parameter $\theta_{a,t}$ with 95% credible intervals represented through shaded areas.

Figure B.5 displays the median posterior estimates for the parameters $\theta_{a,t}$ for all the countries during the historical period 1751 – 1910. The interpretation is similar to ν_t and is country-specific. In general, the representativeness of the expected number of children under 5 per woman aged 15 – 49 decreases from 1751 up to 1850 for all the European countries, while it remains roughly stable for the US. In the subsequent historical period (1851–1910), the representativeness improves for all countries apart from the Netherlands. Additionally, in the last decade of the study period, Sweden is affected by a sudden downward fluctuation.

B.3 Extraction of Birth and Death Countries

Following the procedure by Colasurdo and Omenti (2024), the extraction of the birth and death countries relies on three methods ranked in order of preference:

- *exact matching using the reported country code*: birth and death countries determined from the reported two-digit country code.
- *regular expression matching*: birth and death countries established by a set of text strings, known as regular expressions, that specify a matching pattern for the name of the country of interest.
- *latitude and longitude for a location*: latitude and longitude coordinates inferred by Kaplanis et al. (2018)

To establish the definitive birth and death locations, the country names from inferred latitude and longitude coordinates are determined harnessing a geo-parsing algorithm. Subsequently, the records belonging to the top 20 countries with the highest number of vital events (births and deaths) are identified. Then, the birth and death countries are determined using the country codes and text strings reported in the records from the top 20 countries according to *exact matching using the reported country code* and *regular expression matching*. Then, to each profile, we assign the birth country determined according to the most accurate method. For instance, if a profile presents two different birth locations, one determined by exact matching and the other based on the inferred coordinates, we assign to the individual the birth country identified by means of the exact matching method since we regard this method as more accurate. We prefer methods based on text strings since the inferred latitude and longitude by Kaplanis et al. (2018) may be subject to reporting errors due to historical changes in boundaries between countries.

B.4 Proof for the TFR Decomposition

Following Schmertmann and Hauer (2019), let us assume that the fertility rates in the population are strictly positive over the age interval $[15, 50)$ and zero otherwise.

We define the following quantities.

- F_x is average fertility rates over the age interval $[x, x+5)$
- TFR is the total fertility rates. $TFR = 5 \cdot \sum_{x=15}^{45} F_x$.
- $\phi_x = \frac{5 \cdot F_x}{TFR}$ is the total fraction of fertility occurring in the age group x
- L_x denotes the person-years lived in age group a from an abridged life table with a radix $l_0 = 1$
- W_x is the observed number of women in the age group x
- $W_{15-49} = \sum_{x=15}^{45} W_x$ is the total observed number of women in the age interval $[15, 50)$.
- r is called under-reporting multiplier and indicates the ratio of the child/woman ratio (CWR) in the true population of interest to the CWR estimated from the genealogical sample. This measure can be interpreted by how much the genealogy-based CWR should be inflated to be equal to the CWR calculated in the general population.

$$r = \frac{\frac{C^{\text{true}}}{W^{\text{true}}}}{\frac{C}{W}}$$

Following standard cohort-component projection methods (see Wachter (2014) for details), we are able to calculate the expected number of surviving children aged 0-4 per woman in the age groups x ($x = 15-19, \ldots, 45-49$) at the end of the five-year period, which we call K_x.

$$K_x = \left[\frac{L_{x-5}}{L_x} \cdot F_{x-5} + F_x \right] \cdot \frac{L_0}{2}$$

The mathematical trick is to multiply and divide the right-hand side of the previous equation by $5 \cdot TFR$. This enables us to obtain a new expression that depends on the TFR.

$$K_x = TFR \cdot \frac{L_0}{5} \cdot \frac{1}{2} \left[\frac{L_{x-5}}{L_x} \cdot \phi_{x-5} + \phi_x \right]$$

We can decompose the expected number of children per woman in the age class a into three factors.

$$K_a = TFR \cdot \frac{L_0}{5} \cdot \frac{1}{2} \cdot p_x$$

The first two factors are identical across all the age classes, whereas the last factor p_x varies depending on the age group. The latter can be defined as the average of the fertility proportions experienced by age groups x and $x-5$.

The total number of expected children aged $0-4$ (C) can be calculated as a weighted sum of the expected number of children for each maternal age class using as weights the number of exposed women in each specific age group.

$$C = \sum_{x=15}^{45} K_x W_x$$

It follows

$$C = \sum_{x=15}^{45} TFR \cdot \frac{L_0}{5} \cdot \frac{1}{2} \cdot p_x \cdot W_x = TFR \cdot \frac{L_0}{5} \cdot \sum_{a=15}^{45} p_x \cdot W_x$$

If we divide both sides of the previous equation by the number of women in reproductive ages W, we get the CWR on the left-hand side, which is expressed as a function of the TFR.

$$\frac{C}{W} = \sum_{x=15}^{45} TFR \cdot \frac{L_0}{5} \cdot \frac{1}{2} p_x W_x = TFR \cdot \frac{L_0}{5} \cdot \sum_{x=15}^{45} p_a \frac{W_x}{W} = TFR \cdot \frac{L_0}{5} \cdot p_{\bar{x}}$$

where $p_{\bar{x}}$ can be thought as a weighted average of p_x, which depends both on the age fertility patterns as well as the number of women in maternal ages.

