Combining Survey and Census Data in Time and Space in a Developing World Context

Katherine Wilson

A dissertation submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Jon Wakefield, Chair

Lurdes Inoue

Adam Szpiro

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Combining Survey and Census Data in Time and Space in a Developing World Context

Katherine Wilson

Chair of the Supervisory Committee:
Professor Jon Wakefield
Departments of Biostatistics and Statistics

Obtaining reliable estimates of health indicators at a granular level in space and time is important for informing health intervention and public policy decisions. In low and middle income countries, the data available come from a multitude of sources including vital registration systems, complex surveys, and disease registries. These data sources are often of varying quality. In particular, the information on health outcomes may be aggregated over space and time. Overall, this poses a modeling challenge as there is a mismatch between the underlying process, the observed data, and the inferential resolution desired. This work tackles three main issues that are commonly faced in this setting by developing Bayesian models tailored to the specific problem at hand. In particular, this work considers incorporating data where the outcome has been aggregated over space (common in census data), the outcome is associated with a point in space but the exact location is unknown (common in household survey data), and the outcome has been aggregated over time (common in modeling child mortality using census data).

TABLE OF CONTENTS

LIST OF FIGURES

GLOSSARY

DA: Data augmentation

DHS: Demographic and Health Survey

EB: Empirical Bayes

EA: Enumeration area

FBH: Full birth history

GMRF: Gaussian Markov random field

HMC: Hamiltonian Monte Carlo

ICAR: Intrinsic conditional autoregressive

INLA: Integrated nested Laplace approximation

LA: Laplace approximation

LMIC: Low and middle income countries

MCMC: Markov chain Monte Carlo

MICS: Multiple Indicator Cluster Survey

PC: Penalized complexity

RW: Random walk

SBH: Summary birth history

SPDE: Stochastic partial differential equation

TMB: Template model builder

U5M: Under-five mortality

U5MR: Under-five mortality rate

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Jon Wakefield, for being a tremendous source of guidance, encouragement, and inspiration. Jon was generous in his time and attention. I thoroughly enjoyed our weekly meetings, which struck the perfect balance of feedback on research, professional mentorship, and fun tangents on other aspects of life. I could not be more grateful.

I am very appreciative of my committee members, Adam Szpiro, Lurdes Inoue, Simon Hay, and Steve Hawes, who have provided helpful feedback and valuable insight on this work. Additionally, thank you to the members of working groups that I have been a part of at the University of Washington, as well as Jim Thorson, Dan Simpson, Geir-Arne Fuglstad, and Andrea Riebler for the many enlightening research discussions that have shaped various aspects of this work.

From the University of Washington, I would also like to thank Andrew Zhou, Lyn Brumback, Susanne May, Katie Kerr, Noah Simon, Amy Willis, Mauricio Sadinle, Zaid Harchaoui, Barbara McKnight, Thomas Richardson, Patrick Heagerty, Scott Emerson, and Gitana Garofalo. They have supported me as supervisors, mentors, instructors, and administrators during the course of my graduate studies. I am grateful to my officemates over the years, members of the 2013 biostatistics cohort, and other students for their support, company, and friendship.

I would also like to thank Phil Rosenberg and Bill Anderson who served as my mentors at the National Cancer Institute during a summer internship. Working with them on a research project was my first foray into the field of biostatistics and they inspired me to pursue a graduate degree.

Finally, I would like to extend my gratitude to my family and friends. Especially, my parents Ann and Brian, my brother Sam, and my husband, Matt, who have encouraged and supported me through every step of this process. Thank you, I could not have done this without you.

DEDICATION

To my family.

Chapter 1
INTRODUCTION

In 2015, members of the United Nations adopted the 2030 Agenda for the Sustainable Development, which emphasizes 17 Sustainable Development Goals (SDGs). These goals cover a range of global issues including the improvement of health. Monitoring progress towards the achievement of the goals is of high priority. In particular, the agenda states that the indicators of the SDGs "should be disaggregated, where relevant, by income, sex, age, race, ethnicity, migratory status, disability and geographic location, or other characteristics" (United Nations General Assembly, 2015; United Nations, 2017). In many low and middle income countries (LMIC), vital registration systems to track these indicators are nonexistant or lacking full coverage. Thus, other sources, such as complex surveys and censuses are turned to, to obtain valid estimates at the desired resolutions (i.e., across space and time).

This dissertation focuses on developing methods that can provide estimates and corresponding levels of uncertainty of health indicators when data are of varying levels of quality and resolution across space and time. In particular, we consider combining data from complex household surveys and censuses.

1.1 Motivating Examples

1.1.1 Continuous Spatial Models for Aggregate Data

The motivation for this work comes from wanting to use a continuous spatial model, but with data that are available at discrete levels. In particular, censuses and disease registries often provide aggregate level information, such as the number of individuals with some health indicator living in a particular administrative level. Traditionally, Markov random field spatial models have been employed to acknowledge spatial dependence between administrative areas

and allow for smoothing across space. In the context of an irregular set of areas, these models always have an ad hoc element with respect to the definition of a neighborhood scheme.

In Chapter 3, we exploit recent theoretical and computational advances to carry out modeling at the continuous spatial level, which induces a spatial model for the discrete areas. This approach also allows reconstruction of the continuous underlying surface, but the interpretation of such surfaces is delicate since it depends on the quality, extent and configuration of the observed data. We focus on models based on stochastic partial differential equations (SPDEs; Lindgren et al., 2011). We consider the setting where only aggregate data is available, as motivated by the famous Scotland lip cancer dataset (Clayton and Kaldor, 1987). We also consider when aggregate data are supplemented with point data, as motivated by mapping disease indicators in LMIC where point data in the form of complex household surveys are also available. We carry out Bayesian inference and, in the language of generalized linear mixed models, if the link is linear, an efficient implementation of the model is available via integrated nested Laplace approximations (INLA; Lindgren and Rue, 2015). For nonlinear links, we present two approaches: a fully Bayesian implementation using a Hamiltonian Monte Carlo (HMC) algorithm (Neal, 2011) and an empirical Bayes (EB) implementation, that is much faster, and is based on Laplace approximations (LA) using the `R` package `TMB` (Kristensen, 2014).

1.1.2 Health Indicator Spatial Modeling with Incomplete Location Data

Surveys are a primary tool for obtaining estimates of health indicators in many developing countries. Due to cost, complex household surveys are only taken on a small, random subset of a country's inhabitants. Researchers and policymakers often desire estimates of these health indicators in areas where the survey did not occur. Building a continuous spatial model is a common approach that is used to obtain predictions of risk at unsampled locations. For confidentiality purposes, this information is usually not given exactly, either by randomly displacing the location or by only reporting the administrative area of the point (which we refer to as "masking"). Spatial analyses tend to treat the displaced location as

the truth, though the impact this has on model-based predictions and associated uncertainty intervals has not been extensively explained. We show that using data augmentation (DA), it is possible to account for this additional complexity, given the sampling nature of the surveys.

Our proposed approach is able to account for both types of incompleteness and the method is described in Chapter 4. A simulation is carried out and highlights the degree to which displacement or masking influence the relationship between spatial covariates and the health indicator and the effect on interpolated surfaces.

1.1.3 Under-Five Mortality Estimation with Survey and Census Data

High under-five mortality (U5M) tends to be concentrated in developing regions where much of the information informing estimates comes from surveys and censuses. In both, women are asked about their birth histories, but with varying degrees of detail. Full birth history (FBH) data contain the reported dates of births and deaths of every surveyed woman's children. In contrast, summary birth history (SBH) data contain only the total number of children born and total number of children who died for each woman. Both types contain covariates such as the woman's age. Specialized methods are needed to accommodate this type of data into analyses of child mortality trends.

In Chapter 5 we develop a DA scheme within a Bayesian framework where for SBH data, birth and death dates are introduced as auxiliary variables. In Chapter 6 we propose a computationally efficient procedure based on using several reasonable approximations to the full data generative model. In both cases, we specify a full probability model for the data, so that many of the well-known biases that exist in this data can be accommodated, along with space-time smoothing on the underlying mortality rates. We illustrate our approaches in simulation, showing that uncertainty is reduced when incorporating SBH data over simply analyzing all available FBH data. We also apply our approach to survey and census data from Malawi.

1.2 Methodological Contributions of Dissertation

The methods proposed in this dissertation make it possible to obtain reliable estimates (point and interval) of health indicators across space and time from data that may be aggregated across space and time.

The primary methodological contributions of Chapter 3 are deriving the form of the distribution for spatially aggregate data and proposing several computational strategies that can be employed for inference. In Chapter 4, the primary contribution is a framework for modeling spatial point data, where the true locations are unknown. The primary contributions of Chapter 5 are two-fold. First, we develop a data augmentation scheme that maps between latent fertility and mortality rates and temporally aggregated birth history data. Second, we describe an approach for combining separate estimates made on SBH and FBH data. In Chapter 6, the primary contributions are a sensible approximation to the data augmentation approach of Chapter 5 and a quick implementation using Laplace approximations.

1.3 Organization of Dissertation

Chapter 2 serves as a review of statistical methods that are drawn from in the proceeding chapters. The next four chapters focus on the methodological contributions to the examples described above. Each chapter is organized as follows. First, we begin each chapter with an introduction, where the relevant literature is reviewed and necessary background information is provided. Next, the main statistical methods are described in detail. A simulation study follows and typically an application is also given. Finally, each chapter ends with a discussion. Chapter 7 provides concluding comments and a discussion on a future project aiming to combine the methods proposed in the dissertation.

Chapter 2

BACKGROUND

This chapter provides a brief review of relevant concepts that will be used in multiple chapters throughout the dissertation. Methods that are used for only one chapter are described therein. First, a discussion of Gaussian Markov random fields (GMRFs) is provided. In particular, a focus is on intrinsic GMRFs (IGMRFs), which in many of the applications are used as priors for spatial and temporal random effects. The link between GMRFs and Gaussian random fields (GRFs) is also discussed. Next, the framework of Bayesian hierarchical models and relevant computational tools are described, as they will be featured heavily in all Chapters. The penalized complexity (PC) prior for the precision parameter is also described. An overview of missing data methods, specifically data augmentation is presented. Finally, a brief description of survey sampling is provided. Although survey sampling techniques are not directly used in many of the proposed methods of this dissertation, they are used as a way to assess the performance of these proposed approaches.

2.1 Gaussian Markov Random Fields

A GMRF is a finite-dimensional random vector that follows a multivariate normal distribution. Define $\boldsymbol{x} = [x_1, \ldots, x_n]$ with $\boldsymbol{x} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{x}$ is a n-dimensional GMRF with mean $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$ if and only if the density of $\boldsymbol{x}$ can be written as follows:

$$\pi(\boldsymbol{x}) = (2\pi)^{-n/2}|\boldsymbol{Q}|^{1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

and for $i \neq j$,

$$\boldsymbol{Q}_{ij} = 0 \iff x_i \perp x_j \mid \boldsymbol{x}_{-\{i,j\}},$$

where $\boldsymbol{x}_{-\{i,j\}}$ represents the collection of $\boldsymbol{x}$ without x_i and x_j. The last line describes the notion of *conditional independence* (also known as the Markov property), where for given $\boldsymbol{x}_{-\{i,j\}}$, discovering x_j gives no information on x_i. We focus on cases where $\boldsymbol{Q}$ is sparse.

2.1.1 Second-Order Random Walk Model (RW2)

The second-order random walk model for evenly-spaced increments is constructed assuming independent second-order increments, that is,

$$\Delta^2 x_i = (x_{i+2} - x_{i+1}) - (x_{i+1} - x_i)|\tau \sim N(0, \tau^{-1})$$

where τ is the precision. Thus, the density is,

$$\begin{aligned}\pi(\boldsymbol{x}|\tau) &\propto \tau^{(n-2)/2} \exp\left(-\frac{\tau}{2}(x_{i+2} - 2x_{i+1} + x_i)^2\right) \\ &= \tau^{(n-2)/2} \exp\left(-\frac{\tau}{2}\boldsymbol{x}^\top \boldsymbol{R}\boldsymbol{x}\right)\end{aligned}$$

where the structure matrix,

$$\boldsymbol{R} = \begin{bmatrix} 1 & -2 & 1 & & & & & & \\ -2 & -5 & -4 & 1 & & & & & \\ 1 & -4 & 6 & -4 & 1 & & & & \\ & 1 & -4 & 6 & -4 & 1 & & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & & \\ & & & 1 & -4 & 6 & -4 & 1 & \\ & & & & 1 & -4 & 6 & -4 & 1 \\ & & & & & 1 & -4 & 5 & -2 \\ & & & & & & 1 & -2 & 1 \end{bmatrix}$$

with all other entries being 0. Note that $\boldsymbol{R}$ has rank $n-2$, and hence is an improper GMRF, also known as an intrinsic GMRF (IGMRF).

In Bayesian analyses we will define a hyperprior on the precision parameter, τ. Defining an interpretable hyperprior is made difficult in that the marginal variance depends on the

dimension of the structure matrix and which IGMRF model (e.g. first-order vs second-order RW) is used. We will follow Sørbye and Rue (2014) and scale the model by setting the generalized variance of $\boldsymbol{x}$ to be 1. This is accomplished by finding the generalized inverse of the neighborhood structure map, $\boldsymbol{\Sigma}^*_{\tau=1} = \boldsymbol{R}^{-1}$ with $\sigma^2_{\tau=1}(x_i) = \Sigma^*_{\tau=1,ii}$, and

$$\sigma^2_{GV} = \exp\left(\frac{1}{n}\sum_{i=1}^{n}\log(\sigma^2_{\tau=1}(x_i))\right).$$

Thus, we define the scaled RW2 model to have precision $\tau\boldsymbol{R}^\star$ where $\boldsymbol{R}^\star = \sigma^2_{GV}\boldsymbol{R}$.

2.1.2 Intrinsic Conditional Autoregressive Model (ICAR)

The intrinsic conditional autoregressive (ICAR) model is specified conditionally as,

$$x_i|x_{-i},\tau \sim N\left(\frac{\sum_{j:j\sim i} x_j}{n_i}, \frac{1}{n_i\tau}\right) \tag{2.1}$$

where n_i are the number of neighbors of i, and the notation $j \sim i$ refers to j being a neighbor of i. Alternatively, define $\boldsymbol{R}$ to be a structure matrix that has elements

$$R_{ij} = \begin{cases} n_i & i = j \\ -1 & i \sim j \\ 0 & \text{otherwise} \end{cases}$$

then it follows that the density can be written as,

$$\pi(\boldsymbol{x}|\tau) \propto \tau^{\text{rank}(\boldsymbol{R})/2}\exp\left(-\frac{\tau}{2}\boldsymbol{x}^\top\boldsymbol{R}\boldsymbol{x}\right).$$

As is the case with the RW2 model, different ICAR models have different marginal variances. Thus scaling the neighborhood structure map may be needed when a hyperprior is assigned to the precision parameter, τ.

2.1.3 SPDE Approach

In contrast to GMRFs that are fundamentally discrete, GRFs are continuously indexed. Consider a domain $\mathcal{D} \in \mathbb{R}^d$; we will focus on the case when $d = 2$, as those results are

most relevant for the work presented in this dissertation. Then $S(\boldsymbol{s})$ is a GRF if all finite collections are jointly multivariate normal. That is, for a collection of points $(s_1, s_2, \ldots, s_n)^\top$,

$$\pi(\boldsymbol{S}) = (2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{S}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{S}-\boldsymbol{\mu})\right)$$

where $S_i = S(s_i)$, $\mu_i = \mu(s_i)$ for some mean function $\mu(\cdot)$, and $\Sigma_{ij} = C(s_i, s_j)$ for some covariance function $C(\cdot,\cdot)$. Specifically, we focus on the Matérn covariance function with scaling parameter $\kappa > 0$ and marginal variance λ^2 and smoothness parameter ν,

$$C_\nu(s_i, s_j) = \frac{\lambda^2}{2^{\nu-1}\gamma(\nu)} \left(\kappa||s_i - s_j||\right)^\nu K_\nu \left(\kappa||s_i - s_j||\right) \tag{2.2}$$

where $||\cdot||$ denotes the Euclidean distance in $\mathbb{R}^2$ and K_ν is the modified Bessel function of the second kind and order $\nu > 0$. In general, it is difficult to learn about the smoothness parameter ν, and so we follow convention and fix this parameter to $\nu = 1$ (Simpson et al., 2012a,b). The benefit of this choice is that the field has one continuous derivative while maintaining computational feasibility.

Lindgren et al. (2011) (see Simpson et al. (2012a) and Simpson et al. (2012b) for summaries) draw an elegant connection between GRFs and GMRFs. They consider the following stochastic partial differential equation (SPDE),

$$(\kappa^2 - \Delta)^{\alpha/2} S(s) = \lambda W(s), \qquad s \in \mathbb{R}^2$$

where $\Delta = (\partial^2/\partial s_1^2) + (\partial^2/\partial s_2^2)$ is the Laplacian on $\mathbb{R}^2$, $W(s)$ is Gaussian white noise and $\alpha = \nu + 1$. They show that the solution to the SPDE is a GRF with Matérn covariance,

$$S(s) = \int_{\mathbb{R}^2} k(s, s') dW(s')$$

where $k(s, s') = C_\nu(s, s')$.

Finally, using finite element analysis, a representation to the solution of the SPDE over a triangulation of the domain (called the mesh) is constructed by a weighted sum of basis functions,

$$S(s) \approx \tilde{S}(s) = \sum_{m=1}^{M} w_m \psi_m(s), \tag{2.3}$$

where M is the number of mesh points in the triangulation, $\psi_m(s)$ is a basis function and $\boldsymbol{w} = [w_1, \ldots, w_M]^\top$ is a collection of weights. The weights $\boldsymbol{w}$ are jointly Gaussian with $\boldsymbol{\mu} = \mathbf{0}$ and sparse $m \times m$ precision matrix, $\boldsymbol{Q}$, depending on spatial hyperparameters λ^2 and κ; hence $\boldsymbol{w}$ is a GMRF. The exact form for $\boldsymbol{Q}$ is chosen so that the resulting distribution for $\tilde{S}(s)$ approximates the distribution of the solution to the SPDE, and thus the form will depend on the basis functions. The basis functions are chosen to be piecewise linear functions; that is, $\psi_m(s) = 1$ at the m-th vertex of the mesh and $\psi_m(s) = 0$ at all other vertices, $m = 1, \ldots, M$. This results in a set of pyramids, each with typically a six- or seven-sided base.

2.2 Bayesian Computation

2.2.1 Hierarchical Models

In this work we focus on Bayesian hierarchical models, which are comprised of three levels:

$$\begin{aligned} \text{Stage I:} \quad & \boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\phi} \sim p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\phi}) \\ \text{Stage II:} \quad & \boldsymbol{\theta}|\boldsymbol{\phi} \sim p(\boldsymbol{\theta}|\boldsymbol{\phi}) \\ \text{Stage III:} \quad & \boldsymbol{\phi} \sim p(\boldsymbol{\phi}) \end{aligned}$$

where $\boldsymbol{y}$ refers to the data, $\boldsymbol{\theta}$ are the latent parameters, and $\boldsymbol{\phi}$ are the hyperparameters.

Bayesian inference centers around computing the posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{y})$, the distribution of the unknown parameters given the observed data. Note that by Bayes rule,

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{y}) &= \frac{p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\phi})p(\boldsymbol{\theta}|\boldsymbol{\phi})p(\boldsymbol{\phi})}{p(\boldsymbol{y})} \\ &\propto p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\phi})p(\boldsymbol{\theta}|\boldsymbol{\phi})p(\boldsymbol{\phi}). \end{aligned}$$

2.2.2 Penalized Complexity Priors

Simpson et al. (2017) introduce the concept of the penalized complexity (PC) prior, which is a natural way to specify a prior distribution in nested models. The prior is designed to penalize deviations from a base model, and thus control overfitting. Specifically, we focus on the prior for the precision parameter in Gaussian random effects models.

Let $\boldsymbol{x} \sim N(\mathbf{0}, (\tau \boldsymbol{R})^{-1})$. The base model would be the absence of this random effect, and in other words it would be one that puts all the mass at $\tau^{-1} = 0$. To derive the prior for τ, a function of the Kullback-Leibler divergence between the base model and the more flexible model is used to characterize the "distance" between the two models. A constant rate penalization is used, resulting in an exponential prior for this distance. Ultimately, the PC prior for τ is,

$$p(\tau) = \frac{\lambda}{2} \tau^{-3/2} \exp\left(-\lambda \tau^{-1/2}\right), \tag{2.4}$$

where $\lambda > 0$ is a parameter characterizing the penalty for deviating from the base model. For interpretation purposes, this PC prior can be specified by 2 terms (U, α) and the relationship $\mathrm{P}(1/\sqrt{\tau} > U) = \alpha$ (the probability that the standard deviation exceeds some value U is equal to α) so that $\lambda = -\log(\alpha)/U$.

2.2.3 Integrated Nested Laplace Approximation

When the distribution of the latent effects $p(\boldsymbol{\theta}|\boldsymbol{\phi})$ is a GMRF, the Integrated Nested Laplace Approximation (INLA) can be used to provide fast and accurate approximations to the posterior marginals (Rue et al., 2009). This is especially useful if the posterior is high-dimensional (i.e., has a large number of latent effects) since traditional Markov chain Monte Carlo (MCMC) algorithms can be slow to converge. This is particularly true of Gaussian Process (GP) models (Filippone et al., 2013). The approach of Rue et al. (2009) is to approximate the posterior marginals,

$$\begin{aligned} p(\theta_i|\boldsymbol{y}) &= \int p(\theta_i|\boldsymbol{\phi}, \boldsymbol{y}) p(\boldsymbol{\phi}|\boldsymbol{y}) d\boldsymbol{\phi}, \\ p(\phi_j|\boldsymbol{y}) &= \int p(\boldsymbol{\phi}|\boldsymbol{y}) d\boldsymbol{\phi}_{-j}, \end{aligned}$$

by

$$\begin{aligned} \tilde{p}(\theta_i|\boldsymbol{y}) &= \int \tilde{p}(\theta_i|\boldsymbol{\phi}, \boldsymbol{y}) \tilde{p}(\boldsymbol{\phi}|\boldsymbol{y}) d\boldsymbol{\phi}, \\ \tilde{p}(\phi_j|\boldsymbol{y}) &= \int \tilde{p}(\boldsymbol{\phi}|\boldsymbol{y}) d\boldsymbol{\phi}_{-j}. \end{aligned} \tag{2.5}$$

They use the following approximation for the marginal posterior of the hyperparameters,

$$\tilde{p}(\boldsymbol{\phi}|\boldsymbol{y}) \propto \left.\frac{p(\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\phi})}{\tilde{p}_G(\boldsymbol{\theta}|\boldsymbol{\phi},\boldsymbol{y})}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*(\boldsymbol{\phi})}$$

where $\tilde{p}_G(\boldsymbol{\theta}|\boldsymbol{\phi},\boldsymbol{y})$ is the Gaussian approximation to $p(\boldsymbol{\theta}|\boldsymbol{\phi},\boldsymbol{y})$ and $\boldsymbol{\theta}^*(\boldsymbol{\phi})$ is the mode of $p(\boldsymbol{\theta}|\boldsymbol{\phi},\boldsymbol{y})$ for a given $\boldsymbol{\phi}$.

A grid of hyperparameters $\boldsymbol{\phi}_g$ is selected along with weights Δ_g so that the posterior marginal of the latent effects, equation (2.5), is numerically integrated,

$$\tilde{p}(\theta_i|\boldsymbol{y}) = \sum_g \tilde{p}(\theta_i|\boldsymbol{\phi}_g,\boldsymbol{y})\tilde{p}(\boldsymbol{\phi}_g|\boldsymbol{y})\Delta_g.$$

Further details are in Rue et al. (2009). An `R` package, `R-INLA`, which implements this method is available (Lindgren and Rue, 2015).

2.2.4 Template Model Builder

The `R` package `TMB` can serve as a flexible alternative to `R-INLA` and is able to provide empirical Bayes (EB) type estimates for parameters (Kristensen, 2014; Kristensen et al., 2016). We consider functions of the form,

$$f(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\phi}) = -\log \underbrace{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\phi})}_{\text{likelihood}} - \log \underbrace{p(\boldsymbol{\beta})}_{\substack{\text{prior on}\\ \text{fixed effects}}} - \log \underbrace{p(\boldsymbol{\gamma}|\boldsymbol{\phi})}_{\substack{\text{prior on}\\ \text{random effects}}} - \log \underbrace{p(\boldsymbol{\phi})}_{\text{hyperprior}},$$

i.e., the negative log likelihood (up to a constant). Previously, $\boldsymbol{\theta}$ denoted the latent parameters (which consisted of both fixed and random effects), and now we distinguish between fixed effects $\boldsymbol{\beta}$, and random effects $\boldsymbol{\gamma}$.

`TMB` uses Laplace approximations (LAs) to integrate out the random effects. Specifically,

$$\begin{aligned} L(\boldsymbol{\beta},\boldsymbol{\phi}) &= \int \exp\left[f(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\phi})\right] d\boldsymbol{\gamma} \\ &\approx L^\star(\boldsymbol{\beta},\boldsymbol{\phi}) = \det\{H(\boldsymbol{\beta},\boldsymbol{\phi})\}^{-1/2} \exp\left[f(\boldsymbol{\beta},\hat{\boldsymbol{\gamma}}(\boldsymbol{\beta},\boldsymbol{\phi}),\boldsymbol{\phi})\right], \end{aligned}$$

where $H(\boldsymbol{\beta}, \boldsymbol{\phi}) = -\frac{\partial^2}{\partial \boldsymbol{\gamma}^2} f(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\phi})|_{\boldsymbol{\gamma}=\hat{\boldsymbol{\gamma}}(\boldsymbol{\beta},\boldsymbol{\phi})}$ and $\hat{\boldsymbol{\gamma}}(\boldsymbol{\beta}, \boldsymbol{\phi}) = \text{argmax}_{\boldsymbol{\gamma}} f(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\phi})$. The `TMB` function returns $-\log L^{\star}(\boldsymbol{\beta}, \boldsymbol{\phi})$ and its derivative so that an estimate for $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ can be obtained using nonlinear optimization techniques,

$$\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\phi}} = \text{argmin}_{\boldsymbol{\beta},\boldsymbol{\phi}} - \log L^{\star}(\boldsymbol{\beta}, \boldsymbol{\phi}),$$

and the Hessian can be used to derive an estimate of uncertainty.

2.2.5 Hamiltonian Monte Carlo

Hamiltonian Monte Carlo (HMC), also known as hybrid Monte Carlo, combines MCMC with deterministic simulation methods found in physics (Duane et al., 1987; Neal, 2011). HMC can be especially useful when dealing with a high-dimensional target distribution as it reduces strong dependence in the chain that can happen when using a Gibbs sampler or the Metropolis algorithm.

The first step in implementing HMC is to define a Hamiltonian function in terms of the target distribution ($p(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{y})$). There are two sets of variables that we consider: the variables of interest ($\boldsymbol{q} = (\boldsymbol{\theta}, \boldsymbol{\phi})$) also known as position variables, and auxiliary momentum variables ($\boldsymbol{p}$), which have a Gaussian distribution, $\boldsymbol{p} \sim N(0, \boldsymbol{M})$. The Hamiltonian function is as follows:

$$\begin{aligned} H(\boldsymbol{q}, \boldsymbol{p}) = & \underbrace{U(\boldsymbol{q})}_{\text{potential energy}} + \underbrace{K(\boldsymbol{p})}_{\text{kinetic energy}} \\ K(\boldsymbol{p}) = & \frac{1}{2} \boldsymbol{p}^{\top} \boldsymbol{M}^{-1} \boldsymbol{p} \\ U(\boldsymbol{q}) = & -\log(p(\boldsymbol{q}|\boldsymbol{y})) = \text{const} - \log(p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\phi})) - \log(p(\boldsymbol{\theta}|\boldsymbol{\phi})) - \log(p(\boldsymbol{\phi})). \end{aligned}$$

Hamilton's equations, the partial derivatives of the Hamiltonian, determine how $\boldsymbol{q}$ and $\boldsymbol{p}$ evolve with time t,

$$\frac{d\boldsymbol{q}}{dt} = \frac{\partial H(\boldsymbol{q}, \boldsymbol{p})}{\partial \boldsymbol{p}} = \boldsymbol{M}^{-1} \boldsymbol{p} \tag{2.6}$$

$$\frac{d\boldsymbol{p}}{dt} = \frac{\partial H(\boldsymbol{q}, \boldsymbol{p})}{\partial \boldsymbol{q}} = -\frac{\partial U(\boldsymbol{q})}{\partial \boldsymbol{q}}. \tag{2.7}$$

Note that according to (2.7) we will need the gradient of the negative log-posterior density. The analytical forms for the negative log-posterior density and its gradient will be given in later sections, when the algorithm is used for a particular problem.

In terms of computer implementation, Hamilton's equations (2.6) and (2.7) are approximated by discretizing time, using a small stepsize ϵ. The algorithm is as follows:

1. Simulate $\boldsymbol{p}(t) \sim N(0, \boldsymbol{M})$
2. Using the leapfrog method with L leapfrog steps, update $\boldsymbol{q}$ and $\boldsymbol{p}$:
   (a) $\boldsymbol{p}(t+\epsilon/2) = \boldsymbol{p}(t) - \frac{\epsilon}{2}\frac{\partial U(\boldsymbol{q})}{\partial \boldsymbol{q}}(\boldsymbol{q}(t))$
   (b) $\boldsymbol{q}(t+\epsilon) = \boldsymbol{q}(t) + \epsilon \boldsymbol{M}^{-1}\boldsymbol{p}(t+\epsilon/2)$
   (c) $\boldsymbol{p}(t+\epsilon) = \boldsymbol{p}(t+\epsilon/2) - \frac{\epsilon}{2}\frac{\partial U(\boldsymbol{q})}{\partial \boldsymbol{q}}(\boldsymbol{q}(t+\epsilon))$
3. Define $\boldsymbol{p}^*$ and $\boldsymbol{q}^*$ as the values of the momentum and parameter vectors after the L leapfrog steps (i.e., at time $t + \epsilon L$) and $\boldsymbol{p}^{j-1}$ and $\boldsymbol{q}^{j-1}$ as the momentum and parameter vectors prior to the leapfrog process (i.e., at time t). Calculate the acceptance probability

$$\begin{aligned}\alpha &= \min\left[\frac{p(\boldsymbol{q}^*|\boldsymbol{y})p(\boldsymbol{p}^*)}{p(\boldsymbol{q}^{j-1}|\boldsymbol{y})p(\boldsymbol{p}^{j-1})}, 1\right] \\ &= \min\left[\exp\left(U(\boldsymbol{q}^{j-1}) - U(\boldsymbol{q}^*) + K(\boldsymbol{p}^{j-1}) - K(\boldsymbol{p}^*)\right), 1\right].\end{aligned}$$

4. Set

$$\boldsymbol{q}^j = \begin{cases} \boldsymbol{q}^* & \text{with probability } \alpha \\ \boldsymbol{q}^{j-1} & \text{otherwise.} \end{cases}$$

To make HMC more efficient, it can be useful to take the mass matrix $\boldsymbol{M}$ as proportional to the inverse of the covariance matrix of the posterior distribution. In our applications, we empirically find the covariance matrix $\boldsymbol{D}$ after several iterations, and set $\boldsymbol{D}^{-1} = \boldsymbol{M}$. In Step (1) of the algorithm we will need to sample from $\boldsymbol{p} \sim N(\boldsymbol{0}, \boldsymbol{D}^{-1})$. To efficiently sample from a multivariate normal distribution defined in terms of its precision matrix, we follow Algorithm 2.4 in Rue and Held (2005), which for exposition purposes is:

1. Compute the Cholesky factorization, $\boldsymbol{D} = \boldsymbol{L}\boldsymbol{L}^\top$ where $\boldsymbol{L}$ is a lower triangular matrix,
2. Sample $\boldsymbol{z} \sim N(\boldsymbol{0}, \boldsymbol{I})$,
3. Solve $\boldsymbol{L}^\top \boldsymbol{p} = \boldsymbol{z}$ by back substitution,
4. Return $\boldsymbol{p}$.

HMC has been implemented in the `Stan` computing environment (Carpenter et al., 2017) and an `R` package, `RStan`, exists.

2.3 Data Augmentation

Consider the scenario where we have a data likelihood, $p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{a})$, but $\boldsymbol{a}$ is a collection of parameters that are missing. This could represent, for example, (unknown) locations in space or the birth and date times of children. Ultimately, the posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{y})$ is desired,

$$p(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{y}) = \int p(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{a}|\boldsymbol{y}) d\boldsymbol{a}.$$

We could represent this as,

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{y}) &= \int \frac{p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{a}) p(\boldsymbol{a}|\boldsymbol{\theta}, \boldsymbol{\phi}) p(\boldsymbol{\theta}|\boldsymbol{\phi}) p(\boldsymbol{\phi})}{p(\boldsymbol{y})} d\boldsymbol{a} \\ &\propto p(\boldsymbol{\theta}|\boldsymbol{\phi}) p(\boldsymbol{\phi}) \int p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{a}) p(\boldsymbol{a}|\boldsymbol{\theta}, \boldsymbol{\phi}) d\boldsymbol{a} \end{aligned}$$

One approach is to integrate the complete data likelihood over the distribution of the missing data, resulting in,

$$p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\phi}) = \int p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{a}) p(\boldsymbol{a}|\boldsymbol{\theta}, \boldsymbol{\phi}) d\boldsymbol{a},$$

where $p(\boldsymbol{a}|\boldsymbol{\theta}, \boldsymbol{\phi})$ is the prior distribution for $\boldsymbol{a}$. In some cases, as illustrated in Chapters 3 and 6, this integral is a convolution of distributions. When $p(\boldsymbol{a}|\boldsymbol{\theta}, \boldsymbol{\phi})$ is normal or Poisson (and in the cases we examine $p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{a}) = 1$)), evaluation of this integral (or sum) results in $p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\phi})$ having a normal or Poisson distribution, respectively. In other cases, this integral (or sum) is a mixture distribution, as seen in Chapters 4 and 5. See Table 2.1 for a description

Scenario	$\boldsymbol{y}$	$\boldsymbol{a}$
Chapter 3	Outcome aggregated over areas	Outcome at grid points in areas
Chapter 4	Outcome at unknown location	Outcome at particular location
Chapter 5	Total number of births and deaths of children by mother	Time of births and deaths of children
Chapter 6	Total number of deaths of children by mother	Indicator for whether child died given birth at a particular time

Table 2.1: Brief descriptions of the observed data $\boldsymbol{y}$ and missing data $\boldsymbol{a}$ in the scenarios considered.

of $\boldsymbol{y}$ and $\boldsymbol{a}$ for each of the applications considered in this dissertation. The usual approach when dealing with mixture distributions is to explicitly introduce $\boldsymbol{a}$ as *auxiliary variables*. An iterative procedure can be used to obtain samples from the posterior, $p(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{y})$. This is the idea of the data augmentation algorithm (Tanner and Wong, 1987). The Gibbs sampling algorithm is as follows, for iteration j:

1. Draw $\boldsymbol{a}^{(j)}$ from $p(\boldsymbol{a}|\boldsymbol{y}, \boldsymbol{\theta}^{(j-1)}, \boldsymbol{\phi}^{(j-1)})$,
2. Draw $\boldsymbol{\theta}^{(j)}, \boldsymbol{\phi}^{(j)}$ from $p(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{y}, \boldsymbol{a}^{(j)})$.

Note that,

$$p(\boldsymbol{a}|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\phi}) \propto p(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{\theta}, \boldsymbol{\phi}) p(\boldsymbol{a}|\boldsymbol{\theta}, \boldsymbol{\phi}) \quad \text{and}$$
$$p(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{y}, \boldsymbol{a}) \propto p(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{\theta}, \boldsymbol{\phi}) p(\boldsymbol{\theta}|\boldsymbol{\phi}, \boldsymbol{a}) p(\boldsymbol{\phi}|\boldsymbol{a}).$$

2.4 Survey Sampling

Let Y_{ik} be a binary indicator for the outcome of interest for the kth individual, $k = 1, \ldots, N_i$ in area (or stratum) i, $i = 1, \ldots, I$. In small area estimation, the population mean in area i is of particular interest, $Y_i = \sum_{k=1}^{N_i} Y_{ik}/N_i$. A survey is conducted to obtain a sample of n_i individuals, a subset of the population. Let I_{ik} be an indicator for membership into the sample. Define π_{ik} to be the first-order inclusion probability, which is the probability

that the kth individual in area i is selected (i.e., $P(I_{ik} = 1) = \pi_{ik}$), and y_{ik} is the observed outcome value. Further define π_{ikl} to be the second-order inclusion probability, which is the probability that both individuals k and l are selected in area i.

The Horvitz-Thompson estimator (Horvitz and Thompson, 1952) for the population mean is

$$\hat{Y}_i^{\mathrm{HT}} = \frac{1}{N_i} \sum_{k=1}^{N_i} \frac{Y_{ik} I_{ik}}{\pi_{ik}} = \frac{1}{N_i} \sum_{k=1}^{n_i} \frac{y_{ik}}{\pi_{ik}}$$

with variance and estimated variance,

$$\mathrm{V}(\hat{Y}_i) = \frac{1}{N_i^2} \sum_{k=1}^{N_i} \sum_{l=1}^{N_i} \frac{y_{ik} y_{il}}{\pi_{ik} \pi_{il}} \left(\pi_{ikl} - \pi_{ik} \pi_{il} \right),$$

$$\hat{\mathrm{V}}(\hat{Y}_i) = \frac{1}{N_i^2} \sum_{k=1}^{n_i} \sum_{l=1}^{n_i} \frac{y_{ik} y_{il}}{\pi_{ik} \pi_{il}} \left(\frac{\pi_{ikl} - \pi_{ik} \pi_{il}}{\pi_{ikl}} \right).$$

When population totals, N_i, are unknown the Hájek estimator (Hájek, 1971)

$$\hat{Y}_i^{\mathrm{Háj}} = \frac{\sum_{k=1}^{n_i} y_{ik} / \pi_{ik}}{\sum_{k=1}^{n_i} 1 / \pi_{ik}}$$

For many cluster designs, various approximations are made in order to estimate the variance.

