Enhancing Birth Rate Analysis with Hierarchical Modeling: Addressing Data Sparsity Across Geographic Levels

Yuzhou Peng, Siyang Wu, Amy Sun

April 2024

1 Introduction

Birth rates critically influence population growth and structure, which is essential for demographic planning in areas like education and healthcare. This demographic indicator also affects economic development, as a youthful population requires investments in education and employment opportunities, while an aging population may necessitate a shift in economic strategies and healthcare services. Additionally, understanding birth rates aids in forecasting healthcare needs, particularly in maternal and child health, ensuring that appropriate resources are allocated to meet the demands of different population segments.

However, there are several states that due to some sort of limitation have less sufffcient data related to birth rate, which lead to problems (eg. less informative) when making decisions out of the existing data. The birth rate in a state is not completely independent from each other due to the fact that they come from same country. We are going to propose a way to fully utilize the information in the dataset, and meanwhile conduct comparison between different regions across US (eg, west vs. east) during the process for additional insights.

2 Data

2.1 Dataset Basics

In this study we are going to work on two dataset:

1. Birth dataset from Center for Disease Control and prevention (CDC)
2. County Population Data Dictionary From National Cancer Institute.

2.2 Birth Data

The Birth data in the CDC WONDER database is collected through the National Vital Statistics System (NVSS), which compiles birth data from all birth certificates filed in the United States. This includes detailed information on birth and maternal demographics, collected through a standardized birth certificate form used across the United States. Hospitals, midwives, and other healthcare providers submit this information to state health departments, which in turn report the data to the federal government. The data is not a simple random sample. It is a comprehensive administrative data collection that aims to capture all births in the United States. As such, it is a census of births rather than a sample. This means the data covers the entire population of interest (all births occurring in the U.S. within a given time frame) and is not subject to the same types of sampling errors as survey data might be.

2.3 County Population Data Dictionary

The data for the U.S. Census Bureau's Population and Housing Unit Estimates are collected using a variety of methods including administrative records and data from earlier censuses, specifically the most recent decennial census. This method is not a simple random sample but a comprehensive effort to include all individuals and housing units using established administrative sources to update and adjust the census counts annually.

2.4 Potential Defects

Both dataset are subject to the following points of defects:

- Reporting Errors: Misreporting or inconsistent reporting of data by the individuals filling out birth certificates can introduce inaccuracies.
- Data Entry Errors: Mistakes in the transcription of data from paper forms to digital databases.
- Missing Data: Not all fields on a birth certificate may be completed, leading to missing data issues.

2.5 Data structure

The two dataset are joined together on FIPs code in the form XXYYY, where XX = 2 digit state code and YYY = 3 digit county code. The finalized the dataset has the following structure shown in Table 1. The data contain 5619 data points containing all the available information about the county each year across the United State from 2011 to 2020.

Table 1: Structure of Processed Dataset

Column 1	Column 2	Column 3	Column 4	Column 5
County(name)	FIPs	Birth count	Year	Population

2.6 Explore Data Analysis

In the initial EDA conducted on the integrated dataset, two key visualizations were generated to understand trends in birth rates across the United States from 2011 to 2016. The Figure 1 shows the birth rate per 1,000 people, revealing a general stability in the rates from 2011 to 2016, indicating no significant fluctuations during this period. Subsequently, Figure 2 shows a heat map that was produced to illustrate the spatial distribution of birth rates at the state level. This map was color-coded to reflect the birth rate per 1,000 individuals, facilitating a visual comparison across states. Notably, the heat map identified a data void in Wyoming, suggesting an area for potential data collection improvement. Furthermore, the analysis highlighted a pattern where states in the central region of the U.S. exhibited higher birth rates compared to those on the West and East Coasts, suggesting regional demographic trends that could warrant further investigation to understand underlying factors influencing these disparities.