In addition, if we assume that the sample, from which we retrieve the population counts, is not representative of the whole population, we can add an adjustment factor r. This factor takes into account possible biases in the Child-Woman ratios.

$$\frac{C}{W} \cdot r = TFR \cdot \frac{L_0}{5} \cdot p_{\bar{a}}$$

Rearranging the previous equation for the TFR, the expression becomes

$$TFR = r \cdot \frac{1}{p_{\bar{x}}} \cdot \frac{1}{s} \cdot \frac{C}{W}$$

The previous equation decomposes the TFR as a product of four factors: the CWR, a survival multiplier $\left(\frac{1}{s}\right)$, an age-structure multiplier $\left(\frac{1}{p_{\bar{x}}}\right)$ and an under-reporting multiplier (τ).

B.5 Age-multiplier in $xTFR$, $xTFR^{+}$, $xTFR^{*}$

To estimate the age-multiplier in $xTFR$, following Hauer and Schmertmann (2020), 1,804 fertility schedules in the Human Fertility Database (HFD), for which the true TFR is known, were examined. For each country c and year t, the average $TFR^{*}_{t,c}$ over the previous five years was calculated $\left(TFR^{*}_{t,c} = \frac{1}{5}\sum_{i=t-4}^{t} TFR^{*}_{i,c}\right)$. The empircal values were divided by the observed child-woman ratios $\left(\frac{C_{t,c}}{W_{t,c}}\right)$.

We fit a simple linear regression model using the proportion of women aged 25–34 among those who are aged 15–49 in year t and country c $(\pi_{2534,t,c})$, and $\left(\frac{TFR^{*}_{i,c}}{\frac{C_{t,c}}{W_{t,c}}}\right)$ as a response.

$$\frac{TFR^{*}_{i,c}}{\frac{C_{t,c}}{W_{t,c}}} = \alpha_0 + \alpha_1 \pi_{2534,t,c}$$

After training the model of the HFD schedules, the following estimates were obtained: $\hat{\alpha}_0 = 10.65$ and $\hat{\alpha}_1 = -12.55$.

B.6 Infant Mortality Estimation

The estimation of the probability of death under age 5, denoted by q_{0-4}, depends upon the data availability for the country of interest. In our study, accurate child mortality estimates are available for Sweden, England & Wales, Finland and France for the entire historical period of interest (all the details can be found in Table B.1). However, for the previously mentioned countries, excluding Sweden, these estimates are provided for either

quinquennial or decennial time intervals. To bridge this gap and obtain yearly time series, we use linear interpolation.

In the other countries lacking historical child mortality data before a certain year, we adopt an assumption whereby mortality levels are approximated by the earliest recorded value. For instance, if the initial recorded estimate for child mortality is in 1790, we assume that mortality levels preceding 1790 closely resemble those observed in 1790.

From a statistical point of view, the mortality indicators, before child mortality data for a country are collected, are assumed to be generated from a normal distribution with a mean equal to the estimate during the first year of data collection and with a very small variance (0.01^2). Additionally, in absence of yearly time series of child mortality estimates after the first year of data collection, we again employ linear interpolation.