2.4.1 Accounting for Survey Design in a Discrete Time Survival Analysis

In discrete time survival analysis one divides time into J discrete intervals in which the event of interest (e.g., death) could occur, rather than analyzing time continuously. In the context of under-five mortality modeling, consider dividing the interval $[0, 5)$ years into yearly intervals, i.e., $[0, 1)$, $[1, 2)$, $[2, 3)$, $[3, 4)$, and $[4, 5)$ years. The data consist of sequences of two indicators for each child. The first is an indicator of being at risk in interval $[j, j+1)$, denoted N_j, which would be 1 for each interval the child is alive at the start of. The second is an indicator of death, denoted Y_j, which would be a 1 for the interval they die during and a 0 otherwise. For example, a child that dies at age 3.5 would be at risk during the first 4 intervals, and thus the at risk indicator would be 1 for the first 4 intervals. In terms of the death indicator, the child would be assigned a 0 in the first 3 intervals and assigned a 1 in the last interval, $[3, 4)$.

Associated with each interval is a hazard, denoted ${}_1q_j$, or the probability of the event occurring between age j and age $j+1$ given survival to time j. This can the be used to derive ${}_5q_0$, the probability of death before age 5,

$$ {}_5q_0 = 1 - \prod_{j=0}^{4}(1 - \ {}_1q_j) $$

Logistic regression can be used to obtain estimates of these hazards from the observed data. Specifically,

$$ Y_j | N_j = 1, \ {}_1q_j \sim \text{Binomial}(N_j = 1, \ {}_1q_j) $$
$$ \text{logit}({}_1q_j) = \beta_j. $$

The supplementary of Mercer et al. (2015) detail how to obtain an estimate and associated variance in the context of under-five mortality modeling. The first step is to incorporate design weights from complex surveys into logistic regression. Define $w_{ik} = 1/\pi_{ik}$ to be the reported survey weights for individual k in area i. Define $\boldsymbol{B}_i$ to be the finite population parameter of length 5 with entries B_{ij} in area i (as compared to the superpopulation parameter $\boldsymbol{\beta}_i$). Consider $\hat{\boldsymbol{B}}_i$, where each entry $\hat{B}_{ij}$ is the solution to,

$$ 0 = \sum_{k=1}^{n_i} w_{ik} \left[y_{ikj} - \frac{\exp(\hat{B}_{ij})}{1 + \exp(\hat{B}_{ij})} \right], \quad j = 1, \ldots, 5. $$

This can be fit using `svyglm()` from the `survey` package in `R` (Lumley, 2004, 2018) and will provide an estimate of the variance $\boldsymbol{\Sigma}$; see Mercer et al. (2015) for the derivation of this variance.

The asymptotic distribution of $\hat{\boldsymbol{B}}_i$ is,

$$ \hat{\boldsymbol{B}}_i \sim N(\boldsymbol{B}_i, \boldsymbol{\Sigma}_i) $$

and define $\hat{\boldsymbol{\Sigma}}_i$ to be an estimate of $\boldsymbol{\Sigma}$. Define $\eta(\boldsymbol{B}_i) = \eta_i = \text{logit}({}_5q_0^i) = \log\left\{\prod_{j=1}^{5}(e^{B_{ij}} + 1) - 1\right\}$, then,

$$ \hat{\eta}_i \sim N(\eta_i, \hat{V}_i) $$

where $\hat{V}_i$ is found using the delta method,

$$\hat{V}_i = \frac{\partial \eta_i}{\partial \boldsymbol{B}_i}^\top \hat{\boldsymbol{\Sigma}} \frac{\partial \eta_i}{\partial \boldsymbol{B}_i}.$$

Chapter 3

COMBINING POINT AND AREA-LEVEL DATA

3.1 Introduction

When modeling residual spatial dependence, it is appealing to formulate modeling in terms of an underlying continuous spatial surface, and this is the usual approach for point-referenced data. Continuous modeling becomes more difficult when the data contain regional aggregates at varying spatial resolutions. In epidemiological studies, data is often aggregated for reporting or anonymization. While there exists a wealth of techniques to model regional data at a fixed resolution (Cressie and Wikle, 2011; Banerjee et al., 2014), these models do not extend in a straightforward fashion to situations where more than one resolution is used. In this chapter, we develop methods for dealing with such situations.

We focus on two motivating settings. In the first scenario, we suppose we have areal data only. In epidemiology and the social sciences this situation is the most common, since such data usually satisfy confidentiality constraints, and typically arises from aggregation over a disjoint, irregular partition of the study map, based on administrative boundaries. As an example, we consider incident lip cancer counts observed in 56 counties in Scotland over the years 1975–1980. These data provide a good test case, since they have been extensively analyzed in the literature; see Wakefield (2007) and the references therein. In this setting, we may view the continuous underlying surface as a device to induce a spatial prior for the areas that avoids the usual arbitrary element of defining neighbors over an irregular geography.

The second scenario we consider is one in which one data source is available as area level averages over a set of areas, but is supplemented with data collected from surveys at known locations. Our interest in this problem arises from spatial modeling of demographic indicators in a developing world context. One source of data is the census, which provides

data at the aggregate level, e.g., the average or sum of a variable over an administrative areal unit. In many countries, demographic and health information are also collected via household surveys, such as Demographic and Health Surveys (DHS; Corsi et al., 2012), and these provide a second source of data. These surveys are typically stratified cluster designs with countries being stratified into coarse areas and into urban/rural, with enumeration areas (EAs) sampled within strata, and then households sampled within EAs. In these surveys, the locations of the EAs, i.e., the GPS coordinates, are often available. In Chapter 4 we consider the case where the exact GPS coordinates are unknown. In each of these examples, we assume that there is a latent, continuous Gaussian random field (GRF) that varies in space, $\{S(s) : s \in R \subset \mathbb{R}^2\}$ where R is our study region of interest.

The situation with which we are concerned with in this chapter is closely related to the change of support problem (COSP; Bradley et al., 2016; Cressie and Wikle, 2011; Gelfand, 2010; Gotway and Young, 2002). This problem occurs when one would like to make inference at a particular spatial resolution, but the data are available at another resolution. Much of this work focuses on normal data, with block kriging being used. For example, Fuentes and Raftery (2005) combine point and aggregate pollution data, with the latter consisting of outputs from numerical models, produced over a gridded surface. Berrocal et al. (2010) considered the same class of problem, but added a time dimension and used a regression model with coefficients that varied spatially to relate the observed data to the modeled output. Moraga et al. (2017) develop a similar framework to ours and use a stochastic partial different equation (SPDE) approach in order to relate two levels of pollution data. Specifically, the model they propose relates the continuous surface to the area (grid) level by taking an unweighted average of the surface at various points within each grid. We extend this work in several regards, most importantly, our model can accommodate non-normal outcomes and we also allow for a more complex relationship between the point-level process and the aggregated data. Therefore, we are able to address a wider range of situations.

Diggle et al. (2013) take a different approach for discrete data and model various applications using log-Gaussian Cox processes, including the reconstruction of a continuous spatial

surface from aggregate data. Their approach is based on MCMC and follows Li et al. (2012) in simulating random locations of cases within areas, which is a computationally expensive step. Recently, Taylor et al. (2015, 2018) have focused on improving computation for this scenario and extended the framework to include spatiotemporal models.

A related problem to the COSP, is the modeling of areal data over time, but with boundary changes. Lee et al. (2017) analyze space-time data on male bladder cancer in Nova Scotia; the spatial aggregation changes over time, with the older data tending to be of aggregate form and the point data being the norm in more recent years. Building on previous work (Fan et al., 2011; Li et al., 2012; Nguyen et al., 2012), they use a local EM algorithm in conjunction with a local polynomial to model the risk surface.

We propose a three-stage Bayesian hierarchical model that can combine point and areal data by assuming a common underlying smooth, continuous surface. We use the SPDE approach of Lindgren et al. (2011) to model the latent field, which allows for computationally efficient inference. The chapter is structured as follows. In Section 3.2 we describe the model and in Section 3.3 the computational details. A simulation study in Section 3.4 considers a number of scenarios including point data, areal data, and a combination of the two. In Section 3.5 we illustrate the non-linear areal data only situation for the famous Scotland lip cancer example. Section 3.6 contains concluding remarks.

3.2 Model Description

We propose a general model framework for inference that can be used for data collected at points, over areas, or a combination of the two. We describe the likelihood first for normal and then for Poisson data (as an illustration of a non-normal outcome), before concluding with a discussion of the model for the latent spatial surface.

3.2.1 Normal Responses

In general, models are specified at the point level. We describe the normal model in the context of modeling household wealth over a spatial region. Since we will be concerned with

observations at the area-level we will introduce general notation. The region of interest, R, is divided into n disjoint areas denoted R_i, with N_i households in area R_i, $i = 1, \ldots, n$. Let $Y_{ih} = Y(s_{ih})$ denote the h-th response associated with location s_{ih} (e.g., longitude and latitude), with covariate information $z_{ih} = z(s_{ih})$, $h = 1, \ldots, N_i$; we assume a single covariate only for notational simplicity, with the extension to multiple covariates being straightforward. The household-level model is $Y_{ih} \mid \mu_{ih},\ \sigma^2 \sim_{ind} \mathrm{N}(\mu_{ih}, \sigma^2)$, with $\mu_{ih} = \mu(s_{ih}) = \beta_0 + \beta_1 z_{ih} + S_{ih}$ and $S_{ih} = S(s_{ih})$ being the spatial random effect, where the spatial model is a GRF. We have assumed the measurement error variance σ^2 is the same for each response but this can easily be relaxed. When data are available from a census we observe the average response in each of the areas $\bar{Y}_i = \frac{1}{N_i} \sum_{h=1}^{N_i} Y_{ih}$. The induced area-level model is $\bar{Y}_i \mid \mu_i,\ \sigma^2 \sim \mathrm{N}(\mu_i, \sigma^2/N_i)$ where,

$$\mu_i = \frac{1}{N_i} \sum_{h=1}^{N_i} \{\beta_0 + \beta_1 z_{ih} + S_{ih}\}. \tag{3.1}$$

3.2.2 Poisson Responses

In the second case we consider, we assume that only the sum of all binary events, $Y_{i+} = \sum_{j=1}^{N_i} Y_{ij}$, is observed and recorded in area R_i. The individual-level model is $Y_{ij} \mid p_{ij} \sim_{ind}$ Bernoulli(p_{ij}). We assume a rare event scenario, along with a log-linear model, so that, $p_{ij} = p(s_{ij}) = \exp\{\beta_0 + \beta_1 z(s_{ij}) + S(s_{ij})\}$. We sum over all cases to give, $Y_{i+} \mid \mu_i \sim$ Poisson(μ_i), where,

$$\begin{aligned} \mu_i &= \sum_{j=1}^{N_i} \mathrm{E}[Y_{ij} | \boldsymbol{x}_{ij}] \\ &= \sum_{j=1}^{N_i} \exp\{\beta_0 + \beta_1 z(\boldsymbol{x}_{ij}) + S(\boldsymbol{x}_{ij})\}. \end{aligned} \tag{3.2}$$

If we have non-rare outcomes and only observe the sum then the situation is far more difficult to deal with since the sum of binomials with varying probabilities is a convolution of binomials. If we observe the individual outcomes Y_{ij} (and not just the sum), then we can model each as binomial (so that we do not have to resort to the convolution). The common

situation with disease counts and expected numbers are available across a set of areas is considered in Section 3.5.

3.2.3 Model for the Latent Process

We assume a zero-mean latent GRF. There are many choices for describing how the form of the covariance changes with distance, but we follow Stein (1999) and others who make a strong argument for the Matérn covariance function defined in equation (2.2). We define the practical range $\rho = \sqrt{8\nu}/\kappa$ (where κ is the scaling parameter) as the distance at which the correlation drops to approximately 0.1. The smoothness parameter ν is set to 1 (see Section 2.1.3).

3.3 Computation

There are two steps to the computation, first the continuous latent surface is discretized in a convenient fashion (Section 3.3.1), and second the posterior is approximated. We begin with the normal case (Section 3.3.2) before turning to the more difficult Poisson case (Section 3.3.3).

3.3.1 Approximating the Latent Process

The major hurdle to the more widespread modeling of spatial data with a continuous surface has been the computation. In particular, inverting and finding the determinant of the Matérn covariance matrix, which is in general not sparse, has been a roadblock when the number of points is not small. However, recent work by Lindgren et al. (2011) detail the connection between GRFs and Gaussian Markov random fields (GMRFs) via a stochastic partial differential equation (SPDE). See 2.1.3 for details.

We follow the SPDE approach and approximate the GRF over the mesh. For inference, the discretized version of the spatial prior is combined with the likelihood. In the setting where we have known locations, it follows from (2.3) that the value of the spatial random

effect at an observation point, s_{ij}, can be approximated by a weighted average of the value of the GMRF on the three nearest mesh vertices. We can write, $S(s_{ij}) \approx \tilde{S}(s_{ij}) = \boldsymbol{A}_{ij}^{\top}\boldsymbol{w}$, where $\boldsymbol{A}_{ij}$ is an $M \times 1$-vector of weights that corresponds to the ij-th row of a sparse projection matrix $\boldsymbol{A}$ and $\boldsymbol{w} \sim N\left(0, \boldsymbol{Q}^{-1}\right)$ with the precision matrix $\boldsymbol{Q}$ depending on hyperparameters λ^2 and κ. The nonzero entries of $\boldsymbol{A}_{ij}$, which correspond to the mesh points comprising the triangle containing s_{ij}, are calculated using barycentric interpolation. In the case where the observation location, s_{ij}, is at a mesh vertex, $\boldsymbol{A}_{ij}$ contains one non-zero entry that is equal to one.

For the normal response model with areal data, we use a fully Bayesian approach, since a fast computational strategy is available. Specifically, the integrated nested Laplace approximation (INLA), an approach for analyzing latent Gaussian models (Rue et al., 2009), can be used. INLA works by using a combination of Laplace approximations along with numerical integration to obtain approximations to the posterior marginals. The SPDE approach has also been implemented in the `R` package `R-INLA` (Lindgren and Rue, 2015). For the Poisson response model, `R-INLA` cannot be used for data aggregated over areas; instead, we consider approaches that involve empirical Bayes (EB), Laplace approximations (LA), MCMC, or hybrid combinations of these techniques.

3.3.2 Normal Responses

The likelihood is normal with mean (3.1), and for simplicity we assume no covariates. The key to implementation is to approximate the integrated residual spatial area risk using the mesh. We do not observe the exact locations of all households in each area, so we incorporate population information through gridded population density and weight the approximated spatial surface at the mesh points accordingly. Defining $d(s_{im})$ to be the "relative" population density at location s_{im} satisfying $d(s_{im}) \geq 0$ and $\sum_{m=1}^{M_i} d(s_{im}) = 1$, we obtain,

$$\mu_i \approx \beta_0 + \sum_{m=1}^{M_i} d(s_{im}) x_{im} = \beta_0 + \boldsymbol{D}_i^{\top}\boldsymbol{w}, \tag{3.3}$$

where M_i is the number of mesh points in area R_i and $\boldsymbol{D}_i$ is an $M \times 1$ vector with up to M_i nonzero entries $d(\boldsymbol{x}_{ik})$.

This type of model can be fit using INLA since $\boldsymbol{D}_i^\top \boldsymbol{w}$ is Gaussian.

3.3.3 Poisson Responses

For areal Poisson data, we have the model $Y_{i+} \mid \mu_i \sim_{ind} \text{Poisson}(\mu_i)$. We use a weighted average of the exponentiated spatial random effect at the mesh points contained in the area to form μ_i. That is, and again ignoring covariates, we approximate the integral (3.2), to give

$$\mu_i \approx N_i \exp(\beta_0) \sum_{m=1}^{M_i} d(s_{im}) \exp(x_{im}) = N_i \exp(\beta_0)\ \boldsymbol{D}_i^\top \boldsymbol{T} \tag{3.4}$$

where $\boldsymbol{D}_i$ is an $M \times 1$ vector, as defined in Section 3.3.2 and $\boldsymbol{T} = [\exp(x_1), \ldots, \exp(x_M)]^\top$.

Due to the structure of this model, it is not possible to use the `R-INLA` software for fitting, but we describe three alternatives. First, a quick approximation is offered by EB with a LA being used to integrate out the spatial random effects. To implement this, we use the `R` package `TMB` (which stands for Template Model Builder; Kristensen 2014). This is very efficient and estimates of the spatial hyperparameters and fixed effects can be computed within minutes. See Section 2.2.4 for an overview of this package. Second, we resort to MCMC methods. It is well known that in the Gaussian Process context, MCMC methods can be inefficient (Filippone et al., 2013). We opt to use a Hamiltonian Monte Carlo (HMC; Neal, 2011) transition operator for updating $\boldsymbol{w}$ (see section 2.2.5). Specifically, we first update the spatial hyperparameters $\boldsymbol{\phi}$ using a random walk proposal and then jointly update $\boldsymbol{w}$ and β_0 using HMC. Finally, we consider a hybrid approach where estimates for the spatial hyperparameters $\boldsymbol{\phi}$ are found using the EB approach and then, conditional on these estimates, posteriors for $\boldsymbol{w}$ and any fixed effects are explored using MCMC methods.

3.4 Simulation Study in the Normal Response Case

3.4.1 Set Up

We illustrate the method for normal responses via a simulation considering observations associated with points and observations associated with areas. As a motivating example, we assume the aim is to construct a poverty surface; understanding the spatial structure of poverty and poverty-related factors is of considerable interest (e.g., Gething et al., 2015; Minot and Baulch, 2005; Okwi et al., 2007). Poverty has many different facets, and we take the wealth index (Rutstein and Johnson, 2004) as our measure, which serves as a surrogate for long-term standard of living. The wealth index is comprised of several variables such as household ownership of consumables, access to drinking water, and toilet facilities. The score is then standardized to have mean 0 and standard deviation 1. We simulate a surface of the average wealth index score within households.

We will consider situations in which the wealth index is measured at point locations and we also consider incorporating census data, which provides the average wealth index at the area-level. Observations associated with points are taken from a design that is informed by the Kenya DHS (Kenya National Bureau of Statistics, 2015). It is simplified in that we do not consider stratification or explicit cluster sampling for the 400 locations, which correspond to the centroids of enumeration areas (EAs) from the Kenya 2008 DHS. The dots on the plots in Figure 3.1 indicate the locations of these sampling points. We emphasize that these are point locations. It would be straightforward to extend the simulation and model to acknowledge the complex design, see Wakefield et al. (2018).

Let $i = 1, \ldots, n$ index the administrative areas in Kenya and $j = 1, \ldots, n_i$ represent the EAs in area R_i. Hence, $\sum_{i=1}^{n} n_i = 400$. Furthermore, let $h = 1, \ldots, N_i$ index the households included in the census in area R_i and let N_{ij} be the number of households surveyed at the jth location (EA) in R_i. For our simulation, the number of households participating in a survey, N_{ij}, ranges from 41 to 81, with mean 55 to give 21,946 households in total. We let Y_{ih} be the wealth index of household h in area R_i. As assumed in the DHS, all households

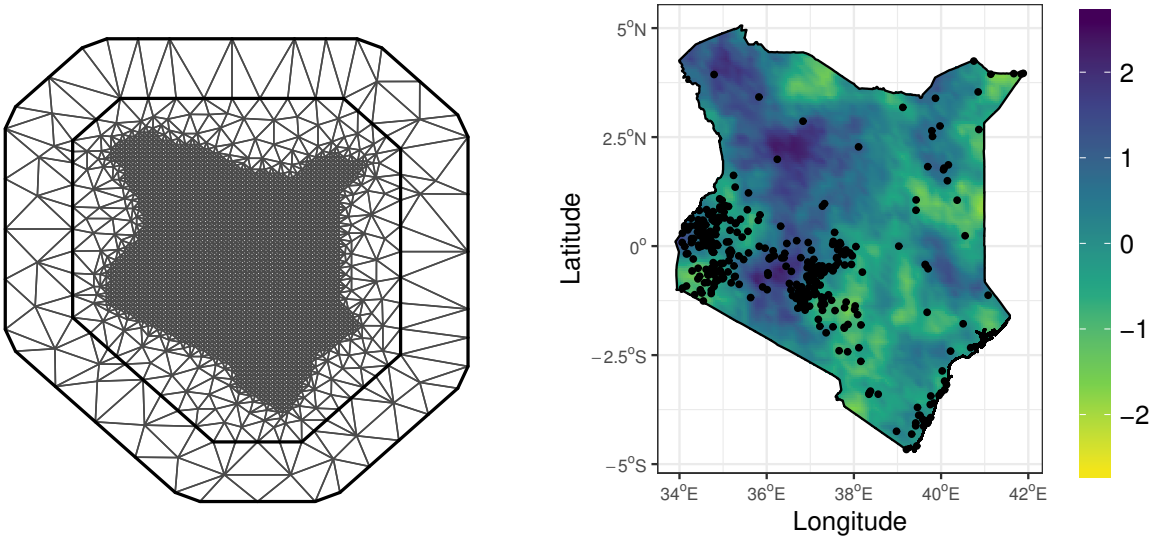

Figure 3.1: Mesh (left) and latent spatial surface (right) used for the simulations. The mesh extends beyond the border of Kenya to avoid boundary effects. The black dots represent the locations of the 400 enumeration areas.

in EA j have the same location.

We consider the data generating mechanism, $Y_{ih} \mid \mu_{ih},\ \sigma^2 \sim \mathrm{N}(\mu_{ih}, \sigma^2)$ for $h = 1, \ldots, N_i$, $i = 1, \ldots, n$. Thus, for census data we assume the following model for the average wealth index in area R_i,

$$\begin{aligned} \bar{Y}_i \mid \mu_i,\ \sigma^2 &\sim \mathrm{N}\left(\mu_i, \frac{\sigma^2}{N_i}\right), \\ \mu_i &= \beta_0 + \frac{1}{N_i}\sum_{h=1}^{N_i} S(s_{ih}) \end{aligned} \tag{3.5}$$

where $S(s)$ is the value of the spatial random effect at geographic location s_{ih}. For survey data we assume the following model for the average household wealth index taken at EA j in area R_i,

$$\begin{aligned} \bar{Y}_{ij} \mid \mu_{ij},\ \sigma^2 &\sim \mathrm{N}\left(\mu_{ij}, \frac{\sigma^2}{N_{ij}}\right), \\ \mu_{ij} &= \beta_0 + S(s_{ij}) \end{aligned} \tag{3.6}$$

where $S(s_{ij})$ is the spatial random effect evaluated at the centroid, s_{ij}, of the EA. In this setting, σ^2 represents measurement error.

We assume that the spatial model is a MGRF with Matérn covariance controlled by variance parameter λ^2 and scale parameter κ. We set $\beta_0 = 0$, $\sigma^2 = 0.25$, and $\lambda^2 = 0.75$. To simulate the spatial surface, we use the SPDE approach, which requires a triangulated mesh. This mesh is shown in the left panel of Figure 3.1 with $M = 2,786$ mesh points; these mesh points are approximately 15 km apart in the interior of Kenya. We set $\kappa = \exp(1/2)$, which corresponds to a practical range of $\rho = \sqrt{8}/\kappa = 1.72$ degrees. The simulated average household wealth index surface, i.e., $\beta_0 + \tilde{S}(\boldsymbol{s})$, is shown in the right panel of Figure 3.1, where the spatial effect $\tilde{S}(\boldsymbol{s})$ approximates $S(\boldsymbol{s})$.

To simulate data at the 400 EAs, we approximate (3.6) by $\mu_{ij} = \beta_0 + \tilde{S}(s_{ij})$ where $\tilde{S}(s_{ij})$ is the simulated spatial effect at EA j in area R_i. To simulate the census data we use gridded population estimates from SEDAC (Center for International Earth Science Information Network - CIESIN - Columbia University, 2016), which are available on an (approximately) 1 km square grid at the equator. The gridded population estimates are then transformed to household estimates by dividing the population estimates by 3.9, the mean size of households in 2014 (Kenya National Bureau of Statistics, 2015). We then approximate (3.5) by $\mu_i = \beta_0 + \frac{1}{N_i}\sum_{g=1}^{G_i} N_{ig}\tilde{S}(s_{ig})$ where N_{ig} is the household estimate for grid cell g in area R_i, $N_i = \sum_{g=1}^{G_i} N_{ig}$ is the household estimate for area R_i, and $\tilde{S}(s_{ig})$ is the simulated spatial effect at the centroid of grid cell g.

We consider five different scenarios with varying levels of information available on location: (1) survey data only, (2) census data up to county level ($n = 47$) only, (3) both survey data and census data up to county level, (4) census data up to provincial level ($n = 8$) only, and (5) both survey data and census data up to provincial level. When we analyze survey and census data together, we assume the two data sources are independent, which in practice means that the surveyed population is only a small fraction of the total population.

To assess accuracy of the reconstruction under each scenario, we compute the mean

squared error (MSE) and mean absolute error (MAE) of the spatial effect surface by

$$\text{MSE} = \left(\sum_{i=1}^{n} M_i\right)^{-1} \sum_{i=1}^{n}\sum_{m=1}^{M_i} (\hat{x}_{im} - x_{im})^2, \qquad \text{MAE} = \left(\sum_{i=1}^{n} M_i\right)^{-1} \sum_{i=1}^{n}\sum_{m=1}^{M_i} |\hat{x}_{im} - x_{im}|$$

respectively, where $\hat{x}_{im}$ is the posterior mean and x_{im} is the "true" value of the spatial effect at mesh point s_{im}. Both the MSE and MAE are given in Table 3.1 for all 5 scenarios. The top row of Figure 3.2 gives the sampling locations/areas, with each column corresponding to a different sampling scheme, and the middle and bottom rows the posterior means and standard deviations of the spatial effect surface.

3.4.2 Computational Details

The household-level model is $Y_{ih} \mid \mu_{ih} \sim \text{N}(\mu_{ih}, \sigma^2)$, with

$$\mu_{ih} = \beta_0 + S(s_{ih}),$$

with unknown variance σ^2. Across all simulation scenarios, we use the priors, $\beta_0 \sim \text{N}(\mu_{\beta_0}, \sigma^2_{\beta_0})$, $\sigma^2 \sim \text{Beta}(2, 5)$, $\boldsymbol{\phi} \sim \text{N}\left(\boldsymbol{\mu}_\phi, \boldsymbol{\Sigma}_\phi\right)$, with $\mu_{\beta_0} = 0$, $\sigma^2_{\beta_0} = 100$,

$$\boldsymbol{\mu}_\phi = \begin{bmatrix} -1.17 \\ -0.0933 \end{bmatrix}, \qquad \boldsymbol{\Sigma}_\phi = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}.$$

The hyperprior for $\boldsymbol{\phi} = [\log\lambda, \ \log\kappa]^\top$ is chosen to be fairly vague. Here, the prior mean for ϕ_1 corresponds to a marginal variance λ^2 of 1. The prior mean for ϕ_2 corresponds to a practical range ρ of roughly 20% of the domain size.

We use the `R` package `R-INLA` for computation. Fitting models involving observations with exact locations is straightforward as there exist functions to define the matrix $\boldsymbol{A}$ used to project the spatial random effect from the mesh vertices to point locations (see Section 3.1). Details of how to specify these models in `R-INLA` using the SPDE approach can be found in Lindgren and Rue (2015). In order to fit the models that involve census data, we adapt $\boldsymbol{A}$ since this matrix can be viewed as a way to average the random effect at mesh points. In these scenarios, we define $\boldsymbol{D}$ to be a matrix with n rows and M columns, made up

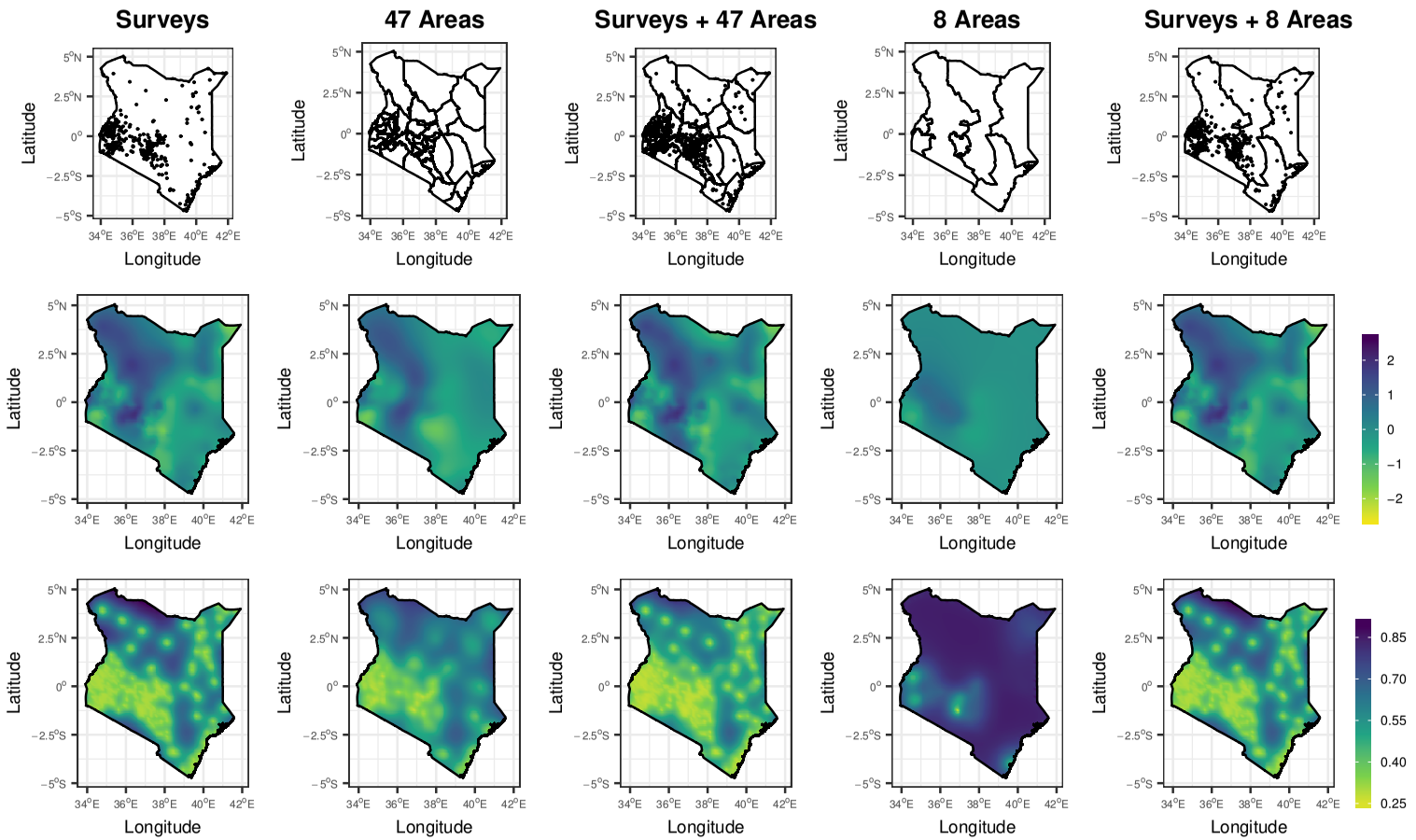

Figure 3.2: Comparison of results under the five scenarios: 400 surveys with exact location (Surveys), census data at the county level (47 Areas), both survey and census data at the county level (Surveys + 47 Areas), census data at the provincial level (8 Areas), and both survey and census data at the provincial level (Surveys + 8 Areas). Top row is the data available under each scenario. Black dots are the locations of the enumeration areas and black borders correspond to the various boundaries (47 counties and 8 provinces). Middle row is the predicted (posterior mean) of the spatial surface. Bottom row is the posterior standard deviation of the predicted surface.

Scenario	β_0: 0	σ^2: 0.25	ρ: 1.72	λ^2: 0.75	MSE	MAE
Surveys	0.0614	0.292	2.16	0.894	0.206	0.333
	(-0.549, 0.635)	(0.235, 0.365)	(1.56, 3.12)	(0.515, 1.65)		
47 Areas	0.198	0.213	2.08	0.687	0.339	0.463
	(-0.334, 0.728)	(0.0618, 0.367)	(1.37, 3.24)	(0.388, 1.27)		
Surveys + 47 Areas	0.0529	0.323	2.11	0.833	0.170	0.316
	(-0.519, 0.575)	(0.262, 0.401)	(1.56, 3.03)	(0.500, 1.50)		
8 Areas	-0.0108	0.211	1.10	0.654	0.551	0.602
	(-0.511, 0.444)	(0.0617, 0.364)	(0.383, 2.30)	(0.317, 1.76)		
Surveys + 8 Areas	0.0571	0.298	2.17	0.879	0.201	0.330
	(-0.552, 0.616)	(0.241, 0.371)	(1.58, 3.15)	(0.516, 1.63)		

Table 3.1: Posterior median and 95% credible intervals (CI) for parameters, mean squared error (MSE) and mean absolute error (MAE) of the surfaces in the simulation under five scenarios: 400 surveys with exact location (Surveys), census data at the county level (47 Areas), both survey and census data at the county level (Surveys + 47 Areas), census data at the provincial level (8 Areas), and both survey and census data at the provincial level (Surveys + 8 Areas).

of row-vectors $\boldsymbol{D}_i^\top$ of length M, where M is the number of mesh points. These row vectors $\boldsymbol{D}_i^\top$ contain up to M_i non-zero entries $d(s_{im})$. In the case of area-level observations, we use $\boldsymbol{D}$ in place of $\boldsymbol{A}$ when fitting the model using `R-INLA`. In scenarios involving a combination of point and areal data, the resulting projection matrix contains rows from both $\boldsymbol{D}$ and $\boldsymbol{A}$.

3.4.3 Survey Data

In the first scenario, we consider the situation in which we have survey data available from 400 EAs. To fit the model using `R-INLA`, we construct the projection matrix $\boldsymbol{A}$ as described in Section 3.3.1. We fit model (3.6) using the SPDE approach.

Posterior medians and 95% credible intervals (CIs) for the parameters are presented in Table 3.1 and the predicted spatial random effect surface is depicted in Figure 3.2 (left column). In general, the posterior medians are relatively close to their true values and all credible intervals cover the true value, though are fairly wide. The predicted spatial surface (posterior mean) over Kenya is visually similar to the true spatial surface, though there is some attenuation. Regions of Kenya that have a higher spatial effect are predicted to be lower and vice versa; this shrinkage to the mean phenomenon is well known in the spatial literature (Section 6.4 of Diggle and Ribeiro, 2007). We also see that the posterior standard deviation of the spatial effect is lower in the vicinity of the 400 enumeration areas and higher elsewhere. The posterior median and 2.5th and 97.5th percentiles of the predicted average household wealth index is depicted in 3.3 (left column) and we see similar patterns.

3.4.4 Census Data (47 Counties)

We next consider a situation in which we have census data for each of the $n = 47$ counties in Kenya. To implement (3.5) we approximate μ_i using (3.3), which requires the population density at the mesh points. To determine the population estimate corresponding to the grid containing the mesh point, we used gridded population estimates from SEDAC.. Figure 3.4 depicts the $n = 47$ counties and mesh points with population density $d_{im} > 1/M_i$ in gray.

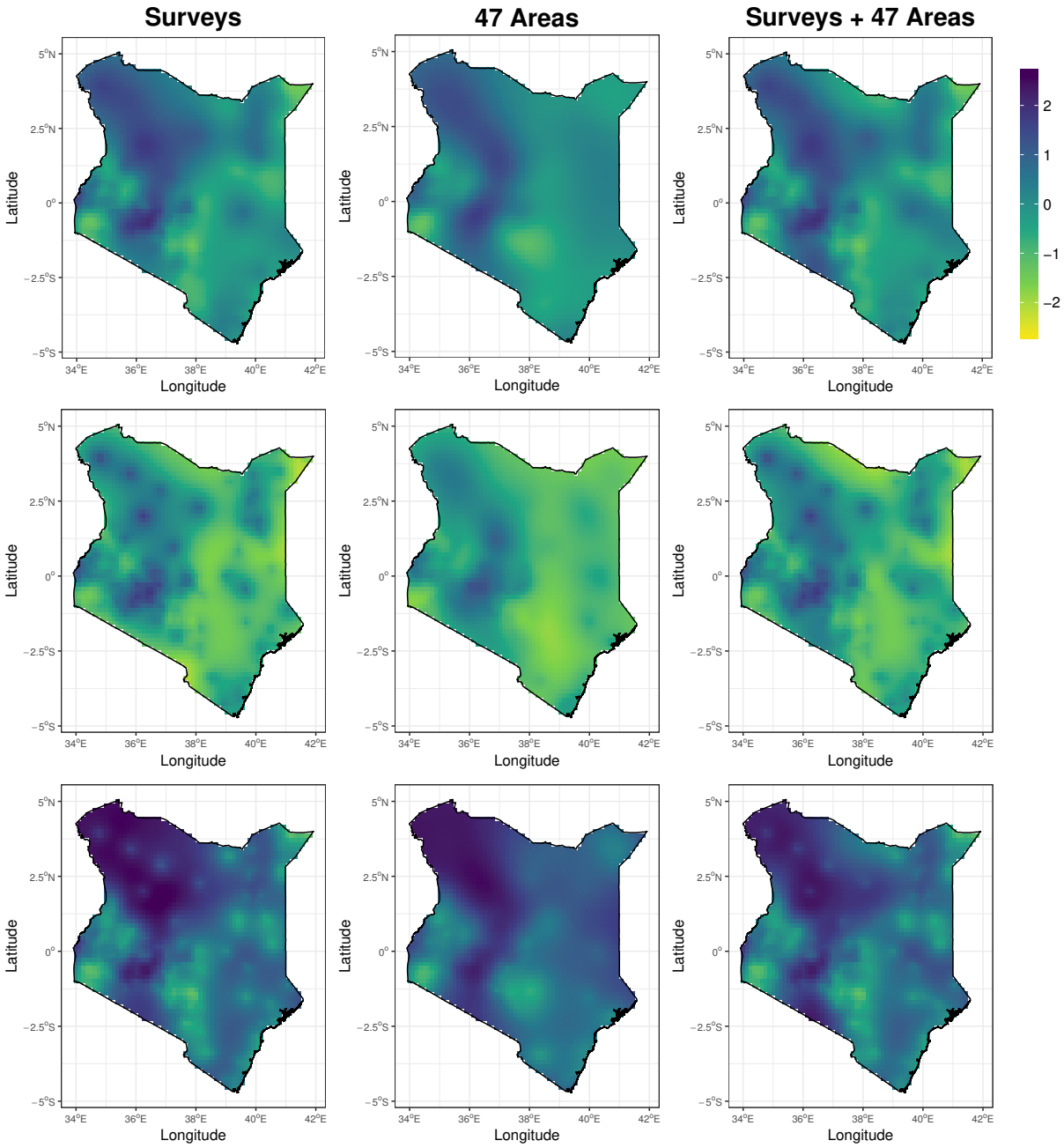

Figure 3.3: Comparison of results when the 400 surveys are used (Left), census data from the 47 counties are used (Middle), and both are used (Right). Top row is the predicted (posterior median) household wealth index surface, middle row is the 2.5 percentile, and bottom row is the 97.5 percentile.

