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

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

[1] *Department of Statistical Sciences, University of Bologna*
[2] *Department of Statistical Sciences, University of Toronto*
[3] *Department of Sociology, University of Toronto*
[*] *Corresponding email:* riccardo.omenti2@unibo.it

October 2023

Short Abstract

Accurate subnational fertility estimates are crucial for shaping policy decisions across diverse sectors, including education, health care, and social welfare. One of the major challenges in producing these estimates is the presence of small populations, in which information about birth counts stratified by the age of the parent at the birth of the child may be lacking or inadequate. In this research 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 pyramids and models age-specific mortality and fertility patterns allowing for uncertainty. We present a real data application focusing on fertility estimation in US counties for the historical period $1982-2019$. Preliminary results reveal distinctive fertility trajectories for men and women across different US counties. Furthermore, the proposed model exhibits significant potential for the examination of male and female fertility behaviors across diverse regions and time frames in multiple countries.

Extended Abstract

1 Introduction

Precise subnational fertility estimates represent an essential tool for analyzing shifts in fertility patterns within a country. Small-area fertility estimates help researchers in the to identify compositional and contextual factors influencing fertility behaviors at a local level.

In this research paper we emphasize the importance of examining not only women's fertility at a subnational level but also that of men. While women's fertility has been well-documented globally, men's fertility tend to be neglected due to a significant lack of high quality information (Coleman (1995)). This bias mirrors the focus on women's fertility in data collection efforts. Collecting data on women's fertility behavior is comparatively easier due to a more precise definition of the childbearing age interval and superior information quality in surveys (Greene and Biddlecom (2000)). Nonetheless, confining fertility studies exclusively to women lead researchers to neglect the distinctive aspects associated with men's fertility behaviors (Schoumaker (2019)). Previous studies (Dudel and Klüsener (2016);Ratcliffe et al. (2000);Tragaki and Bagavos (2014)) have revealed significant differences in male and female fertility age patterns. Men have been found to have a broader reproductive age span, displaying age-specific fertility rates that are lower at younger ages and higher at older ages (Schoumaker (2017)). Furthermore, the literature reports substantial heterogeneity in the disparities between male and female TFRs. In low fertility settings, female fertility has been found to be slightly higher than that of men (Dudel and Klüsener (2016)). Conversely, in high fertility societies research has demonstrated notably higher fertility rates among men than among women (Tragaki and Bagavos (2014)). While variations in male and female total and age-specific fertility rates have been extensively examined at a national level (see Schoumaker (2019)), less attention has been devoted to the evaluation of male and female fertility at a subnational level. Previous research works by Schoumaker (2019) and by Dudel and Klüsener (2016) proposed a methodological framework to calculate indicators of male age-specific and total fertility rates at a national level exploiting survey data.

This research paper aims to propose a model for estimating the period Total Fertility Rates (TFR) across multiple geographical areas characterized by disparate age-sex population structures. Specifically, our modelling approach builds upon the framework proposed by Schmertmann and Hauer (2019). This model enables to estimate period TFRs using minimal input data, namely the number of children aged $0-4$ and the number of women in the reproductive age interval $15-49$, and does not require the knowledge of the number of births disaggragated by maternal age classes. Whereas the model by Schmertmann and Hauer (2019) focuses solely on the estimation of female Total Fertility Rates, our objective is to expand this modelling framework to the estimation of both male and female fertility at a subnational level. To the best of our knowledge, this research paper represents the first attempt to develop a model for estimating male Total Fertility Rates at a subnational level. Our model relies on a Bayesian hierarchical structure which enables to incorporate prior information about standard mortality schedules and common age-specific fertility patterns. In addition, it allows to build uncertainty intervals around TFR estimates. For this reason, our modelling approach helps to inform fertility patterns in smaller geographical areas, in which the uncertainty around the data is higher. A Bayesian hierarchical approach has been implemented by Ševčíková et al. (2018) for predicting female TFRs across multiple subnational regions. The former model builds upon the Bayesian method developed by the United Nations to predict national TFRs (see Alkema et al. (2011)) and is informed by standard fertility transition theory. Our modelling approach estimates subnational TFRs for both men and women using limited input data coupled with demographic knowledge about common mortality schedules and age-specific fertility patterns.

The forthcoming sections of this research paper will be structured around three main sections, namely Method section, Data Application section, and Concluding Remarks and Future Developments section. The Method section will provide a detailed description of our proposed model. The Data Application section will illustrate an empirical application of the proposed Bayesian model using population counts from US counties. The Preliminary Results and Future Developments section will offer some preliminary insights, along with directions for future developments.