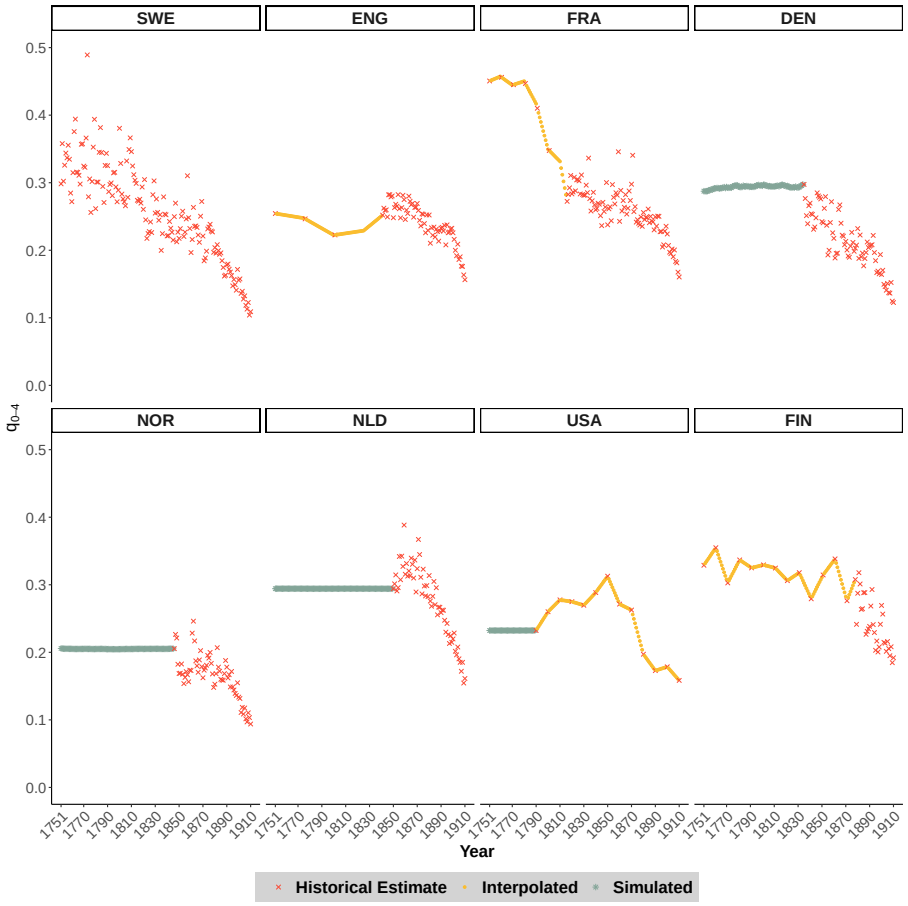


Figure B.6: Probability of death under age 5 (q_{0-4}) in the selected countries during the historical period 1750 – 1910. Distinct point shapes and colors are employ to distinguish how the estimates were obtained.

Appendix C

Appendix C: Supplemental Information

C.1 Simulating Age-specific Fertility Patterns

By simulating age-specific proportions of life-time fertility (see Figure C.1) for six hypothetical regions over a 10-year time window, we are able to capture a wide variety of age-specific fertility patterns. In order to obtain the trajectories in Figure C.1, we generated values for the model parameters according to the probability distributions specified above and employed the simulated values of the parameters to obtain numerical values for $\gamma^s_{x,a,t}$ according to Equation 4.7. The simulated numerical values for $\phi^s_{x,a,t}$ can be obtained by applying the following transformation to the simulated values of $\gamma^s_{x,a,t}$.

$$\phi^s_{x,a,t} = \frac{\exp\left(\gamma^s_{x,a,t}\right)}{\sum_{x=15}^{\omega^s} \exp\left(\gamma^s_{x,a,t}\right)} \tag{C.1}$$

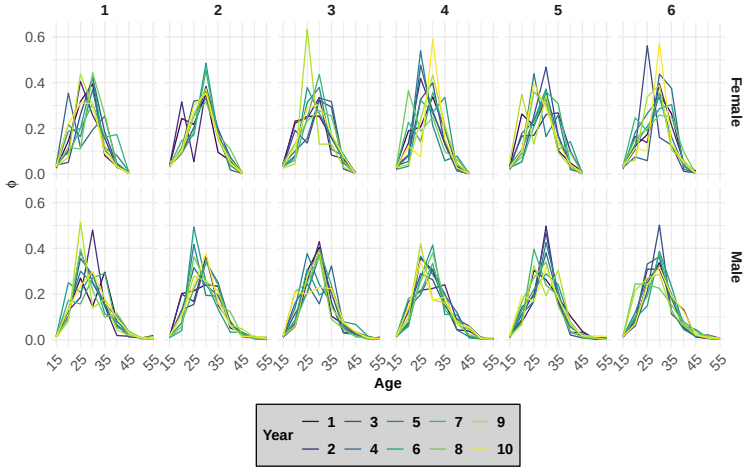


Figure C.1: The figure illustrates simulated age-specific fertility proportions ($\phi^s_{x,a,t}$) for both women and men from 6 hypothetical regions over a 10-year period.

C.2 Variance Estimation for Subnational Mortality Data

In the US data example, we used the estimates of the variances of the subnational age-specific death probabilities $\hat{q}^s_{x,a,t}$ to estimate the variances of the subnational age-specific person-years estimates. Specifically, for each combination of age group x, area a, time t and sex s, we transform the reported death probability estimates ($\hat{q}^s_{x,a,t}$) in the logit scale $\text{logit}(\hat{q}^s_{x,a,t})$ and simulate the transformed death probabilities from a normal distribution centered around the subnational estimate on the logit scale reported in the life table from the US Mortality Database. We draw a random sample of J logit-transformed death probabilities.