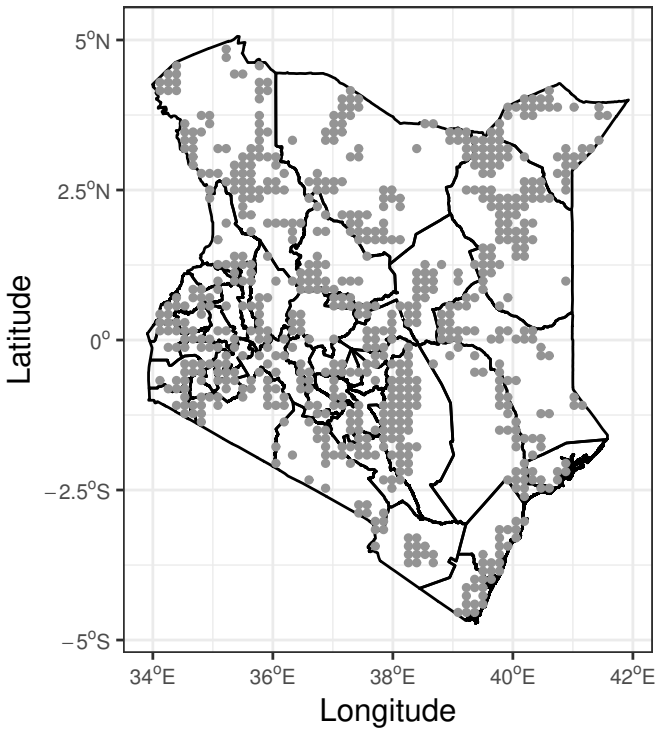

Figure 3.4: Distribution of population in the 47 counties of Kenya. The gray circles represent mesh points where the population density is larger than average for that area.

The results are presented in Table 3.1 and depicted in Figure 3.2 (second column). Again, we see that the posterior medians are relatively close to the true values. The predicted spatial surface is similar to the truth and is very similar to the predicted surface estimated for the point data. In general, the posterior standard deviation of the spatial effect is higher under this scenario than when we had information from 400 surveys. This is also evident when comparing the 2.5th and 97.5th percentile of the predicted average household wealth index, displayed in 3.3 (middle column).

3.4.5 Survey and Census Data (47 Counties)

Another scenario that might arise is one in which we have both survey data at 400 EAs and census data available for 47 counties. Thus, we simply combine the methods from Sections 3.4.3 and 3.4.4. The results are presented in Table 3.1 and displayed in Figures 3.2 (third column) and 3.3 (right column). Overall, there is a slight improvement over the survey information only case. We note that there are some identifiability problems when estimating the two variance parameters, which manifests itself here with σ^2 being overestimated and λ^2 underestimated.

3.4.6 Census Data (8 Provinces)

In order to evaluate the effect when area information is known at a greater aggregate level than previously considered, we examine a situation where we only have census data available for each of the $n = 8$ provinces in Kenya. Implementation-wise, this scenario is analogous to the one previously described in Section 3.4.4. Results are presented in Table 3.1 and a depiction of the posterior mean and standard deviation along with a map displaying the 8 provinces is in Figure 3.2 (fourth column). In this scenario, inference for the parameters is severely deteriorated when compared to the previous cases. In particular, the credible intervals are much wider than in the previous scenarios and the MSE and MAE are substantially larger.

3.4.7 Survey and Census Data (8 Provinces)

The last scenario we consider is similar to that in Section 3.4.5, where we have survey and census data available (at the provincial level). Parameter estimates are presented in Table 3.1 and the posterior mean and standard deviation of the random effect is depicted in Figure 3.2 (last column). Again, identifiability issues are evident in inference for the variances. The spatial effect surface is similar to the surveys-only scenario.

In terms of the MSEs, the values are 0.206, 0.339, and 0.170 when we have survey data with geographic coordinates, census data at the county-level ($n = 47$), and a combination. In this simulation, there is a loss of accuracy when we only have census data, but it is not dramatic. However, when we aggregate at the provincial-level ($n = 8$) the MSE is 0.551 and when we additionally incorporate survey data the MSE is 0.201. In general, we see a modest improvement when incorporating the census data over just using survey data. The improvement is significantly better when we used county-level census data rather than provincial-level census data. Similar trends hold for the mean absolute errors.

3.5 Application to Scottish Lip Cancer Data

We use the Scotland lip cancer data as an illustrative example of how the method can be applied to areal data. A common model for spatial smoothing for such data is the Besag-York-Mollié (BYM) model (Besag et al., 1991). They propose a discrete spatial model where S_i, the spatial random effect in area i, is decomposed into two components, $S_i = U_i + V_i$. Here, U_i is assigned an intrinsic conditional autoregressive (ICAR) prior; see equation (2.1). The other component V_i allows for independent shocks in each area, $V_i \sim_{iid} \mathrm{N}(0, \tau_v^{-1})$. Unfortunately, this specification for the random effects depends on defining a somewhat arbitrary neighborhood structure. As an alternative, we consider spatial modeling via an underlying latent GRF. This may be viewed simply as a mechanism to induce spatial dependence between the areas, and then report the aggregate estimates only. More optimistically, one may report the continuously indexed surface, but this is an intrinsically dangerous endeavor.

Let R_i denote county i, $i = 1, \ldots, n = 56$ and let Y_{iaj} be the binary male lip cancer indicator in stratum (age-band) a of county i at location s_{ij}, $j = 1, \ldots, N_{ia}$ where N_{ia} is the male population in county R_i age group a. In the usual case, the available data correspond to summed disease counts $Y_{i++} = \sum_{a=1}^{A} \sum_{j=1}^{N_{ia}} Y_{iaj}$ and expected numbers E_i; these expected numbers are often pre-calculated as $E_i = \sum_{a=1}^{A} N_{ia} q_a$, where q_a is a reference risk for stratum a. The q_a may be taken from a previous time period or calculated (via internal standardization) in advance. The rarity of many diseases, and the lack of stratum-specific information, means that simplifying modeling assumptions are needed, as we now describe.

We proceed as in the no strata case and assume for a rare disease $Y_{iaj} \mid p_{iaj} \sim_{ind} \text{Poisson}(p_{iaj})$, for $j = 1, \ldots, N_{ia}$, individuals in strata a, county i, where $p_{iaj} = \exp\{\beta_0 + \beta_a + S_a(s_{ij})\} = q_a \exp\{\beta_0 + S_a(s_{ij})\}$, with $S_a(s_{ij})$ representing the spatial random effect for strata a at location s_{ij}. This leads to $Y_{i++}|\mu_i \sim \text{Poisson}(\mu_i)$, and, proceeding as before,

$$\begin{aligned}
\mu_i &= \sum_{a=1}^{A} \sum_{j=1}^{N_{ia}} \mathrm{E}_a\left[Y_{iaj}|s_{ij}\right] = \sum_{a=1}^{A} q_a \sum_{j=1}^{N_{ia}} \exp\{\beta_0 + S_a(s_{ij})\} \\
&\approx \sum_{a=1}^{A} N_{ia} q_a \exp(\beta_0) \sum_{m=1}^{m_i} d_a(s_{im}) \exp\{S_a(s_{im})\} \\
&= E_i \exp(\beta_0) \sum_{m=1}^{M_i} d(s_{im}) \exp\{S(s_{im})\} = E_i \theta_i
\end{aligned}$$

where the first equality on the last line follows from assuming a common residual spatial risk surface across stratum ($S_a(\boldsymbol{s}) = S(\boldsymbol{s})$) and common population density across stratum ($d_a(\boldsymbol{s}) = d(\boldsymbol{s})$). This allows us to separate the age-standardization from the risk surface estimation to give the data model. Standardization in this fashion leads to the spatial modeling of the *relative risk*, θ_i, an aggregate summary. The standardized incidence ratio (SIR) is $\text{SIR}_i = Y_i/E_i$ and is the MLE of θ_i from the Poisson model with mean $E_i\theta_i$. The SIRs are depicted in the top left hand panel of Figure 3.5. It is evident from the map that there is large variability in the area relative risks, with apparent strong spatial dependence.

Inference for this model proceeds as discussed in Section 3.3.3. The mesh used is shown

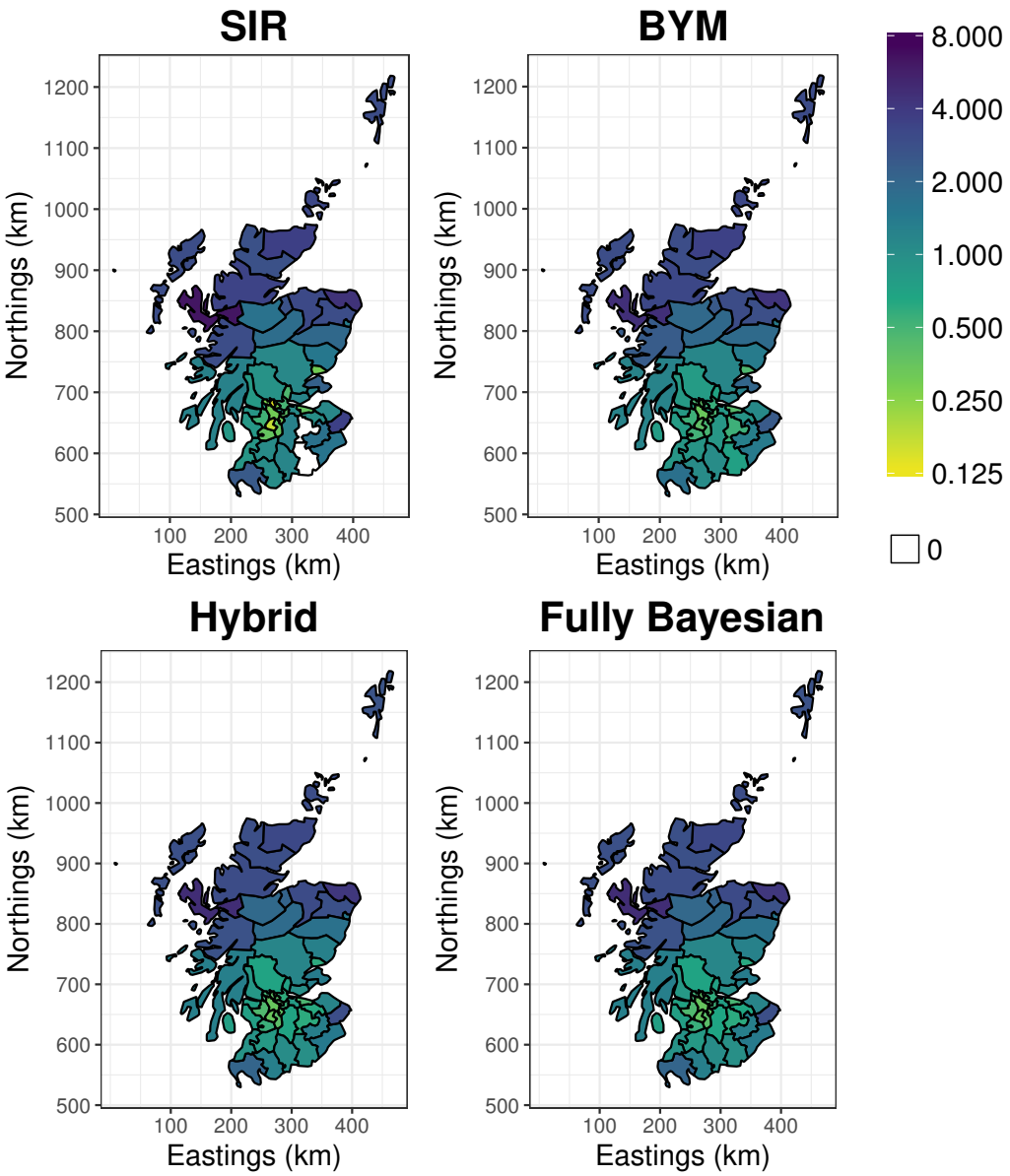

Figure 3.5: Top left: the SIR estimates of the relative risks of lip cancer in 56 counties of Scotland. Top right: relative risk estimates (posterior medians) from the BYM model. Bottom left: relative risk estimates (posterior medians) from aggregating results from the hybrid EB/MCMC approach. Bottom right: relative risk estimates (posterior medians) from aggregating results from the fully Bayesian approach.

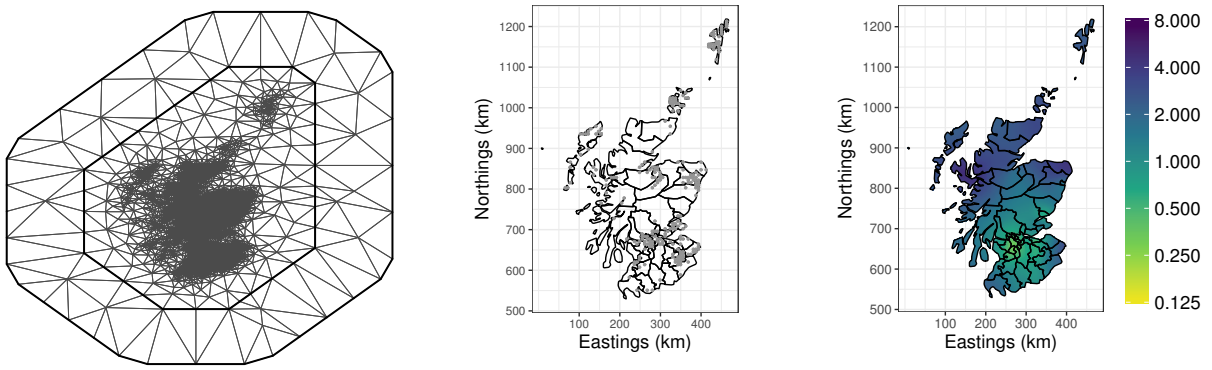

Figure 3.6: Left: mesh used for Scotland analysis, consisting of $M = 2,417$ mesh points. Middle: distribution of population in Scotland. The gray circles represent mesh points where the population density is larger than average for that area. Right: the predicted continuous relative risk surface from using the fully Bayesian approach.

in Figure 3.6 with $M = 2,417$ mesh points, which results in mesh points that are ≈ 6.3 km apart. We determined the relative population density for each R_i in the same manner as we did for the Kenya simulation and mesh points associated with higher relative population densities are shown in Figure 3.6.

3.5.1 Computational Details

We considered several different computational strategies. We implemented an EB approach in which the spatial random effects were integrated out using Laplace approximations, a fully Bayesian approach (using HMC), and a hybrid of these two in which HMC was used, with the spatial hyperparameters $\boldsymbol{\phi}$ fixed at the EB estimates. For the fully Bayesian approach, we initialized 4 chains, used a burn in of 10,000 iterations, ran an additional 1,000,000 iterations and thinned them to ultimately save 1,000 iterations from each chain. For the hybrid approach, we also initialized 4 chains, used a burn in of 500 iterations, and ran an additional 1,000 iterations for each chain.

Empirical Bayes

As fast alternatives to a completely Bayesian approach, we consider two strategies, both based on empirical Bayes (EB) estimation. In the first, we use EB estimation to obtain estimates for the spatial hyperparameters $\boldsymbol{\phi}$ and the fixed effect β_0. In the second, we use a hybrid approach where we first use EB estimation to estimate $\boldsymbol{\phi}$, and then proceed, conditional on these values.

For the first, strictly EB, approach, the EB estimates are defined as

$$\begin{aligned}\left(\hat{\boldsymbol{\phi}}^{\text{EB}},\ \hat{\beta}_0^{\text{EB}}\right) &= \text{argmax}_{\phi,\beta_0}\ p(\boldsymbol{\phi}, \beta_0|\boldsymbol{y}) \\ &= \text{argmax}_{\phi,\beta_0} \int_x p(\boldsymbol{\phi}, \boldsymbol{w}, \beta_0|\boldsymbol{y})\ d\boldsymbol{w}, \\ &= \text{argmax}_{\phi,\beta_0} \int_x f(\boldsymbol{y}|\beta_0, \boldsymbol{w}) p(\boldsymbol{w}|\boldsymbol{\phi}) \tilde{p}(\boldsymbol{\phi}) \tilde{p}(\beta_0)\ d\boldsymbol{w},\end{aligned}$$

where we use $\tilde{p}(\cdot)$ to denote a flat prior, that is $\tilde{p}(\cdot) \propto 1$. For the EB approach we use uninformative, flat priors for both the fixed effect β_0 and spatial parameters $\boldsymbol{\phi}$. Therefore, the EB estimates, the posterior modes, are also maximum likelihood estimates (MLEs). We then use the invariance of MLEs and the delta-method to obtain estimates for $\exp(\beta_0)$ and functions of the spatial hyperparameters.

For the second, "hybrid", approach, the EB estimates are defined as

$$\begin{aligned}\hat{\boldsymbol{\phi}}^{\text{Hybrid}} &= \text{argmax}_{\phi}\ p(\boldsymbol{\phi}|\boldsymbol{y}) \\ &= \text{argmax}_{\phi} \int_x \int_{\beta_0} p(\boldsymbol{\phi}, \boldsymbol{w}, \beta_0|\boldsymbol{y})\ d\beta_0\ d\boldsymbol{w}, \\ &= \text{argmax}_{\phi} \int_x \int_{\beta_0} f(\boldsymbol{y}|\beta_0, \boldsymbol{w}) p(\boldsymbol{w}|\boldsymbol{\phi}) \tilde{p}(\boldsymbol{\phi}) p(\beta_0)\ d\beta_0\ d\boldsymbol{w}.\end{aligned}$$

We then use these estimates in the second, MCMC-based, step, which is described in the following section. In the hybrid approach, we use a normal prior for the intercept, $\beta_0 \sim \text{N}(\mu_{\beta_0}, \sigma^2_{\beta_0})$ with $\mu_{\beta_0} = 0, \sigma^2_{\beta_0} = 100$, and place uninformative, flat priors on the spatial parameters, $\tilde{p}(\boldsymbol{\phi}) \propto 1$.

In both cases, to implement EB estimation we use the `R` package `TMB` (Kristensen, 2014;

Kristensen et al., 2016), which is computationally efficient since it uses sparse matrix operators. See Thorson et al. (2015) for a discussion on using `TMB` with the SPDE approach.

We briefly summarize how to implement this approach. We first construct a so-called template file that contains the joint distribution, which is the product of the likelihood $f(\boldsymbol{y}|\beta_0, \boldsymbol{w})$ and priors $p(\boldsymbol{w}|\boldsymbol{\phi})$, $\tilde{p}(\beta_0)$ or $p(\beta_0)$, and $\tilde{p}(\boldsymbol{\phi})$. To obtain the densities that we would like to optimize, $p(\boldsymbol{\phi}, \beta_0|\boldsymbol{y})$ and $p(\boldsymbol{\phi}|\boldsymbol{y})$, we use `TMB` to integrate out $\boldsymbol{w}$ and, optionally, β_0. This integration is carried out using Laplace approximations. We then numerically optimize the density using gradients to obtain the EB estimates denoted $\hat{\boldsymbol{\phi}}^{\text{EB}}$ and $\hat{\beta}_0^{\text{EB}}$ (or $\hat{\boldsymbol{\phi}}^{\text{Hybrid}}$ for the hybrid approach) and the associated variance-covariance matrix (based on the Hessian), denoted $\hat{\boldsymbol{\Sigma}}^{\text{EB}}$ (for the strictly EB approach).

MCMC

For the fully Bayesian approach we use as priors, $\beta_0 \sim \text{N}(\mu_{\beta_0}, \sigma^2_{\beta_0})$ and $\boldsymbol{\phi} \sim \text{N}\left(\boldsymbol{\mu}_\phi, \boldsymbol{\Sigma}_\phi\right)$, with $\mu_{\beta_0} = 0, \sigma^2_{\beta_0} = 100$,

$$\boldsymbol{\mu}_\phi = \begin{bmatrix} 3.24 \\ -4.51 \end{bmatrix}, \qquad \boldsymbol{\Sigma}_\phi = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}.$$

As in the Kenya simulation example, the priors for the hyperparameters $\boldsymbol{\phi}$ are vague and $\boldsymbol{\mu}_\phi$ corresponds to a marginal variance λ^2 of 1 and practical range ρ of roughly 20% of the domain size.

In the fully Bayesian approach, we first begin by updating $\boldsymbol{\phi}$ conditional on $\boldsymbol{w}$, β_0, and $\boldsymbol{y}$ using a random-walk proposal. The proposal distribution we use is,

$$\boldsymbol{\phi}^{(t+1)} \sim \text{N}(\boldsymbol{\phi}^{(t)}, c \times \hat{\boldsymbol{\Sigma}}_\phi^{\text{EB}}),$$

where $\hat{\boldsymbol{\Sigma}}_\phi^{\text{EB}}$ is the inverse Hessian corresponding to the estimates for $\boldsymbol{\phi}$ obtained from EB estimation described in the preceding section.

The second step is then similar for both the hybrid and fully Bayesian approach. We update $\boldsymbol{w}$ and β_0 conditional on $\boldsymbol{\phi}$ and $\boldsymbol{y}$, by using Hamiltonian Monte Carlo (HMC; Neal, 2011). In the hybrid approach, $\boldsymbol{\phi}$ is taken to be $\hat{\boldsymbol{\phi}}^{\text{Hybrid}}$. The negative log posterior U

(modulo a constant term), is found to be

$$U = -\beta_0 \boldsymbol{y}^\top \mathbf{1}_n - \boldsymbol{y}^\top \log(\boldsymbol{DT}) + \exp(\beta_0)\boldsymbol{E}^\top \boldsymbol{DT} + \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{Q}\boldsymbol{w} + \frac{1}{2\sigma_{\beta_0}^2}\beta_0^2,$$

where $\mathbf{1}_n$ is an $n \times 1$ vector of all ones, $\boldsymbol{D}$ is an $n \times M$ matrix where each row $\boldsymbol{D}_i^\top$ contains up to m_i nonzero weights $d(\boldsymbol{x}_{ik})$, $\boldsymbol{T} = [\ \exp(x_1), \ldots, \exp(x_M)\ ]^\top$, and $\boldsymbol{E} = [\ E_1, \ldots, E_n\ ]^\top$. We also compute the derivatives to be,

$$\begin{aligned}
\frac{\partial U}{\partial \beta_0} &= -\boldsymbol{y}^\top \mathbf{1}_n + \exp(\beta_0)\boldsymbol{E}^\top \boldsymbol{DT} + \frac{\beta_0}{\sigma_{\beta_0}^2}, \\
\frac{\partial U}{\partial \boldsymbol{w}} &= (\boldsymbol{D}\ \mathrm{diag}(\boldsymbol{T}))^\top \left[\exp(\beta_0)\boldsymbol{E} - (\mathrm{diag}(\boldsymbol{DT}))^{-1}\,\boldsymbol{y}\right] + \boldsymbol{w}^\top \boldsymbol{Q},
\end{aligned}$$

where $\mathrm{diag}(\boldsymbol{T})$ is a diagonal matrix with entries $T_1, \ldots, T_M$ along the diagonal. Parameters that are tuned for desired acceptance are c, the step size, and number of leapfrog steps for each HMC iteration.

In exploratory runs, it appeared that computation could be improved by defining $\boldsymbol{w}^* = \boldsymbol{w} + \beta_0$, in which case, $\boldsymbol{w}^* \mid \beta_0, \boldsymbol{\phi} \sim N(\beta_0, \mathbf{Q}^{-1})$. Under this parameterization,

$$\begin{aligned}
U &= -\boldsymbol{y}^\top \log(\boldsymbol{DT}^*) + \boldsymbol{E}^\top \boldsymbol{DT}^* + \frac{1}{2}(\boldsymbol{w}^* - \beta_0 \mathbf{1}_M)^\top \boldsymbol{Q}(\boldsymbol{w}^* - \beta_0 \mathbf{1}_M) + \frac{1}{2\sigma_{\beta_0}^2}\beta_0^2, \\
\frac{\partial U}{\partial \beta_0} &= \beta_0 \mathbf{1}_M^\top \boldsymbol{Q}\mathbf{1}_M - \mathbf{1}_M^\top \boldsymbol{Q}\boldsymbol{w}^* + \frac{\beta_0}{\sigma_{\beta_0}^2}, \\
\frac{\partial U}{\partial \boldsymbol{w}^*} &= (\boldsymbol{D}\ \mathrm{diag}(\boldsymbol{T}^*))^\top \left[\boldsymbol{E} - (\mathrm{diag}(\boldsymbol{DT}^*))^{-1}\,\boldsymbol{y}\right] + \boldsymbol{w}^{*\mathrm{T}} \boldsymbol{Q} - \beta_0 \boldsymbol{Q}\mathbf{1}_M,
\end{aligned}$$

where $\boldsymbol{T}^* = [\ \exp(x_1^*), \ldots, \exp(x_M^*)\ ]^\top$. This alternative parameterization is implemented for the hybrid approach.

Further gains in speed can be found by specifying a better scaling of the "mass matrix" (the covariance of the momentum variables used in the HMC algorithm). In the most simple case, the mass matrix is chosen to be the identity matrix, which corresponds to i.i.d. momentum variables. For the hybrid approach, we adapt this matrix to again be diagonal, but with entries along the diagonal corresponding to the inverse posterior variance. We empirically estimate this after running several hundred iterations of the algorithm using an identity mass matrix.

Parameter	Empirical Bayes	Hybrid	Fully Bayesian
$\exp(\beta_0)$	1.99 (1.35, 2.94)	2.03 (1.37, 3.04)	2.03 (1.36, 3.25)
ρ	71.9 (35.2, 147)	80.3 (39.0, 166)	88.6 (45.4, 232)
λ^2	0.516 (0.279, 0.952)	0.533 (0.283, 1.01)	0.580 (0.304, 1.31)

Table 3.2: Estimates and 95% credible intervals (CIs) for parameters in the Scotland example under the three different computational approaches. For the empirical Bayes approach, estimates are based on transformed MLEs and CIs are based on the delta-method. For the hybrid approach, estimates are based on transformed MLEs for the spatial parameters and the posterior median and 2.5th and 97.5th percentile are presented for the intercept. For the fully Bayesian approach, estimates reported are posterior medians and CIs are based on the 2.5th and 97.5th percentiles.

Trace plots and histograms for the fully Bayesian and hybrid approaches are shown in Appendix A.1. Trace plots and calculated $\hat{R}$ (which were all less than 1.05) suggested convergence (Gelman and Rubin, 1992).

ICAR Model

We follow Fong et al. (2010) and use the following priors, $p(\tau_u) \sim \text{Gamma}(1, 0.2/0.59)$, $p(\tau_v) \sim \text{Gamma}(1, 0.14)$.

3.5.2 Results

Estimates and 95% CIs for the parameters $\exp(\beta_0)$, ϕ, and λ^2 are presented in Table 3.2 . There is good agreement across the three approaches, though we notice that the posterior credible intervals tend to be wider when using the fully Bayesian computational strategy, which is not surprising given the use of the delta-method to calculate the CIs in the EB approach.

We also obtain predictions and posterior standard deviations of $\tilde{S}(\boldsymbol{x})$, displayed in Figure 3.7. We note that the posterior standard deviation of the surface is smallest in regions of Scotland where the population is greatest, and larger elsewhere. Furthermore, the posterior

standard deviation tends to be a little lower for the hybrid approach than for the fully Bayesian approach, which is not surprising given that the spatial hyperparameters were fixed in the hybrid approach. The predicted continuous relative risk surface using the fully Bayesian approach (posterior median) is presented in Figure 3.6. As expected, we see that the continuous relative risk surface is largest in the counties with higher SIRs, and lowest in the counties with the smallest SIRs.

We obtain relative risk estimates (posterior medians), as well as 95% CIs for each of the 56 counties from this model by aggregating (with respect to the population density) the continuous relative risk surface within each county. To obtain estimates of the desired quantiles in both the fully Bayesian and hybrid approach, for each county R_i we obtain $b = 1, \ldots, B = 4,000$ draws from the "aggregated" relative risk surface, $\theta_i^{(b)} = \exp\left(\beta_0^{(b)}\right) \boldsymbol{D}_i^\top \boldsymbol{T}^{(b)}$, where $\boldsymbol{T}^{(b)} = \left[\exp\left(x_1^{(b)}\right), \ldots, \exp\left(x_M^{(b)}\right) \right]^\top$ (see (3.4)). As before, $\boldsymbol{D}_i$ has at most m_i nonzero entries that correspond to the population density estimates, $d(s_{im})$. From here, we can obtain the desired summary measures. Posterior medians are presented in Figure 3.5 and the 95% CIs are displayed in Figure 3.8. We see that the relative risk estimates are nearly identical for both computational strategies and are similar to the SIRs, but that the estimates are shrunk towards the overall mean, which is as expected.

We also compare our results to those obtained using the BYM model. We obtained a posterior median for $\exp(\beta_0)$ of 1.09 (95% CI: 0.962, 1.24), τ_u of 3.01 (95% CI: 1.56, 5.81), and τ_v of 15.5 (95% CI: 5.56, 40.6). The predicted relative risks (posterior medians) for each county are presented in Figure 3.5, and the 95% CIs are presented in Figure 3.8 The results are very similar to the continuous model.

For the fully Bayesian approach using HMC (using our own code), it took approximately a week to fit the model using a computing cluster. This can be improved tremendously by using the hybrid approach. It takes on the order of minutes to obtain the EB estimates and about 10 minutes to run the HMC.

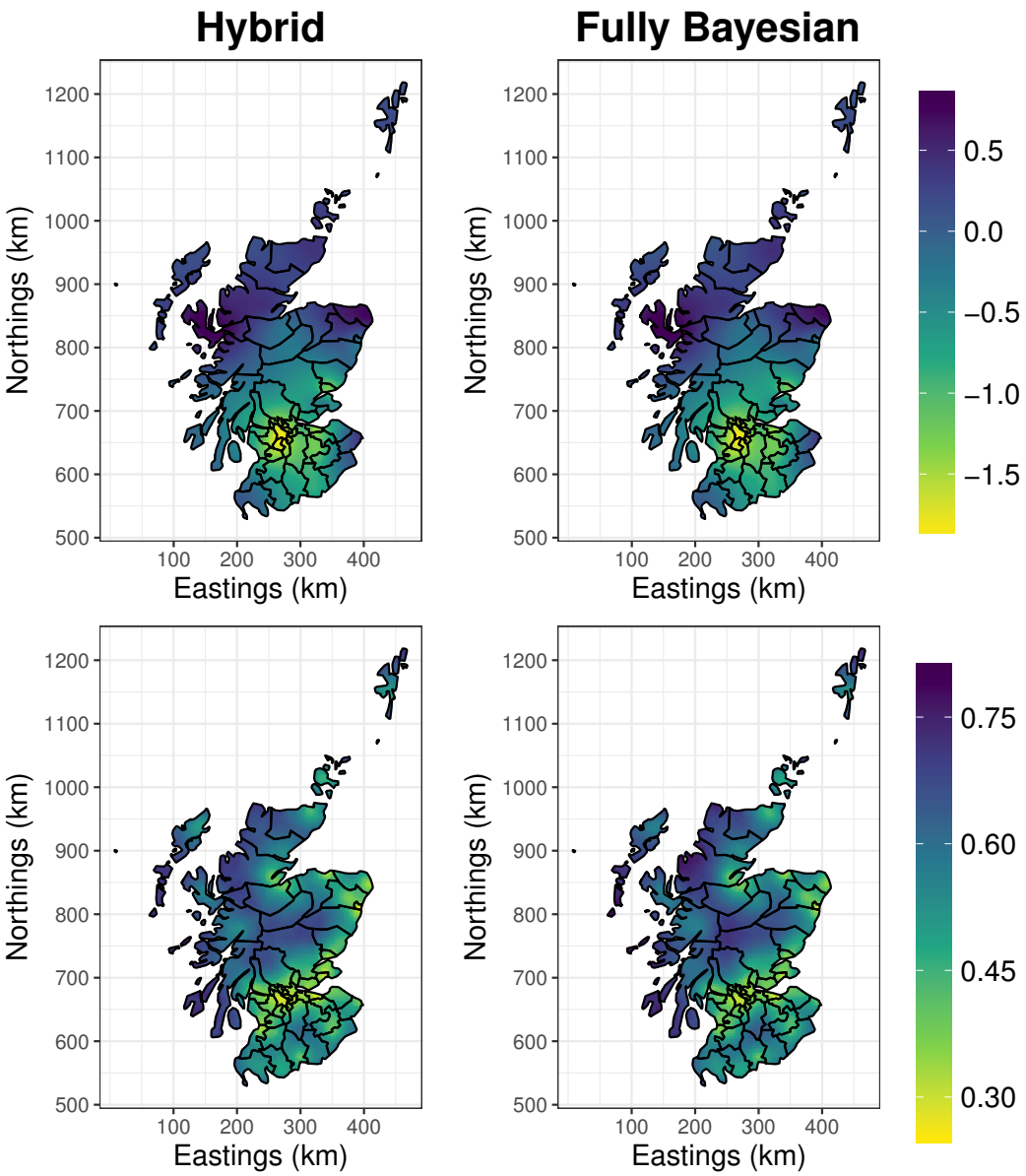

Figure 3.7: Top: predicted (posterior mean) spatial surface. Bottom: posterior standard deviation of the spatial surface. Left column are results from the hybrid approach and right column are results from the fully Bayesian approach.

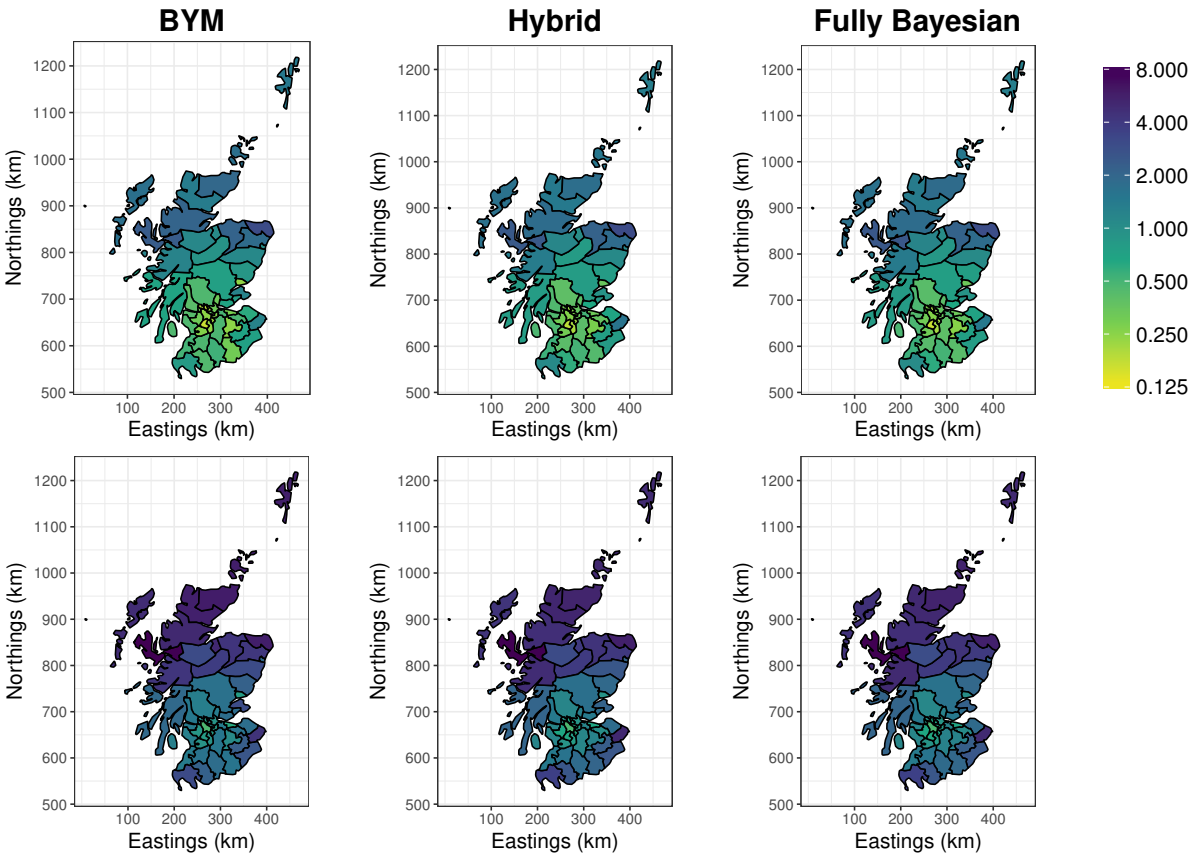

Figure 3.8: Comparison of the 95% CI of relative risk under different models for the spatial surface and computational strategies. Left column are estimates from the BYM model, middle column are estimates from using the hybrid approach, and right column are estimates from the fully Bayesian approach. Top row is 2.5th percentile, and bottom row is 97.5th percentile.

3.6 Discussion

In this chapter, we propose a Bayesian hierarchical model that can accommodate observations taken at different spatial resolutions. To this end, we assume a continuous spatial surface, which we model using the SPDE approach.

In the simulation example, we considered surveys taken at point locations and census data associated with areas. When the only data available was census data at the county-level, there was not a substantial loss in accuracy when comparing it to a situation in which we had survey (point) data. In general, we would not expect this to hold when comparing point data to areal data as the loss of information depends on the strength of the spatial dependence, the number and geographical configuration of the areas, and on the amount and quality of the survey data. When the size of the areas increased (comparing 47 counties to 8 provinces), estimates for the spatial parameters and the overall household wealth index were highly variable and the predicted spatial surface was much less nuanced.