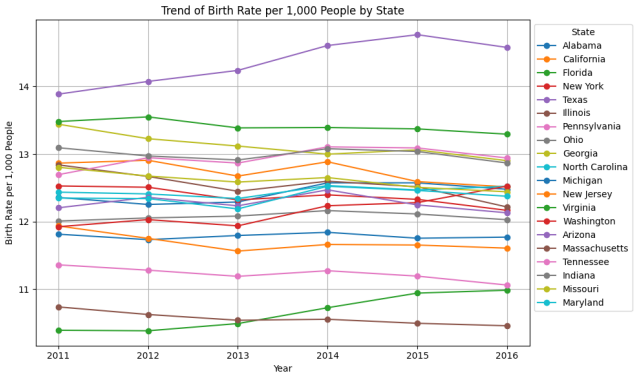

Figure 1: Birth rate per 1000 people from 2011 to 2016

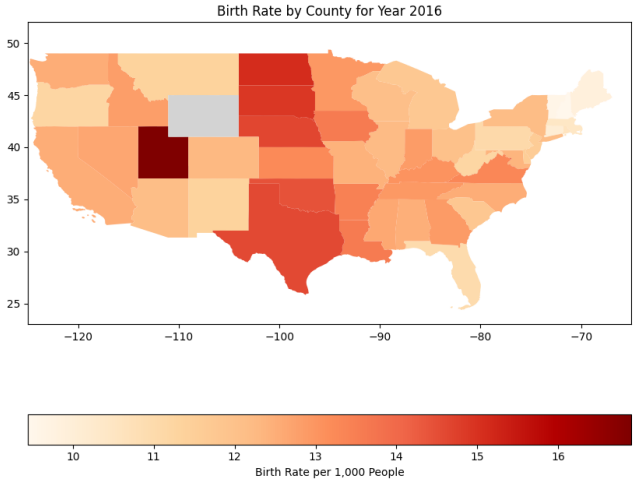

Figure 2: US heatmap

3 Method

3.1 Modeling

To deal with data sparsity and preserve the dependency structure within birth rate distributions between counties, we propose a Bayesian hierarchical model to solve our problems.

Before we construct the the model, there are a few assumptions to be declared:

- Birth rates of counties in the same state conditionally independently follow the same distribution.
- Parameters of state-level distributions conditionally independently follow the same distribution parameterized by some hyper parameters.
- Birth rate of year t is dependent on that of year t-1

3.1.1 Notation

We define our notations to be used as follow:

- z_{ij} is the new born population of county i in state j
- N_{ij} is the total population if county i in state j
- λ_{ij} is the estimated birth rates of county i in state j.
- ω_j, κ_j are the parameters of underlying distribution of birth rate in state j.
- ω, κ are the hyper parameters that follow a pre-designed prior distribution

  λ_{ij} and ω_j are our quantities of interest, representing birth rates level of counties and states respectively.

3.1.2 Model Structure

The basic structure of our proposed Bayesian hierarchical model is described in the figure (Figure 3) below.

In particular,

$$z_{ij} \sim Poisson(N_{ij}\lambda_{ij})$$

$$\lambda_{ij} \sim Beta(\omega_j(\kappa_j - 2) + 1, (1 - \omega_j)(\kappa_j - 2) + 1);$$

$$\omega_j \sim Beta(\omega(\kappa - 2) + 1, (1 - \omega)(\kappa - 2) + 1) \text{ and } \kappa_j - 2 \sim Gamma(S_\kappa, R_\kappa)$$

$$\omega \sim Beta(A_\omega, B_\omega) \text{ and } \kappa - 2 \sim Gamma(S_\kappa, R_\kappa)$$

Where A_ω, B_ω, S_κ and R_κ are pre-designed parameters. $A_\omega = B_\omega = 1$ and $S_\kappa = R_\kappa = 0.01$ if there is no prior knowledge