$$\text{logit}(q_{0,a,t})^{(j)} \sim \mathcal{N}\left(\text{logit}(\hat{q}_{0,a,t}), \left[\frac{\hat{\sigma}_{q_{0,a,t}}}{\hat{q}_{0,a,t}\cdot(1-\hat{q}_{0,a,t})}\right]^2\right) \quad \text{with} \quad j = 1,\ldots,J \tag{C.2}$$

$$\text{logit}(q^s_{x,a,t})^{(j)} \sim \mathcal{N}\left(\text{logit}(\hat{q}^s_{x,a,t}), \left[\frac{\hat{\sigma}_{q^s_{x,a,t}}}{\hat{q}^s_{x,a,t}\cdot(1-\hat{q}^s_{x,a,t})}\right]^2\right) \quad \text{with} \quad j = 1,\ldots,J \tag{C.3}$$

where J denotes the number of simulated probabilities from the statistical distribution, $\text{logit}(q^s_{x,a,t})^j$ indicates the $j-th$ simulated value in the random sample and $\left[\frac{\hat{\sigma}_{q^s_{x,a,t}}}{\hat{q}^s_{x,a,t}\cdot(1-\hat{q}^s_{x,a,t})}\right]^2$ is the variance of $\text{logit}(\hat{q}^s_{x,a,t})$ calculated using the Delta Method (see Van der Vaart (2000) for details). The quantity $\hat{\sigma}^2_{q^s_{x,a,t}}$ indicates the variance of the subnational death probability estimates from US mortality database. An identical notation is employed for the subnational death probability for children aged $0-4$ ($q_{0,a,t}$).

Afterwards, we transform the previous simulated values into the original scale. From the simulated samples of death probabilities by age group x, time t, county a and sex s

$$\begin{aligned} &q^{(1)}_{0,a,t},\ldots,q^{(j)}_{0,a,t},\ldots,q^{(J)}_{0,a,t} \\ &q^{s(1)}_{x,a,t},\ldots,q^{s(j)}_{x,a,t},\ldots,q^{s(J)}_{x,a,t} \end{aligned} \tag{C.4}$$

we employ standard life table relationships to obtain samples of the simulated person-years by age group x, sex s, area a and time t.

$$\begin{aligned} &\tilde{L}^{(1)}_{0,a,t}, \ldots, \tilde{L}^{(j)}_{0,a,t}, \ldots, \tilde{L}^{(J)}_{0,a,t} \\ &\tilde{L}^{s(1)}_{x,a,t}, \ldots, \tilde{L}^{s(j)}_{x,a,t}, \ldots, \tilde{L}^{s(J)}_{x,a,t} \end{aligned} \tag{C.5}$$

In order to estimate the uncertainty around the person-years estimates, we calculate the empirical variance of the person-years values for each combination of time t, sex s, age group x and area a. Hence, $\hat{\sigma}^2_{\hat{L}_{0,a,t}}$ and $\hat{\sigma}^2_{\hat{L}^s_{x,a,t}}$ are estimated as

$$\hat{\sigma}^2_{\hat{L}_{0,a,t}} = \frac{\sum_{j=1}^{J} \left(\tilde{L}^{(j)}_{0,a,t} - \bar{\tilde{L}}_{0,a,t} \right)^2}{J} \tag{C.6}$$

$$\hat{\sigma}^2_{\hat{L}^s_{x,a,t}} = \frac{\sum_{j=1}^{J} \left(\tilde{L}^{s(j)}_{x,a,t} - \bar{\tilde{L}}^s_{x,a,t} \right)^2}{J} \tag{C.7}$$

where $\bar{\tilde{L}}_{0,a,t} = \frac{\sum_{j=1}^{J} \tilde{L}^{(j)}_{0,a,t}}{J}$ and $\bar{\tilde{L}}^s_{x,a,t} = \frac{\sum_{j=1}^{J} \tilde{L}^{s(j)}_{x,a,t}}{J}$ are the empirical averages of the simulated person-years estimates.

For our US example, we simulate $J = 1,000$ values for the person-years for each combination of age group x, sex s, county a and time t. In Figure C.2, we show 50 simulated person-years estimates for Calaveras County in California in year 2000 by age group x and sex s. Given that this county is characterized by a relatively small population (roughly 46,000 in 2019), we observe some variations in the simulated person-years trajectories. Counties with high population sizes present almost no variation in the simulated person-years trajectories. As a consequence, the uncertainty around the Person-Year estimates from these counties will be very tiny.

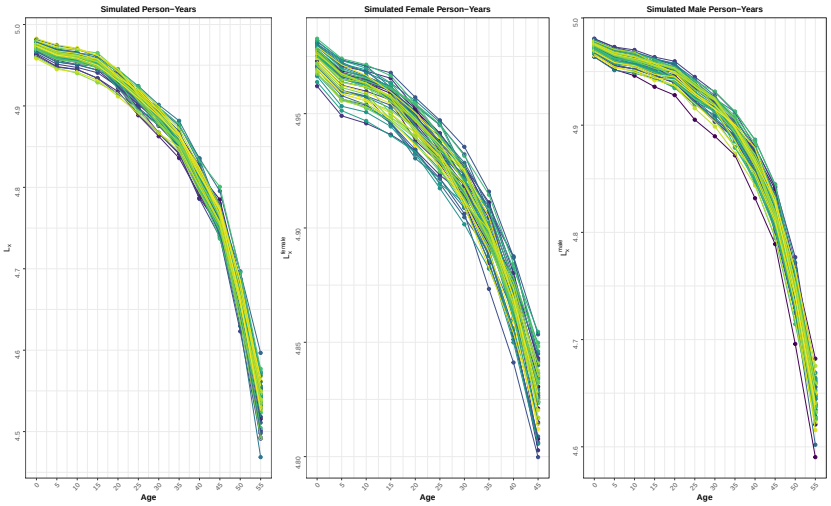


Figure C.2: $J = 50$ simulated person-years trajectories for Calaveras County in California in year 2000. Simulations are reported for the whole population of the county and by sex.