With point data we are not able to learn anything about the surface at a spatial resolution which is less than the distance between the two closest points. If we only have areal data, the situation is far worse. Hence, one should not over-interpret fine spatial scale structure. In both the point and areal data cases, model checking is difficult. For areal data, one may simply view the model as a method to induce an area-level spatial prior. The results can be presented at the area-level, as we did with the Scottish lip cancer example; with area-level data only, we can only check the model at the aggregate level.

We also applied our method to the Scotland lip cancer dataset. Overall, there were very minor differences in the relative risk estimates for each county when comparing our continuous spatial model to a discrete spatial model (i.e., the BYM model) but again there is strong spatial dependence in these data (which explains in part the popularity of these data). However, we note that modeling a continuous surface is particularly attractive in that it is not subject to definitions of administrative boundaries, which can often be arbitrary, and a continuous risk surface, in general, more accurately reflects disease etiology. Furthermore,

it can easily be adapted to situations where we might also have point level covariate data. In the latter case, the use of the models we have described avoids the ecological fallacy which occurs when area-level associations differ from the individual-level counterparts. To avoid ecological bias, one requires point level covariates but the availability of such data is increasing (Gething et al., 2015). If the administrative level is an important component (rather than solely an arbitrary construct), a random effect for the administrative level or administrative level covariates could be included in the model. One important consideration here is that placing a spatial structure on administrative level random effects could result in identifiability issues when there is also a spatially continuous random effect in the model. Thus, the priors would need to be carefully defined. This is likely less of a concern if the administrative level random effects are unstructured (i.e., iid error).

For normal outcomes, all computation can be performed quickly using `R-INLA`. In the simulation example, it took about 2 minutes to fit each of the models on a standard laptop. For Poisson outcomes, computation is much more difficult. There has been an increased interest in implementing sparse matrix operations in `Stan`, which would improve usability of this method. In general, there is still a need for computationally efficient MCMC schemes for Gaussian process data.

In the simulation study we considered, we looked at combining survey (point) data and census (areal) data. However, with older DHS surveys exact geographic coordinates are not available. Instead, it is only known in which area of the country the survey was taken. This is different from the problem we considered in that, instead of observing outcomes that are associated with an entire area (census data), outcomes are for a specific point in the area, but that exact location is unknown. Therefore, the methods proposed here would need to be altered to accommodate this type of situation.

All code and input data used in the simulation and application is available on github (`https://github.com/wilsonka/pointless-spatial-modeling`).

Chapter 4

INCORPORATING POINT DATA WITH INCOMPLETE GEOGRAPHIC LOCATIONS

4.1 Introduction

Effective implementation of health programs requires information on where unmet need exists within countries. In many low and middle income countries (LMIC), health data can come from a variety of sources, and in many cases the sources are an incomplete representation of all of the country's inhabitants. Surveys are a primary tool for obtaining vital information. One example of surveys that are commonly used, especially in developing countries, are the Demographic and Health Surveys (DHS). Typically, the DHS employs stratified, cluster sampling, where the clusters represent a small area of the country. Within clusters, households are randomly selected. They are designed to provide reliable estimates at a prespecified, and usually geogaphically large, administrative level. However, policymakers and researchers are often interested in modeling and understanding health indicators at lower levels, e.g., at the district or on a 10km x 10km grid level.

To protect respondent confidentiality, survey data from households within the same cluster are aggregated to a single point, the centroid of the cluster. This cluster location information can be used for spatial modeling. However, the geographic identifiers available in the DHS can vary and typically the precise coordinates of the cluster centers are not publicly available (Burgert et al., 2013). Recently, the geographic locations of the clusters (i.e., the centroids) are provided, but they are displaced up to 10km. For example in Kenya displaced GPS data are available for DHS completed on or after 2003. In older DHS and other surveys such as the Multiple Indicator Clusters Surveys (MICS) only the larger administrative area within which the cluster resides is reported. We will refer to this procedure as "masking".

In the literature this is sometimes referred to "aggregation", though not to be confused with the aggregation procedure that was the focus of Chapter 3. In that chapter, we considered censuses, which provided outcomes that are aggregated over an entire administrative area. Here, we consider point data, but where the geographic location of the point is assigned to the administrative area within which the point belongs. For the purposes of this chapter, we will ignore issues involving the first step of aggregating household data to a single point, though an approach similar to that from Chapter 3 could be used. Instead, we focus on issues surrounding displacement and masking of the cluster locations.

First, we consider the displacement scenario. A naive analysis would ignore the jittering of the cluster centroids and fit a continuous spatial model using the displaced location. The effect of doing this in spatial analyses has been studied by Gething et al. (2015), Chapter 4. Using real data, they simulated 100 datasets with jittered location information and assessed the impact on analyses involving several indicators of interest. They investigated the effect of the displacement on spatial correlation, spatial covariate associations, and model derived surfaces. In their example, using empirical variograms, they found that there was not a large impact on spatial correlation. However, they did find some differences in the relationship between spatial covariates and the outcomes; models naively using the spatial covariate value at the displaced value tended to have lower R^2 although this was not always the case. They also observed some inaccuracies in predicted surfaces when using displaced data, and these differences tended to be exacerbated when the spatial covariate changed quickly in space. Based on work by Perez-Heydrick et al. (2013), the DHS have proposed guidelines that when using spatial covariate raster data, the average value of the covariate in cells within a specified buffer (10km for rural clusters and 2km for urban clusters) of the reported cluster location should be used in analyses.

Having location information subject to positional error can be thought of as an error-in-variables problem. In particular, let $\{s_1, \ldots, s_n\}$ denote the set of true, unobserved, covariates (locations) that give rise to the set of observed outcomes $\{y_1, \ldots, y_n\}$. Denote the measured (reported) covariates (locations) associated with the outcomes as $\{u_1, \ldots, u_n\}$.

The Berkson measurement error model would be, $s_i = u_i + \epsilon_i^\star$ where $\epsilon_i^\star$ is an error term, whereas the classical measurement error model would be, $u_i = s_i + \epsilon_i$ where ϵ_i is an error term (Carroll et al., 2006). The Berkson model is particular appealing when there are a set of desired locations that outcomes should be collected at, but the actual location has been perturbed. This could be a result of using an imprecise positional instrument. On the other hand, the classical measurement error model would arise when outcomes are collected at a particular location, but the reported location has been perturbed. Many proposed approaches seeking to address the positional error issue appear to focus on normally-distributed outcomes under a Berkson measurement error model (Gabrosek and Cressie, 2002; Cressie and Kornak, 2003; Fanshawe and Diggle, 2011). To overcome the computationally expensive Monte Carlo integration from earlier work using a Berkson measurement error model, Fronterrè et al. (2018) use an approximate composite likelihood for inference. In their application to DHS data from Senegal, the displacement mechanism is approximated and the stratified sampling nature of the data is ignored. In this chapter, we develop a method under the classical measurement error model and the outcome distribution can be non-normal.

Now, we turn to the issue of masking. To incorporate masked data reported at the administrative area, one solution is to use a discrete spatial model, such as the ICAR model. Using this type of model would not allow for higher spatial resolution maps than the broadest administrative level available, and more recent surveys provide (jittered) geographic coordinates. Further issues could arise if the divisions of regions change over time. Additionally, these boundaries are often arbitrary and using a discrete model can be difficult to interpret if regions differ substantially in size and shape (see Chapter 3). Golding et al. (2017) fit a continuous spatial model and develop an approximate strategy for including data associated with areas in the context of modeling the under-five mortality rate, and do not seem to distinguish between aggregate and point-level data with missing coordinates. To do this, they randomly generated points in an area according to the population density. Points nearby were grouped together to form "pseudo-clusters" and assigned a weight based on the population that each "pseudo-cluster" represented. These weights then essentially partition the

observed data to each of the "pseudo-clusters."

The organization of this chapter is as follows. In Sections 4.2 and 4.3, we make explicit the problem and propose a solution that can accommodate masked (only administrative area available) or displaced (jittered coordinates) data. In Sections 4.4 and 4.5, we conduct a simulation study to assess the impact of each of these problems on spatial modeling. We also consider disclosure risk, which, in this case, refers to the ability to identify the true cluster location from the reported cluster location. Finally, we conclude with a discussion in Section 4.6, which includes information on future work.

4.2 Method

For the purposes of this chapter, we will suppose that the cluster data is associated with a true point location, namely the centroid of the cluster. Let $i = 1, \ldots, I$ index the administrative areas, $j = 1, \ldots, J$ index the strata (typically $J = 2$ for urban/rural), and $k = 1, \ldots, K_{ij}$ index the clusters within administrative area i and strata j. Consider cluster k in strata j and administrative area D_i, denote the true cluster centroid location by s_{ijk} and available location information by u_{ijk}. Suppose the set of all possible cluster locations (i.e., the sampling frame) is known and denote the set of the potential locations in area i, strata j by $E_{ij} = \{E_{ije}\}$, $e = 1, \ldots, m_{ij}$.

In the *masking scenario*, only the area in which the cluster is located is reported, which we will denote by $u_{ijk} = \{s_{ijk} \in E_{ij}\}$. That is, the prior on the location is

$$p(s_{ijk} = E_{ije} | u_{ijk}) = d_{ije}, \quad e = 1, \ldots, m_{ij}, \tag{4.1}$$

where d_{ije} is the probability potential location E_{ije} was selected. If proportional to size sampling was undertaken, then $d_{ije} \propto N_{ije}$ where N_{ije} is the population size of enumeration area E_{ije}. If random sampling was undertaken, then $d_{ije} \propto 1$.

In the *displacement scenario*, a jittered version of the true location is reported, which we will denote by $u_{ijk} = s_{ijk} + \epsilon_{ijk}$ where ϵ_{ijk} is the jittering probability density function. Specifically, we consider the DHS jittering algorithm in which the true location is randomly dis-

placed according to the distribution (in polar coordinates), $p(r,\theta) = (2\pi R)^{-1} I_{0<r<R} I_{0<\theta<2\pi}$ where $R = 2$km for urban clusters and $R = 5$km for 99% of rural clusters and $R = 10$km for the remaining 1% of rural clusters. That is,

$$p(s_{ijk} = E_{ije}|u_{ijk}) \propto p(u_{ijk}|s_{ijk} = E_{ije}) \times p(s_{ijk} = E_{ije}), \quad e = 1, \ldots, m_{ij},$$

where $p(s_{ijk})$ corresponds to (4.1). To derive the first term, we will need to marginalize over possible values of R. First note that for a given R,

$$p(u_{ijk}|s_{ijk} = E_{ije}, R) = [2\pi R d(u_{ijk}, E_{ije})]^{-1} C_{ije,R} I_{0<d(u_{ijk},E_{ije})<R}$$

where $d(u_{ijk}, E_{ije}) = [(E_{ije1} - u_{ijk1})^2 + (E_{ije2} - u_{ijk2})^2]^{1/2}$ is the distance between the possible location $E_{ije} = (E_{ije1}, E_{ije2})$ and the reported location $u_{ijk} = (u_{ijk1}, u_{ijk2})$ and

$$C_{ije,R} = \left[\int_{u \in \text{Area } i} [2\pi R d(u, E_{ije})]^{-1} I_{0<d(u,E_{ije})<R} du \right]^{-1} \tag{4.2}$$

is the normalizing constant that accounts for the jittered point being restricted to stay within admin area D_i. Therefore,

$$p(s_{ijk} = E_{ije}|u_{ijk}) \propto \begin{cases} d_{ije}[4\pi d(u_{ijk}, E_{ije})]^{-1} C_{ije,R=2} I_{0<d(u_{ijk},E_{ije})<2km} & \text{if urban} \\ d_{ije}\{0.99 \times [10\pi d(u_{ijk}, E_{ije})]^{-1} C_{ije,R=5} I_{0<d(u_{ijk},E_{ije})<5km} & \\ +0.01 \times [20\pi d(u_{ijk}, E_{ije})]^{-1} C_{ije,R=10} I_{0<d(u_{ijk},E_{ije})<10km}\} & \text{if rural} \end{cases} \tag{4.3}$$

Denote the outcome data measured at each cluster as $\boldsymbol{y}$, the available location information as , the true (unobserved) location information as $\boldsymbol{s}$, non-spatial covariates as $\boldsymbol{X}$, spatial covariates as $\boldsymbol{Z}$ where $\boldsymbol{z}_{ijk} = \boldsymbol{z}(s_{ijk})$ is the vector of spatial covariates associated with the true location of cluster k in strata j and administrative area i. Following Chapter 3, we use the SPDE approach. Thus, we assume there is a latent GRF with a Matérn covariance function, $\{S(s) : s \in D \subset \mathbb{R}^2\}$ where D is the region of interest. To use the approximation, a mesh is first created. For the simulation we considered, the mesh consisted of $M = 2,765$ mesh points and is shown in Figure 4.1. Let $\boldsymbol{\phi}$ represent the parameters of the Gaussian

Process model, $\boldsymbol{w}$ be a vector of the weights (i.e., the value of the spatial random effect at the mesh points), and $\boldsymbol{\psi}$ be a vector of basis functions (see equation (2.3)). Let $\boldsymbol{\beta}$ be a vector of the fixed effects, and define $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{w})$. We could proceed using a data augmentation (DA) algorithm as described in Chapter 2, based on the factorizations:

$$p(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{y}, \boldsymbol{s},) \propto p(\boldsymbol{y}|\boldsymbol{s}, \boldsymbol{\theta}, \boldsymbol{\phi}) \times p(\boldsymbol{\theta}, \boldsymbol{\phi}), \tag{4.4}$$

$$p(\boldsymbol{s}|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\phi},) \propto p(\boldsymbol{y}|\boldsymbol{s}, \boldsymbol{\theta}, \boldsymbol{\phi}) \times p(\boldsymbol{s}|). \tag{4.5}$$

4.3 INLA within MCMC

For inference, we propose using an approximate Gibbs sampling strategy, known as "INLA within MCMC" (Gómez-Rubio and Rue, 2017; Gómez-Rubio and Palmí-Perales, 2017). The motivation for doing this is that the model cannot be fit with INLA given that it is a mixture distribution. Therefore, one could use MCMC for inference (Marin et al., 2005). However, such algorithms can be inefficient for Gaussian processes as seen in Chapter 3 and by Filippone et al. (2013). The key for our implementation is to note that if the locations of the clusters are fixed, the conditional models can be fit using INLA.

Bivand et al. (2014, 2015) recognized that some models could be fit in `R-INLA` once certain parameters in the model were fixed. Specifically, they defined a grid of values for the "problem parameters" and used `R-INLA` to fit a conditional model. The reported marginal likelihood from `R-INLA` is then used to obtain the posterior distributions of these "problem parameters". Lastly, Bayesian modeling averaging (Hoeting et al., 1998) is used to derive the posterior distribution for the other parameters.

In the setting that we consider, it is not straightforward to derive a grid of values with high posterior probability for the "problem parameters" (in our case, unknown locations). A variation on the approach of Bivand et al. (2014, 2015) is to use a Metropolis-Hastings algorithm for the "problem parameters". Gómez-Rubio and Rue (2017); Gómez-Rubio and Palmí-Perales (2017) propose doing just this. They use the marginal likelihood from fitting conditional models in `R-INLA` to determine the acceptance probability for the Metropolis-

Hastings step. They note that the marginal likelihood reported by `R-INLA` is an estimate, and the limiting distribution is not exactly the desired stationary distribution. For their purposes, they argue and show that the difference is not significant.

We take this approach one step further and propose a new algorithm using `R-INLA` to fit (4.4) and generate a sample for $(\boldsymbol{\theta}, \boldsymbol{\phi})$ based on the posterior (INLA) approximation. This sample is then used to generate a sample for $\boldsymbol{s}$. This avoids needing a proposal distribution, as the posterior conditional distribution is available exactly. Therefore, the "INLA within MCMC" algorithm can be summarized as:

1. Initialize $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\beta}^{(0)}, \boldsymbol{w}^{(0)})$.
2. Iterate:

   (a) Sample $s_{ijk}^{(t+1)}$ using Gibbs sampling,

   $$p(s_{ijk} = E_{ije}|y_{ijk}, u_{ijk}, \boldsymbol{\theta}^{(t)}) \propto p(s_{ijk} = E_{ije}|u_{ijk}) \times p(y_{ijk}|s_{ijk} = E_{ije}, \boldsymbol{\theta}^{(t)}) \quad (4.6)$$

   where the first term on the right corresponds to (4.1) for the masking scenario and (4.3) for the displacement scenario. The second term on the right corresponds to the complete data likelihood.

   (b) Use INLA to obtain the approximate conditional posterior, denoted $\tilde{p}(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{y}, \boldsymbol{s}^{(t+1)})$. Sample $\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\phi}^{(t+1)}$ from the approximate posterior.

We note that the implementation of `R-INLA` means that the hyperparameters, ϕ, are defined on a grid meaning that when a joint sample is drawn from the approximation in step (b), the hyperparameters can only fall on the grid. By default, a new grid for the hyperparameters is constructed during each iteration (since the cluster locations changes). This grid could be fixed ahead of time if this is desired; however, we allow the grid to change from one iteration to the other in our examples.

How accurate the results are relies on the accuracy of INLA and the joint posterior sampling algorithm used in `R-INLA`.

	Rural	Urban
Central	7,816	4,192
Coast	4,268	3,569
Eastern	12,396	3,234
Nairobi	0	10,394
North Eastern	2,230	433
Nyanza	9,787	3,041
Rift Valley	19,097	6,051
Western	7,383	1,419

Table 4.1: Number of potential clusters in each administrative area and strata.

4.4 Simulation Setup

To investigate the impact of masking and displacement of cluster centroids, a masterframe of all sampling locations approximately representing the true masterframe from the 2009 Kenya census was created based on population density retrieved from WorldPop (2016). This was done by first dividing the gridded population density into two zones: urban and rural within each county. To identify these zones, thresholding was used so that the proportion exceeding the threshold amount matched the proportion urban in the 2014 Kenya DHS (Kenya National Bureau of Statistics, 2015). The 1km by 1km grids that exceeded the threshold were labeled as urban and otherwise labeled rural. The masterframe of all sampling locations was then created by randomly drawing coordinates proportional to population density within each strata (urban/rural crossed with county) to obtain 95,310 enumeration areas; see Figure 4.1 for locations and Table 4.1 for counts by province and strata. Finally, 398 clusters (right panel in Figure 4.1) were then randomly sampled (uniformly, not proportional to size), stratified by province and urban/rural. The number of clusters within each sampling strata were chosen to match the 2008 DHS.

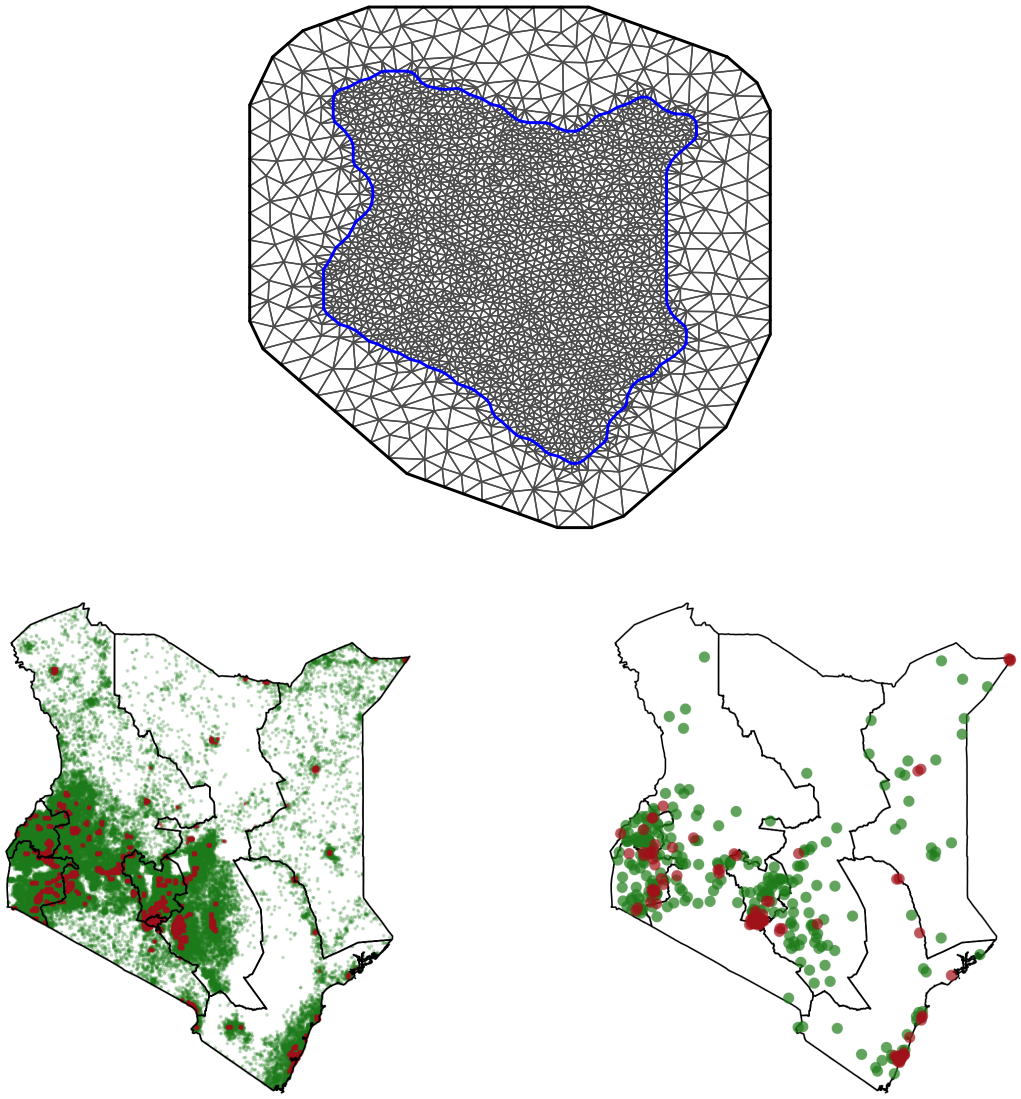

Figure 4.1: Top: Mesh over the geography of Kenya. Bottom left: Kenya provinces with locations of centroids. Bottom right: Kenya provinces with true locations of the 398 clusters. Red: urban. Green: rural.

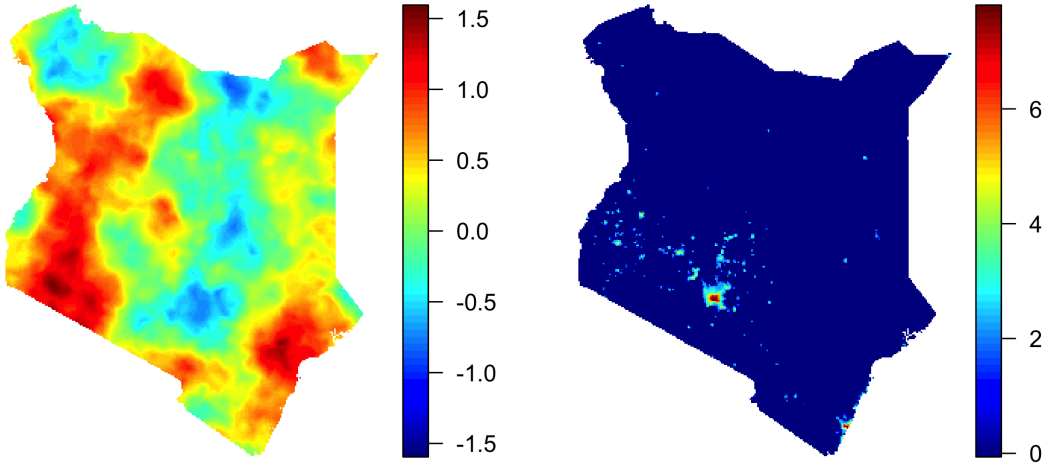

Figure 4.2: Left: Spatial surface $\tilde{S}(\cdot)$ used in the simulation. Right: square root of nighttime lights surface.

4.4.1 Model

Data is generated via the following model,

$$Y_{ijk}|p(s_{ijk}) \sim \text{Binomial}(25, p(s_{ijk}))$$
$$\text{logit}(p(s_{ijk})) = \beta_0 + \beta_1 \boldsymbol{z}_{ijk} + \tilde{S}(s_{ijk})$$

where $\tilde{S}(\cdot)$ is the SPDE approximation to the Gaussian process spatial random effect surface; see (2.3) in Chapter 2. In our simulation, we used the square-root of nighttime lights (NOAA nighttime lights series) as the spatial covariate, $\boldsymbol{z}$. The spatial surface $\tilde{S}(\cdot)$ and nighttime lights surface are plotted in Figure 4.2.

4.4.2 Scenarios

We will consider several simulation scenarios where centroid locations are jittered or masked, described in Table 4.2. Row 1a corresponds to fitting the "gold standard" model, which is if all cluster centroids are available exactly. Row 2a corresponds to fitting the model using

		Centroid Locations	Spatial Covariate
1a	INLA	100% exact	
2a	INLA naive	100% jittered	
3a	INLA within MCMC	100% jittered	
4a	INLA	50% exact	
5a	INLA	50% exact, 50% at centroids	
6a	INLA within MCMC	50% exact, 50% masked	
1b	INLA	100% exact	✓
2b	INLA naive	100% jittered	✓
3b	INLA within MCMC	100% jittered	✓
4b	INLA	50% exact	✓
5b	INLA	50% exact, 50% at centroids	✓
6b	INLA within MCMC	50% exact, 50% masked	✓

Table 4.2: Simulation scenarios considered

the jittered locations of the centroids. Figure 4.3 shows the true and jittered locations for clusters in the Western province. Row 3a corresponds to using our proposed approach to accommodate the jittered nature of the locations. Rows 4a–6a refer to a masking scenario where we will mask 50% of the centroids (that is, only the strata and administrative area are known); Figure 4.4 shows the location of the clusters where the true centroids are known and only the administrative area and strata are known. Row 4a corresponds to fitting the model only to the data where the cluster information is known exactly. Row 5a corresponds to also incorporating the data from the masked cluster locations. Here, we use the centroid of all potential cluster locations. Row 6a corresponds to using our proposed approach. We repeat these scenarios when a covariate is included (rows 1b–6b). In each case, we investigate the effect on surface reconstruction and covariate associations.

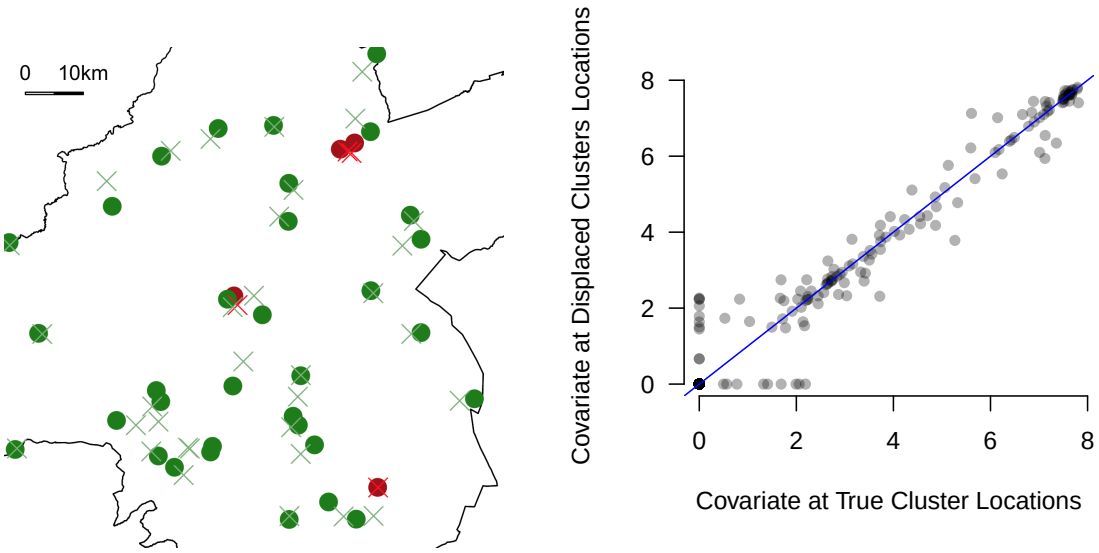

Figure 4.3: Left: true and jittered locations in simulation, zoomed in on the Western province. Solid points: true locations of clusters. ×: displaced locations. Red: urban clusters. Green: rural clusters. Right: value of covariate at true and jittered locations.

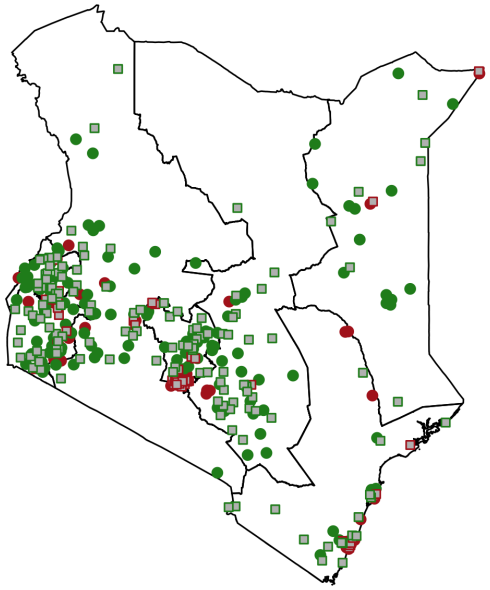

Figure 4.4: Solid points: GPS locations of clusters known. Grey squares: only admin area of clusters known. Red: urban clusters. Green: rural clusters.

4.4.3 Computation

For the "INLA within MCMC" algorithm, (4.6) is as follows:

$$p(s_{ijk}^{(t+1)} = E_{ije}|y_{ijk}, u_{ijk}, \boldsymbol{\theta}^{(t)}) \propto p(s_{ijk}^{(t+1)} = E_{ije}|u_{ijk}) \times \left\{\text{expit}\left(\beta_0^{(t)} + \beta_1^{(t)}\boldsymbol{z}(E_{ije}) + \tilde{S}(E_{ije})^{(t)}\right)\right\}^{y_{ijk}} \times \left\{1 - \text{expit}\left(\beta_0^{(t)} + \beta_1^{(t)}\boldsymbol{z}(E_{ije}) + \tilde{S}(E_{ije})^{(t)}\right)\right\}^{25-y_{ijk}}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{w})$, s_{ijk} is the true location of cluster k in strata j and administrative area i, u_{ijk} is the available location information for the cluster, and E_{ije} is a potential cluster location in strata j and area i, $e = 1, \ldots, m_{ij}$. The number of potential locations, m_{ij}, ranged from 1 to 1,015 for the jittering scenario (median 142) and 433 to 19,097 for the masking scenario. To obtain the normalization factors (4.2), for each possible enumeration area, we simulate 1,000 jitterings of the point following the DHS jittering algorithm and determine the proportion of realizations that fall within the administrative area.

The priors for the fixed effects (intercept and covariate association) were $N(0, 100)$. The effective range was set to 255 km and variance of the spatial process was set to 0.25. The priors for these spatial hyperparameters were analogous to the ones given in Section 3.4.2.

4.5 Simulation Results

Trace plots (see Appendix A.2) and calculated $\hat{R}$ (Gelman and Rubin, 1992) (which were all less than 1.05) suggested convergence for all scenarios and approaches. Posterior medians and 95% credible intervals (CIs) for the fixed effects and spatial hyperparameters are presented in Table 4.3. Figures 4.5–4.8 show the predicted latent surface $\tilde{S}(\cdot)$ (posterior median) and posterior standard deviation for the jittering and masking scenarios, respectively. First, we consider jittering, where the "best case" scenarios are 1a and 1b, where the true cluster locations are available. In general, using the jittered coordinates does not significantly impact the results, except for perhaps the uncertainty in the spatial surface (bottom rows of Figures 4.5 and 4.6). Additionally, differences seem to be larger when a spatial covariate is involved

Model	β_0	β_1	ϕ_1	ϕ_2
Truth	-1.5	0.15	3.93	-4.5
1a	-1.24 (-1.75, -0.85)	-	3.97 (3.60, 4.36)	-4.55 (-5.09, -4.05)
2a	-1.22 (-1.64, -0.84)	-	3.97 (3.60, 4.36)	-4.55 (-5.09, -4.05)
3a	-1.23 (-1.69, -0.81)	-	3.98 (3.61, 4.35)	-4.58 (-5.22, -4.10)
4a	-1.09 (-1.43, -0.50)	-	3.81 (3.33, 4.31)	-4.45 (-5.08, -3.86)
5a	-1.07 (-1.58, -0.55)	-	4.03 (3.57, 4.51)	-4.61 (-5.25, -4.00)
6a	-1.13 (-1.59, -0.63)	-	3.94 (3.44, 4.40)	-4.56 (-5.35, -4.00)
1b	-1.10 (-1.67, -0.48)	0.16 (0.12, 0.19)	4.00 (3.63, 4.40)	-4.71 (-5.36, -4.13)
2b	-1.08 (-1.58, -0.59)	0.14 (0.11, 0.18)	3.98 (3.61, 4.39)	-4.69 (-5.34, -4.11)
3b	-1.14 (-1.81, -0.48)	0.16 (0.12, 0.19)	3.99 (3.60, 4.37)	-4.71 (-5.52, -4.16)
4b	-1.06 (-1.99, 0.05)	0.18 (0.13, 0.23)	4.02 (3.56, 4.52)	-4.82 (-5.59, -4.13)
5b	-1.00 (-1.89, 0.27)	0.19 (0.15, 0.23)	4.05 (3.59, 4.54)	-4.85 (-5.62, -4.15)
6b	-1.05 (-1.81, -0.27)	0.18 (0.14, 0.23)	4.05 (3.57, 4.51)	-4.79 (-5.74, -4.17)

Table 4.3: Posterior medians (95% CIs) for parameters in the simulation scenarios considered (see Table 4.2)

in our simulation (1b and 2b). When using the DA approach for jittered data (3a and 3b), we also see some minor differences, and perhaps some "recovery" of the best case scenario.

Next, we consider the masking scenario, where the "best case" scenarios are again 1a and 1b, where true cluster locations are available for all clusters. Across the board, posteriors tend to be wider when we consider cases where only 50% of the clusters with GPS coordinates (4a and 4b). The approach that uses the centroid for the masked data (5a and 5b) gives slightly different results. Noticeably, these results tend to be worse when a spatial covariate is involved (5b). In this scenario, we had take the location for the masked data to be the centroid location of the potential locations and thuse used the value of the spatial covariate at that centroid location (rather than averaging the covariate from the potential locations). The DA approach (6a and 6b), where the other 50% of the clusters with only the admin area known are also included, show similar results. We find a more noticeable narrowing of the

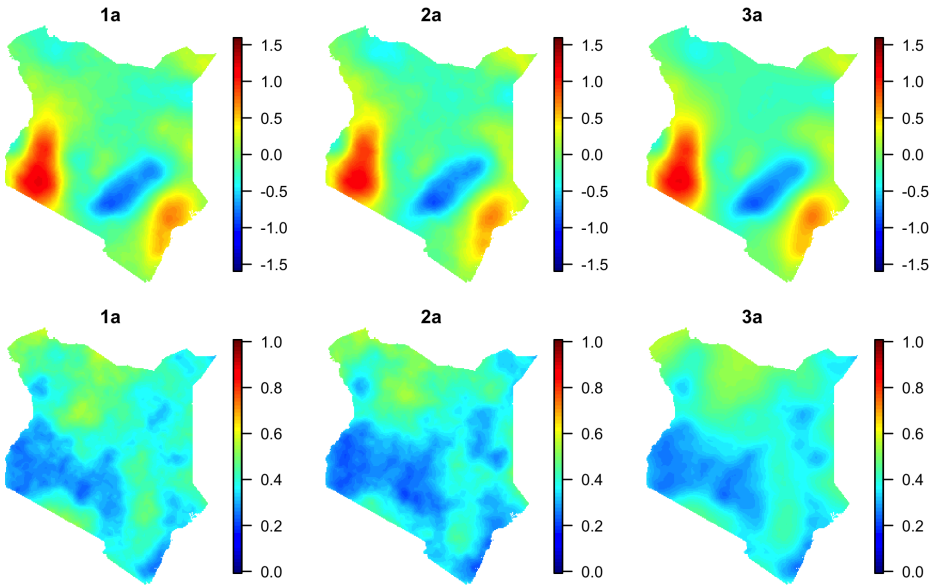

Figure 4.5: Top row: posterior medians of latent spatial surface. Bottom row: posterior standard deviations of latent spatial surface for the jittering scenario without a spatial covariate.

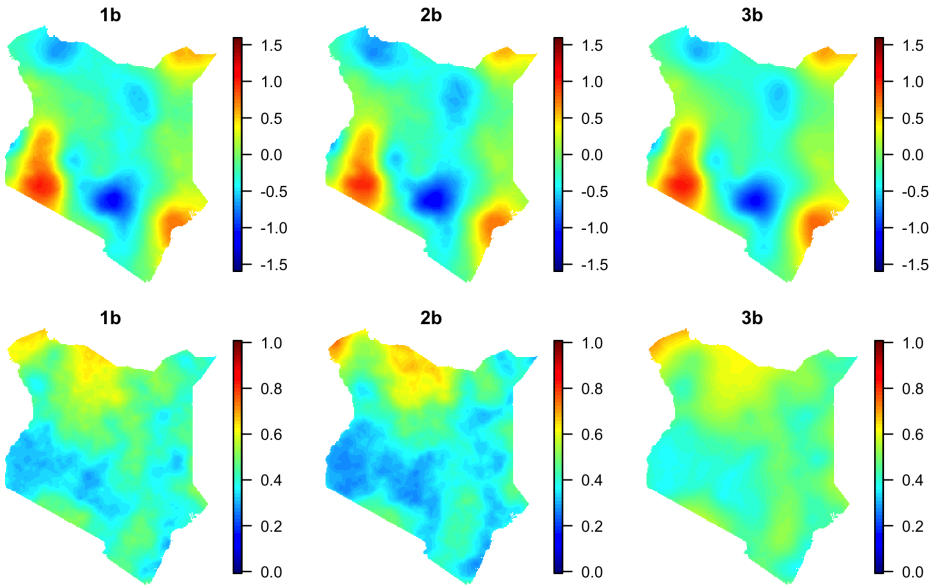

Figure 4.6: Top row: posterior medians of latent spatial surface. Bottom row: posterior standard deviations of latent spatial surface for the jittering scenario with spatial covariate.