2 Method

2.1 Model Setup

Let $C_{a,t}$ be the observed number of children in area a at time t. We assume that it can be modelled as a Poisson distribution:

$$C_{a,t}|K^{s}_{x,a,t} \sim \text{Pois}\Big(\sum_{x=15}^{\omega^s} K^{s}_{x,a,t}E^{s}_{x,a,t}\Big) \tag{1}$$

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

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

$$K^{s}_{x,a,t} = \Bigg[\frac{L^{s}_{x-5,a,t}}{L^{s}_{x,a,t}} \cdot F^{s}_{x-5,a,t} + F^{s}_{x,a,t}\Bigg] \cdot \frac{L_{0,a,t}}{2} \tag{2}$$

where $L^{s}_{x,a,t}$ denotes the expected person-years lived by individuals of sex s in age group x in area a at time t. $L_{0,a,t}$ denotes the person-years lived by individuals in the age group $0-4$ in area a at time 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 at time 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 2, which can be rewritten as

$$\begin{aligned} K^{s}_{x,t,c} &= TFR^{s}_{a,t} \cdot \frac{L_{0,a,t}}{5} \cdot \frac{1}{2}\Bigg[\frac{L^{s}_{x-5,a,t}}{L^{s}_{x,a,t}} \cdot \phi^{s}_{x-5,a,t} + \phi^{s}_{x,a,t}\Bigg] \\ &= TFR^{s}_{a,t} \cdot \frac{L_{0,a,t}}{5} \cdot p^{s}_{x,a,t} \end{aligned} \tag{3}$$

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 at time 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 at time t. $TFR^{s}_{a,t}$ is the period Total Fertility Rates experienced by either men or women in area a at time t. This demographic measure can be interpreted as the expected number of children per man (woman) in area a at time t if he (she) is subject to the current period age-specific fertility rates throughout his (her) reproductive ages. $p^{s}_{x,a,t}$ can be defined as an average of the life-time fertility experienced by either men or women in the age groups x and $x-5$. Note that we are assuming that fertility does not begin until age 15; hence, we set $\phi^{s}_{10,a,t} = 0$.

Equation 3 decomposes the expected number children per man (woman) in age group x in area a at time t as a product of three multiplicative factors. The first two factors, i.e., $TFR^{s}_{a,t}$ and $\frac{L_{0,a,t}}{5}$, remain constant across the reproductive age groups, while the third factor $p^{s}_{x,a,t}$ varies with age.

Our proposed Bayesian model 3 incorporates demographic knowledge and uncertainty about demographic quantities by placing models and priors on mortality and fertility parameters in equation 2.

2.2 Model for Age-specific Fertility

To incorporate prior knowledge about age fertility schedules, we model the share of life time fertility in a age group x on the log scale as

$$\boldsymbol{\gamma}^{s}_{a,t} = \boldsymbol{m}^{s} + \boldsymbol{y}^{s}_{1}\beta^{s}_{1,a,t} + \boldsymbol{y}^{s}_{2}\beta^{s}_{2,a,t} \tag{4}$$

where $\gamma^{s}_{x,a,t} = \log\left(\frac{\phi^{s}_{x,a,t}}{\phi^{s}_{15,a,t}}\right)$ 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. $\boldsymbol{\gamma}^{s}_{a,t}$ is vector whose elements are $\gamma^{s}_{x,a,t}$ at the distinct reproductive age groups. The size of such vector is 9 for men and 7 for women. The transformation of $\phi^{s}_{x,a,t}$ ensures that the elements of the vector $\boldsymbol{\gamma}^{s}_{a,t}$ on the right hand-side of equation 4 can

assume both positive and negative values.

$\boldsymbol{m}^s$, $\boldsymbol{y}_1^s$ and $\boldsymbol{y}_2^s$ 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 fertility schedules ($\gamma_{x,a,t}^s$), while $\boldsymbol{y}_1^s$ and $\boldsymbol{y}_2^s$ are the first and second left-singular vectors which are obtained via a Singular Value Decomposition on the matrix $\boldsymbol{Y}^s$ whose columns are log-transformed male (female) age-specific fertility schedules.
For example, in our application to US counties we employ the US national age-specific female fertility curves for the historical period $1933 - 2021$ retrieved from the Human Fertility Database (Jasilioniene et al. (2015)). US age-specific male fertility schedules are derived from the Human Fertility Collection (Grigorieva et al. (2015)) and cover the historical period $1969 - 2015$.