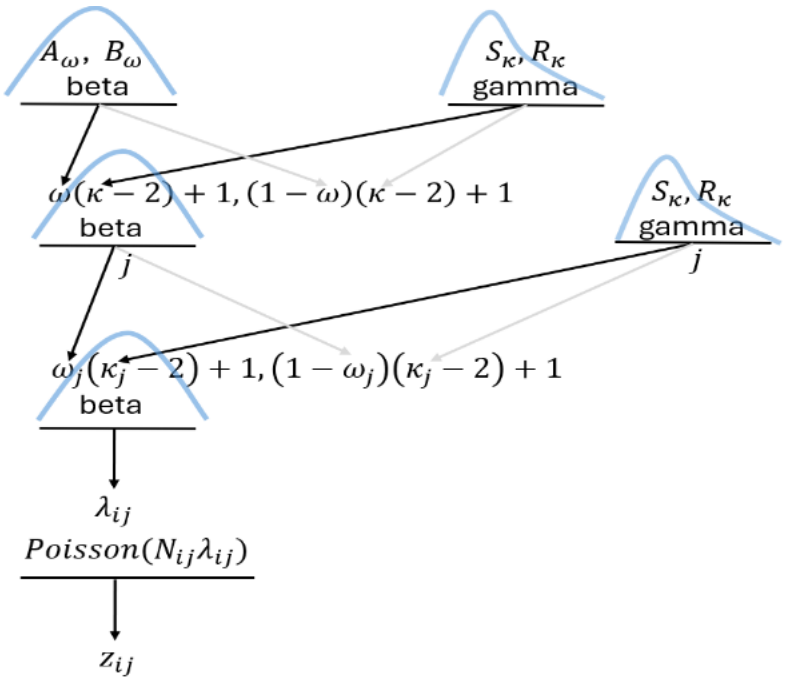

Figure 3: Hierarchical Model Structure

3.2 Experiment Schema

3.2.1 Preliminary Analysis

To avoid under-fitting, before drawing any conclusions from our model, we need to justify whether it is appropriate to only build one model that includes all birth rates data (referred to as "full model" in the following discussion). For comparison, we firstly categorize our data into a few divisions (as shown in the Figure 4). We then build several models where each model fits only birth rates data from one division (referred to as "division-based models" in the following discussion). Finally, we make inference on the difference between samples drawn from full model and division models to see if there is any significant difference.

For simplicity in demonstration, we only display the comparison results of county birth rate in Cumberland, ME. Since ME belongs to division of New-England, we draw samples of Cumberland from full model and New-England division model.

The histogram of lambda differences (Figure 5) indicates the distribution of these disparities over many samples. The results suggested the mean to be around zero, which means there is little differences between the two models. In addition, a 95% credible interval is [-0.0025, 0.0024]. This indicates that the mean difference is not statistically significant. In conclusion, dividing the data by divisions for the purpose of birth rate estimation is not necessary. Therefore, we are going analyze the data based on full model in the following discussions.

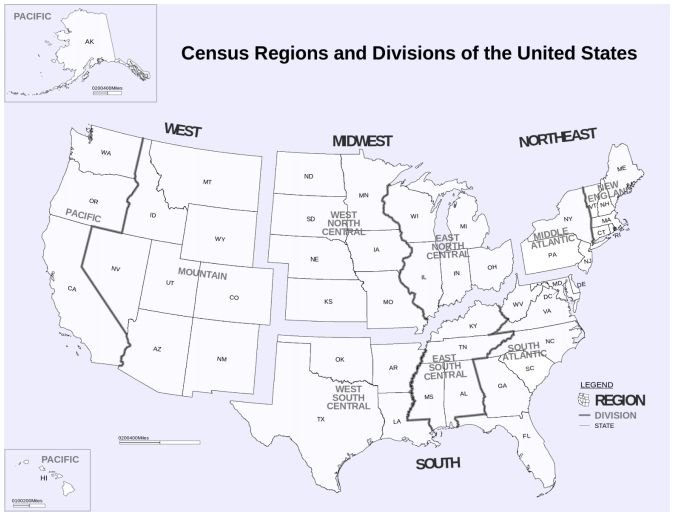

Figure 4: The U.S. Census Region

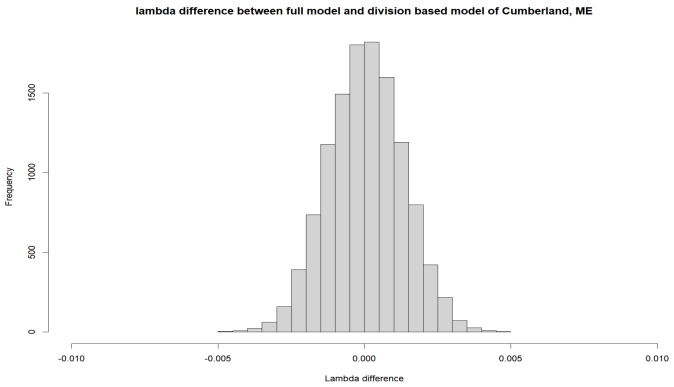

Figure 5: Lambda Difference Between Full Model and Division Based Model of Cumberland, ME