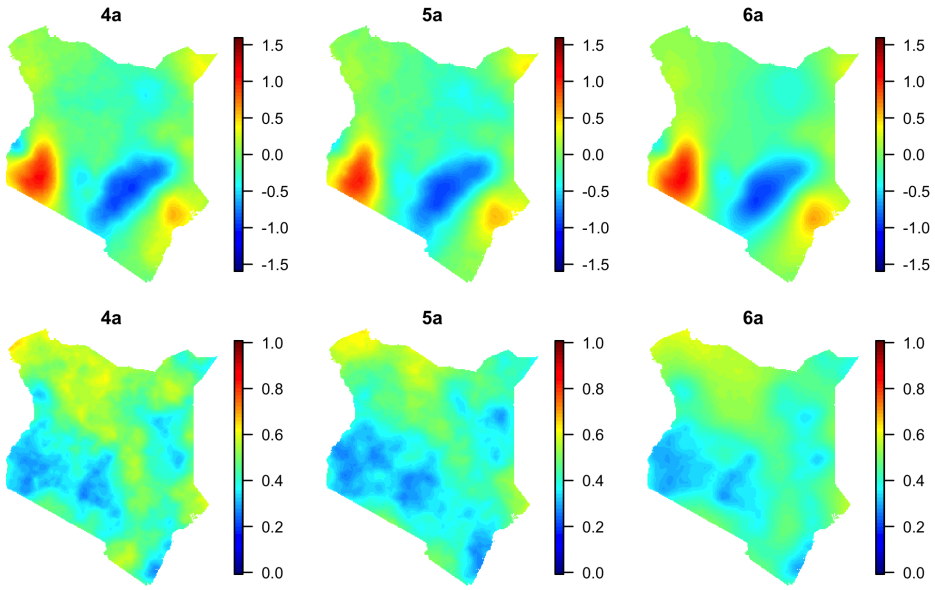

Figure 4.7: Top row: posterior medians of latent spatial surface. Bottom row: posterior standard deviations of latent spatial surface for the masking scenario without spatial covariate.

95% CIs in the scenario involving the covariate (6b).

The predicted probability surfaces are in Figures 4.9 and 4.10. Plotted are the posterior medians and 95% CIs. The posterior medians tend to be similar within the jittering scenarios and within the masking scenarios. It appears there is some overall reduction in uncertainty when using DA for the masking scenario, though this varies significantly spatially (Figure 4.11).

Additionally, we consider the mean squared error (MSE) of the predicted latent surface $S(s)$ and the probability surface (on the logit scale) over Kenya,

$$\text{MSE}^{(M)} = \frac{1}{G}\sum_{g=1}^{G}\left\{E(Y_g^{(M)} - y_g)\right\}^2 + \frac{1}{G}\sum_{g=1}^{G}\text{Var}(Y_g^{(M)})$$

with g indexing points, y_g being the true value of the surface at location s_g, and $Y_g^{(M)}$ being the estimate for the surface for model M at location g. We consider 2 different resolutions. In the first, predictions are made on a 1km $\times$ 1km grid, i.e., the grid points are 1km apart.

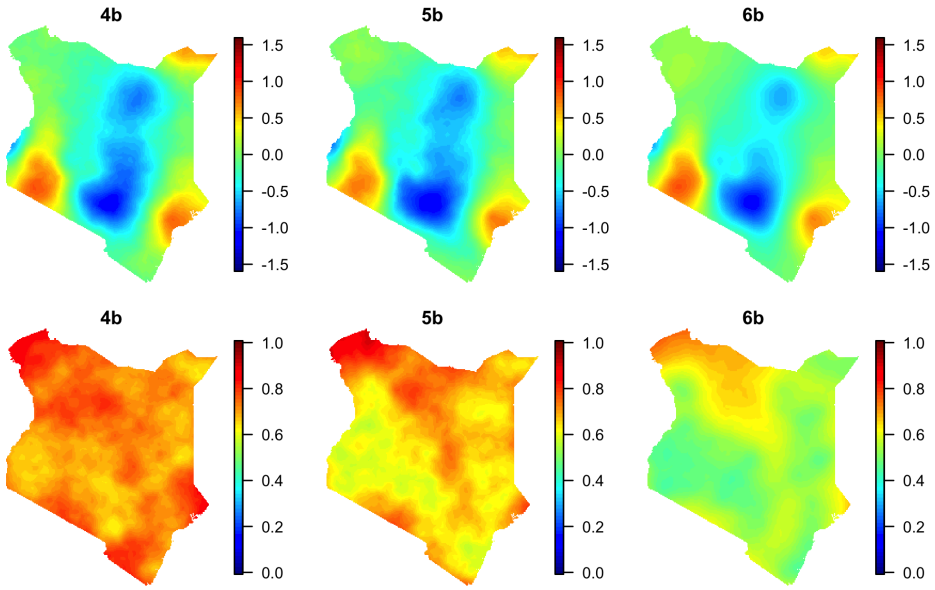

Figure 4.8: Top row: posterior medians of latent spatial surface. Bottom row: posterior standard deviations of latent spatial surface for the masking scenario with spatial covariate.

In the second, the predicted probability surface on the 1km × 1km grid is aggregated up to obtain predictions on a 5km × 5km grid. The values for the 1km × 1km grid, including average squared bias are in Table 4.4 (results were similar for 5km × 5km) and we can see that using the reported locations (the naive approach) tends to result in more bias as compared to using the correct approach, though this does not always hold. There also seems to be little to no benefit in using the DA approach in this setting. However, when we consider the masking scenario, we find a benefit in using DA over including the masked data via the centroid approach or not including the masked data at all.

An important consideration is the disclosure risk, or ability to identify the enumeration area a particular set of data arose from. Exact identification in the jittering case would be possible if there is only 1 possible EA within 2km for urban coordinates or within 10km for rural coordinates. In our example, 1 cluster could be exactly identified with another 10 having only at most 5 possible EAs; see Figure 4.12.

Another potential avenue for disclosure risk is if the posterior probability is significantly

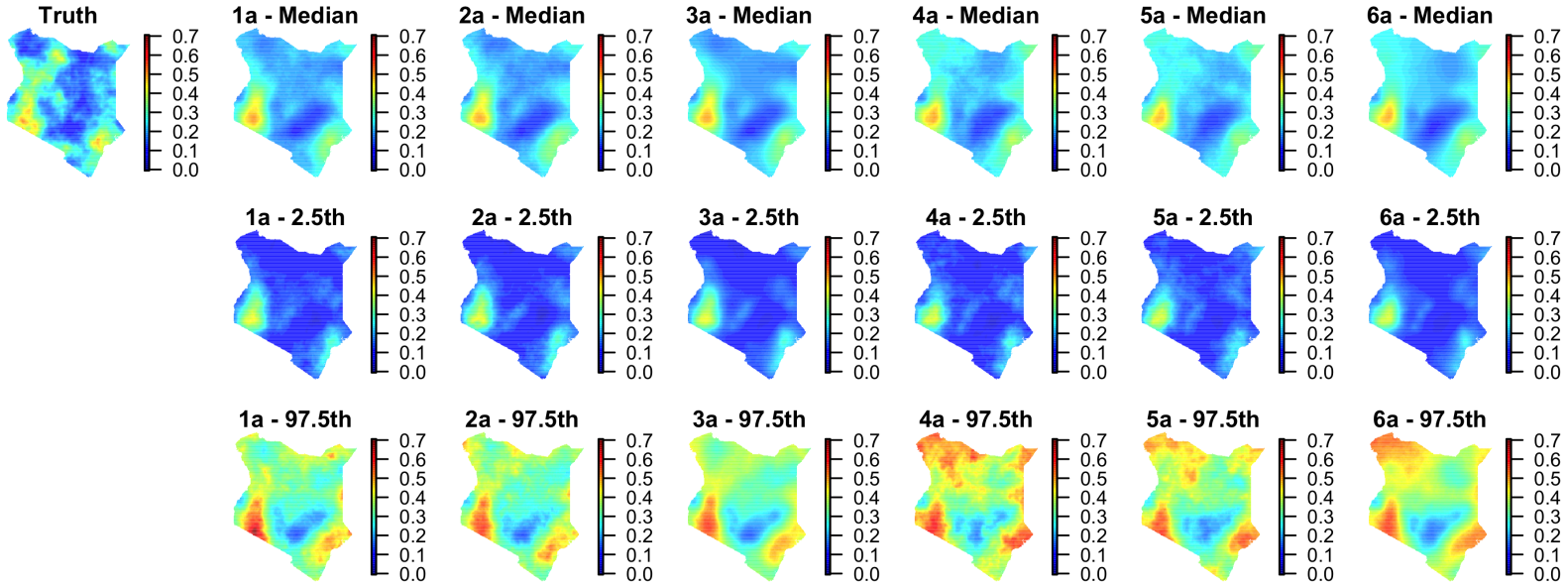

Figure 4.9: Predicted probability surface for simulation without spatial covariate. Top row: posterior median. Middle row: 2.5th percentile. Bottom row: 97.5th percentile.

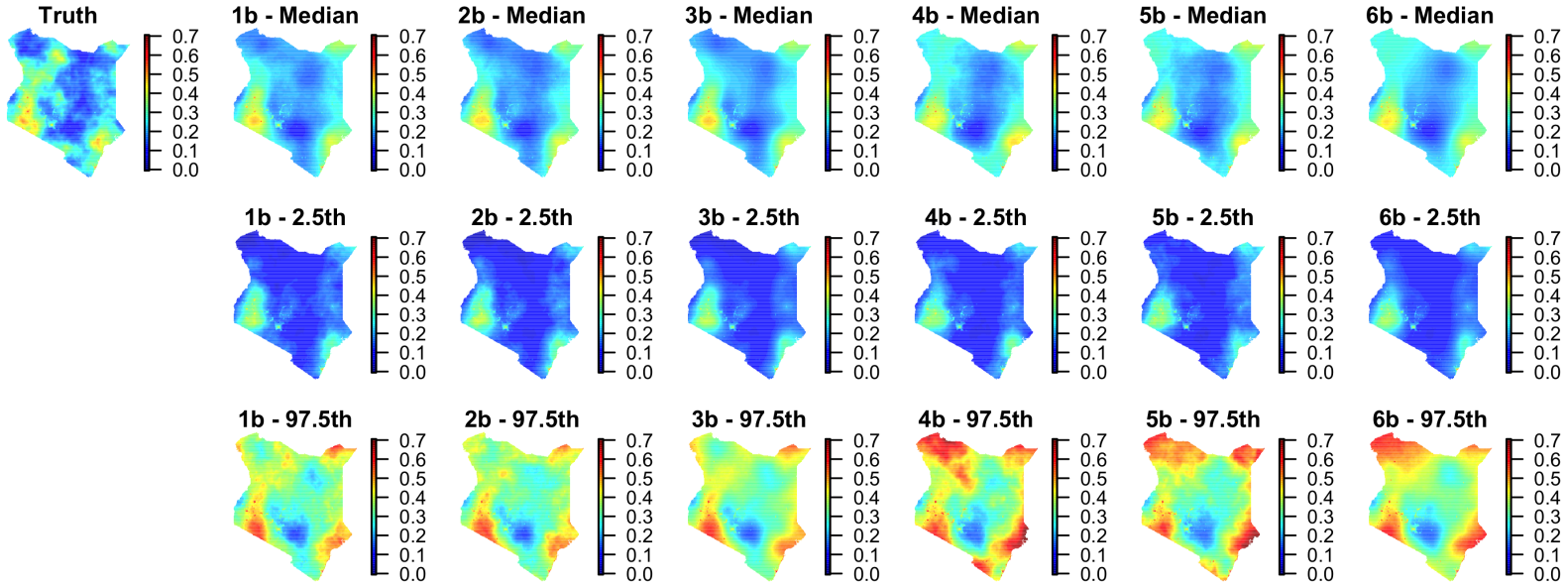

Figure 4.10: Predicted probability surface for simulation with spatial covariate. Top row: posterior median. Middle row: 2.5th percentile. Bottom row: 97.5th percentile.

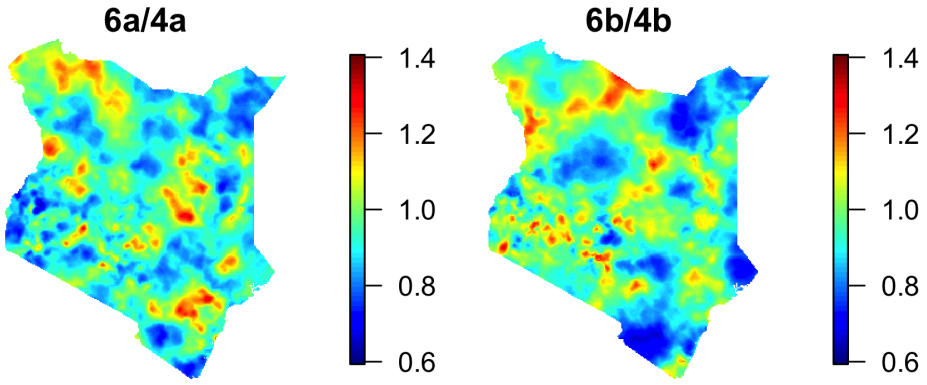

Figure 4.11: Ratio of posterior standard deviation of logit(p) in DA approach to 50% only approach for masking scenario. Values less than 1 indicates that the posterior standard deviation is lower when the data with masked location information is incorporated over not including it.

	$\tilde{S}(s)$	$p(s)$
1a	30.2 (16.3)	19.0 (8.81)
2a	30.5 (17.5)	20.2 (9.82)
3a	33.6 (19.0)	20.3 (9.73)
1b	45.4 (26.1)	20.7 (9.79)
2b	43.8 (26.4)	20.9 (9.66)
3b	45.8 (23.9)	21.0 (9.73)
4a	46.2 (26.0)	29.0 (14.0)
5a	43.6 (26.3)	25.2 (13.2)
6a	38.5 (22.1)	24.9 (12.4)
4b	65.4 (28.9)	27.7 (12.4)
5b	85.7 (35.0)	27.4 (13.5)
6b	56.2 (25.8)	26.3 (12.7)

Table 4.4: MSE (bias^2) of the probability surface from the various models on a 1km $\times$ 1km grids. All values have been multiplied by 100.

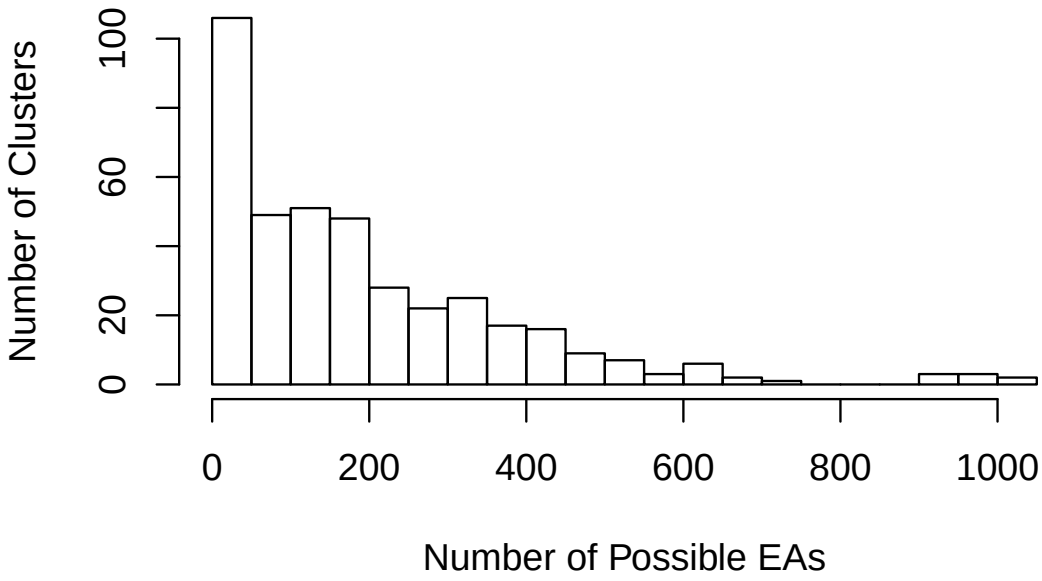

Figure 4.12: Histogram of potential disclosure risk.

larger for one particular EA than for the other potential ones. To establish this, the posterior probability of the possible EAs is calculated and the largest and second largest are compared. We do this for scenario 3a. First, we note that for 6 clusters, there was 1 possible EA with posterior probability > 0.95, meaning that for those clusters disclosure risk is highly probable. Additionally, for 26 (105) clusters the most likely EA had a posterior probability that was more than 5 (2) times higher than the second most likely.

This is less of a concern for the masking procedure as the number of possible EAs for each cluster range from 433 to 19,097 and the posterior probabilities were fairly uniform. Figure 4.13 shows the prior and posterior probabilities for one cluster that was known to be from a rural EA from the Coast province with the outcome $y = 5$ for 5a (no spatial covariate) and $y = 2$ for 5b (spatial covariate). Noticeably, the posterior probability is lower than the prior probability in the central eastern region where the latent spatial surface is highest.

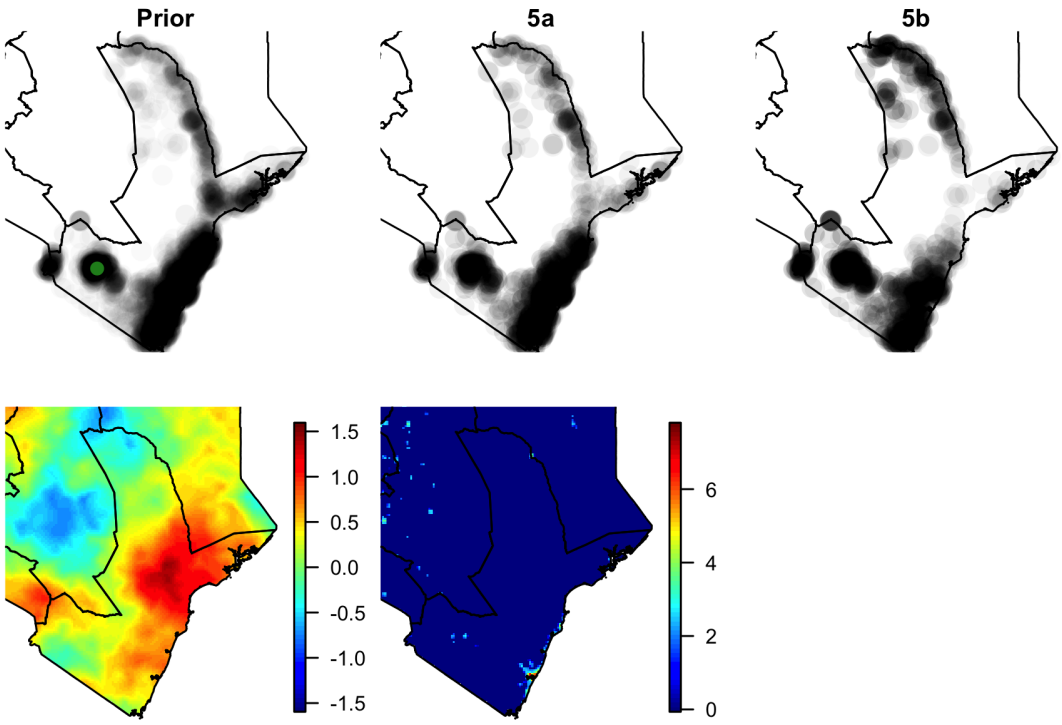

Figure 4.13: Prior and posterior probability of EA location for masking scenario. Darker indicates higher probability. Green point is the true location of the cluster. Also shown is the latent spatial surface (bottom left) and light surface (bottom middle).

4.6 Discussion

In this chapter, we propose an approach for incorporating data with missing or jittered GPS location information. We develop an "INLA within MCMC" approach, where we alternate between (1) updating the location of clusters by sampling from the full conditional posterior and (2) fitting conditional models using INLA and sampling from the approximated conditional posterior. In terms of computation time, it took about 52 hours to run 1,000 iterations for each scenario. The main computational burden comes from fitting the 1,000 `R-INLA` models.

We show that inference tends to be improved when the procedure that results in missing location information is taken into account through a simulation. Jittering of the coordinates did not have a significant impact on the results and one could argue that the more complicated DA procedure is not warranted. Further, there is a very real risk of identifying the true locations of (some of) the clusters, which is a privacy concern. From the DHS Terms of Use, users of geographic data "agree to treat all data as confidential, and to make no effort to identify any individual, household, or enumeration area in the survey" (`https://dhsprogram.com/data/terms-of-use.cfm`). We illustrate that in our simulation this is possible if the potential sampling locations (i.e., the masterframe) are available. Ideally, the jittering should provide a balance between accuracy in inference if the (incorrect) geographic coordinates are used and confidentiality in terms of not being able to uniquely identify the enumeration area that the cluster comes from. Assuming that there should be at least two enumeration areas possible for a reported location to have come from, one approach would be to calculate the distance between all of the points and their closest point within each strata (e.g., administrative area and urban/rural). The minimum radius needed to guarantee two possible enumeration areas is double the maximum distance to the closest point. This would not necessarily guarantee posterior probabilities that are not very close to 1, but is a potential strategy for reducing some disclosure risk. Notably, this is less of a concern when the data is masked and this is the setting where we saw greatest improvements

in inference.

In our model formulation, access to a masterframe was assumed. However, in many cases a masterframe is not available; therefore, the possible cluster locations are unknown. In this case, one could be created as it was done for the simulation and assumed to be correct or the grid cells of a population density raster could be used as a surrogate, with the population density value of the grid cell being used in (4.1).

Validity of the "INLA within MCMC" computational approach we employ relies on the validity of the INLA approximation. In our simulation, this approach seems to be accurate enough (in comparison to the models that could be fit solely in INLA) based on the posteriors obtained. To fully evaluate this approach, it would be best to compare results to only using MCMC. More practically, another option includes trying cruder approximations (i.e., a Gaussian and simplified Laplace approximation) in the INLA step to see if the results seem stable, as suggested by Rue et al. (2009). Another strategy could be to refine the hyperparameter grid that is used in the approximation. One could also validate overall results by holding out data and then comparing results on a large administrative area level.

Future work involves expanding the scope of the simulation study to cases where the masterframe is not available, investigating the impact of other covariates that are smoother in space, altering the spatial range of the underlying process, and including different levels of masking. In the simulation, we supposed that the the available location information was the provincial level, but other geographies exist such as the county level. Additionally, we plan to apply this method to survey data from Kenya.

Chapter 5

CHILD MORTALITY ESTIMATION INCORPORATING SUMMARY BIRTH HISTORY DATA

5.1 Introduction

An estimate of the under-five mortality rate (U5MR) over time is crucial for determining effectiveness of public health interventions and for better allocating resources. In countries without vital registration systems that track births and deaths, the data often come from surveys that are typically of one of two types: full birth history (FBH) and summary birth history (SBH). In surveys that collect FBH data, women are asked to recall the birth and, if applicable, death dates for each of their children. In contrast, in surveys that collect SBH data, women are only asked for the total number of children they have had and the number of those children that died. Given the temporal information contained in FBH data, it is relatively straightforward to obtain estimates of child mortality and U5MR using survey or model-based approaches (Mercer et al., 2015; Pezzulo et al., 2016; Rutstein and Rojas, 2006; Wakefield et al., 2018). Unlike FBH data, SBH data do not directly provide temporal information on when births and deaths occurred, and thus require more specialized methods. Overall, SBH data are easier and faster to collect, and thus widely available. In a recent study of U5MR in Africa, approximately 44% of the surveys contained only SBH data and 90% of the births were from SBH (Golding et al., 2017).

5.1.1 Existing SBH Methods

A common approach to obtaining U5MR estimates based on SBH data is to use variations of the Brass method (Brass, 1964; Feeney, 1980; Trussell, 1975), which involve model life tables. This approach provides an estimate of U5MR by five-year age groups of women and

additionally assigns each estimate a reference time in the past using a series of complex formulas. These formulas make use of quantities that are observed in SBH data: the number of children that died, the number of children born, and number of women surveyed along with women's age. Since this approach does not use a statistical model, it is not immediately clear how to obtain estimates of uncertainty. One proposed approach is to use the jackknife (Pedersen and Liu, 2012). A major shortcoming of the Brass method is that the mortality rates in the most recent time periods are based on the youngest women. Children born to younger mothers tend to have worse outcomes, resulting in bias. Therefore, these estimates are typically excluded from analyses. Alternatively, Brass-type methods using time since first birth or marriage are used instead of age of mother (Hill and Figueroa, 1999; United Nations, 1983). Currently, the United Nations produce national estimates of U5MR that incorporate SBH data via indirect estimates obtained from the time since first birth variant of the Brass method (Alkema et al., 2014; Hill et al., 2012). Hill et al. (2012) summarizes the historical approach of combining the estimates using a loess smoother in time. The bootstrap can be used to obtain uncertainty intervals (Alkema and New, 2012). A more recent approach, proposed by Alkema et al. (2014) and Alkema and New (2014), combines the different sources of data and uses a Bayesian penalized regression spline to model trends over time. Uncertainty in the SBH data is incorporated by using the jackknife, or by using survey weights if microlevel data is available, and otherwise fix the uncertainty to a predetermined number.

Other methods (Hill et al., 2015; Rajaratnam et al., 2010) that do not rely on the demographic models used in the Brass method have been proposed, but critically do not assume a full probability model. Rajaratnam et al. (2010) describe a number of methods, including one that uses FBH data to derive empirical distributions of births and deaths prior to the survey and then matches SBH women to the relevant empirical distribution. This provides a yearly estimate of the ratio of children that died to children ever born. The ratio is then related via a logistic regression model to the probability that a child dies within five years, calculated using FBH data. In practice, FBH data from surveys in different countries and

time periods are pooled together to build the regression model and empirical distributions. Hill et al. (2015) propose a birth history imputation approach in which SBH women are matched to FBH women, who are typically available from an earlier survey in that country. Women are matched by age, number of births, and number of deaths. The FBH data are then used to impute births and deaths to the SBH women. This approach gave disappointing results when validating mortality estimates obtained from this method and comparing to estimates computed from a later FBH survey (Brady and Hill, 2017; Hill et al., 2015). The authors attribute this to incompatibility of the SBH and FBH data, stemming from data quality issues with SBH data.

5.1.2 Malawi Data Available

As our motivating example, we consider retrospective birth history data taken between 2004 and 2015 in Malawi, from five surveys and a census. Specifically, we use FBH data from the 2004 Malawi Demographic and Health Survey (DHS), the 2010 Malawi DHS, the 2015 Malawi DHS, the 2006 Multiple Indicator Cluster Survey (MICS), and the 2013 MICS. We include surveyed women who are aged 15–49 at time of the survey. We use SBH data from the 2008 Malawi Census. From the census, we include women who are aged 25–49 at time of the survey. Excluding women under 25 who provide SBH data is consistent with other analyses that incorporate SBH data (Hill et al., 2012). The DHS and MICS use 2-stage stratified cluster sampling designs. The 2004 DHS was stratified by urban-rural classification with certain districts being oversampled. The 2006 MICS was stratified by district. The 2010 DHS, 2013 MICS, and 2015 DHS were stratified by urban-rural classification crossed with district. We use available microdata from the census, which is a 10% sample.

We focus on the 9 districts in the Central region: Dedza, Dowa, Kasungu, Lilongwe, Mchinji, Nkhota Kota, Ntcheu, Ntchisi, and Salima. In this region, birth history information is available from 109,713 women (43% of women had SBH information only). The number of clusters, women surveyed, births, and deaths for each of the surveys and census is in Table 5.1. Across all surveys and the census, the age of the mother at time of survey, district, and

Table 5.1: Data summaries, by survey.

Survey		No. Clusters	No. Women	No. Births	No. Deaths
Census 2008			71,618	332,055	62,884
DHS 2004		186	4,199	13,394	2,532
DHS 2010	Training	140	3,879	12,626	2,125
	Holdout	145	3,983	12,873	2,066
DHS 2015		284	8,417	23,240	2,621
MICS 2006	Training	174	4,518	14,505	2,448
	Holdout	186	4,832	15,069	2,397
MICS 2013		381	8,267	24,798	3,028

strata are available. The 5 surveys that contain FBH information also provide reported birth and death dates for each child. The census contains only the reported number of births and deaths for each woman.

Results from an exploratory analysis, similar to the diagnostic measures reported by Hill et al. (2015), where FBH information was transformed to SBH information are depicted in Figure 5.1. The proportion of women interviewed tends to be similar across surveys. There appears to be some age heaping present, where women report their age as ending in "5" or "0." There also do not appear to be systematic differences between the number of total births women report having had across surveys. However, there appear to be differences in the ratio of children dead (CD) to children ever born (CEB). Importantly, the differences tend to vary in rural and urban areas. We would expect the proportion CD/CEB among women in the census to be similar to the proportion CD/CEB among women in the 2006 MICS and 2010 DHS since these surveys were taken closest in time. These surveys are much noisier, but we can see that in rural areas the proportion CD is higher than both the 2006 MICS and 2010 DHS across almost all women. This is not the case in urban areas. This

trend motivates including a term in our model that can capture this pattern.

Weighted estimates (Horvitz and Thompson, 1952) and 95% confidence intervals (CIs) going back to 1980 were computed for each of the FBH surveys and are shown in Figure 5.2. Overall, under-five mortality is decreasing, but there is significant heterogeneity over districts and time periods.

The organization of this chapter is as follows. In Section 5.2 we describe our method, which is based on data augmentation (DA). In Section 5.3 we describe an alternative approach based on combining estimates from using weighted, also known as direct, estimators on FBH data, and indirect estimates from using the Brass method on SBH data. In Section 5.4 we show the usefulness of the approaches in a simulation. We apply our approach to the survey and census data from Malawi in Section 5.5. Finally, we conclude with a discussion in Section 5.6.

5.2 Data Augmentation Method

We propose using a DA approach within a Bayesian framework. This is implemented via a Markov chain Monte Carlo (MCMC) algorithm, where each iteration is divided into two major steps. In the first step, the missing birth and death dates (available for FBH data) are introduced as auxiliary variables for the SBH data. In the second step, mortality and fertility parameters are then updated conditional on this imputed FBH data and combined with the existing FBH data.

In modeling SBH data, we will model fertility rates and mortality rates, which we specify as probabilities. The forms for these models are heavily driven by context. We define the fertility rate as the probability a woman gives birth at age m and time t and denote it as $f(m, \boldsymbol{x}(t))$, where $\boldsymbol{x}(t)$ contains the covariates at time t associated with fertility. The mortality rate, or probability a child dies between age a and $a+1$, is denoted by ${}_1q_a(\boldsymbol{x}(t)) = q_a(1, \boldsymbol{x}(t))$, where $\boldsymbol{x}(t)$ contains covariates at time t associated with mortality.

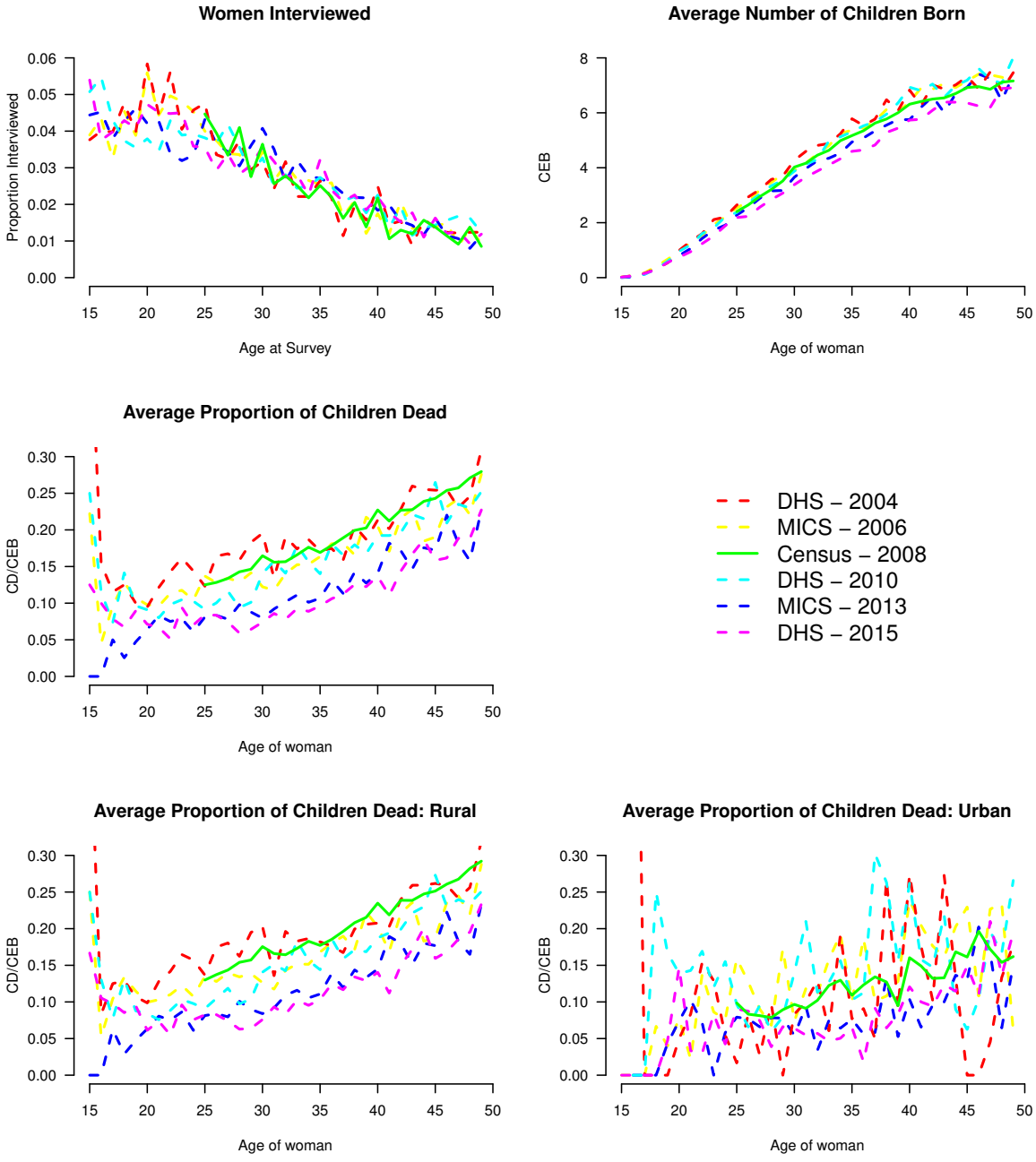

Figure 5.1: Top left: Proportion of women between ages 15–49 interviewed. Top right: Average number of reported children ever born (CEB). Middle left: Average number of children dead (CD) to CEB. Bottom row: Average number of CD to CEB by rural and urban strata.

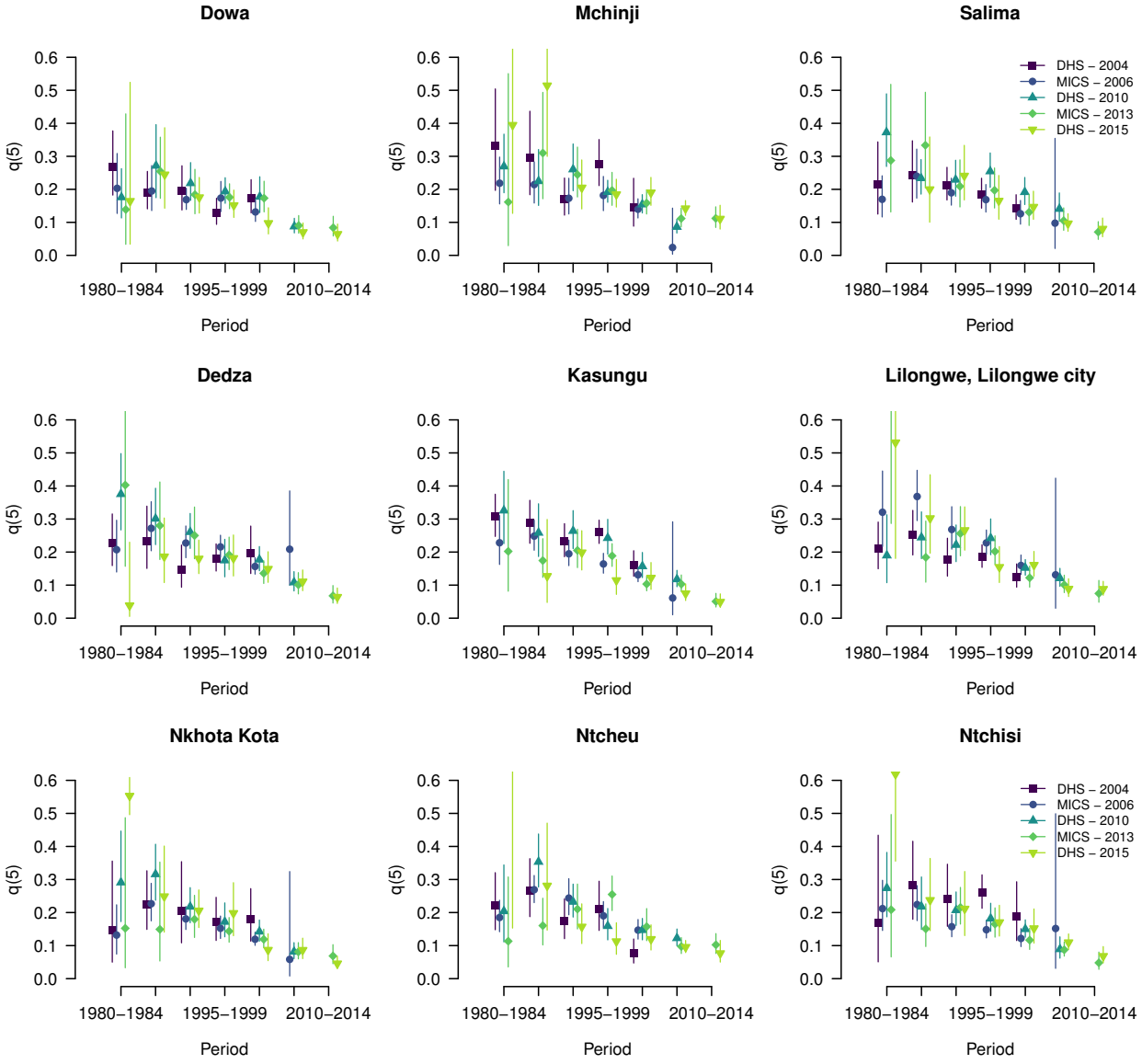

Figure 5.2: HIV-adjusted weighted estimates of under-five mortality ($q(5)$) over time for the 9 districts in central Malawi, by survey. Points are the estimates and lines are the 95% Wald-based confidence intervals computed on the logit scale.