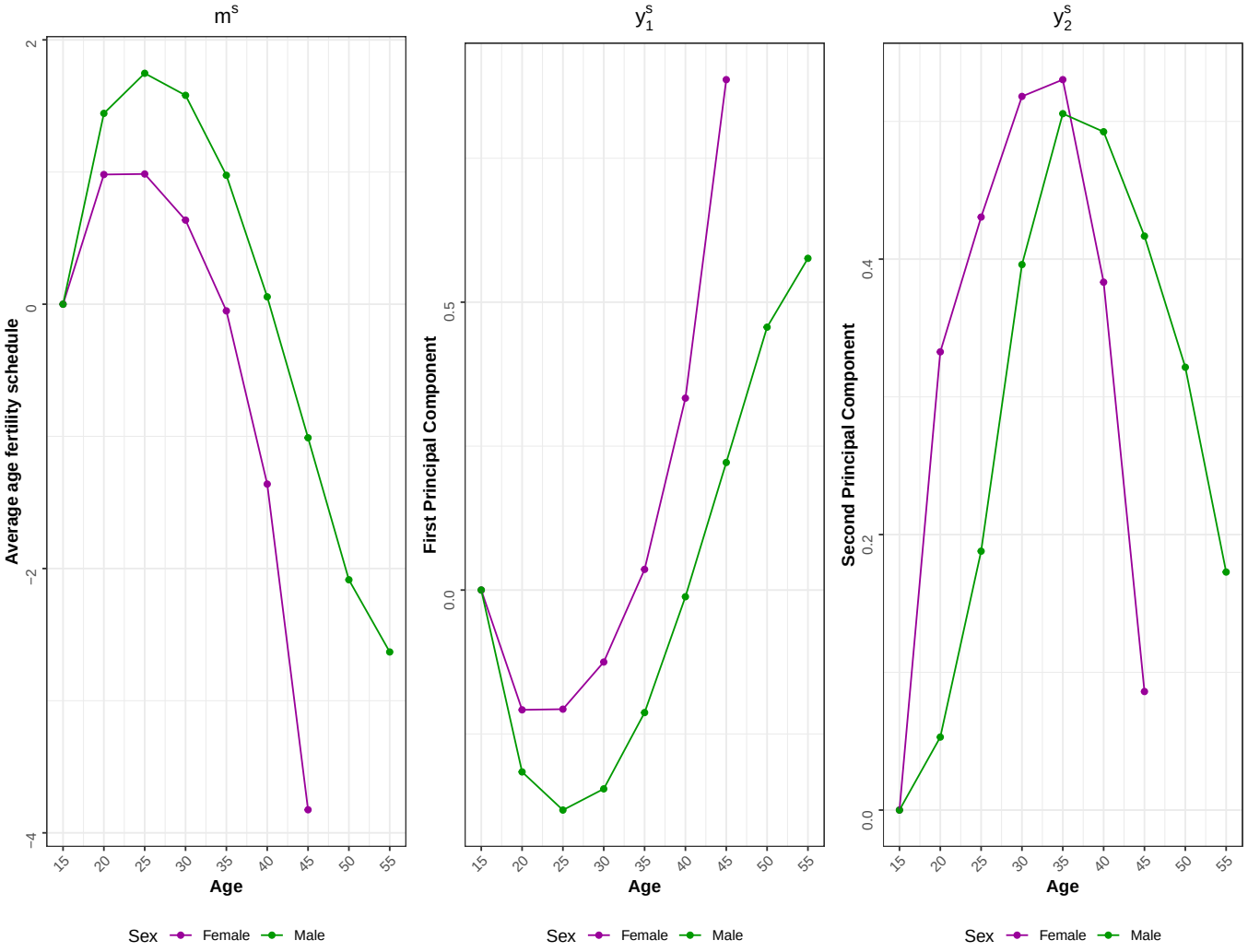

Figure 1: The figure provides an example of the Principal Component Analysis applied to logged U.S. age- and sex-specific fertility proportions ($\gamma_{x,a,t}^s$). The first plot illustrates the average logged fertility rates proportions across the distinct reproductive age classes by sex. The second and third plots display the values of the first ($\boldsymbol{y}_1^s$) and second ($\boldsymbol{y}_2^s$) left-singular vectors of the matrix ($\boldsymbol{Y}^s$) separately for men and women.

The mean $\boldsymbol{m}^s$ describes the overall age-specific fertility curve. As expected, both men's and women's 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 narrower female reproductive age span. For both men and women, the first principal component seems to allow for the postponement of the mean age at childbearing as its components are strictly increasing for both men and women starting from the age class $25 - 29$. On the other hand, the second principal component enables higher fertility levels in the reproductive age classes $30 - 34$ for women and $35 - 39$ for men.

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

$$\beta^s_{1,a,t}, \beta^s_{2,a,t} \sim \mathcal{N}(0,1) \tag{5}$$

Setting the distribution of such parameters to a standard normal ensures that their support is restricted to the interval $[-2,2]$ in order to better mimic the national age-specific fertility curves.
Overall, we believe the inclusion of the first two principal components is sufficient to capture the most common age-specific fertility schedules. Indeed, by simulating multiple trajectories for the age-specific proportions of life-time fertility (see figure 2), we are able to capture a wide variety of age-specific fertility patterns. In order to obtain the trajectories in figure 2, we generated values for the coefficients $\beta^s_{1,a,t}$ and $\beta^s_{2,a,t}$ from the standard normal distribution and employed the simulated values of the parameters to obtain numerical values for $\gamma^s_{x,a,t}$. The simulated numerical values for $\phi^s_{x,a,t}$ can be obtained by applying the following transformation

$$\phi^s_{x,a,t} = \frac{\exp\left(\gamma^s_{x,a,t}\right)}{\sum_{x=15}^{\omega^s} \exp\left(\gamma^s_{x,a,t}\right)} \tag{6}$$

