Estimating subnational fertility rates with a principal component-based Bayesian spatiotemporal model

Ameer Dharamshi[1], Celeste Winant[2], Magali Barbieri[2,3], and Monica Alexander[4,5]

[1]Department of Biostatistics, University of Washington
[2]Department of Demography, University of California Berkeley
[3]Institut National d'Études Démographiques
[4]Department of Statistics, University of Toronto
[5]Department of Sociology, University of Toronto

September 15, 2024

Abstract

Understanding the economic, social, and cultural consequences of declining fertility rates requires high-quality fertility data over long periods of time and at fine geographic scales. This is a challenging task as at subnational levels, small population counts lead to high stochasticity; consequently, zero birth observations are a frequent occurrence. In this paper, we propose a principal component-based Bayesian spatiotemporal model to estimate age-specific fertility rates in small subnational areas. The model exploits structural patterns in fertility through the principal components, and stabilizes estimation by pooling information on the local manifestation of these patterns in space and time. We apply our model to county-level fertility data from California for 1982-2022 to estimate age-specific fertility rates and downstream fertility indicators. The model appears to perform well in these initial experiments and we identify several interesting patterns in Californian fertility for further study.

1 Introduction

Recent decades have seen consistent declines in fertility in most countries across the globe. Highly industrialized countries in particular are experiencing fertility rates well below the replacement level of 2.1 children per woman. The United States was an outlier for many years, maintaining higher fertility relative to peer countries, though it too has experienced sharp declines over the past 15 years, reaching a record low of 1.6 in 2023 (National Center for Health Statistics, 2024).

Despite the unprecedented nature of this phenomenon, in the United States context, relatively little research has been conducted studying the causes and consequences of low fertility. Yet the importance of this research is quite clear: studies of European and East Asian experiences have found that changes in the demographic structure triggered by low fertility have far-reaching consequences ranging from instability in social services to challenging economic environments due to a shrinking labour force (see for example Bloom et al. 2008). One of the primary barriers impeding research in this area has been the lack of high-quality easily accessible fertility data extending over long enough time periods to study trends, and disaggregated at small enough geographic levels to study differential patterns by social, policy, economic, or other contexts.

In this abstract, we propose a method to produce small-area fertility estimates in order to fill this gap. In small populations, classical demographic methods cannot be used due to high stochasticity in observed demographic counts. We thus propose a new principal component-based Bayesian spatiotemporal model to estimate age-specific fertility rates for all counties in the United States for the years 1982-2022. The model stabilizes estimation in small populations through two mechanisms. First, we exploit the fact that there

are known regularities in fertility schedules (Schmertmann et al., 2014). We extract such patterns from aggregate large population reference data for use as building blocks when constructing small population fertility curves. This approach has seen much success in subnational mortality rate estimation (Alexander et al., 2017; Dharamshi et al., 2024). We then introduce spatiotemporal structures that allow information to be shared across space and time in recognition of the fact that fertility patterns are expected to change gradually along these dimensions.

The estimates produced by our model will form the first complete series of county-level age-specific fertility rates in the United States. These results will be made freely available as part of the new United States Fertility Database (USFDB), and will be a useful resource for the demographic community to answer pressing questions regarding the future of fertility in the United States.

2 Data

The proposed model is applied to official county-level birth and population statistics (National Center for Health Statistics, 2024; Census Bureau, 2024). The birth data are processed from the restricted-use Natality files published by the National Center for Health Statistics (NCHS) at the Centers for Disease Control which we have obtained through a Data User Agreement (DUA). The DUA was necessary because the publicly available files do not include information on the mother's county of residence. The Natality files include individual birth records for all births that occurred in the United States. From these, we tabulate births by calendar year (1982-2022), sex, and age of the mother in the 5-year groups. The birth tabulations are then combined with the annual female population estimates by county, sex, and age group published by the Census Bureau to calculate all fertility indicators.

3 Methods

3.1 Principal components

Principal component models have enjoyed much success in demographic modelling as they offer a means of capturing the strong regularities observed in age-specific demographic rates across varying populations (Alexander et al., 2017; Dharamshi et al., 2024).

This approach begins with a matrix of large population log-fertility curves, $\mathbf{X}$, and computes the singular value decomposition $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}'$ where $\mathbf{U}$ is the matrix of left singular vectors, $\mathbf{\Sigma}$ is the diagonal matrix of singular values, and $\mathbf{V}$ is the matrix of right singular vectors, which we refer to as "principal components". The principal components capture structural patterns in log-fertility over age.