5.2.1 Simple Example

To motivate our approach, we first consider a simple scenario where we suppose that fertility and mortality do not change with time. Suppose we survey a woman who is age 18 at the time of the survey and has had two children, one of whom has died. Further suppose that the youngest age she could have given birth was 15 and that it is not possible to have multiple births. Additionally, we will discretize time and work on a yearly scale. Therefore, assuming the woman could have given birth when she was 15, 16, or 17, there are 3 options for when the children were born relative to the woman's age (i.e., at ages 15 and 16, 15 and 17, or 16 and 17). We also know that one of the children has died. This means there are 6 options for when the children were born and which of the children died (e.g., the child born when the mother was 15 died and the child born at 16 survived). The probabilities of each of these 6 options will depend on the fertility rates and the mortality rates. A visual depiction of these scenarios is given in Figure 5.3.

Now suppose we knew when the children were born and which of those children died (i.e., we selected one of the 6 scenarios). For the child who died, we can enumerate the options for age of death. In this case, suppose the woman was 15 when the child who died was born. Again considering a yearly time scale, there are 3 scenarios (i.e., the child died between ages 0 and 1, ages 1 and 2, or ages 2 and 3). The probability associated with each of these options is based on the mortality rates (see Figure 5.3).

We have given an overview of the DA that links SBH information to FBH information for a simple example. This forms the backbone of the DA step of our method, and we now give details of the probability model that is used in this step.

5.2.2 Step 1: Data Augmentation

Let t denote calendar time in years. Consider a woman with B births, D deaths, and covariates $\boldsymbol{x}(t)$. Let m_s be the age of the woman at time of survey and t_s be the year the survey was completed. Define $\boldsymbol{t}_b = \{t_{b_1}, \ldots, t_{b_B}\}$ to be a vector of length B which contains

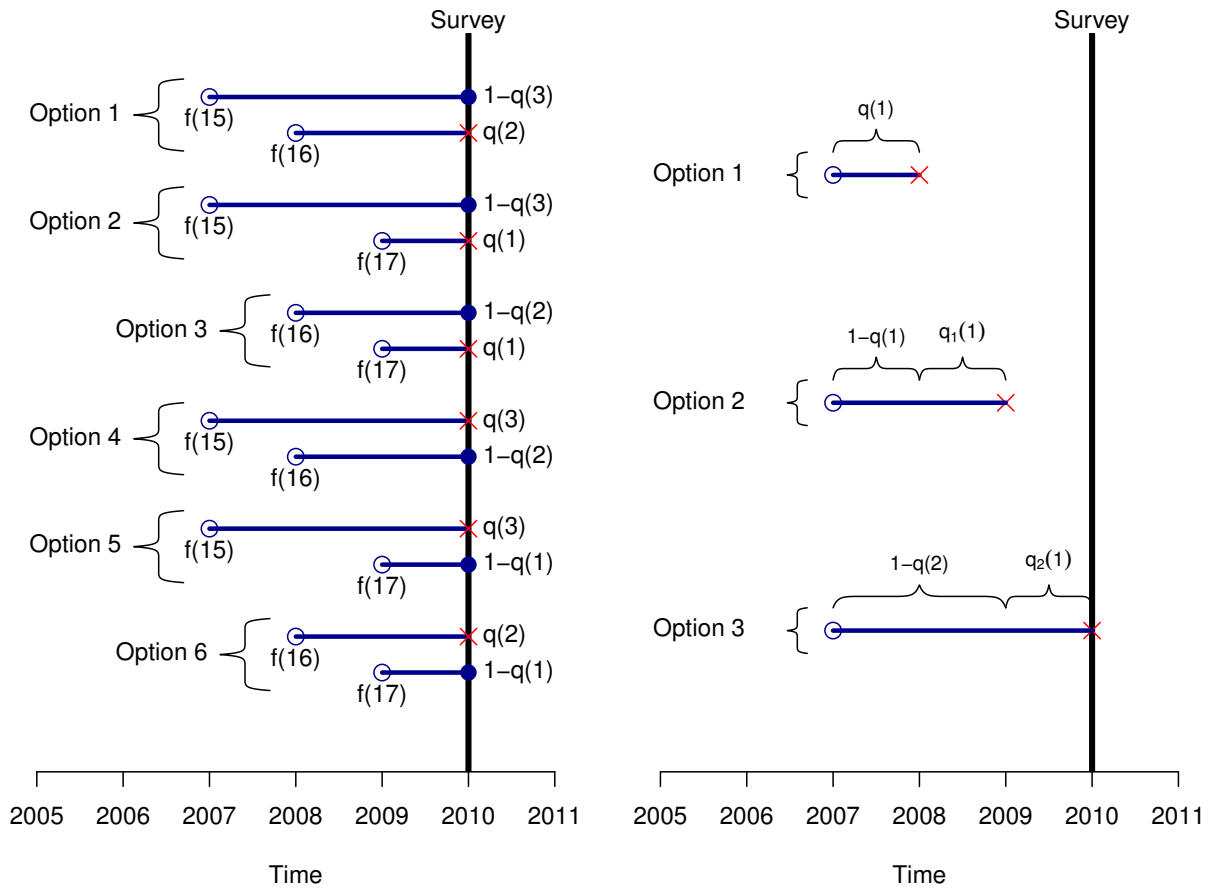

Figure 5.3: Left: Possible birth years and death indicators for 2 children with mother's age $m_s = 18$, assuming 1 death. Fertility $f(m) = 0$ for $m < 15$. Right: Possible death ages for a child born to a mother of age $m_b = 15$. The $\times$'s represent a death. Here, $q(a)$ is the probability of death by age a.

the (unknown) years of birth of the B children. Define $\boldsymbol{d} = \{d_1, \ldots, d_B\}$ to be a vector of length B which contains the "death indicators" for the children (i.e., $d_i = 1$ if child i dies). Therefore, $\sum_{i=1}^{B} d_i = D$. It will also be helpful to define m_{b_i} to be the age of the woman when she gave birth to child i. Note that $m_{b_i} = m_s - (t_s - t_{b_i})$.

The probability of children being born in particular years and surviving or dying by the time of the survey is proportional to the probability of giving birth to the children in those years, observing them to survive through the survey or die at some point prior to the survey,

and not having children born in the other years:

$$\Pr\left(\boldsymbol{T}_b = \boldsymbol{t}_b,\ \boldsymbol{D} = \boldsymbol{d}\right) \propto \left(\prod_{i=1}^{B} f(m_{b_i}, \boldsymbol{x}(t_{b_i})) \times \left[1 - \prod_{a=1}^{t_s - t_{b_i}} \{1 - {}_1q_{a-1}(\boldsymbol{x}(t_{b_i} + a - 1))\} \right]^{d_i} \right.$$
$$\left. \times \left[\prod_{a=1}^{t_s - t_{b_i}} \{1 - {}_1q_{a-1}(\boldsymbol{x}(t_{b_i} + a - 1))\} \right]^{1-d_i} \right) \times \prod_{\substack{t \notin \{t_{b_1}, \ldots, t_{b_B}\}, \\ t < t_s}} \{1 - f(m, \boldsymbol{x}(t))\}. \tag{5.1}$$

This probability is conditional on B births, D deaths, covariates $\boldsymbol{x}(t)$, survey taken in year t_s, a woman of age m_s at time of survey, and functions $f(\cdot)$ and $q(\cdot)$, but for notational ease we do not explicitly state this. As expressed, this probability does not allow for multiple births and assumes that births in separate years are independent. However, these assumptions could be relaxed.

Given birth in year t_{b_i}, $d_i = 1$, covariates $\boldsymbol{x}(t)$, and the mortality function $q(\cdot)$, the probability that child i dies in a particular year t_{d_i} is proportional to the probability of surviving to year $t_{d_i} - 1$ and subsequently dying in the next year,

$$\Pr(T_{d_i} = t_{d_i}) \propto {}_1q_{t_{d_i} - t_{b_i} - 1}(\boldsymbol{x}(t_{d_i} - 1)) \prod_{a=1}^{t_{d_i} - t_{b_i} - 1} \{1 - {}_1q_{a-1}(\boldsymbol{x}(t_{b_i} + a - 1))\}. \tag{5.2}$$

These are the necessary probabilities for updating the birth and death years for all SBH women's children (an example of which is shown in Figure 5.3). In terms of computation, it is possible to enumerate all options thus computing the denominators for (5.1) and (5.2), and sample from the full posterior conditional distribution. This can be efficiently achieved by grouping women together that have the same age, number of births, number of deaths, and covariates so that the desired probabilities need only be computed once. However, computational gains can be made using a Metropolis-Hastings algorithm. This is especially useful for older women, who tend to have more children. One simple approach is to iterate through all of a woman's children and sample the birth years one-by-one, where $q(\cdot)$ on the

left hand side represents the proposal probability,

$$q(T_{b_i} = t_{b_i}, d_i = 1 \mid i \leq D) \propto f(m_{b_i}, \boldsymbol{x}(t_{b_i})) \times \left[1 - \prod_{a=1}^{t_s - t_{b_i}} \{1 - {}_1q_{a-1}(\boldsymbol{x}(t_{b_i} + a - 1))\}\right],$$

$$q(T_{b_i} = t_{b_i}, d_i = 0 \mid i > D) \propto f(m_{b_i}, \boldsymbol{x}(t_{b_i})) \times \left[\prod_{a=1}^{t_s - t_{b_i}} \{1 - {}_1q_{a-1}(\boldsymbol{x}(t_{b_i} + a - 1))\}\right].$$

In the data analysis, this independence chain proposal resulted in acceptance ratios > 0.6.

5.2.3 Step 2: Parameter Updates

After performing the DA step, the imputed FBH from the SBH is combined with available FBH and the fertility and mortality probabilities are updated. Let $\mathrm{Y}(m, t)$ be an indicator for birth at woman's age m and year t, then $\mathrm{Y}(m, t) \mid f(m, \boldsymbol{x}(t)) \sim \text{Bernoulli}\,(f(m, \boldsymbol{x}(t)))$.

For the mortality model, we will use a discrete hazards model. Let $\mathrm{Z}_a(t)$ be an indicator for death between age a and $a + 1$ in year t. The likelihood is $\mathrm{Z}_a(t) \mid {}_1q_a(\boldsymbol{x}(t)) \sim \text{Bernoulli}\,({}_1q_a(\boldsymbol{x}(t)))$. The U5MR for time t and with covariates $\boldsymbol{x}(t)$ is,

$${}_5q_0(\boldsymbol{x}(t)) = 1 - \prod_{a=0}^{4} \{1 - {}_1q_a(\boldsymbol{x}(t))\}.$$

Forms for $f(\cdot)$ and ${}_1q_a(\cdot)$ are given in Sections 5.4 and 5.5 for the simulation and data analysis, respectively.

5.3 Weighted Estimation with the Brass Method

First, we briefly provide a review of the Brass method as it pertains to obtaining estimates of the U5MR and a description of how to obtain estimates of uncertainty. Lastly, we describe how to combine Brass and weighted estimates and incorporate an adjustment for HIV.

5.3.1 The Brass Method

Define D_{m_s}, B_{m_s} and $d_{m_s} = D_{m_s}/B_{m_s}$ to be, respectively, the total number of children dead, the total number of children born and the proportion that died, to women aged m_s at the

time of survey. Then

$$E(d_{m_s}|B_{m_s}) = \int_0^{m_s} c_{m_s}(a)q(a)\mathrm{d}a \tag{5.3}$$

where $c_{m_s}(a)$ is the proportion of births to women who are m_s at the time of the survey a years prior to the survey and $q(a)$ is the probability that a child born a years before the survey dies before the survey. The Brass method treats (5.3) as deterministic, replacing the expectation on the left side with the observed proportion, d_{m_s}. By the mean value theorem, there exists an $a^* \in (0, m_s)$ such that,

$$d_{m_s} = q(a^*) \int_0^{m_s} c_{m_s}(a)\mathrm{d}a = q(a^*).$$

The key idea of the Brass method is to identify a^* and thus use the observed proportion of children dead to find $q(a^*)$ (Chapter 11 of Preston et al., 2000).

Brass (1964) achieved this by using model life tables and a polynomial fertility model to numerically integrate (5.3) using 5-year age groups of women, $i = 15-19, 20-24, \ldots, 45-49$. The proportion dead in age group i, denoted $\tilde{d}_i$, is then compared to a $\tilde{q}(a)$ curve obtained from model life tables to find the a_i^* such that $\tilde{d}_i = \tilde{q}(a_i^*)$. These times, a_i^*, were then adjusted to whole number of years. For example, $i = 15-19,\ 20-24,\ 25-29,\ 30-34$ correspond approximately to $q(1),\ q(2),\ q(3),\ q(5)$ (see Table 4 of Brass, 1964). These correspondences are not exact; therefore, adjustment terms are needed. Ideally, $c_i(a)$ (the proportion of births to women in age group i) would be used. However, with those proportions unknown, fertility information from cohorts of women are used instead. The Brass method makes use of observed parity measures, taken to be the mean number of children born to women in each age group. Comparing parity measures across age groups can provide a sense of the earliness of fertility. Define P_1, P_2, and P_3 to be the mean number of children ever born to all women in age groups $15-19$, $20-24$, and $25-29$, respectively. Trussell (1975) developed the model that is commonly used to adjust the observed proportion of children dead d_i in age group i to $q(x)$,

$$q(x) = d_i \left(a_{1i} + a_{2i}\frac{P_1}{P_2} + a_{3i}\frac{P_2}{P_3} \right),$$

where the set of coefficients, a_{1i}, a_{2i}, and a_{3i} were estimated via simulation. In fact, 4 sets of coefficients were derived, one for each of the Coale and Demeny (1983) regional model life tables: "North", "West", "South", and "East".

To acknowledge changing mortality, Coale and Trussell (1977) assume mortality declines linearly. Using (5.3), the reference time a^* would equal the mean length of time since birth of children. Again using simulation, Coale and Trussell (1977) develop another formula involving parity measures and coefficients, b_{1i}, b_{2i}, and b_{3i}, to identify a time for which the Brass estimate is most relevant for age group i,

$$t(x) = b_{1i} + b_{2i}\frac{P_1}{P_2} + b_{3i}\frac{P_2}{P_3}.$$

Finally, to obtain estimates of $q(5)$, the Brass estimates $q(x)$ are converted using numbers based on life tables. To determine which model life table to follow (and thus which set of coefficients to use), we follow the suggestion of Hill (2013) and first obtain direct estimates of ${}_1q_0$ and ${}_4q_1$ from the 2010 DHS and 2006 MICS (FBH surveys taken near to the time of the census). We compare the direct estimates at the regional level with the 4 Coale–Demeny regional model life tables and select the model life table that the direct estimates most closely follow (see Figure 5.4). Although none of the models fit the observed values exactly, the North model appeared to fit best. The coefficients for both formulas and adjustment procedure can be found in the United Nations' *Manual X* (United Nations, 1983).

5.3.2 Estimates of Uncertainty for the Brass Method

Define $\hat{\theta}(t)^{\mathrm{B},\star} = \text{logit}\left({}_5\hat{q}_0(t)^\star\right)$ where ${}_5\hat{q}_0(t)^\star$ are the Brass (indirect) estimates. We assume,

$$\hat{\theta}(t)^{\mathrm{B},\star} \sim N\left(\theta(p)^\star, V(t)^{\mathrm{B},\star}\right)$$

where $\theta(p)^\star$ is the true (unadjusted for HIV) logit U5MR in time period p containing the reference time t and $V(t)^{\mathrm{B},\star}$ is the variance. Note that there are seven estimates, as there are seven five-year age groups of women. We will assume they are independently distributed.

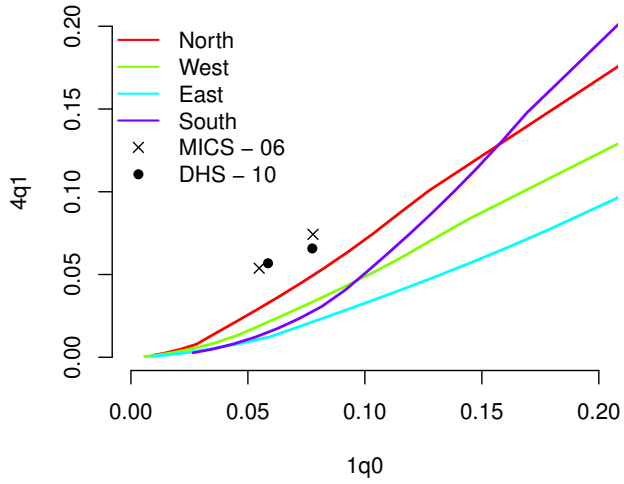

Figure 5.4: Direct estimates of ${}_1q_0$ and ${}_4q_1$ obtained from 2006 MICS and 2010 DHS for the time periods 2000–2004 and 2005–2009 compared to the 4 Coale-Demeny regional model life tables.

To obtain variance estimates when using the Brass Method we use a technique similar to Pedersen and Liu (2012) based on the jackknife. We adapt their approach, which is based on deleting clusters, since cluster level information is not available for the census SBH data, and instead define jackknife samples based on women. For the jth sample (where the jth woman is removed from the dataset), we compute $\hat{\theta}_j(t)^\star = \text{logit}\left({}_5\hat{q}_{0,j}(t)^\star\right)$ where ${}_5\hat{q}_{0,j}(t)^\star$ is computed using the Brass method on this reduced sample. For n total women, we calculate,

$$\bar{\theta}(t)^\star = \frac{1}{n}\sum_{j=1}^{n}\hat{\theta}_j(t)^\star, \quad \text{and}$$

$$\hat{V}(t)^{\text{B},\star} = \widehat{\text{var}}\left(\hat{\theta}(t)^{\text{B},\star}\right) = \frac{n-1}{n}\sum_{j=1}^{n}\left\{\hat{\theta}_j(t)^\star - \bar{\theta}(t)^\star\right\}^2,$$

where the reference time t is computed using the reference time Brass equation and is held fixed during this procedure. Theoretically, there is uncertainty in the reference time, but as with all previous implementations of Brass we do not consider that here.

5.3.3 Combining Brass and Weighted Estimates

Using FBH data, we can construct weighted (direct) estimates and associated variances for each FBH survey on the 5-year period scale. For survey s in period p, denote the estimate of the logit U5MR as $\hat{\theta}(p;s)^{\mathrm{W},\star}$ and estimated variance $\hat{V}(p;s)^{\mathrm{W},\star}$. Again we assume,

$$\hat{\theta}(p;s)^{\mathrm{W},\star} \sim N\left(\theta(p)^{\star}, V(p;s)^{\mathrm{W},\star}\right).$$

We follow Wakefield et al. (2018) and use HIV multiplicative correction factors on the U5MR, obtained using the approach of Walker et al. (2012) to adjust for HIV bias. This correction is needed to adjust for women who died during the HIV epidemic in Malawi and are therefore not included in the survey. These women would tend to have children with worse outcomes, thus mortality would be biased downwards and the aim of the adjustment terms is to correct for this. They are survey specific and the HIV correction factors are depicted in Figure 5.5. For an unadjusted estimate and associated variance, we sample 100,000 realizations and transform them using the HIV correction factors, denoted $k(p;s)$,

$$\begin{aligned}\phi^{\star}(p;s) &\sim N\left(\hat{\theta}(p;s)^{\mathrm{W},\star}, \hat{V}(p;s)^{\mathrm{W},\star}\right),\\ \theta(p;s) &= \text{expit}\left\{\frac{\phi^{\star}(p;s)}{k(p;s)}\right\}.\end{aligned}$$

The means and variances of the HIV adjusted samples are computed and denoted $\hat{\theta}(p;s)^{\mathrm{W}}$, and $\hat{V}(p;s)^{\mathrm{W}}$, respectively. The same procedure is used for the Brass estimates and variances.

These corrected estimates and variances are then combined over all surveys and times t that fall in period p via inverse variance weighting. Therefore, the overall HIV-corrected estimate and HIV-corrected variance are respectively,

$$\begin{aligned}\hat{\theta}(p) &= \hat{V}(p) \times \left[\sum_{s}\left\{\hat{V}(p;s)^{\mathrm{W}}\right\}^{-1}\hat{\theta}(p;s)^{\mathrm{W}} + \sum_{\mathrm{t}\in p}\left\{\hat{V}(t)^{\mathrm{B}}\right\}^{-1}\hat{\theta}(t)^{\mathrm{B}}\right],\\ \hat{V}(p) &= \left[\sum_{s}\left\{\hat{V}(p;s)^{\mathrm{W}}\right\}^{-1} + \sum_{\mathrm{t}\in p}\left\{\hat{V}(t)^{\mathrm{B}}\right\}^{-1}\right]^{-1}.\end{aligned}$$

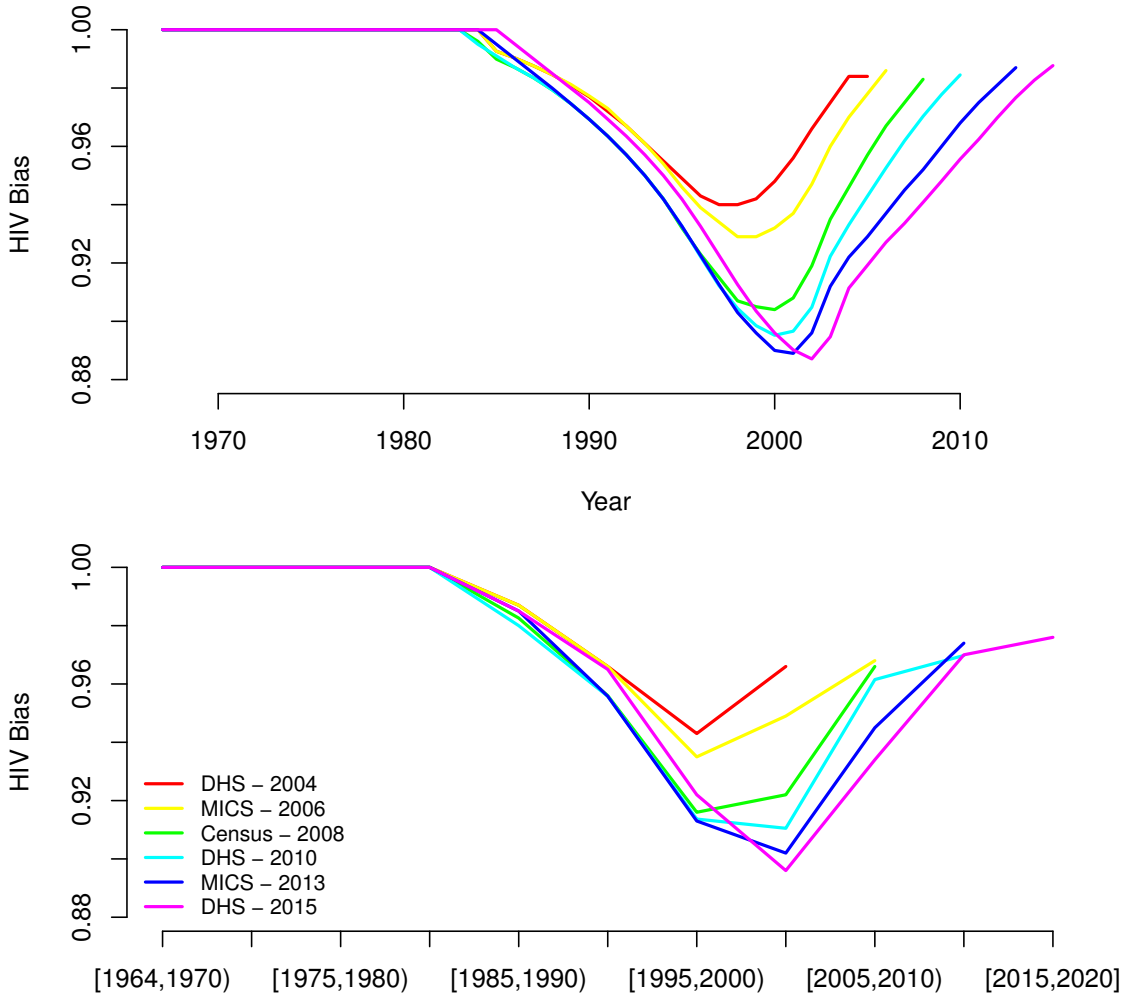

Figure 5.5: Top: yearly HIV bias color-coded by survey in Malawi. Bottom: 5-year HIV bias color-coded by survey in Malawi.

5.4 Simulation Study

We illustrate the gains that can be made by incorporating SBH data using our approach in a simulation study. We hold fertility fixed over time, but set mortality to vary in 5-year increments. Fertility probabilities and discrete hazards for each period are presented as horizontal lines in Figures 5.6 and 5.7, respectively. As evident from the figures, fertility was set to be constant over five-year age groups of women and closely resembles fertility patterns observed in the 2010 Malawi DHS (National Statistical Office - NSO/Malawi and ICF Macro, 2011). Additionally, we use three distinct discrete hazards for each five-year time period with ${}_1q_0$ and ${}_1q_a$ for $a = 1, \ldots, 4$ roughly following the North regional model life table. Birth histories for a total of 5,000 women aged 15 to 49 were simulated. The distribution of ages roughly followed the 2010 Malawi DHS. We will consider two surveys taken in 2010, with one survey containing FBH information for 1,000 women and the other survey containing SBH information for the remaining 4,000 women.

Birth histories for 5,000 women were simulated on a discrete, yearly time scale. For simplicity, we allowed the year prior to the survey to be completely observed and did not allow for births during the survey year, which follows the simple example provided in Section 2.1. Thus, children could be born at any point prior to and including $t_s - 1$ (when the woman was aged $m_s - 1$), and could die in t_s where t_s is the year of the survey and m_s is the age of the woman at time of survey.

Figure 5.8 illustrates how FBH data were simulated for a woman who was 25 at the time of the survey. In the top left panel, when the woman is 15, the probability she gives birth is $f(15)$ (fertility does not change over time). In this example, she does not give birth. In the middle top panel, when the woman is 16, the probability she gives birth is $f(16)$. Here, she does give birth. In the following year (top right panel), the probability the woman gives birth is $f(17)$, and the probability the child dies is ${}_1q_0 = q_0(1)$. As the woman and her children age, we observe her to have 3 children at ages 16, 18, and 23. One child dies between age 1 and 2 and another dies between age 2 and 3. Her other child survives through the time of

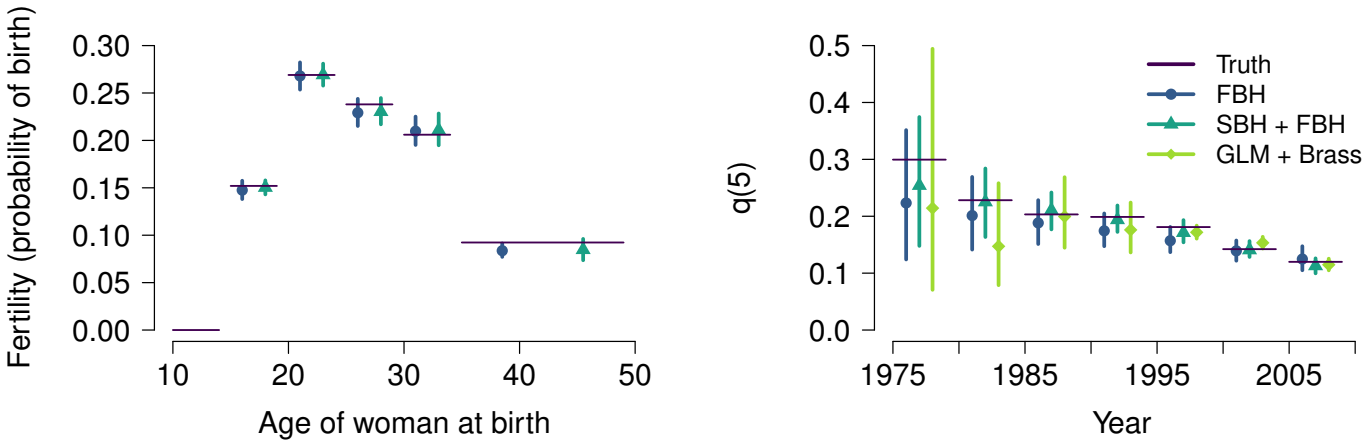

Figure 5.6: Fertility probabilities (left), and U5MR (q(5)) over time (right) used in the simulation. Horizontal solid lines indicate the truth. Points indicate posterior medians and vertical lines indicate 95% credible intervals using only FBH data and both FBH and SBH data using our model (circles and triangles, respectively). Diamonds are estimates of q(5) obtained by combining the Brass method fit on SBH data and a binomial logistic (GLM) fit to FBH data by period.

the survey. For women with FBH data, this information is completely observed. For women with SBH data, we only observe the total number of children the woman had and the number of those children that died (in this example, 3 births and 2 deaths).

We take as our model for fertility, $\text{logit}(f(m, \boldsymbol{x}(t))) = \beta_{c[m]}$, where the probability of birth depends on the mother's age, via 5 factors,

$$c[m] = \begin{cases} 1 & m = 15, \ldots, 19 \\ \vdots & \vdots \\ 4 & m = 30, \ldots, 34 \\ 5 & m = 35, \ldots, 49 \end{cases} \tag{5.4}$$

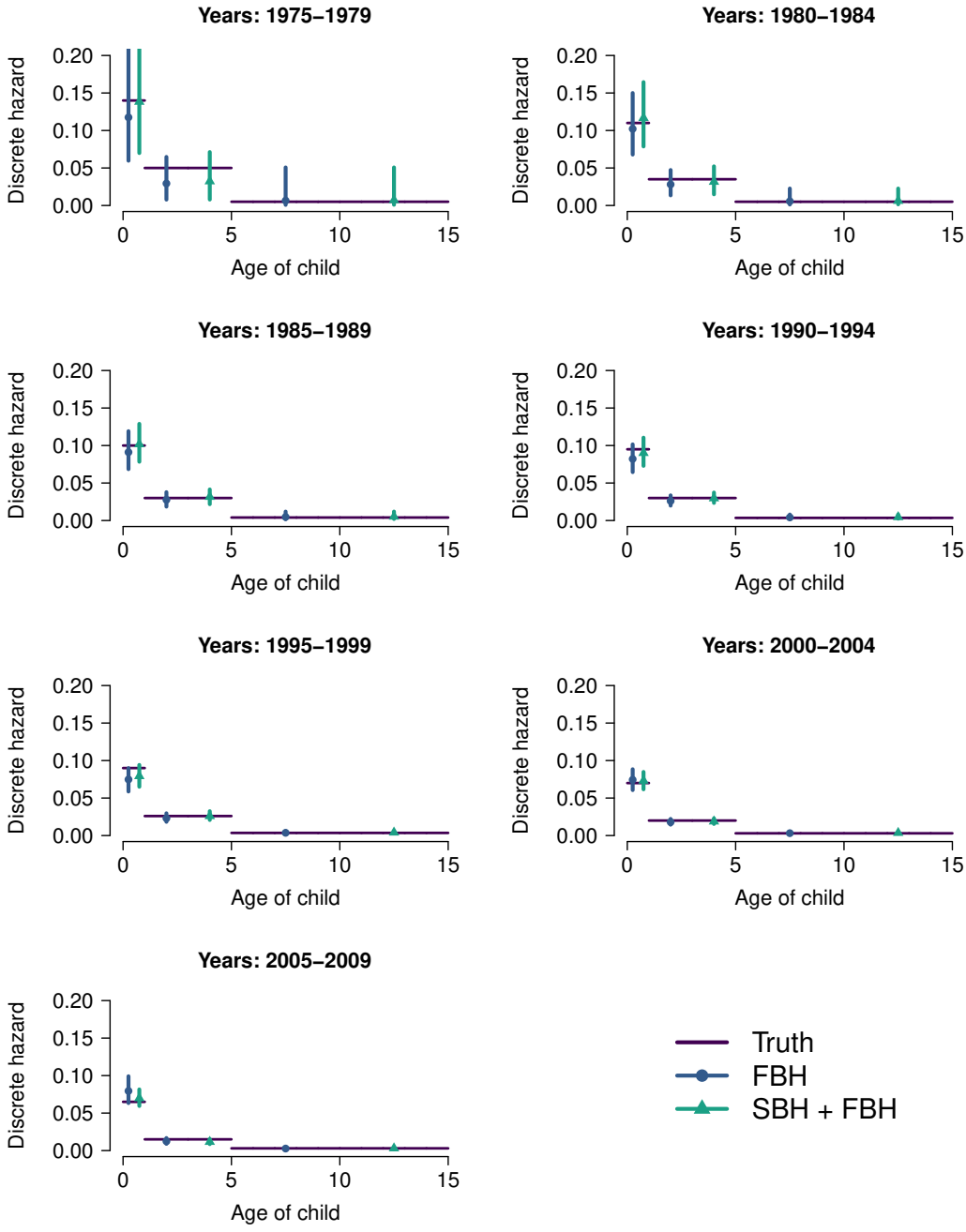

Figure 5.7: Discrete hazards by time period. Note: the probability of death within one year after age 5 was considered to be constant (only up to age 15 is plotted). Horizontal solid lines indicate the truth. Points indicate posterior medians and vertical lines indicate 95% CIs using only FBH data and both FBH and SBH data.

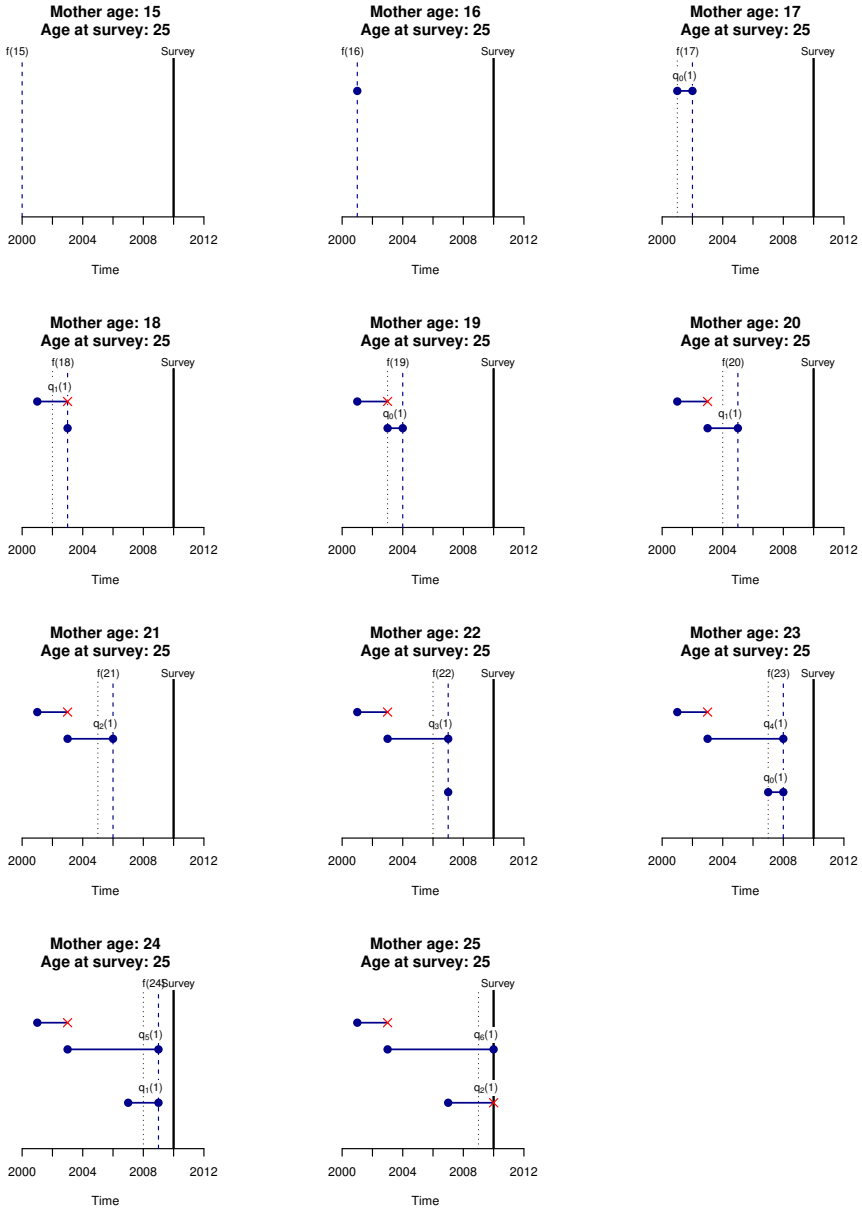

Figure 5.8: Illustration of the data generating mechanism along with relevant probabilities. Suppose a woman is 25 at the time of the survey in 2010 and suppose $f(m) > 0$ for $m \geq 15$. Starting at the top right and proceeding left and down are panels following her and any children she has forward through time starting at age 15. The blue dashed line represents the current year and black dotted line represents the prior year. Blue circles represent births and survival, red "x"s represent deaths.

and assign independent priors $\beta_{c[m]} \sim \mathrm{N}(0, 10^2)$. For mortality, we define 3 age groups

$$b[a] = \begin{cases} 0 & a = 0 \\ 1 & a = 1, \ldots, 4 \\ 2 & a = 5, \ldots \end{cases} \tag{5.5}$$

and set $\mathrm{logit}({}_1q_a(\boldsymbol{x}(t))) = \xi_{b[a]}(p) + \beta_{b[a]}$ where $\xi_{b[a]}(p)$ is a temporal random effect that follows a random walk of order 2 (RW2) model that depends on a precision κ. We take the penalized complexity prior for precision as a hyperprior for κ (Simpson et al., 2017).