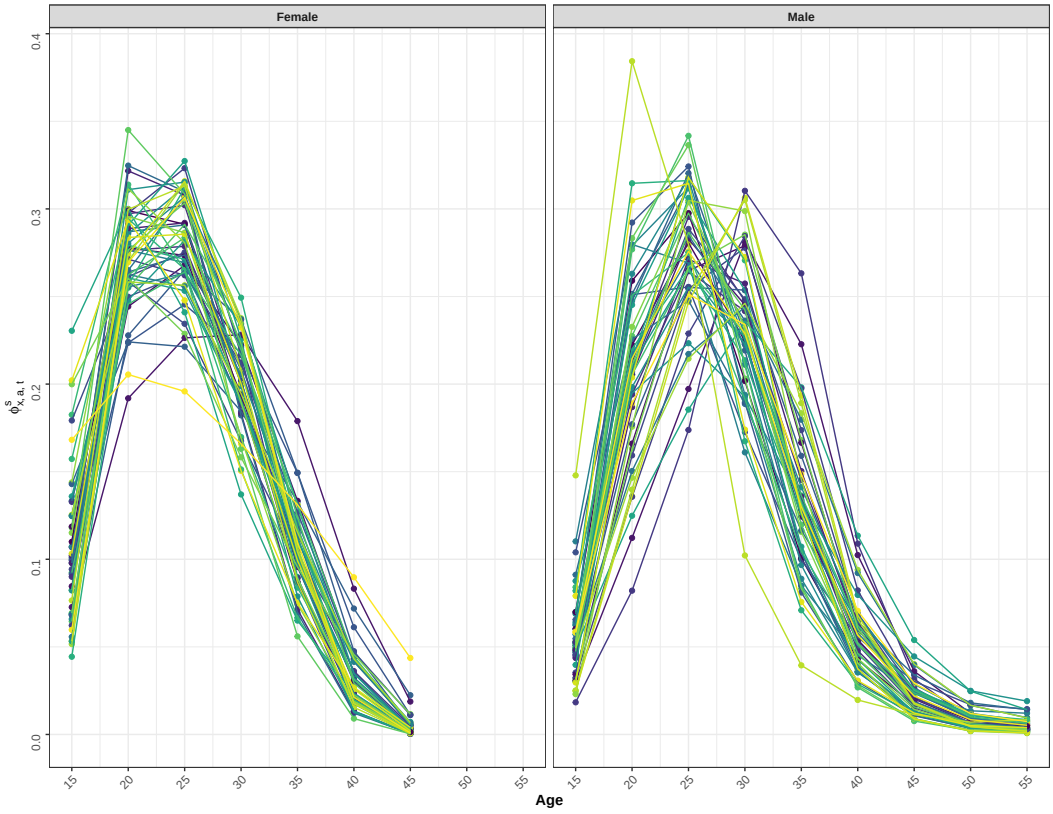

Figure 2: The figure illustrates 50 possible patterns for the age-specific fertility proportions ($\phi^s_{x,a,t}$) for both women and men.

Priors on Total Fertility Rates

The Total fertility rate $TFR^s_{t,a}$ from equation 3 is modelled through a normal distribution, whose mean is assumed to be equal to the national Total Fertility Rate observed at time t ($TFR^{nat,s}_t$) for either men or women.

$$TFR^s_{t,a} \sim \mathcal{N}(TFR^{nat,s}_t, \sigma^2_{TFR^s_{t,a}}) \tag{7}$$

The standard deviation parameter $\sigma_{TFR^s_{t,a}}$ is assigned a non-informative prior. In practice, we assign to such parameter a normal distribution with mean equal to 0 and variance equal to 2.

$$\sigma_{TFR^{s}_{t,a}} \sim \mathcal{N}(0, 2)$$

The practical implication of centering the prior distribution of $TFR^{s}_{t,a}$ to the national value $TFR^{nat,s}_{t}$ is to shrink counties with a small population towards the national average. In this manner, fertility levels in small counties are partially informed by the national fertility level. On the contrary, fertility levels in larger counties are primarily determined by their population counts.

2.3 Priors on Mortality Parameters

Schmertmann and Hauer (2019) modelled 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 ($L_{0,a,t}$ and $L^{s}_{x,a,t}$) of equation 3. Specifically, we can take advantage of the availability of US subnational life table estimates. In our data example, we used county-specific life tables for the historical period $1982-2019$ from the US Mortality Database (visit website usa.mortality.org for downloading the data and see Alexander et al. (2017) for the methodological details). However, we acknowledge that the direct modelling of the Person-Years is feasible provided that subnational mortality life tables are available for the country of interest. In case we lack detailed subregional mortality data, we can resort to other mortality models such as the log-quadratic model that only requires the knowledge of child mortality.