The first three principal components derived from United States state-level log-fertility data are given in Figure 1. The first principal component captures the characteristic inverted-"U" shape of log-fertility, where fertility is generally the highest between ages 20 and 35. The second and third principal components represent changes in the shape of the base fertility curve, allowing for delayed childbearing (PC2) and increased concentration around peak ages (PC3).

Equipped with these foundational patterns, we can build models that target the contribution of each principal component in each location and time point of interest rather than the age-specific fertility rates directly. This efficient use of demographic knowledge dramatically reduces the number of parameters that must be estimated, leading to more precise final estimates.

3.2 Model overview

We begin by introducing the notation for our model. Let $B_{a,c,t}$ and $P_{a,c,t}$ denote the number of births and the population of women observed in age group a, county c, and year t. Then, let $\lambda_{a,c,t}$ denote the corresponding fertility rate. This is our target of inference. We assume a Poisson model for these quantities:

$$B_{a,c,t} \sim \text{Poisson}(P_{a,c,t}\lambda_{a,c,t}).$$

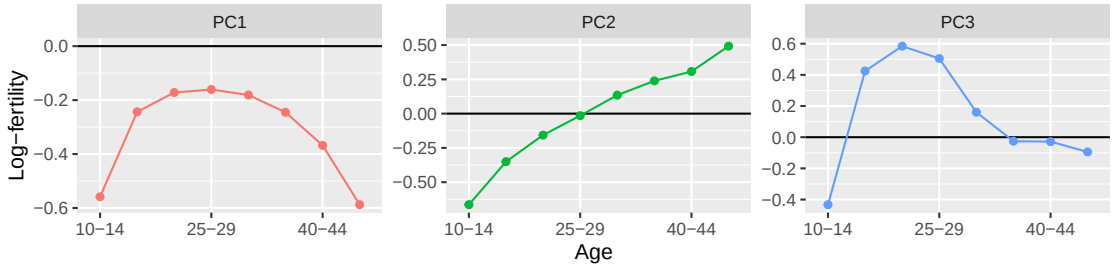

Figure 1: The panels plot the first through third principal components extracted from the singular value decomposition of the state-level log-fertility matrix.

As discussed, our latent model for the fertility rates is constructed on the log-scale using the principal components derived from aggregate state-level data. Specifically,

$$\log \lambda_{a,c,t} = \sum_{j=1}^{3} \beta_{j,c,t} V_{j,a}$$

where $V_{j,\cdot}$ is the jth principal component and $\beta_{j,c,t}$ are the county-year specific coefficients to be estimated. Intuitively, this model constructs log-fertility curves as linear combinations of established fertility patterns encoded by the principal components. Our task is then to estimate the contribution of each component.

The number of principal components included is an important decision in principal component regression models. Here we use the three plotted in Figure 1 as they represent over 99% of the variation in log-mortality rates though we intend to investigate sensitivity to this decision in the full paper.

3.3 Spatiotemporal smoothing

In practice, estimating the $\beta_{j,c,t}$ coefficients independently for each county-year is not practical as the small birth and population counts observed in some counties will lead to imprecise estimates. We thus propose a spatiotemporal model that borrows strength across space and time to stabilize estimation of the log-fertility rates. This model is not, however, assigned to the log-fertility rates themselves. Rather, we build a spatiotemporal model for each of the three collections of $\beta_{j,c,t}$ terms. This serves two purposes: 1. we expect that spatiotemporal dependence in age-specific log-fertility rates are a consequence of spatiotemporal dependence in the high-level patterns captured by the principal components and therefore our proposed model targets the underlying mechanism at play, and 2. from a practical perspective, attempting to alternatively construct one unified age-space-time process is not a computationally practical endeavour.

Formally, our proposed model decomposes $\beta_{j,c,t}$ into the sum of several terms:

$$\beta_{j,c,t} = \mu_j + \alpha_j(t) + \omega_j(c) + \delta_j(c,t),$$

where μ_j is the overall mean for the jth principal component coefficients, $\alpha_j(t)$ is the temporal component, $\omega_j(c)$ is the spatial component, and $\delta_j(c,t)$ is the spatiotemporal interaction term. The first terms, μ_j, are specified as fixed effects on the principal component vectors. For the temporal, spatial, and spatiotemporal terms, we consider the Type III Knorr-Held random effect model (Knorr-Held, 2000; Blangiardo et al., 2013; Blangiardo and Cameletti, 2015). This model assumes that both the temporal components, $\alpha_j(t)$, and spatial components, $\omega_j(c)$, decompose into the sum of two terms: a structured model and unstructured Gaussian noise. For $\alpha_j(t)$, we take the structured term to be a random walk 2 to promote gradual changes in the overall time trend. For $\omega_j(c)$, we model both terms together using the BYM2 parameterization of Riebler et al. (2016) where the structured spatial term is an areal random effect where neighbours are defined by counties that share a border, allowing for local information pooling. Finally $\delta_{j(c,t)}$ is defined as an interaction between the unstructured temporal term and the structured spatial term which can be intuitively understood

as allowing an independent perturbation to the structured spatial field in $\omega_j(c)$ at each time point (for further technical details see Knorr-Held 2000). This interaction allows the model to adapt the spatial field as county interrelationships evolve over time.