3.2.2 Year Progression Modeling

In this part of experiment, we want to take advantages of the inductive property of Bayesian models so that we may manage to gain an overview of how birth rate from 2011 to 2016. Moreover, the inductive procedure includes information from previous data which may yield more confident inferences (i.e. tighter credible intervals of parameters.)

Unlike single year model where we assume a non-informative and vague prior distribution ($A_\omega = B_\omega = 1$ and $S_\kappa = R_\kappa = 0.01$), we estimate prior parameters through samples drawn from last year's model via method of moments (referred to as year progression models in the following discussions).

In particular, suppose we have hyper parameter samples of year t: $\omega_i^{(t)}$ and $\kappa_i^{(t)} - 2$, i = 1,2...N. We estimate prior distribution of year t+1 by following procedures:

$$\omega^{(t+1)} \sim Beta(A_\omega^{(t+1)}, B_\omega^{(t+1)});$$

$$\kappa^{(t+1)} - 2 \sim Gamma(S_{\kappa}^{(t+1)}, R_{\kappa}^{(t+1)});$$

$$\hat{A}_{\omega}^{(t+1)} = \bar{\omega}^{(t)}[\frac{\bar{\omega}^{(t)}(1-\bar{\omega}^{(t)})}{D^2_{\omega^{(t)}}} - 1];$$

$$\hat{B}_{\omega}^{(t+1)} = (1-\bar{\omega}^{(t)})[\frac{\bar{\omega}^{(t)}(1-\bar{\omega}^{(t)})}{D^2_{\omega^{(t)}}} - 1];$$

$$\hat{R}_{\kappa}^{(t+1)} = \frac{\bar{\kappa}^{(t)} - 2}{D^2_{\kappa^{(t)}}};$$

$$\hat{S}_{\kappa}^{(t+1)} = \frac{(\bar{\kappa}^{(t)} - 2)^2}{D^2_{\kappa^{(t)}}}$$

where $\bar{\omega}^{(t)}$ and $\bar{\kappa}^{(t)} - 2$ are sample means, $D^2_{\omega^{(t)}}$ and $D^2_{\kappa^{(t)}}$ are sample variances.

4 Results

4.1 MCMC Convergence Results

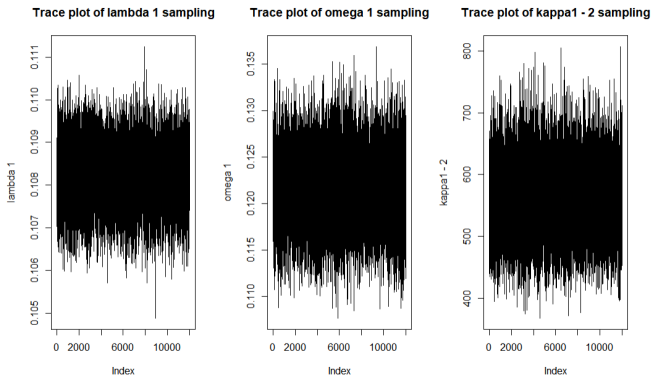

Figure 6: Trace Plots of MCMC Sampling Results

The sampling procedure is implemented via Markov Chain Monte Carlo (MCMC) algorithm to estimate the posterior distributions of parameters in our model. In general, the trace plots (For simplicity, only display a few plots. Figure 6) suggest that our sampling process reached a stationary distribution. In addition, there are no signs of 'sticking' at certain values, indicating MCMC iterations have sufficiently explored the whole parameter spaces. Thus we may conclude MCMC algorithm has reached a success convergence.