Operationally, in order to include the uncertainty associated to the subnational Person-Years estimates in 3, we assume that the Person-Years $L^{s}_{x,a,t}$ for individuals of sex s in an age group x in area a at time 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.

$$\tilde{L}_{0,a,t} \sim \mathcal{N}\left(\hat{L}_{0,a,t}, \hat{\sigma}^{2}_{\hat{L}_{0,a,t}}\right) \tag{8}$$

$$\tilde{L}^{s}_{x,a,t} \sim \mathcal{N}\left(\hat{L}^{s}_{x,a,t}, \hat{\sigma}^{2}_{\hat{L}^{s}_{x,a,t}}\right) \tag{9}$$

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. $\hat{\sigma}^{2}_{\hat{L}_{0,a,t}}$ is the corresponding variance calculated by means of simulations. This parameter is included to incorporate information about child mortality. Similarly, $\hat{L}^{s}_{x,a,t}$ denotes the estimated sex-specific Person-Years for the age classes $x = 10, 15, \ldots, 45$ for women and $x = 10, 15, \ldots, 55$ for men. Such estimates are obtained from a sex-specific subnational life table referring to area a and time t. $\hat{\sigma}^{2}_{\hat{L}^{s}_{x,a,t}}$ is the corresponding variance of the estimates.

In our US example, we use the estimates of the variances of the subnational death probabilities $\hat{q}^{s}_{x,a,t}$ to estimate the variances of the subnational 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

$$\text{logit}(q_{0,a,t})^{(j)} \sim \mathcal{N}\left(\text{logit}(\hat{q}_{0,a,t}), \left[\frac{\hat{\sigma}_{q_{0,a,t}}}{\hat{q}_{0,a,t} \cdot (1 - \hat{q}_{0,a,t})}\right]^{2}\right) \quad \text{with} \quad j = 1, \ldots, J \tag{10}$$

$$\text{logit}(q^{s}_{x,a,t})^{(j)} \sim \mathcal{N}\left(\text{logit}(\hat{q}^{s}_{x,a,t}), \left[\frac{\hat{\sigma}_{q^{s}_{x,a,t}}}{\hat{q}^{s}_{x,a,t} \cdot (1 - \hat{q}^{s}_{x,a,t})}\right]^{2}\right) \quad \text{with} \quad j = 1, \ldots, J \tag{11}$$

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

$$\begin{matrix} q^{(1)}_{0,a,t}, \ldots, q^{(j)}_{0,a,t}, \ldots, q^{(J)}_{0,a,t} \\ q^{s(1)}_{x,a,t}, \ldots, q^{s(j)}_{x,a,t}, \ldots, q^{s(J)}_{x,a,t} \end{matrix} \tag{12}$$

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.

$$\begin{matrix} \tilde{L}^{(1)}_{0,a,t}, \ldots, \tilde{L}^{(j)}_{0,a,t}, \ldots, \tilde{L}^{(J)}_{0,a,t} \\ \tilde{L}^{s(1)}_{x,a,t}, \ldots, \tilde{L}^{s(j)}_{x,a,t}, \ldots, \tilde{L}^{s(J)}_{x,a,t} \end{matrix} \tag{13}$$

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

$$\hat{\sigma}^2_{\hat{L}_{0,a,t}} = \frac{\sum_{j=1}^{J} \left(\tilde{L}^{(j)}_{0,a,t} - \bar{\tilde{L}}_{0,a,t} \right)^2}{J} \tag{14}$$

$$\hat{\sigma}^2_{\hat{L}^s_{x,a,t}} = \frac{\sum_{j=1}^{J} \left(\tilde{L}^{s(j)}_{x,a,t} - \bar{\tilde{L}}^s_{x,a,t} \right)^2}{J} \tag{15}$$

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 U.S. 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.