C.3 Details on Data Sources

Information	Content	Period	Geographical Detail	Source
Female fertility	Fertility indicators	1933-2021	US country	Human Fertility Database (link)
Male fertility	Fertility indicators	1969-2015	US country	Human Fertility Collection
Mortality	Life tables	1982-2019	US counties	US Mortality Database (link)
Fertility	Birth records	1969-2004	US states	US National Bureau of Economic Research (link)
Population	Population counts by age and sex	1969-2021	US counties	National Cancer Institute (link)
Female fertility	Births by maternal ages	1975-2023	Australian regions	Australian Bureau of Statistics (link)
Male fertility	Births by paternal ages	1975-2023	Australian regions	Australian Bureau of Statistics (link)
Population	Population counts by age and sex	1975-2023	Australian regions	Australian Bureau of Statistics (link)
Mortality	Life tables	2001-2020	Australian regions	Human Life-Table Database (link)

Table C.1: Summary of the data sources

C.4 Additional Figures

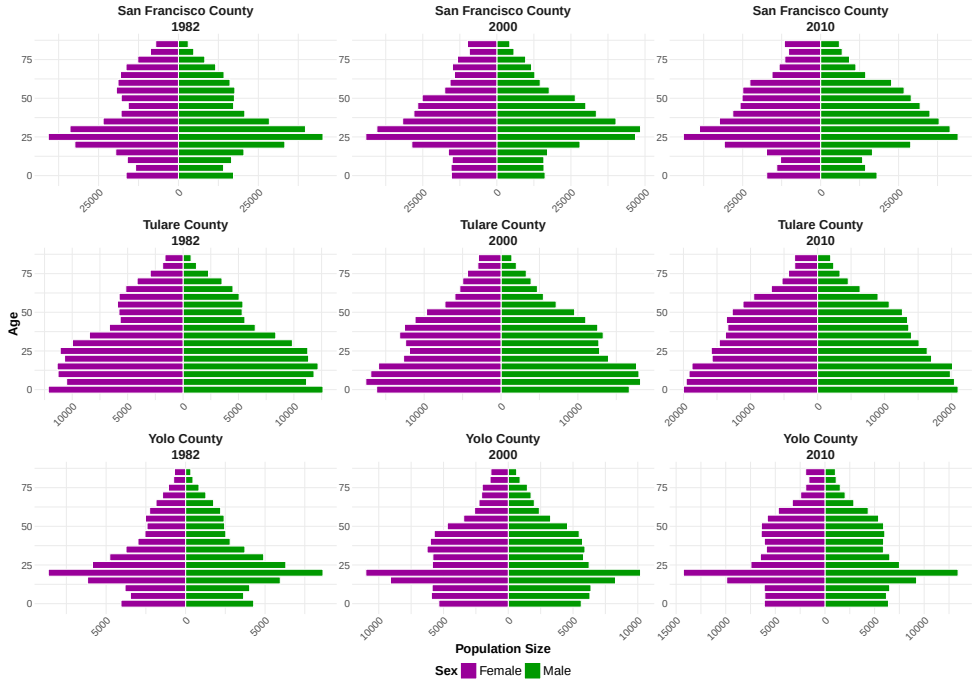


Figure C.3: Population pyramids for three selected California counties in years 1982, 2000, 2010.