4.2 Birth Rate Trends: Hierarchical Model vs. MLE, 2011-2016

Figure 7 illustrates the comparison of birth rate trends per 1,000 people among a subset of states, contrasting the birth rates estimated by maximum likelihood estimation (MLE) from the dataset with the mean values derived from a hierarchical omega model, spanning the years 2011 to 2016. The results indicate that the hierarchical model

effectively captures the overall trend observed in the MLE birth rates. Furthermore, the hierarchical model tends to consolidate the estimates by adjusting individual state rates towards the overall mode, thereby reducing variability and aligning state-specific estimates more closely with the central tendency.

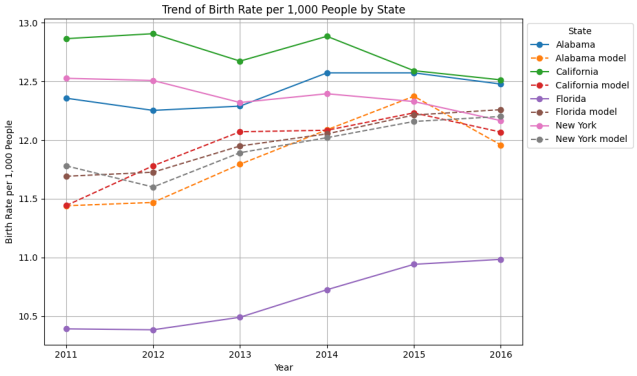

Figure 7: Comparison between birth rate trend from maximum likelihood estimation and model sampling

4.3 Performance on Sparse Data

The hierarchical model demonstrates particularly valuable properties when analyzing states with sparse data (e.g., only one county reporting). In such cases, employing MLE for inference could introduce significant biases and uncertainties. Figure 8 illustrates this by comparing MLE results from Alaska, with only one reported observation, to outcomes derived from the hierarchical omega model. The hierarchical approach is advantageous in these contexts as it leverages information from the entire dataset to improve estimates in data-sparse regions, thereby enhancing stability and reliability of the statistical inference

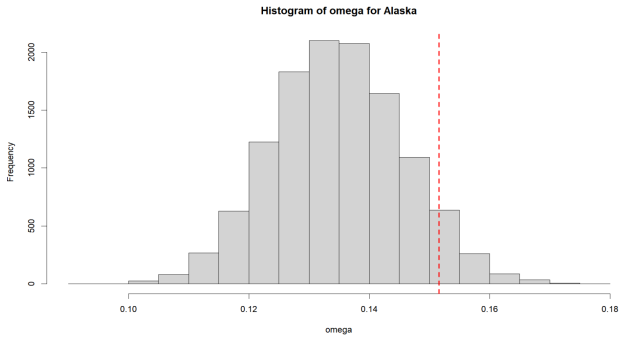

Figure 8: Comparison between birth rate trend from maximum likelihood estimation and model sampling

4.4 Year Progression Model vs. Single Year Model

4.4.1 Run Time Efficiency

During the process of sampling, we find out that after we specify the prior distribution parameters by steps described in section 3.2.2, the sampling time has largely decreased compared to a single year model with non-informative prior distribution. Taking year of 2016 as an example, the run time for single year model of 2016 is around 28 minutes while year progression model of 2016 runs for only around 2 minutes 30 seconds.

4.4.2 More Confident Inference

Another advantage of year progression model is that with each year passed by, the inductive nature allows the model to combine previous information to produce more concentrated samples compared to single year model and thus yields tighter credible intervals. As seen in Figure 9.

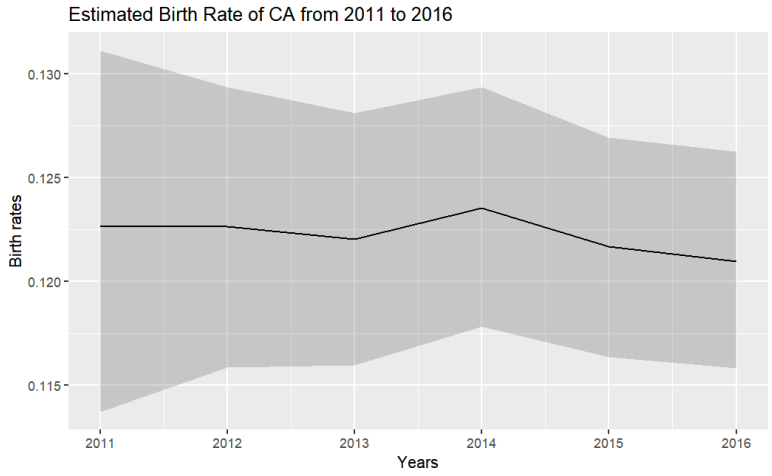

Figure 9: Ribbon Plots of Estimated Birth Rate in CA

More explicitly, as shown in Figure 10, year progression model present a more significant disparity in birth rates among states MA, CA and TX.