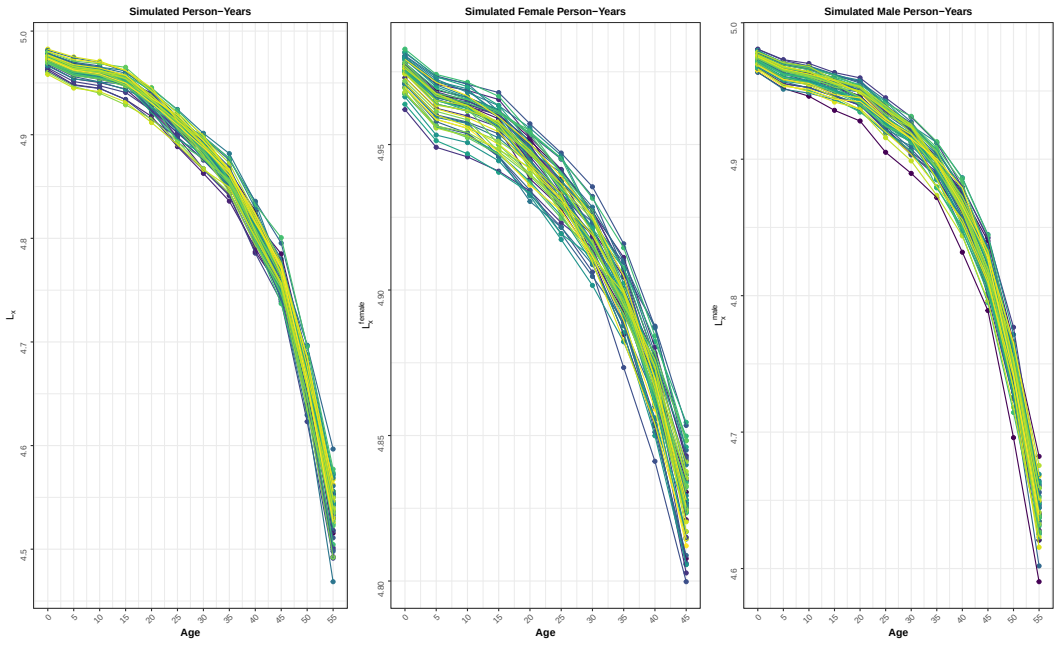

Figure 3: $J = 50$ simulated Person-Years trajectories for Calaveras County in California in year 2000. Simulations are reported for the whole population of the county and by sex.

In figure 3, 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.

2.4 Model summary

The following set of equations summarizes the whole model set up. The number of children aged $0-4$ are assumed to be Poisson-distributed (equation 16). The fertility and mortality parameters that govern the Poisson-process (equation 17) are specified by imposing a hierarchical structure. Preliminary information about the overall fertility level is included by placing a probability distribution on the parameter $TFR^{s}_{t,a}$ (equation 21) and on its standard deviation $\sigma_{TFR^{s}_{t,a}}$ (equation 22). Prior knowledge about the age-specific fertility patterns is incorporated by specifying model 18 and a probability distribution on the parameters $\beta^{s}_{1,a,t}$ and $\beta^{s}_{2,a,t}$ (20). Finally, we include prior knowledge about child and adult mortality by placing a prior probability distributions directly on the Person-Years parameters (23 and 24)

$$C_{a,t}|K^{s}_{x,a,t} \sim \text{Pois}\left(\sum_{x=15}^{\omega^{s}} K^{s}_{x,a,t}E^{s}_{x,a,t}\right) \tag{16}$$

$$K^{s}_{x,a,t} = TFR^{s}_{a,t} \cdot \frac{L_{0,a,t}}{5} \cdot \frac{1}{2}\left[\frac{L^{s}_{x-5,a,t}}{L^{s}_{x,a,t}} \cdot \phi^{s}_{x-5,a,t} + \phi^{s}_{x,a,t}\right] \tag{17}$$

$$\begin{aligned} \boldsymbol{\gamma}^{s}_{a,t} &= \boldsymbol{m}^{s} + \boldsymbol{y}^{s}_{1} \cdot \beta^{s}_{1,a,t} + \boldsymbol{y}^{s}_{2} \cdot \beta^{s}_{2,a,t} \\ \boldsymbol{\gamma}^{s}_{a,t} &= [\gamma^{s}_{15,a,t}, \ldots, \gamma^{s}_{\omega^{s},a,t}] \end{aligned} \tag{18}$$

$$\phi^{s}_{x,a,t} = \frac{\exp\left(\gamma^{s}_{x,a,t}\right)}{\sum_{x=15}^{\omega^{s}} \exp\left(\gamma^{s}_{x,a,t}\right)} \tag{19}$$

$$\beta^{s}_{1,a,t}, \beta^{s}_{2,a,t} \sim \mathcal{N}(0,1) \tag{20}$$

$$TFR^{s}_{t,a} \sim \mathcal{N}(TFR^{nat,s}_{t}, \sigma^{2}_{TFR^{s}_{t,a}}) \tag{21}$$

$$\sigma_{TFR^{s}_{t,a}} \sim \mathcal{N}(0,2) \tag{22}$$

$$\tilde{L}_{0,a,t} \sim \mathcal{N}\left(\hat{L}_{0,a,t}, \hat{\sigma}^{2}_{\hat{L}_{0,a,t}}\right) \tag{23}$$

$$\tilde{L}^{s}_{x,a,t} \sim \mathcal{N}\left(\hat{L}^{s}_{x,a,t}, \hat{\sigma}^{2}_{\hat{L}^{s}_{x,a,t}}\right) \tag{24}$$

The subscript s refers to sex, the subscript t refers to time and the subscript a denotes the subregion. The subscript x denotes the age with ω^{s} being the last reproductive age group. ω^{s} is set to be $45-49$ for women and $55-59$ for men.