Several different analyses are considered. First, a complete case analysis is performed (i.e., using the 1,000 women with FBH data). Second, we use our proposed DA approach and include the SBH data on 4,000 women and combine with the FBH data. In both analyses, we use the aforementioned models and use a Hamiltonian Monte Carlo (HMC) algorithm (Neal, 2011) to sample from the posterior during the parameter update step. This approach is explained further for the Malawi example in Section 5.5, specifically Section 5.5.2. Finally, we follow Section 5.3 and combine the Brass method fit on SBH data with estimates from fitting a logistic regression model to the FBH data; that is,

$$\begin{aligned} Z_a(t)|{}_1q_a(t) &\sim \text{Bernoulli}({}_1q_a(t)), \\ \mathrm{logit}({}_1q_a(t)) &= \beta_{c[a],p}, \end{aligned}$$

where a separate model is fit for each 5-year time period p.

Results are depicted in Figures 5.6 and 5.7. Uncertainty from the method that uses Brass is substantially greater in earlier periods. This is because the SBH data only contributes to the 3 most recent time periods. However, this approach tends to produce reasonable estimates in the more recent time periods (diamond points). This is interesting since the Brass method is based on assuming a demographic life model, which may not hold in this simulation. However, since many of the numbers used in the simulation are based on observed fertility and mortality data, this likely explains the good performance of Brass.

Table 5.2: Summary measures for fertilities, in the simulation. Ratio of CIs are relative to the width of the FBH intervals.

	Absolute Bias ×100		Width of CIs ×100		Ratio of CIs
	FBH	SBH + FBH	FBH	SBH + FBH	
15–19	0.44	0.16	1.9	1.5	0.77
20–24	0.10	0.015	2.9	2.3	0.81
25–29	0.87	0.76	2.9	2.8	0.97
30–34	0.38	0.50	3.0	3.4	1.12
35–49	0.85	0.76	1.4	2.3	1.59

To assess accuracy we use Absolute Bias $= |\hat{\psi}^{(M)} - \psi|$ where ψ is the truth for a generic parameter (i.e., $f(m)$ or ${}_5q_0(t)$) and $\hat{\psi}^{(M)}$ is an estimate of ψ (e.g. posterior median) obtained from approach/model M. To compare measures of uncertainty we take the ratio of the 95% credible/confidence intervals (CIs). Results are in Tables 5.2 and 5.3. In general, absolute bias and the width of the uncertainty intervals is lower when we include SBH data.

5.5 Application to Central Malawi

5.5.1 Model

Based on exploratory analysis and well known differences in fertility across different covariate groups (National Statistical Office - NSO/Malawi and ICF Macro, 2011), we propose the following model for fertility,

$$\text{logit}\left(f(m, \boldsymbol{x}(t))\right) = \xi_{c[m]}(p) + \beta_m + \beta_{strata} \tag{5.6}$$

where $\xi_{c[m]}(p)$ is a mother's age group specific RW2 in (roughly) 5-year time periods p. We take $c[m]$ to be the same as in the simulation (see Equation (5.4)). However, since we observe women in the available FBH data to have given birth between ages $9 - 48$, we adjust $c[m]$ accordingly: another group for women aged $9 - 14$ is added and the oldest age

Table 5.3: Summary measures for hazards, in the simulation. Ratio of CIs are relative to the width of the FBH intervals.

	Absolute Bias ×100			Width of CIs ×100	(Ratio of CIs)	
	FBH	SBH + FBH	GLM + Brass	FBH	SBH + FBH	GLM + Brass
1975–79	7.3	4.3	8.5	22	23 (1.0)	42 (1.9)
1980–84	2.6	0.29	8.1	13	12 (0.94)	18 (1.4)
1985–89	1.5	0.72	0.37	7.7	6.5 (0.84)	12 (1.6)
1990–94	2.4	0.45	2.3	5.8	4.6 (0.81)	8.7 (1.5)
1995–99	2.3	0.93	0.92	4.4	3.9 (0.88)	2.3 (0.52)
2000–04	0.29	0.088	1.1	3.6	2.8 (0.78)	2.1 (0.58)
2005–09	0.54	0.70	0.50	4.2	2.6 (0.62)	2.1 (0.49)

group is truncated at age 48. Since older women are not observed in earlier time periods, women over age 35 are grouped together for the RW2 model. There are 11 time periods p: $1964-1969, 1970-1974, 1975-1979, \ldots, 2010-2014, 2015-2019$. We include a fixed effect for each age m (ages $9-11$ were grouped together because of many 0 counts). Thus, women in age groups defined by $c[m]$ have the same trend in fertilities, but different overall levels of fertility by age. A fixed effect for strata (urban and rural) is also included.

For the mortality model, we will assume for simplicity that the probability of death within one year is the same for children ages 5 and older. We propose the following model for mortality,

$$\begin{aligned} Z_a(t) \mid {}_1q_a^\star(\boldsymbol{x}(t)) &\sim \text{Bernoulli}({}_1q_a^\star(\boldsymbol{x}(t))) \\ {}_1q_a^\star(\boldsymbol{x}(t)) &= k(t;s) \times {}_1q_a(\boldsymbol{x}(t)) \\ \text{logit}\left({}_1q_a(\boldsymbol{x}(t))\right) &= \beta_{SBH,strata} + \beta(c[a], \boldsymbol{x}(t)) \\ \beta(c[a], \boldsymbol{x}(t)) &= \xi_{b[a]}(t) + \beta_{c[a]} + \beta_{district} + \beta_{strata} \end{aligned} \tag{5.7}$$

where $k(t;s)$ is a term for HIV bias that varies by year t and depends on the year the survey

was conducted. The term $\beta_{SBH,strata}$ allows for bias in SBH (census data) by urban/rural. In an exploratory analysis, we found that women living in rural areas in the census reported a higher proportion of their children dead than similarly aged women in the surveys taken either side of that time period (see Figure 5.1). When making predictions, we do not include the HIV or SBH bias terms. We include a RW2 in years, $\xi_{b[a]}(t)$, by age group $b[a]$ (see Equation (5.5)). We include fixed effects for strata and district and also for age group with

$$c[a] = \begin{cases} 0 & a = 0 \\ 1 & a = 1 \\ \vdots & \vdots \\ 4 & a = 4 \\ 5 & a = 5, \ldots \end{cases}.$$

Therefore, the trend in logit hazards is the same for ages 1–4, but each age has its own level of mortality.

5.5.2 Computation

In the simulation, we had supposed that women could not give birth in the year of the survey, thus a woman's fertility history and child's life trajectory are fully observed up to the time of the survey. Clearly, this is not a reasonable simplification for applying our method to the Malawi data. We now describe the special considerations regarding the year of the survey.

Calendar time is first defined on yearly scales, based on the month and year of the survey. In order to more accurately align with the woman's age and to reduce the number of corrections (as described later), we divide time into the most recent 6 months prior to the survey and 12 month intervals preceding that. For example, if a survey was taken in February 2009, the time intervals would look like the following:

- September - December 2008, January and February 2009

- September - December 2007, January - August 2008
- ...

Based on the month and year reported for the birth of any children, we determine in what time interval the child was born.

The next step is to align the age of the woman during each interval. This is based on assuming that, on average, the woman will be the age in years reported plus six months. For example, if a woman reports that she is 35 at the time of survey, on average she will be 35.5. Thus, the most recent interval will correspond to when she is 35, second most recent when she was 34, etc.

Finally, we discretize the intervals by assigning them a year according to the year of survey. The most recent time period would correspond to the year of the survey and each subsequent interval would be one year prior. Since we smooth over time, the misalignment between the time intervals and assigned year is unlikely to be a major issue. Note that if the survey was taken in June there would be no misalignment.

We will need to make corrections to the model based on the yearly discretization of time. Suppose we are in the most recent time interval (which corresponds to the year of the survey). Since this interval only corresponds to 6 months, the fertility, or probability of birth should be, approximately half of what it normally is. This is assuming constant fertility throughout the year. Mathematically, if $m = m_s$ and $t = t_s$ then $f(m_s, t_s) = 0.5f(m, t)$. All other years will not require this correction.

Finally, we need to consider a hazard correction. Suppose a child is born in the most recent time period. At most, they have 6 months of exposure time. On average, we might expect them to be born at the midpoint of that time interval, meaning we are interested in the probability they die within 3 months. Clearly, this is less than the probability of dying within 12 months, but the factor is likely to be $> 1/4$ since children are at highest risk in the first month of life. We base the correction on what is observed in the FBH data and use 0.65.

Now, consider a child born in the second most recent interval. On average, the child will be born halfway through the interval. Since the last interval is only 6 months, their exposure time is, on average, one year. We will not use any adjustments for the earlier periods. However, for a small portion of women they have children that, in terms of the newly defined discretized time, die after the survey. For example, consider a child born 1.5 years before the survey, which using our discretized intervals is 1 year before the survey. In our setup, they are only at risk for one year, but could theoretically die after 1 year and we would observe this death in the data. To balance the fact that some children in the interval may only be observed for 6 months, we will count this as a death within the first year. While the approach is not perfect and will result in slight biases, the impact should be minimal and relatively restricted to the most recent time periods.

We again implement an HMC algorithm (Neal, 2011) for the parameter update step, which requires the negative log posterior and corresponding gradient for our fertility and mortality models. For the fertility model, define $\boldsymbol{X}_f^* = \left[\boldsymbol{X}_f, \boldsymbol{Z}_f\right]$ where $\boldsymbol{X}_f$ is a $n \times p_{X_f}$ matrix (n is the number of observations) and $\boldsymbol{Z}_f$ is a $n \times p_{Z_f}$ matrix, and $\boldsymbol{\theta}^f = \left[\boldsymbol{\beta}_f^\top, \boldsymbol{\xi}_f^\top\right]^\top$ where $\boldsymbol{\beta}_f$ is a column vector of length p_{X_f} and $\boldsymbol{\xi}_f$ is a column vector of length p_{Z_f}. In the simulation, we do not have $\boldsymbol{Z}_f$ or $\boldsymbol{\xi}_f$. For the hazard model, define $\boldsymbol{X}_h^* = \left[\boldsymbol{X}_h, \boldsymbol{Z}_h\right]$ where $\boldsymbol{X}_h$ is a $n \times p_{X_h}$ matrix, and $\boldsymbol{Z}_h$ is a $n \times p_{Z_h}$ matrix, and $\boldsymbol{\theta}^h = \left[1, \boldsymbol{\beta}_h^\top, \boldsymbol{\xi}_h^\top\right]^\top$ where $\boldsymbol{\beta}_h$ is a column vector of length p_{X_h} and $\boldsymbol{\xi}_h$ is a column vector of length p_{Z_h}. Since we will be using a random walk model in time, which does not specify an overall level, we opt to drop the unidentifiable terms (the age group specific intercepts) from Models (6) and (7) instead of imposing a constraint.

For general $\boldsymbol{X}^*$ and $\boldsymbol{\theta}$, with data model $Y_i \mid p_i \sim \text{Binomial}(N_i, k_i p_i)$ where k_i is known (this includes the HIV bias in the mortality model and an adjustment for the year of the

survey) and $\text{logit}(p_i) = \boldsymbol{X}_i^{*\top}\boldsymbol{\theta}$, we have

$$\begin{aligned} f(\boldsymbol{y}) &\propto \prod_{i=1}^{n}(k_i p_i)^{y_i}(1-k_i p_i)^{N_i - y_i} \\ \log f(\boldsymbol{y}) &= \text{const} + \sum_{i=1}^{n}\{y_i \times \text{logit}(k_i p_i) + N_i \times \log(1-k_i p_i)\} \\ &= \text{const} + \sum_{i=1}^{n}\Big[y_i \times \left(\boldsymbol{X}_i^{*\top}\boldsymbol{\theta}\right) + (N_i - y_i) \times \log\{1 + (1-k_i)\exp(\boldsymbol{X}_i^{*\top}\boldsymbol{\theta})\} \\ &\qquad - N_i \times \log\left\{1 + \exp(\boldsymbol{X}_i^{*\top}\boldsymbol{\theta})\right\}\Big]. \end{aligned}$$

We assign independent priors,

$$\begin{aligned} \boldsymbol{\beta}_f &\sim N\big(0, \sigma_\beta^2 \mathbf{I}_{p_{X_f}}\big), & \boldsymbol{\beta}_h &\sim N\big(0, \sigma_\beta^2 \mathbf{I}_{p_{X_h}}\big), \\ \boldsymbol{\xi}_f &\sim N\left(0, (\kappa_f \boldsymbol{K})^{-1}\right), & \boldsymbol{\xi}_h &\sim N\left(0, (\kappa_h \boldsymbol{K})^{-1}\right), \end{aligned}$$

where $\mathbf{I}_p$ is taken to be a $p \times p$ identity matrix, $\sigma_\beta^2 = 100$, and $\boldsymbol{K}$ is the random walk of order 2 (RW2) precision matrix, scaled such that the generalized variance of $\boldsymbol{\xi}$ is 1, following Sørbye and Rue (2014).

We use independent penalized complexity priors for κ_f and κ_h, that is for general κ,

$$\begin{aligned} \pi(\kappa) &= \frac{\lambda}{2}\kappa^{-3/2}\exp(-\lambda\kappa^{-1/2}), \\ \lambda &= -\frac{\log(\alpha)}{2} \end{aligned}$$

where $P(\sigma > u) = \alpha$ with $\sigma = 1/\sqrt{\kappa}$ (Simpson et al., 2017). We set $\alpha = 0.01$ and $u = 0.5$.

Therefore, the negative log posterior (up to a constant) for the hazard model is

$$\begin{aligned} U_h = &-\boldsymbol{y}_h^\top\left(\boldsymbol{X}_h^*\boldsymbol{\theta}_h\right) + \boldsymbol{N}_h^\top \log\left\{\mathbf{1} + \exp(\boldsymbol{X}_h^*\boldsymbol{\theta}_h)\right\} - (\boldsymbol{N}_h - \boldsymbol{y}_h)^\top \log\{\mathbf{1} + (\mathbf{1} - \boldsymbol{k}_h) \circ \exp(\boldsymbol{X}_h^*\boldsymbol{\theta}_h)\} \\ &+ \frac{1}{2\sigma_\beta^2}\boldsymbol{\beta}_h^\top\boldsymbol{\beta}_h + \frac{\kappa_h}{2}\boldsymbol{\xi}_h^\top\boldsymbol{K}\boldsymbol{\xi}_h - \frac{\text{rank}(\boldsymbol{K})}{2}\log(\kappa_h) + \frac{3}{2}\log(\kappa_h) + \lambda\kappa_h^{-1/2} \end{aligned}$$

where $\boldsymbol{y}_h$ is a vector containing the number of children that died, $\boldsymbol{N}_h$ is a vector containing the number of children at risk, $\boldsymbol{k}_h$ is a vector containing the multiplication factors k_{hi}, $\text{rank}(\boldsymbol{K})$ is the rank of matrix $\boldsymbol{K}$, and $\circ$ denotes element-wise multiplication. Define $\eta_{\kappa,h} =$

$\log(\kappa_h)$ so that,

$$U_h = -\boldsymbol{y}_h^\top (\boldsymbol{X}_h^*\boldsymbol{\theta}_h) + \boldsymbol{N}_h^\top \log\{\mathbf{1} + \exp(\boldsymbol{X}_h^*\boldsymbol{\theta}_h)\} - (\boldsymbol{N}_h - \boldsymbol{y}_h)^\top \log\{\mathbf{1} + (\mathbf{1} - \boldsymbol{k}_h) \circ \exp(\boldsymbol{X}_h^*\boldsymbol{\theta}_h)\} + \frac{1}{2\sigma_\beta^2}\boldsymbol{\beta}_h^\top\boldsymbol{\beta}_h + \frac{\exp(\eta_{\kappa,h})}{2}\boldsymbol{\xi}_h^\top\boldsymbol{K}\boldsymbol{\xi}_h - \frac{\text{rank}(\boldsymbol{K})}{2}\eta_{\kappa,h} + \frac{1}{2}\eta_{\kappa,h} + \lambda\exp\left(-\frac{1}{2}\eta_{\kappa,h}\right).$$

The gradient is then,

$$\frac{\partial U_h}{\partial \boldsymbol{\beta}_h} = -\boldsymbol{y}_h^\top\boldsymbol{X}_h + \boldsymbol{N}_h^\top\{(\boldsymbol{G}_h\mathbf{1}_{p_{X_h}}^\top) \circ \boldsymbol{X}_h\} - (\boldsymbol{N}_h - \boldsymbol{y}_h)^\top\{(\boldsymbol{G}_{h,k}\mathbf{1}_{p_{X_h}}^\top) \circ \boldsymbol{X}_h\} + \frac{1}{\sigma_\beta^2}\boldsymbol{\beta}_h^\top$$

$$\frac{\partial U_h}{\partial \boldsymbol{\xi}_h} = -\boldsymbol{y}_h^\top\boldsymbol{Z}_h + \boldsymbol{N}_h^\top\{(\boldsymbol{G}_h\mathbf{1}_{p_{Z_h}}^\top) \circ \boldsymbol{Z}_h\} - (\boldsymbol{N}_h - \boldsymbol{y}_h)^\top\{(\boldsymbol{G}_{h,k}\mathbf{1}_{p_{Z_h}}^\top) \circ \boldsymbol{Z}_h\} + \exp(\eta_{\kappa,h})\boldsymbol{\xi}_h^\top\boldsymbol{K}$$

$$\frac{\partial U_h}{\partial \eta_{\kappa,h}} = \frac{\exp(\eta_{\kappa,h})}{2}\boldsymbol{\xi}_h^\top\boldsymbol{K}\boldsymbol{\xi}_h - \frac{\text{rank}(\boldsymbol{K})}{2} + \frac{1}{2} - \frac{1}{2}\lambda\exp\left(-\frac{1}{2}\eta_{\kappa,h}\right)$$

where $\boldsymbol{G}_h = \text{expit}(\boldsymbol{X}_h^*\boldsymbol{\theta}_h)$ (a column vector of length n) and $\boldsymbol{G}_{h,k} = \frac{(\mathbf{1}-\boldsymbol{k}_h) \circ \exp(\boldsymbol{X}_h^*\boldsymbol{\theta}_h)}{\mathbf{1}+(\mathbf{1}-\boldsymbol{k}_h) \circ \exp(\boldsymbol{X}_h^*\boldsymbol{\theta}_h)}$ (also a column vector of length n).

The negative log posterior and gradient is similar for the fertility model. In the simulation the vector $\boldsymbol{k}$ is set to $\mathbf{1}$.

5.5.3 Results

We fit four different models to the data. First, we computed weighted estimates based only on FBH data. Second, we used the approach described in Section 5.3 to combine the direct estimates with indirect estimates obtained using the Brass method at the district-level on SBH data. Both of these were done on the period (5-year time scale) since there were insufficient data to support separate hazards on the yearly time scale. Third, we fitted (5.6) and (5.7) to all available FBH data (DHS and MICS). Finally, we used our DA approach to incorporate SBH data and fitted models (5.6) and (5.7) to all data.

Trace plots for the mortality model suggest convergence (see Appendix A.3). In general, posteriors for the hazards look similar for the age of the child, $\beta_{c[m]}$. However, we do see some differences in posteriors for the district-level effects. In terms of the urban effect, the posterior medians are -0.397 (95% CI: -0.458, -0.338) for FBH data only and -0.377 (95%

CI: -0.458, -0.338) for FBH + SBH. In terms of the bias terms, the posterior median for the indicator for SBH data is 0.171 (95% CI: 0.153, 0.191) and for the indicator for SBH and urban data is -0.228 (95% CI: -0.294, -0.162). The interpretation is that the SBH data result in an increase in the yearly hazard odds of 18.6% in rural areas and a decrease in the yearly hazard odds of 5.5%.

Results for fertility are found in Figures 5.9–5.11 and show similar results between FBH only and FBH and SBH analyses. Fertility tends to be higher in rural areas, to be highest among women in their twenties, and to decrease over time for all age groups.

Figures 5.12–5.14 (left and middle panels) compare posterior medians and 95% credible intervals for the U5MR. The posteriors tend to be similar when we add in the census data and use our proposed model. We aggregate to five year time periods in order to compare with the other methods. To do this, we sample births using the fertility samples for our FBH only and FBH and SBH models. We then use them to average the samples of $q(5)$ over strata and years within 5-year time periods.

Define m to be woman's age, str to be strata (rural, urban), d to be district, t to be the year, and p to be the 5-year time period. Define $\text{FP}(m;t;d;str)$ to be the total number of women of age m at time t in district d and strata s. Denote the number of births these women experience (children ever born) as $\text{CEB}(m;t;d;str)$. Let $j = 1, \ldots, J$ index the samples. To obtain samples of $q(5)$ on the 5-year time period and district level, we employ the following steps:

1. Simulate $\text{CEB}(m;t;d;s)^{(j)} \sim \text{Binomial}(\text{FP}(m;t;d;s), f(m;p;d;s)^{(j)})$
2. Transform:

$$ {}_5q_0(p;d)^{(j)} = \sum_{t \in p} \sum_{str} {}_5q_0(t;d;str)^{(j)} \left\{ \frac{\sum_m \text{CEB}(m;t;d;str)^{(j)}}{\sum_{t \in p} \sum_{str} \sum_m \text{CEB}(m;t;d;str)^{(j)}} \right\}. $$

In the right panel of Figures 5.12–5.14, we see that the indirect estimates from Brass contribute to 3 time periods: 1990–1994, 1995–1999, and 2000–2004 and often result in higher $q(5)$. In general, the estimates are similar across the other three methods. Uncertainty is much higher when using the weighted estimates on FBH data as compared to using our

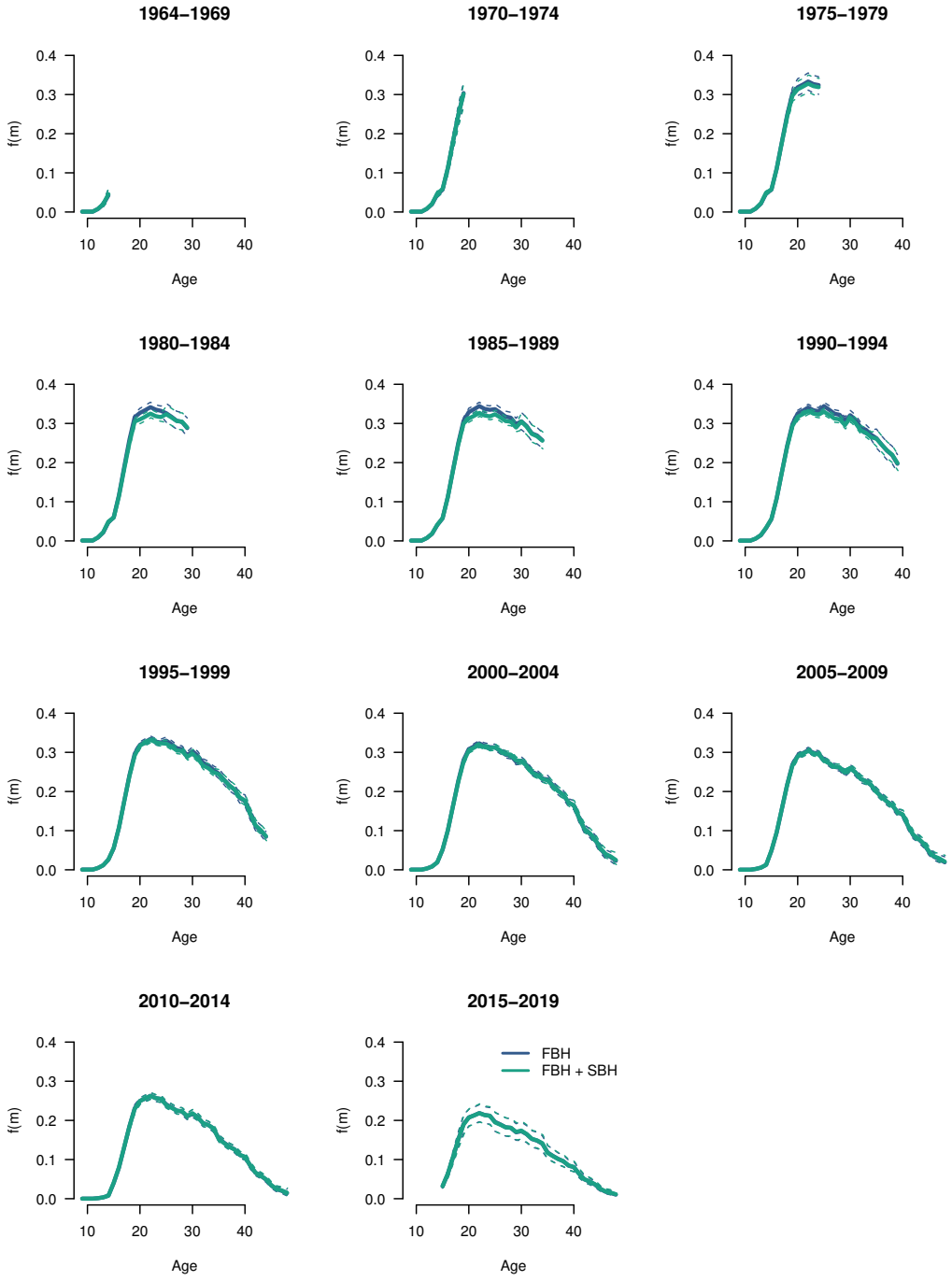

Figure 5.9: Fertility probabilities in rural areas by age of woman across 5-year time periods. Solid lines: posterior medians. Dashed lines: 95% credible intervals. Ages where there are no observations are excluded.

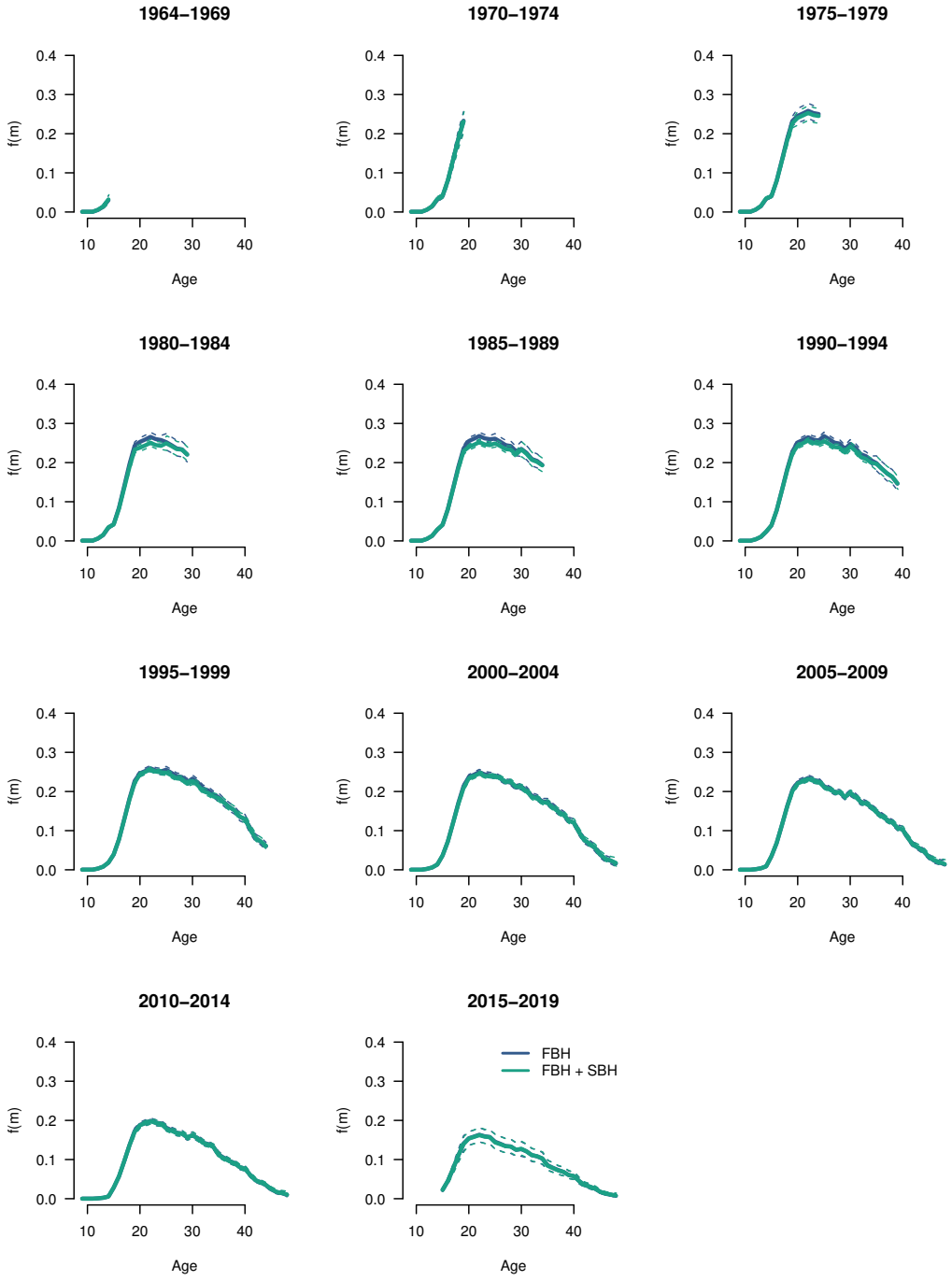

Figure 5.10: Fertility probabilities in urban areas by age of woman across 5-year time periods. Solid lines: posterior medians. Dashed lines: 95% credible intervals. Ages where there are no observations are excluded.

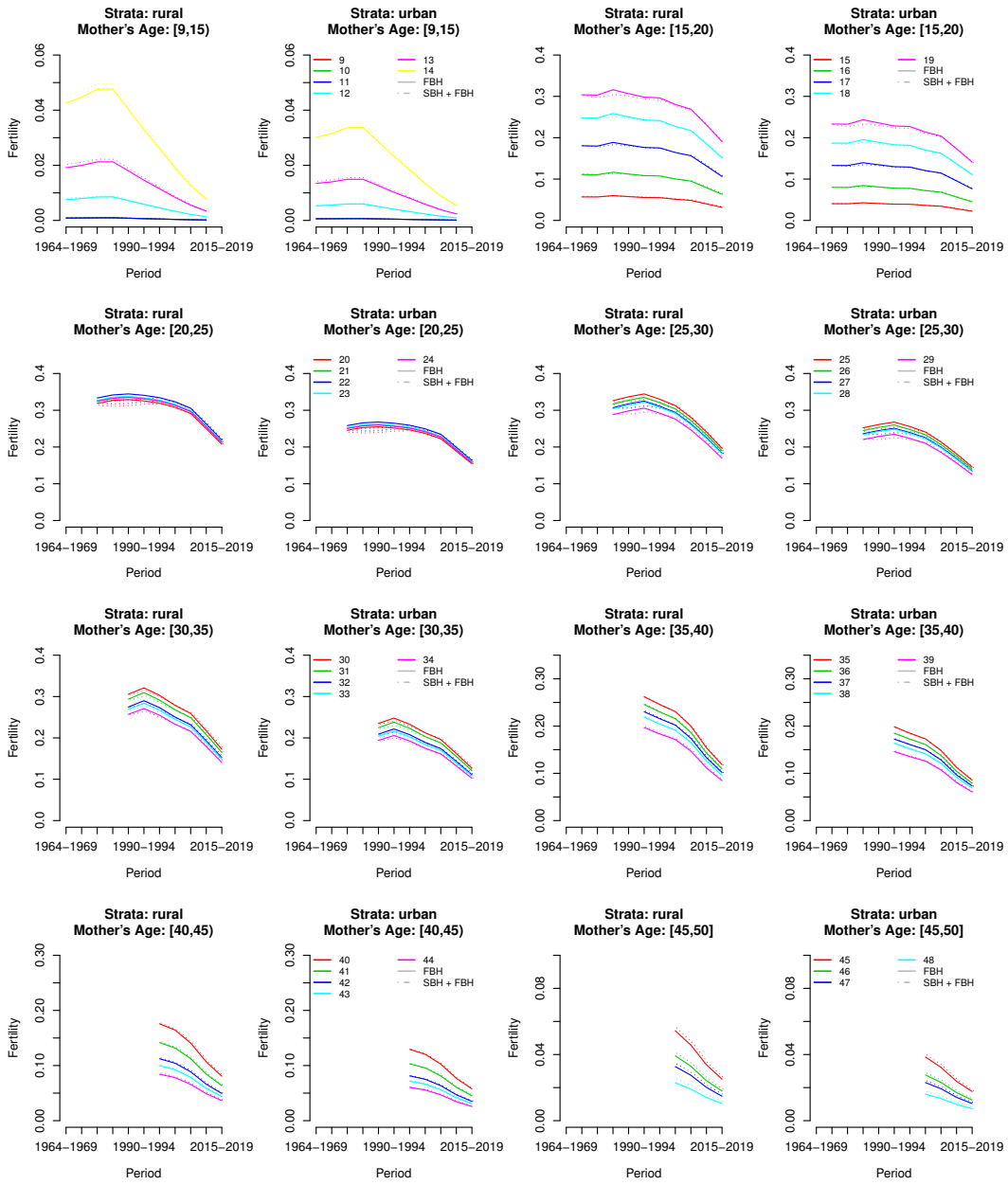

Figure 5.11: Fertility probabilities (posterior medians) by 5-year time periods across woman's age. Ages where there are no observations are excluded.

proposed model.

To assess the accuracy of the four models, we divide the FBH data into training and holdout data. We split the 2006 MICS and 2010 DHS into two roughly equal-sized groups: a training and validation set. The validation set consisted of 331 clusters and the remaining 314 were used in the training set (see Table 5.1). We refit the models to the training data (2004 DHS, 2008 Census, 2013 MICS, 2015 DHS, and training sets from the 2006 MICS and 2010 DHS) and assess the accuracy of estimation using a weighted mean squared error (MSE). Denote the estimate of logit of U5MR in district d time period p for model M, $Y_{dp}^{(M)}$. For the weighted and weighted and Brass estimates, we use the asymptotic variance of the estimate and for the Bayesian analyses, the posterior variances. These four estimates and variances are compared with the weighted estimates of the logit of U5MR from the holdout clusters, y_{dp} taken to be the truth. We assess the accuracy by time period over all 9 districts by using a weighted MSE:

$$\text{MSE}(p)^{(M)} = \sum_{d=1}^{9} w_{dp} \left\{ E\left(Y_{dp}^{(M)} - y_{dp}\right)\right\}^2 + \sum_{d=1}^{9} w_{dp} \text{Var}\left(Y_{dp}^{(M)}\right)$$

where $p = \{1985{-}1989, 1990{-}1994, 2000{-}2004, 2005{-}2009\}$ and $w_{dp} = \hat{V}_d(p)^{-1} / \sum_{d=1}^{9} \hat{V}_d(p)^{-1}$, with $\hat{V}_d(p)^{-1}$ denoting the variance of the weighted estimates in district d and time period p. This allows us to upweight the MSE in districts and periods where the "truth" is more certain.

Models involving direct estimates tend to perform worse (Table 5.4 and Figure 5.15). Figure 5.15 also provides a sense of the uncertainty in the numbers used as the "truth" and displays the estimate and 95% uncertainty interval of $Y_{dp}^{(M)}$ for each model M on the probability scale, where we approximate the distribution of $Y_{dp}^{(M)}$ with a normal distribution. Adding SBH via the Brass method or the DA approach tends to improve the predictions. In our smoothed model, the variance component of the MSE is 55% lower when SBH data are included; however, the bias tends to be higher.

We also calculate the percentage relative error (PARE; Table 5.5) of the probabilities

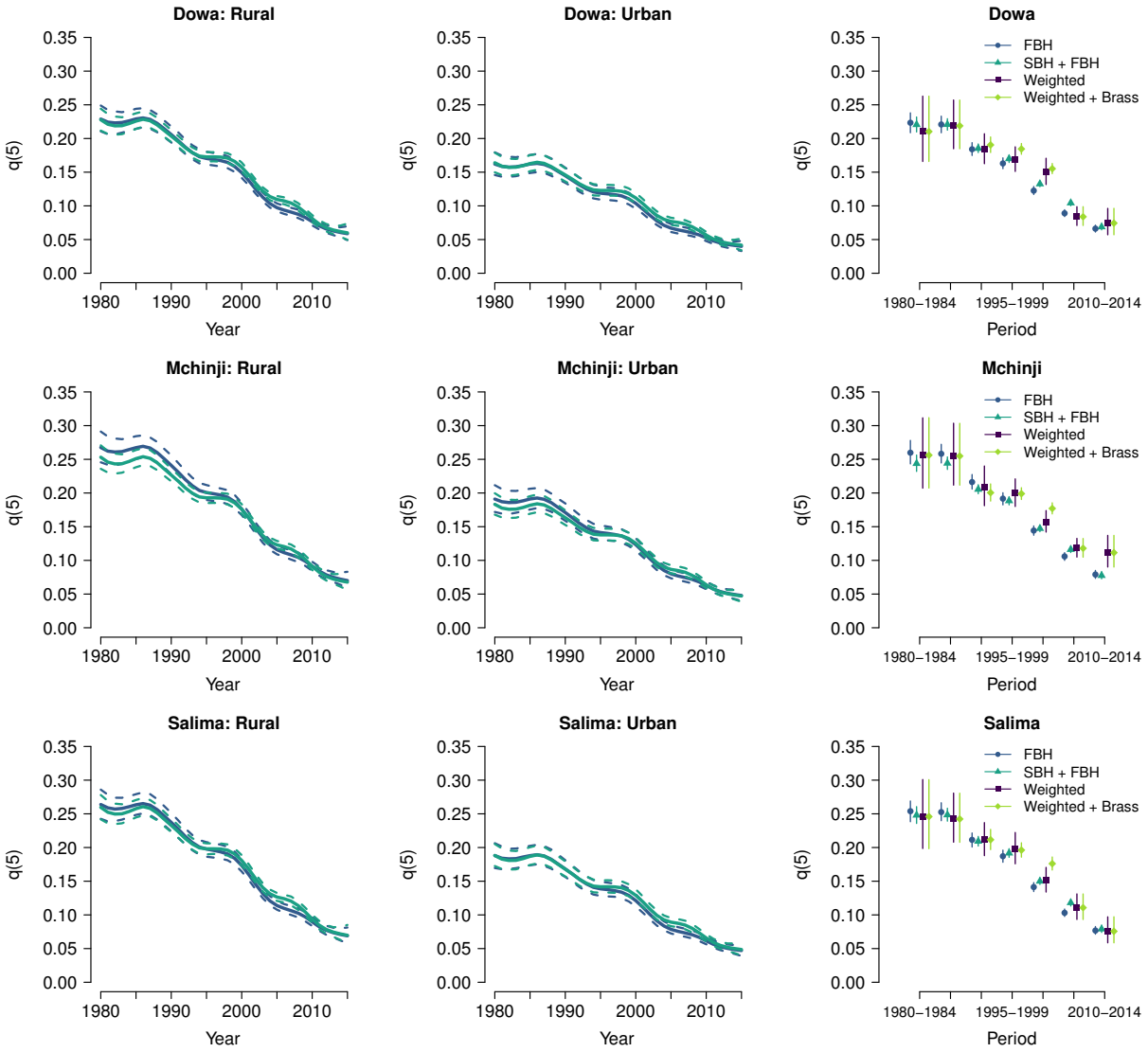

Figure 5.12: Left and middle panel: posterior medians (points and solid lines) and 95% credible intervals (dashed lines) for U5MR ($q(5)$). Right panel: estimates of U5MR (points) and 95% uncertainty interval (lines).