We conclude by noting that unlike pure hierarchical specifications (such as those used in Alexander et al. (2017) and Dharamshi et al. (2024)), our proposed spatiotemporal model is less susceptible to overreliance on large urban centres to determine patterns, and instead draws on local strength through the spatial terms. We intend on performing a comprehensive comparison between our proposed spatiotemporal model and a simpler hierarchical model in the full paper.

3.4 Computation

The proposed model is fit using the integrated nested Laplace approximation (INLA) method, implemented in the `INLA` R package, using the default `INLA` priors for all parameters (Rue et al., 2009). For inference on downstream functions of the log-fertility rates, we generate 1,000 samples of the linear predictors from the approximate posterior distribution.

4 Results

We apply our proposed model to the births and population data described in Section 2 subset to Californian counties to estimate fertility rates in each county-year-age group for the period of 1982-2022. We focus our attention here on California as its counties act as a microcosm of the United States as a whole, and will study the entire country in the full paper. At this time, we note that county-year-age group observations with between one and nine births, representing approximately 22% of the data, have been censored to protect privacy, though we will gain access to this data through our DUA in the near future.

To illustrate the output of the model, in Figure 2 we select three counties of varying population sizes and plot the observed and estimated fertility rates for five evenly spaced years. Observed data are given by the black points, posterior median estimates are given by the blue line, and 90% credible intervals are given by the blue bands. County names are suppressed in order to comply with our data agreement.

Figure 2 offers several interesting substantive and methodological insights. On the former, we see the increase in fertility from 1982 to 1992 followed by sharp declines in subsequent decades. We also see that the two smaller counties have higher and earlier fertility than the large urban county. Regarding the timing of fertility, we note that the deferral in fertility to later ages is on full display in the large county: the estimated fertility curves for 1982 and 1992 peak in the 20-24 and 25-29 age groups whereas they peak in the 30-34 age group in 2012 and 2022. Methodologically, we see that as expected, our estimates are more uncertain the smaller the county, though we comment that the intervals are not prohibitively large. We note that in 2022, the estimated curve for the large county does not track the observed data exactly despite the large population. We plan to investigate this further and explore whether the model requires additional allowances for overdispersion in log-fertility beyond the principal components.

In Figure 3, we plot maps of our estimates of the total fertility rate for all of California for the same set of years as in Figure 2. The first row displays the median posterior point estimates and the second row displays the posterior standard deviations. Similar patterns emerge in the total fertility rate: there is a clear increase in total fertility in the first decade under study followed by a sharp decline in subsequent years. Interesting spatial patterns can be seen in the maps. The interior and eastern regions generally have higher fertility though the gap narrows in more recent years. We comment that one county in the east-central region has persistently low total fertility. An investigation into the data reveals that the observed data are all zero births - all other observations have been censored. After we gain access to the full data, we plan to investigate how the estimates for this county change.

Turning to the second row of Figure 3, we note that the standard deviations are generally quite small though some of the smaller eastern counties do have elevated standard errors as expected. Most interesting is the northeastern-most county which has consistently elevated standard deviations. We intend to investigate this county further and understand whether this is a function of the population size or the fact that as a