2.5 Model implementation

Our final goal is to draw sample from the posterior distribution of the $TFR^{s}_{a,t}$ in an area a at time t conditional on the observed number of children aged 0-4 $C_{a,t}$ and the counts of men (women) $E^{s}_{15,a,t}, \ldots, E^{s}_{\omega^{s},a,t}$. Note that we do not place any probability distribution on the number of exposed men and women.
If we let $\boldsymbol{\theta}^{s}_{t,a} = (TFR^{s}_{a,t}, \sigma_{TFR^{s}_{t,a}}, \boldsymbol{\beta}^{s}_{a,t}, \tilde{L}_{0,a,t}, \tilde{\boldsymbol{L}}^{s}_{a,t})$ be a vector containing the model parameters with $\tilde{\boldsymbol{L}}^{s}_{a,t} = [\tilde{L}^{s}_{15,a,t}, \ldots, \tilde{L}^{s}_{\omega^{s},a,t}]$ and $\boldsymbol{\beta}^{s}_{a,t} = [\beta^{s}_{1,a,t}, \beta^{s}_{2,a,t}]$, the posterior distribution for the model parameters conditional on the observed data is given by

$$\mathcal{P}(\boldsymbol{\theta}^{s}_{t,a}|C_{a,t},\boldsymbol{E}^{s}_{a,t}) \propto \mathcal{L}(C_{a,t}|\boldsymbol{\theta}^{s}_{t,a}) \cdot f_{TFR}(TFR^{s}_{a,t}) \cdot f_{\sigma}(\sigma_{TFR^{s}_{t,a}}) \cdot \prod_{i=1}^{2} f_{\beta}(\beta^{s}_{i,a,t}) \cdot f_{L_0}(\tilde{L}_{0,a,t}) \cdot \prod_{j=1}^{\omega^s} f_{L^s}(\tilde{L}^{s}_{j,a,t}) \quad (25)$$

where $\mathcal{L}$ refers to the likelihood of the Poisson and the functions f denote the prior densities for the unknown parameters.

The marginal posterior for $TFR^{s}_{a,t}$ is obtained by integration.

$$\mathcal{P}(TFR^{s}_{a,t}|C_{a,t}) = \int \mathcal{P}(\boldsymbol{\theta}^{s}_{t,a}|C_{a,t},\boldsymbol{E}^{s}_{a,t}) d\sigma_{TFR^{s}_{t,a}} d\boldsymbol{\beta}^{s}_{a,t} d\tilde{L}_{0,a,t} d\tilde{\boldsymbol{L}}^{s}_{a,t} \quad (26)$$

Operationally, posterior samples were obtained using Markov Chain Monte Carlo (MCMC) algorithm implemented in Stan by means of the package *rstan* available in the Statistical Software R (see Carpenter et al. (2017) for details).

3 Data Application

3.1 Application to US counties

We applied our proposed model to estimate period TFRs by US counties for the historical period $1982-2019$. We used population data from the US censuses (downloaded from the website www.census.gov/en.html). National TFR values were retrieved from the Human Fertility Database for women and from the Human Fertility Collection for men. $1,000$ simulations were carried out for the estimation of the variances of the Person-Years.