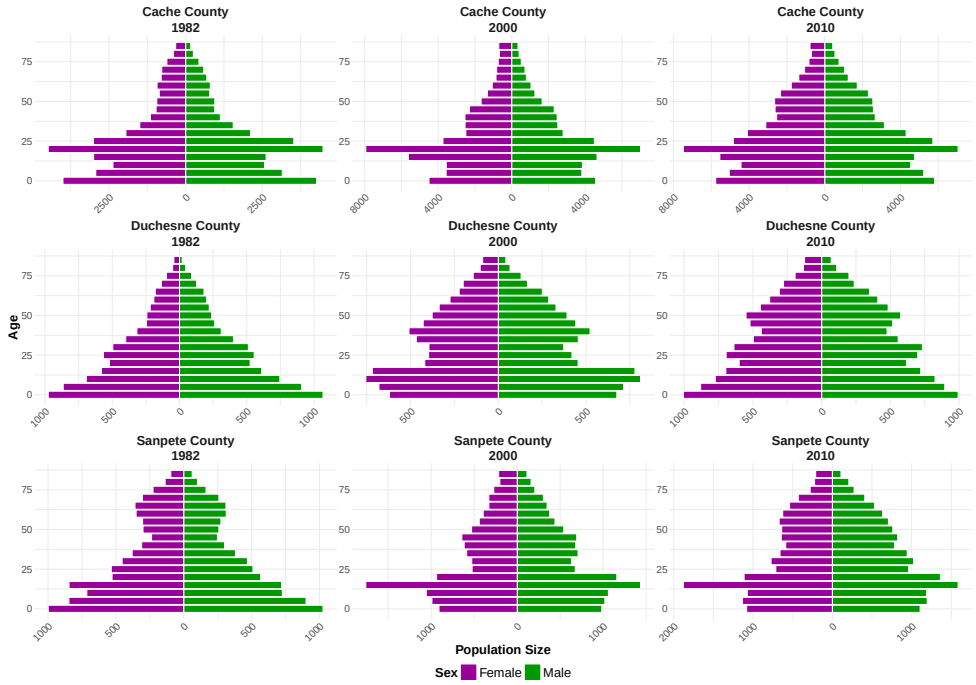


Figure C.4: Population pyramids for three selected Utah counties in years 1982, 2000, 2010.

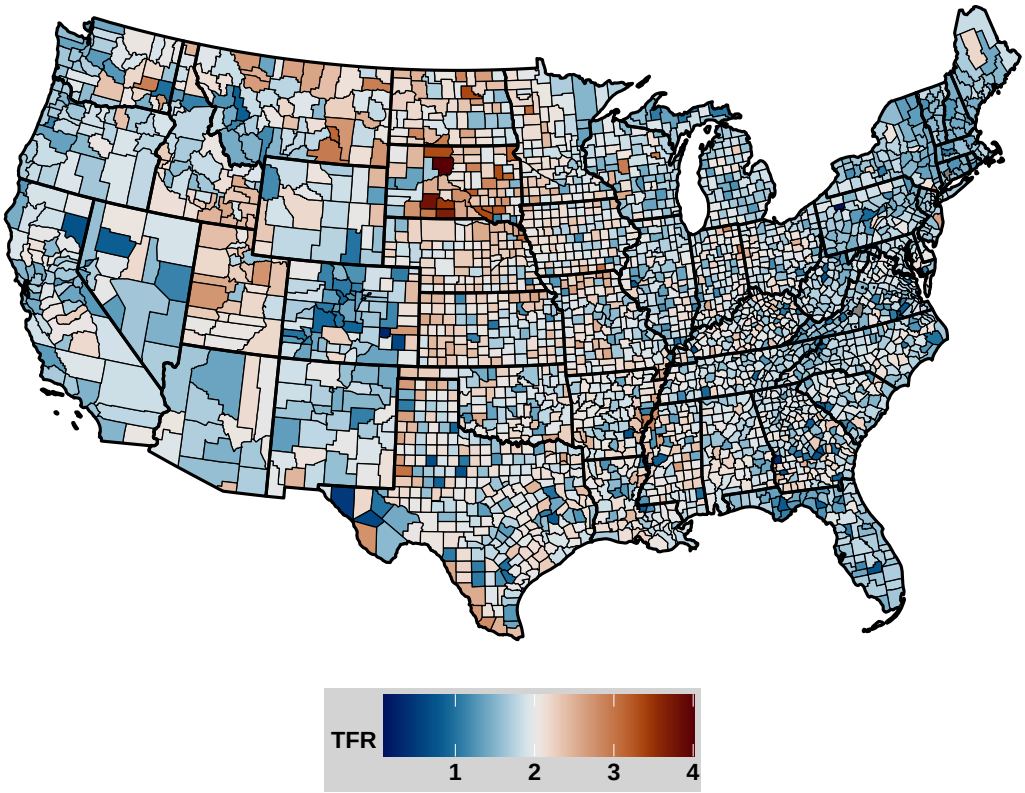


Figure C.5: Spatial distribution of median male TFR in 2019 for continental US.