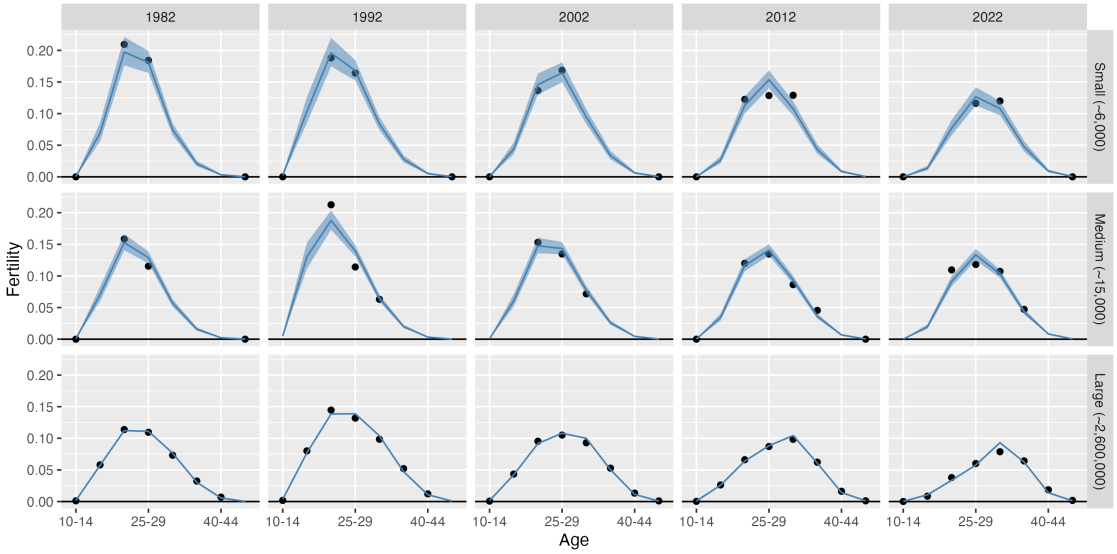

Figure 2: Three examples of county-level fertility rate plots by county-year three counties of varying populations. The black dots represent observed values, and the blue lines and regions indicate posterior medians and 90% credible intervals respectively

border community there is limited opportunities for sharing information across space. It will be interesting to explore whether extending to all US counties will improve the precision of this county's estimates.

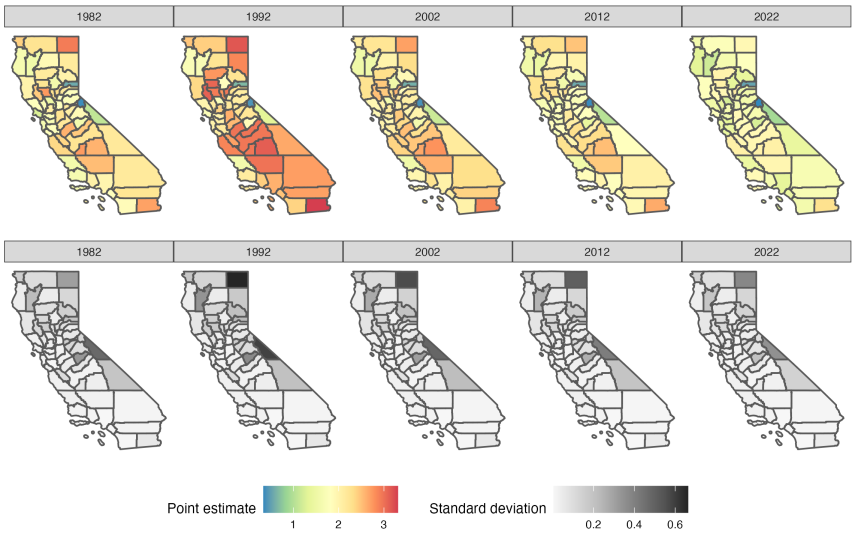

Figure 3: Maps of posterior median estimates of total fertility rates and corresponding standard deviations overlaid with county borders.

5 Discussion

In this abstract, we have proposed a new model for estimating subnational fertility rates. The model combines the parsimony of principal component models with the flexibility of spatiotemporal models such that established fertility patterns are respected while still allowing for local variability. Our initial application of the model to California's fertility data has led to promising results. In the main paper, we intend to further refine our model by conducting comprehensive model validation exercises where we compare it against several simpler (and perhaps a couple of slightly more flexible) alternatives on both simulated and real data. We then intend to apply the final model to the entire United States as a whole and study the resulting fertility rate estimates in detail.

References

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

Blangiardo, M. and Cameletti, M. (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons.

Blangiardo, M., Cameletti, M., Baio, G., and Rue, H. (2013). Spatial and spatio-temporal models with R-INLA. *Spatial and spatio-temporal epidemiology*, 4:33–49.

Bloom, D. E., Canning, D., Fink, G., Finlay, J. E., et al. (2008). The high cost of low fertility in europe. Technical report, Program on the Global Demography of Aging.

Census Bureau (2024). Annual County and Puerto Rico Municipio Resident Population Estimates by Single Year of Age and Sex: April 1, 2020 to July 1, 2023 (CC-EST2023-SYASEX), Vintage 2023.

Dharamshi, A., Alexander, M., Winant, C., and Barbieri, M. (2024). Jointly estimating subnational mortality for multiple populations. *Demographic Research*, (just-accepted):1–22.

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

National Center for Health Statistics (2024). Natality Files 1982-2022, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program.

Riebler, A., Sørbye, S. H., Simpson, D., and Rue, H. (2016). An intuitive bayesian spatial model for disease mapping that accounts for scaling. *Statistical methods in medical research*, 25(4):1145–1165.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(2):319–392.

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.