5 Conclusion

This study has validated the effectiveness of the hierarchical model for analyzing birth rate trends, particularly in states with sparse data. Demonstrating superior performance, the hierarchical model accurately aligns state-specific estimates with broader trends, enhancing reliability and reducing variability. The Year Progression Model, compared to the Single Year Model, has proven more efficient and provided tighter credible intervals, reflecting its capability to incorporate past data for more accurate future predictions. These findings highlight the significant potential of advanced statistical methodologies to inform public health policy and resource allocation, suggesting

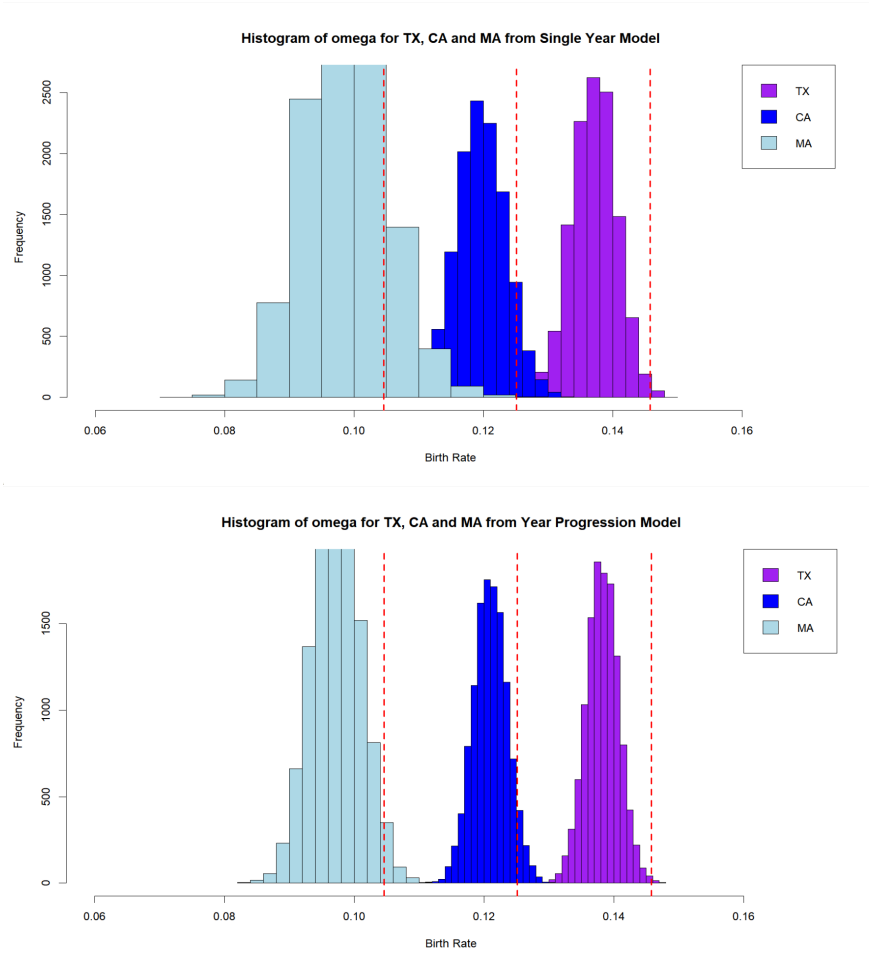

Figure 10: Histogram of 2016 Birth Rates in MA, CA, and TX from Year Progression Model and Single Year Model

that further research should extend these models to a broader range of demographic studies such as having temporal data for mortality rates and etc.