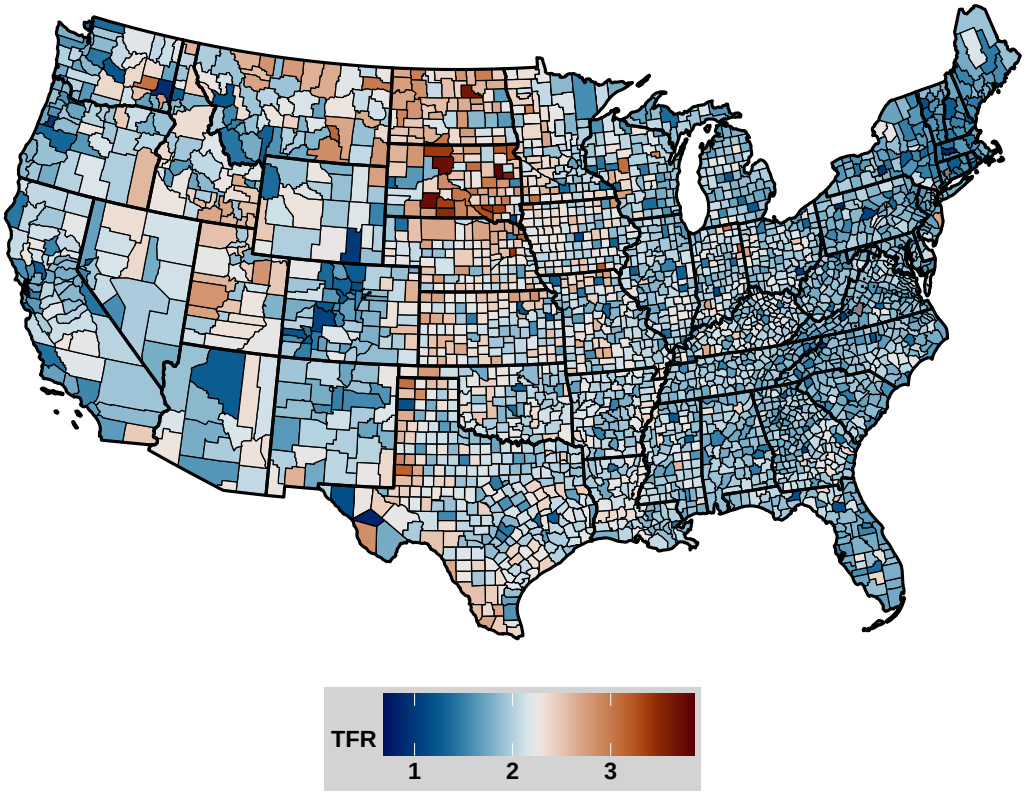


Figure C.6: Spatial distribution of median female TFR in 2019 for continental US.

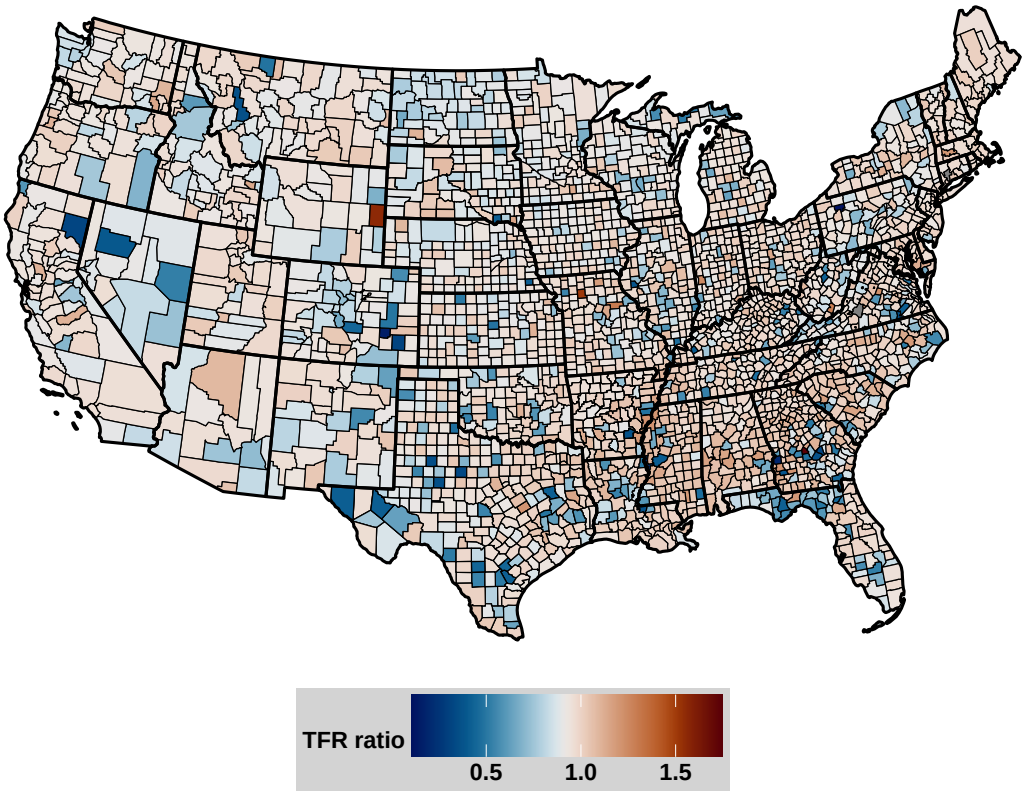


Figure C.7: Spatial distribution of male to female TFR ratio in 2019 for continental US.