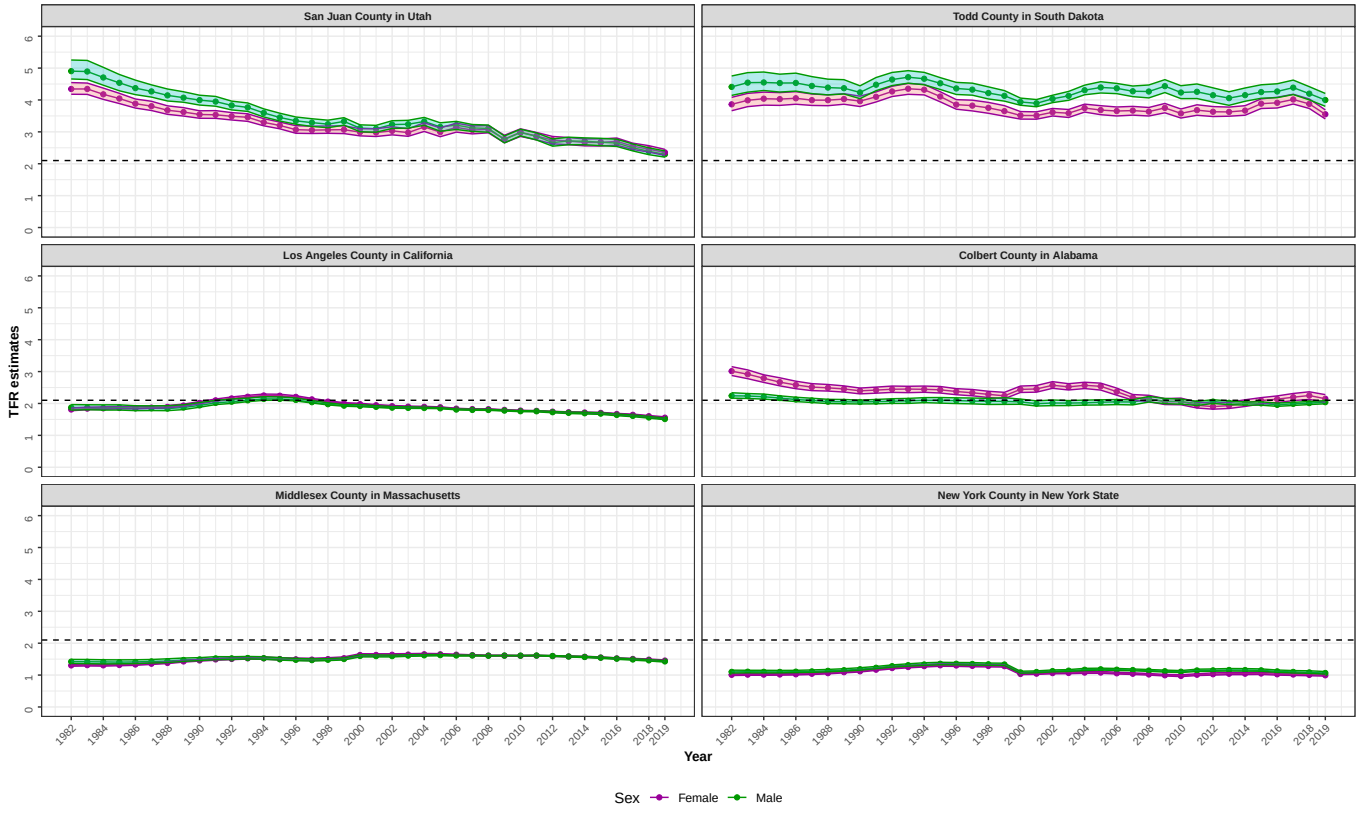

Figure 4: Times series of men's and women's period TFR estimates for six U.S. counties across the period $1982-2019$. 90% credible intervals are also provided. The dotted line represents the replacement fertility level of 2.1.

Figure 4 displays TFR estimates for six US counties characterized by distinct age-sex population structures and from states exhibiting different fertility levels for the period $1982-2019$. The top two graphs show the TFRs

for two rural counties with small population sizes (less than $10,000$ in 2019), one in Utah and one in South Dakota. In the Utah county, we observe that as the TFR declines for both the men and women towards the replacement value of 2.1, their fertility levels seem to overlap. Conversely, in the South Dakota county the TFR remains well above 3, the TFR of men surpasses that of women for the whole historical period. The middle two graphs show the results for two counties with TFRs close to the replacement levels, namely Los Angeles county in California and Colbert County in Alabama. The former does not display significant differences between men's and women's fertility. On the contrary, the latter shows a convergence of the fertility behaviors of men and women. Nonetheless, up to the early 2000s, this county clearly display higher fertility levels among women than among men. The third row display the results for two counties characterized by low fertility levels, New York City county in New York and Middlesex County in Massachusetts. New York City county presents no significant differences between men and women across time and fertility levels below 1.5. Note that in this county we may underestimate the TFRs of both man and women as many individuals in the age range $20-30$ may live New York City for working reasons and may have their own family elsewhere. Middlesex County is characterized by low fertility levels throughout the historical period in examination and does not display any noteworthy differences between men and women.

4 Concluding Remarks and Future Developments

4.1 Concluding Remarks

Building on previous work by Schmertmann and Hauer (2019), we have proposed a Bayesian hierarchical model to estimate subregional TFRs for both men and women using minimal data from age-sex population pyramids. Our proposed model relies on standard age-specific fertility schedules, on national fertility levels as well as on subnational mortality estimates to capture the inherent uncertainty in the demographic parameters that govern the Poisson-process for modelling number of children. Age-specific fertility schedule are informed by characteristic age-specific fertility curves for both men and women. Preliminary knowledge on the TFR parameter is partially informed by the national TFR. The availability of detailed subregional mortality data for the US enables us to place a prior probability distribution directly on the Person-Years without the necessity to employ more sophisticated mortality models. The application of the proposed model to US counties yielded distinctive fertility trajectories for men and women, prompting the need for further investigation.

4.2 Future Developments

In the upcoming months, we wish to refine our model so that it is flexible enough to allow the estimation of subnational TFR levels for multiple countries. This might involve employing different statistical models and priors for the fertility and mortality parameters as well as enlarging the male reproductive age span. Additionally, we recognize the importance of understanding the impact of the number of exposed men and women in the estimation of the TFR. In subregions experiencing high female or male migration, disparities between male and female TFRs may be biased. For this reason, we aim to examine the possible effect of migration on the TFR estimates of our proposed model.
Overall, we argue that this research paper could provide significant contributions to the field of small-area fertility estimation. By estimating subnational fertility levels for both men and women in a wide variety of countries, we intend to shed new light on our understanding of subregional fertility trends in contemporary societies.

References

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

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.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, 76.

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

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

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.

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

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*.

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

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*, 78:570–579.

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). Across the world, is men's fertility different from that of women? *Population Societies*, 548(9):1–4.

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

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

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

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

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.