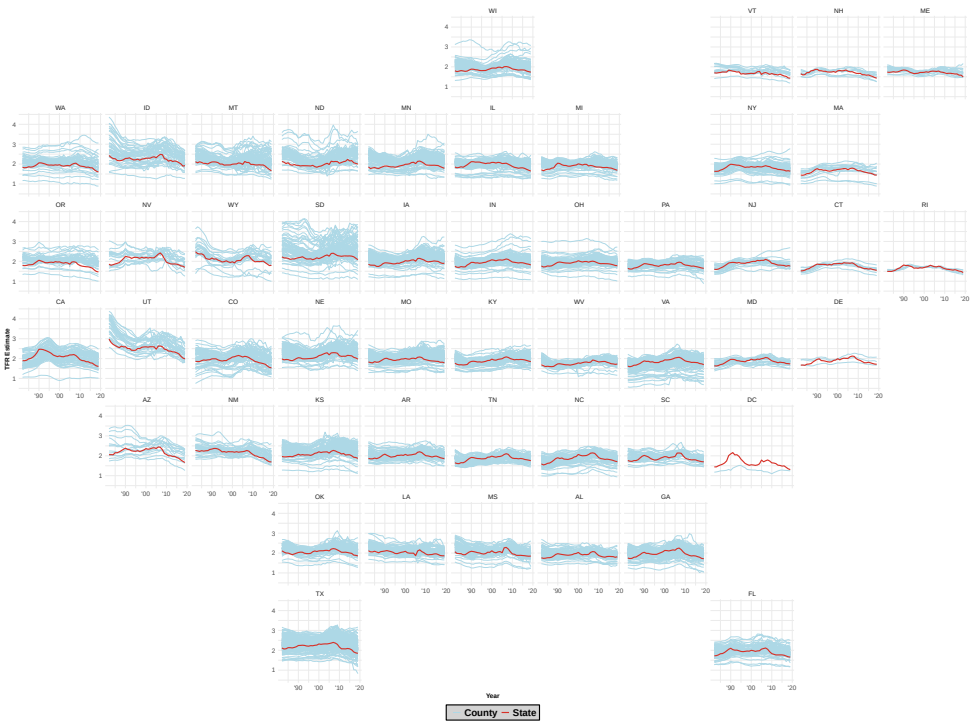


Figure C.8: County-specific time series for the median female TFR estimates (light blue lines) by state. Red lines denote the state-specific TFR values computed from birth registers.

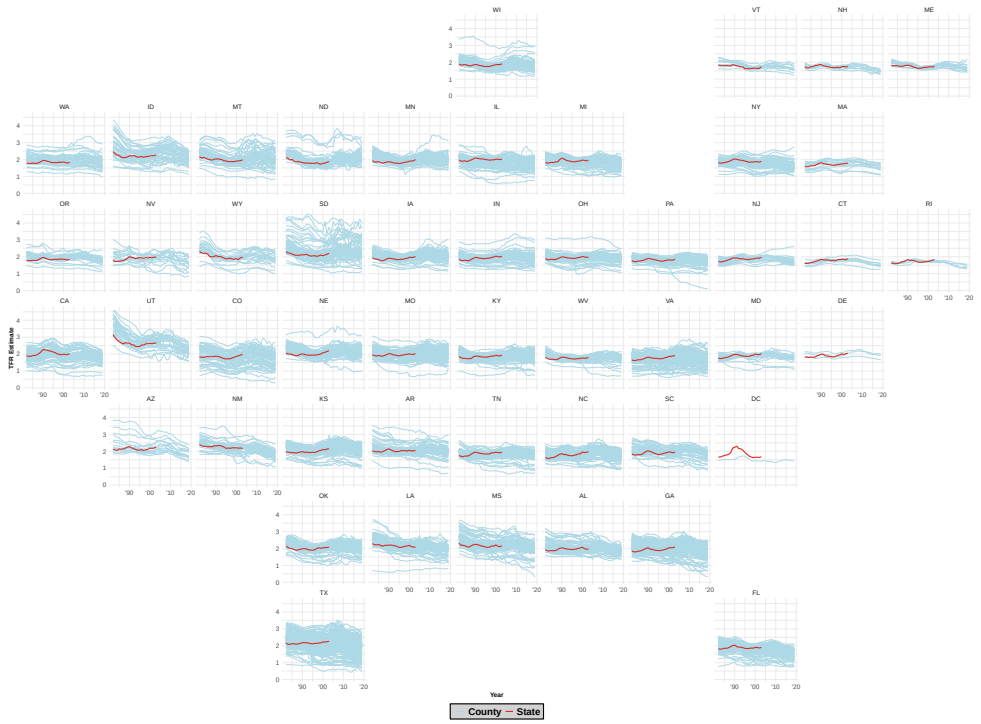


Figure C.9: County-specific time series for the median male TFR estimates (light blue lines) by state. Red lines denote the state-specific TFR values computed from birth registers (available until 2004).