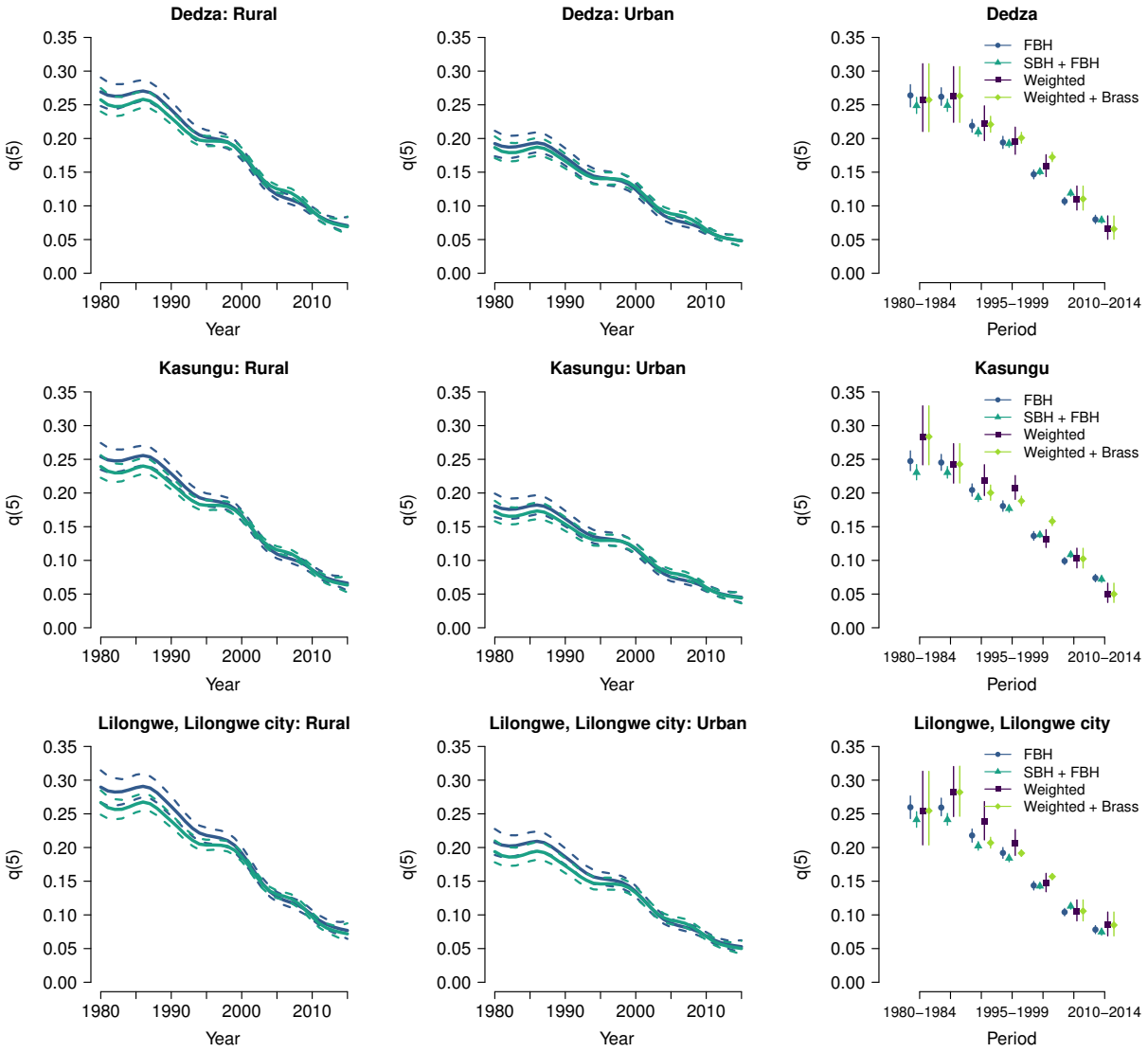

Figure 5.13: Left and middle panel: posterior medians (points and solid lines) and 95% credible intervals (dashed lines) for U5MR ($q(5)$). Right panel: estimates of U5MR (points) and 95% uncertainty interval (lines).

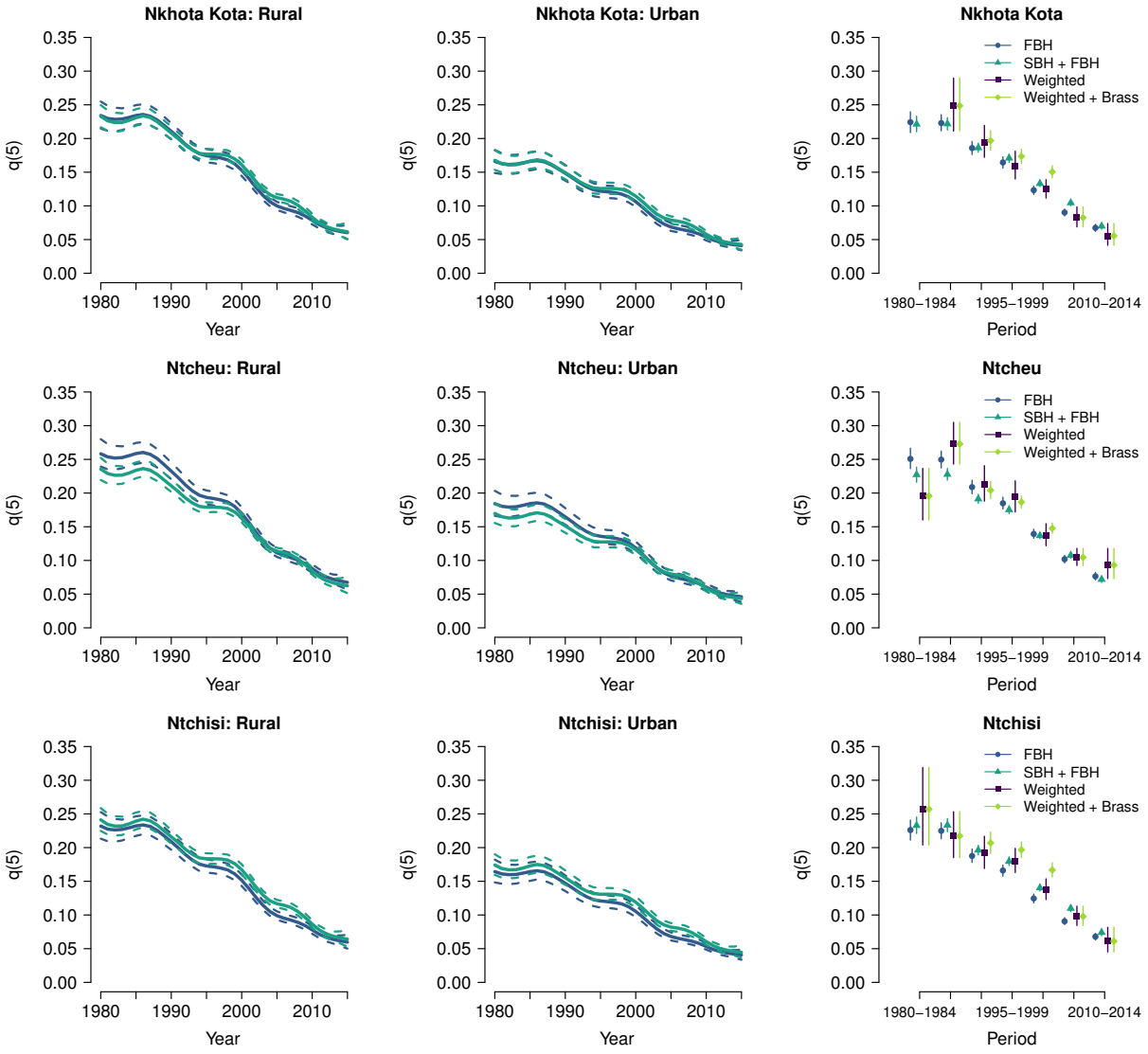

Figure 5.14: Left and middle panel: posterior medians (points and solid lines) and 95% credible intervals (dashed lines) for U5MR ($q(5)$). Right panel: estimates of U5MR (points) and 95% uncertainty interval (lines).

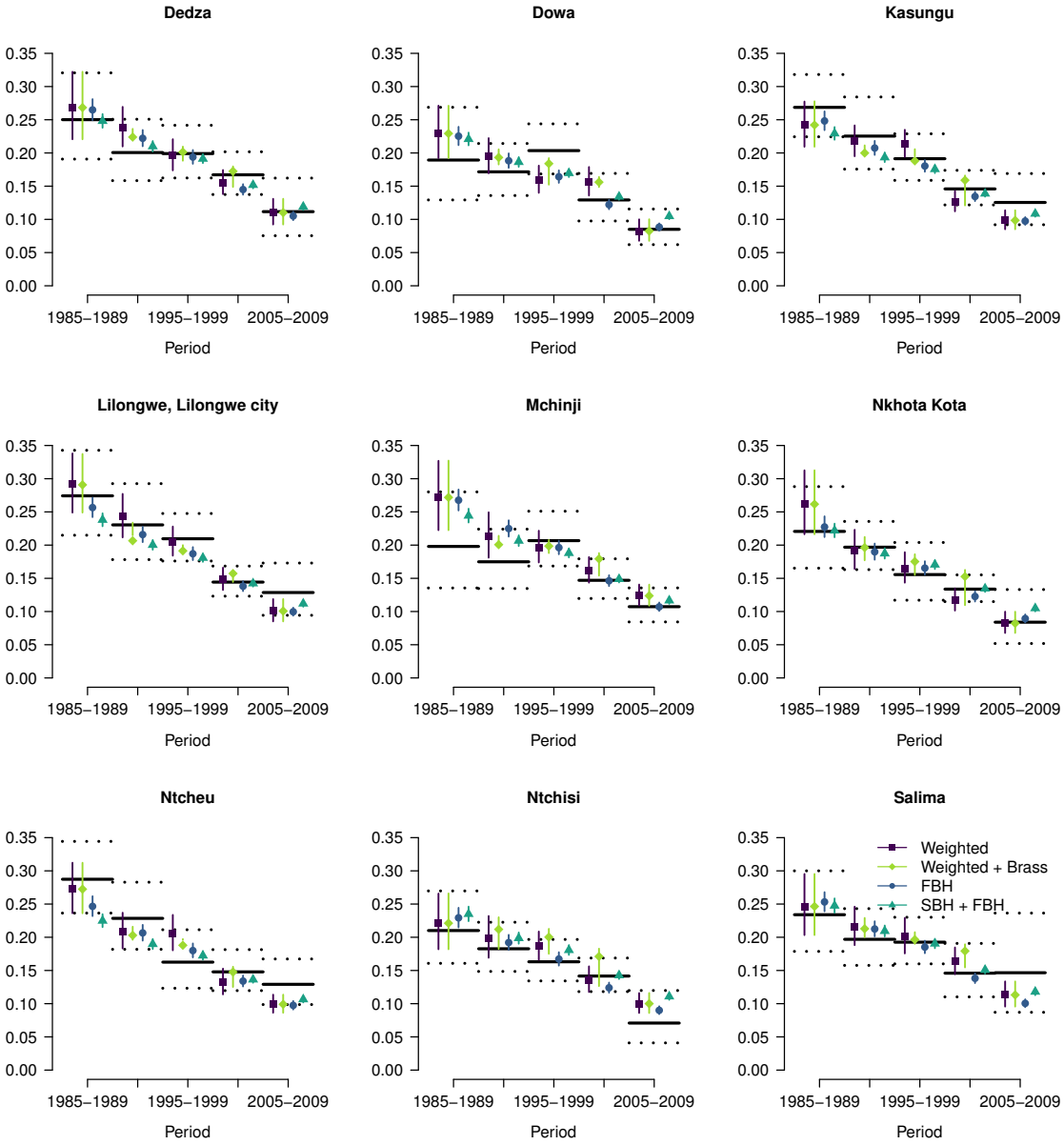

Figure 5.15: Solid black lines are the direct estimates obtained from the holdout data and the dashed black lines are the corresponding 95% confidence intervals. Points are expit $\left(E[Y_{dp}^{(M)}]\right)$ and the vertical lines are the corresponding 95% uncertainty intervals constructed using $\text{Var}(Y_{dp}^{(M)})$ and the quantiles of a normal distribution and then transformed to the probability scale.

Period	Weighted Estimates	Weighted Estimates + Brass	Smoothed Model: FBH	Smoothed Model: FBH + SBH
1985–1989	3.84	3.83	2.69	4.13
1990–1994	2.65	2.08	1.86	2.04
1995–1999	2.92	1.64	1.56	1.78
2000–2004	2.13	2.62	1.21	0.33
2005–2009	5.83	5.82	5.10	4.19
Average	3.07	2.79	2.06	1.99

Table 5.4: Mean Squared Error ×100 by Model.

$q(5)$.

$$\text{PARE}(p)^{(M)} = \frac{1}{9}\sum_{d=1}^{9} w_{dp}\frac{|\hat{q}_{dp}(5)^{(M)} - q_{dp}(5)|}{q_{dp}(5)}$$

where $w_{dp} = \hat{V}_d(p)^{-1}/\sum_{d=1}^{9}\hat{V}_d(p)^{-1}$ and $\hat{V}_d(p)^{-1}$ is the variance of the direct estimates. These (unweighted) values are shown for each district d and period p in Figure 5.16. The substantive results are changed slightly with the DA procedure producing a PARE of 8.8%, compared to 9.0% for the FBH analysis only, the direct estimate PARE was 10.5% and the direct and indirect combined was 11.1%.

5.6 Discussion

We have presented a novel framework for analyzing SBH data in the context of U5MR estimation. Using our approach, data from a variety of sources that contain birth history information at varying degrees of detail can be analyzed together to produce an estimate of U5MR over time with a measure of uncertainty. Our method falls under the umbrella of data augmentation, where we impute FBH information for the SBH data.

We make some simplifying assumptions since we opt to work on a discrete time scale;

Table 5.5: Percent Absolute Relative Error of $q(5)$ estimates ($\times 100\%$) by Model.

Period	Weighted Estimates	Weighted Estimates + Brass	Smoothed Model: FBH	Smoothed Model: FBH + SBH
1985–1989	10.5	10.4	10.4	12.4
1990–1994	9.82	10.2	9.21	9.91
1995–1999	9.69	8.18	7.42	9.03
2000–2004	9.41	12.7	7.62	3.24
2005–2009	16.2	16.2	13.8	16.7
Average	10.5	11.1	8.96	8.76

however, modifications to our approach can be made. For example, multiple births and having a lag time between births could be accommodated by adjusting equation (5.1) and the model for fertility. Further, the approach could be extended to a monthly time scale or continuous time. We also define the fertility patterns by the age of the women rather than time since fist birth or marriage; however, both of these variations are possible with our framework.

In our application, we included fixed effects for strata and districts and did not have these vary in time. One future consideration is using information on migration and allowing the district of the children and women to change over time. Other terms, such as a spatial random effect could be included. Additionally, terms for mother's age, number of births, and if available, sex of the child, which may be informative of U5MR, could also be included. However, producing an overall estimate of U5MR at an administrative level would be more complicated since the estimate would need to be marginalized over the age, birth number, and sex population distributions. For understanding individual-level associations, including these terms into the hazard model would be useful and straightforward to implement in our

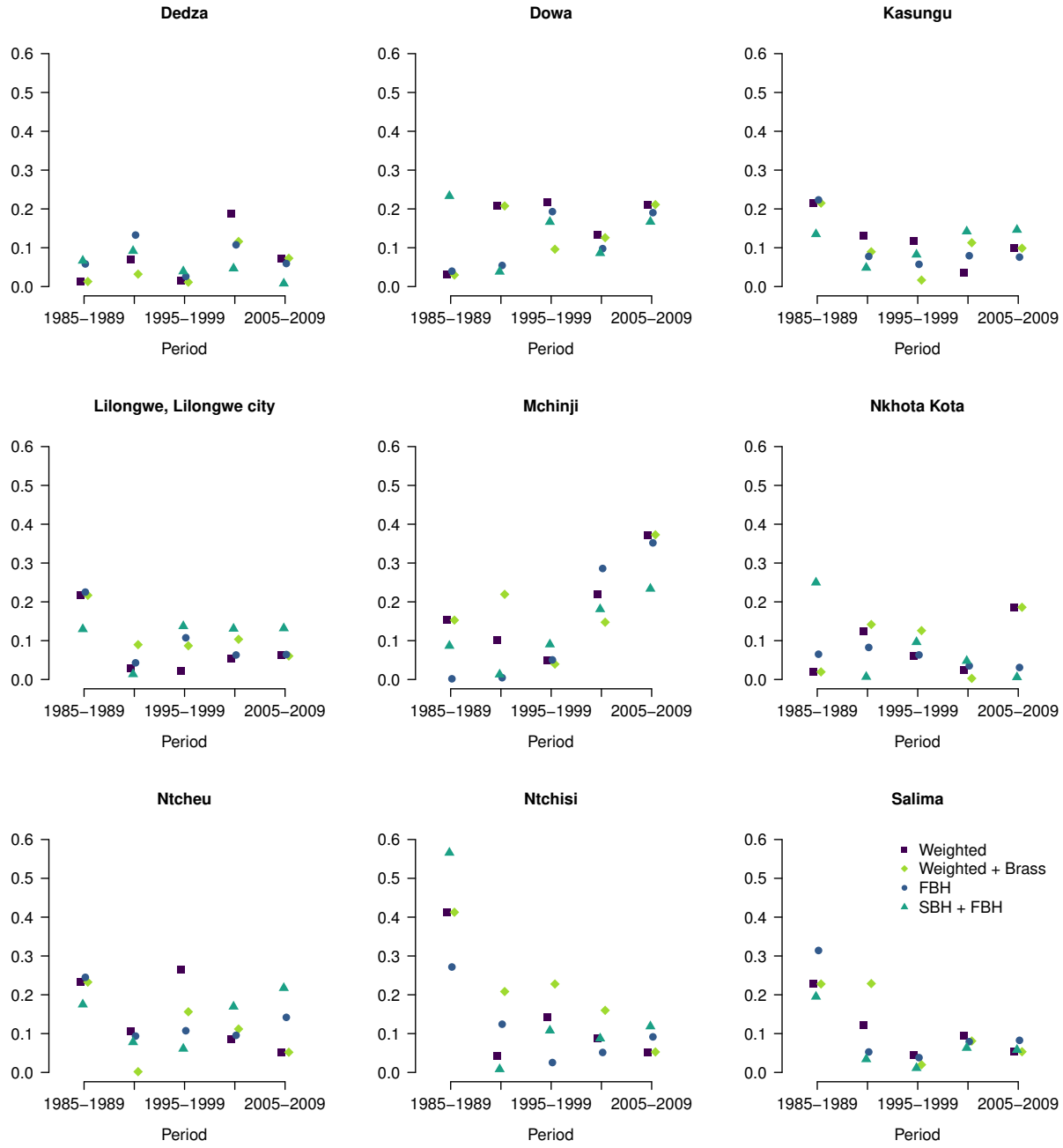

Figure 5.16: Percent absolute relative error by district and period.

approach.

In our simulation, we found that by incorporating SBH data uncertainty is reduced. When

we applied our method to actual data, adjustments to our model were made to accommodate differences between the SBH census data and FBH survey data. In general, we assumed that FBH data were the "gold standard" and adapted our model to reflect this. A fixed effect term for SBH data by strata was added into the mortality model to adjust for the proportion of deaths as reported by SBH women being higher in rural areas than expected (compared to FBH women). We also used a portion of the FBH data as holdout data to validate the models. It is not immediately clear what the truth is in this scenario, however. There are well known biases in both types of data, making model assessment difficult (Silva, 2012). For SBHs, women may omit live births since they are not asked to systematically recall every birth. In the DHS and MICS, women are asked detailed questions about all of their children (including those that no longer live with them) to limit this type of bias. In FBHs, women may misreport when births and deaths occurred. Furthermore, additional questions on pregnancy and postnatal care are asked if a woman has had recent births, thus displacement of births or omission of recent births may occur. Special consideration for the data sources and outcome of interest is needed when developing the fertility and hazard models, and this is highly context specific.

Chapter 6

A MODEL-BASED VARIANT OF THE BRASS METHOD

6.1 Introduction

This chapter is a continuation of Chapter 5. The reader is referred to Section 5.1 for a more thorough background on modeling the under-five mortality rate (U5MR) in countries where data is derived from surveys and censuses.

The United Nations' Sustainable Developmental Goals emphasize the use of subnational estimates when it comes to the U5MR. Subnational estimates of U5MR are an important measure of health across a country and are used to inform public policy. Unfortunately, countries where U5MR is highest often lack vital registration systems that track births and deaths. Instead, data is derived from household surveys and censuses. Many surveys, such as the Demographic and Health Surveys (DHS), provide full birth history (FBH) data, which contains the dates of birth and death (if applicable) of the children of surveyed women. FBH data can be used to obtain time-varying subnational estimates of U5MR (in regions where enough FBH data is collected). Many censuses (and some surveys) collect summary birth history (SBH) data. Here, the available birth history data is the number of births and deaths of surveyed women. In both FBH and SBH data, various covariates are provided, such as the age of the woman. No temporal information on when the births and deaths occurred is provided in SBH data; therefore, inclusion of these data is more challenging.

Historically, variants of the Brass method (Brass, 1964; Coale and Trussell, 1977; Feeney, 1980; Trussell, 1975) are used to incorporate SBH data, with uncertainty being incorporated via the jackknife (Pedersen and Liu, 2012). The Brass method utilizes model life tables and simulation to obtain an equation that maps observed death and birth counts present in SBH data to an estimate of under-five mortality with a corresponding reference time. Rajaratnam

et al. (2010) refine this method and use observed FBH data in place of model life tables. Hill et al. (2015) propose 2 alternatives, including the Birth History Imputation method, where SBH women are matched to FBH women who are of approximately the same age, number of births, and number of deaths. They also describe the Cohort Change method, which requires data from SBH surveys taken 1-2 years apart. They posit that the change in number of children that died and number of children born will largely be driven by the U5MR and thus leverage these observed quantities in the SBH data to derive an estimate of U5MR. In Chapter 5, we propose a model-based approach where the unknown birth and death times in SBH data are simulated and then combined with FBH data to obtain estimates of U5MR in a data augmentation (DA) framework. While this method allows for flexible inclusion of covariates in a model and can accommodate bias in surveys, it is computationally demanding.

In this chapter, we propose a computationally efficient approach using a Poisson approximation and remove the simulation step in the DA procedure. We fit the model using the `R` package Template Model Builder (Kristensen, 2014; Kristensen et al., 2016), which allows one to quickly obtain empirical Bayes (EB) type estimates and a measure of uncertainty. We begin with Sections 6.2 and 6.3 by describing the method, relevant notation, and computational aspects. Via a simulation in Section 6.4, we study the performance of our method. In Section 6.5, we apply the method to data from Malawi. Finally, we end with a conclusion and a discussion of future work in Section 6.6.

6.2 Poisson Approximation

We adopt notation similar to Chapter 5. We again work on a discretized, yearly scale, letting t be calendar time in years and t_s be the year the survey was taken.

Although we are ultimately interested in modeling mortality (death) rates, in our method it is important to consider fertility (birth) rates for reasons that will become clear. Let $f(m, \boldsymbol{x}(t))$ denote the fertility rate, or the probability a woman gives birth at age m and year t with $\boldsymbol{x}(t)$ containing the covariates at time t associated with fertility. For mortality, we use a discrete hazards model. Let ${}_1q_a(\boldsymbol{x}(t)) = q_a(1, \boldsymbol{x}(t))$ denote the mortality rate at age

a, which is technically a probability, that is, the probability a child dies between age a and $a+1$ given survival to age a with $\boldsymbol{x}(t)$ containing the covariates at time t associated with mortality. The parameter of interest is $q(5, \boldsymbol{x}(t))$, the probability of death within 5 years *at time* t and covariates $\boldsymbol{x}(t)$. Here,

$$q(5, \boldsymbol{x}(t)) = 1 - \prod_{i=0}^{4}(1 - {}_1q_i(\boldsymbol{x}(t))). \tag{6.1}$$

Let $q^\star(a, \boldsymbol{x}(t))$ be the probability of dying within a years *given birth at time* t and covariates $\boldsymbol{x}(t)$; thus,

$$q^\star(a, \boldsymbol{x}(t)) = 1 - \prod_{i=0}^{a-1}(1 - {}_1q_i(\boldsymbol{x}(t+i))). \tag{6.2}$$

The difference between (6.1) and (6.2) for $a = 5$ is subtle, but crucial. In (6.1), we imagine a synthetic cohort of children that are born in year t and repeat year t for a total of 5 times. In (6.2), we can imagine a real cohort of children that are born in year t and followed up to year $t+5$.

For FBH data, where information is available on when births and deaths occurred, denote $Y(m,t)$ as an indicator for birth in year t to a woman of age m years. Let $Z_a(t)$ be an indicator that a child dies between ages a and $a+1$ and years t to $t+1$. A reasonable model for FBH data and the one used in Chapter 5 would be as follows:

$$Y(m,t)|f(m,\boldsymbol{x}(t)) \sim \text{Bernoulli}(f(m,\boldsymbol{x}(t))),$$
$$Z_a(t)|{}_1q_a(\boldsymbol{x}(t)) \sim \text{Bernoulli}({}_1q_a(\boldsymbol{x}(t))).$$

Now consider SBH data. For women who are m_s at the time of the survey, define $B_{m_s}(\boldsymbol{x}(t))$ to be the total number of children ever born to those women with covariates $\boldsymbol{x}(t)$, defined for all t. Further, for these women define the (unobserved) number of children born a years prior to the survey to be $B_{m_s}(a, \boldsymbol{x}(t))$. Note that $\sum_{a=0}^{m_s} B_{m_s}(a, \boldsymbol{x}(t)) = B_{m_s}(\boldsymbol{x}(t))$. Similarly, define $D_{m_s}(\boldsymbol{x}(t))$ to be the total number of children that ever died to women who are m_s at the time of the survey who have covariates $\boldsymbol{x}(t)$ for all t. Define the (unobserved) number of

children that were born a years prior to the survey *and* died by the time of the survey (i.e., died within a years) to be $D_{m_s}(a, \boldsymbol{x}(t))$. Again, $\sum_{a=0}^{m_s} D_{m_s}(a, \boldsymbol{x}(t))) = D_{m_s}(\boldsymbol{x}(t))$. Therefore, a reasonable model for SBH data using the unobserved data is,

$$D_{m_s}(a, \boldsymbol{x}(t))|B_{m_s}(a, \boldsymbol{x}(t)), q^\star(a, \boldsymbol{x}(t)) \sim \text{Binomial}(B_{m_s}(a, \boldsymbol{x}(t)), q^\star(a, \boldsymbol{x}(t))).$$

Approximating the Binomial with a Poisson,

$$D_{m_s}(a, \boldsymbol{x}(t))|B_{m_s}(a, \boldsymbol{x}(t)), q^\star(a, \boldsymbol{x}(t)) \sim \text{Poisson}(B_{m_s}(a, \boldsymbol{x}(t))q^\star(a, \boldsymbol{x}(t))).$$

Suppressing the notation $\boldsymbol{x}(t)$, therefore,

$$D_{m_s}|B_{m_s}(a), q^\star(a) \sim \text{Poisson}\left(\sum_{a=0}^{m_s} B_{m_s}(a)q^\star(a)\right). \tag{6.3}$$

Finally, we sum over all $(m_s + 1)^{B_{m_s}}$ possible combinations of when births occurred for a given B_{m_s} to obtain a mixture distribution,

$$D_{m_s}|B_{m_s}, f(m), q^\star(a) \sim \sum_{\sum_{a=0}^{m_s} x(a)=B_{m_s}} \left[P(\boldsymbol{x}|B_{m_s}, f(m)) \quad \times \text{Poisson}\left(\sum_{a=0}^{m_s} x(a)q^\star(a)\right)\right]. \tag{6.4}$$

We approximate (6.4) with,

$$D_{m_s}|B_{m_s}, c_{m_s}(a), q^\star(a) \sim \text{Poisson}\left(B_{m_s}\sum_{a=0}^{m_s} c_{m_s}(a)q^\star(a)\right), \tag{6.5}$$

$$c_{m_s}(a, \boldsymbol{x}(t_s - a)) = \frac{f(m_s - a, \boldsymbol{x}(t_s - a))}{\sum_{a=0}^{m_s} f(m_s - a, \boldsymbol{x}(t_s - a))}.$$

In the simulation, we investigate the implications of using (6.5) by comparing to (6.3) when the distributions of SBH births prior to the survey are available. Further, since $c_{m_s}(a)$ is not often known ahead of time, we propose first fitting a fertility model to FBH data. From this, an estimate of $f(m, \boldsymbol{x}(t))$ can be determined, i.e., $\widehat{f}(m, \boldsymbol{x}(t))$, which can be transformed to $\widehat{c}_{m_s}(a, \boldsymbol{x}(t))$. We use this latter approach in our application to data from Malawi.

Note that

$$\begin{aligned} E[D_{m_s}|B_{m_s}, \boldsymbol{c}_{m_s}\boldsymbol{q}^\star] &= E[E\{D_{m_s}|B_{m_s}(a), \boldsymbol{c}_{m_s}, \boldsymbol{q}^*\}] \\ &= E\left[\sum_{a=0}^{m_s} B_{m_s}(a)q^\star(a)\Big|B_{m_s}, \boldsymbol{c}_{m_s}, \boldsymbol{q}^*\right] \\ &= B_{m_s}\sum_{a=0}^{m_s} c_{m_s}(a)q^\star(a) \end{aligned} \tag{6.6}$$

where $E[B_{m_s}(a)|B_{m_s}, \boldsymbol{c}_{m_s}] = B_{m_s}c_{m_s}(a)$. Thus, we have returned to a familiar equation, namely equation (11.2) in Preston et al. (2000), which forms the basis of the Brass approach. As noted in Chapter 5, the Brass method essentially treats (6.6) as deterministic, replacing the left side with the observed number of deaths. In our approach, we use a Poisson distribution for the total number of deaths conditional on births.

6.3 Computation

We use the `R` package, `TMB` (Kristensen, 2014; Kristensen et al., 2016) for computation. An overview of `TMB` can be found in Section 2.2.4. To use `TMB`, one must first specify the objective function in a C++ template file, which in our case, is the negative log posterior.

Let $\boldsymbol{z}$ denote the vector containing the realized values of $Z_a(t)$, $\mathbf{d}$ the vector containing the realized values of D_{m_s}, $\mathbf{b}$ the vector containing B_{m_s}, $\boldsymbol{c}$ the vector containing $c_{m_s}(a, \boldsymbol{x}(t))$, $\boldsymbol{\beta}$ the vector containing latent mortality fixed effects, $\boldsymbol{\gamma}$ the vector containing latent mortality random effects, and $\boldsymbol{\phi}$ the vector containing hyperparameters for the random effects. In our proposed approach, the negative log posterior is

$$f(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\phi}) = -\underbrace{\log p(\boldsymbol{z}|\boldsymbol{\beta}, \boldsymbol{\gamma})}_{\text{FBH contribution}} - \underbrace{\log p(\mathbf{d}|\mathbf{b}, \boldsymbol{c}, \boldsymbol{\beta}, \boldsymbol{\gamma})}_{\text{SBH contribution}} - \underbrace{\log p(\boldsymbol{\beta}) - \log p(\boldsymbol{\gamma}|\boldsymbol{\phi}) - \log(\boldsymbol{\phi})}_{\text{prior}}.$$

In our models, we consider several random effects where it will be necessary to impose sum-to-zero constraints, e.g., $\boldsymbol{\gamma}^\top \mathbf{1} = 0$. This is done by using equation (2.30) of Rue and Held (2005). Define $\boldsymbol{\gamma}^\star$ to be the unconstrained versions of the random effects with distribution, $\boldsymbol{\gamma}^\star \sim N(0, \boldsymbol{Q}^{-1})$. In particular, we will consider cases when $\boldsymbol{Q}$ refers to the precision matrix

of a second-order random walk (RW2) model and the intrinsic conditional autoregressive (ICAR) model. Since $\boldsymbol{Q}$ is rank deficient, a small offset is added to the diagonal to make it invertible and maintain sparsity of $\boldsymbol{Q}$, as suggested in Chapter 3.3 of Rue and Held (2005). Then,

$$\boldsymbol{\gamma} = \boldsymbol{\gamma}^\star - \boldsymbol{Q}^{-1}\mathbf{1}(\mathbf{1}^\top \boldsymbol{Q}^{-1}\mathbf{1})^{-1}(\mathbf{1}^\top \boldsymbol{\gamma}^\star). \tag{6.7}$$

Therefore,

$$f(\boldsymbol{\beta}, \boldsymbol{\gamma}^\star, \boldsymbol{\phi}) = \underbrace{-\log p(\boldsymbol{z}|\boldsymbol{\beta}, \boldsymbol{\gamma})}_{\text{FBH contribution}} - \underbrace{\log p(\mathbf{d}|\mathbf{B}, \boldsymbol{c}, \boldsymbol{\beta}, \boldsymbol{\gamma})}_{\text{SBH contribution}} - \underbrace{\log p(\boldsymbol{\beta}) - \log p(\boldsymbol{\gamma}^\star|\boldsymbol{\phi}) - \log(\boldsymbol{\phi})}_{\text{prior}},$$

with $\boldsymbol{\gamma}$ being a function of $\boldsymbol{\gamma}^\star$, defined in (6.7).

6.4 Simulation Study

6.4.1 Set-up

In our simulation, we suppose that there were two surveys providing birth history information, both taken in 2010. One survey provided FBH and a much larger survey (analogous to the census in our application to Malawi data) provided SBH. We used the geography of Kenya, which comprises 47 districts. In total, we simulated FBH for $47 \times 4,000$ women and SBH for $47 \times 20,000$ women with equal numbers in each region. A description of how to simulate birth histories can be found in Chapter 5.

The fertility rates, $f(m)$, used in the simulation are shown in purple in Figure 6.1. Fertility was assumed constant over space and time. As can be seen from the figure, fertility was also set to be constant over five-year age groups of women and resembles patterns observed in the 2010 Malawi DHS (National Statistical Office - NSO/Malawi and ICF Macro, 2011).

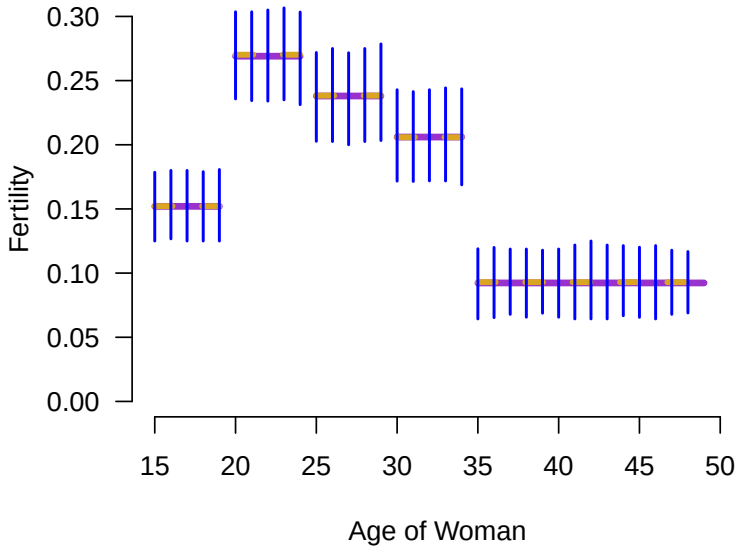

Figure 6.1: Blue: 5th and 95th percentile of observed fertility probabilities in SBH data (by region and surveyed woman's age). Purple: true underlying fertility probabilities. Yellow: estimated fertility probabilities from FBH data.

That is,

$$\text{logit}(f(m, \boldsymbol{x}(t))) = \beta_{c[m]},$$

$$c[m] = \begin{cases} 1 & m = 15, \ldots, 19 \\ 2 & m = 20, \ldots, 24 \\ \vdots & \\ 5 & m = 35, \ldots, 49 \end{cases} \tag{6.8}$$

We simulate deaths using 3 distinct discrete hazards, ${}_1q_0$, ${}_1q_a$ for $a = 1, \ldots, 4$, and ${}_1q_a$ for

$a = 5, \ldots$. These are a function of time p (5-year periods) and region r,

$$\text{logit}({}_1q_a(t;r)) = \beta_{c[a]} + \xi_{c[a]}(p) + U(r) + \epsilon(r),$$

$$c[a] = \begin{cases} 1 & a = 0 \\ 2 & a = 1, \ldots, 4 \\ 3 & a = 5, \ldots \end{cases}$$

where $\beta_{c[a]}$ are fixed effects and $\xi_{c[a]}(p)$, $U(r)$, and $\epsilon(r)$ are random effects. Specifically, $\xi_{c[a]}(p)$ follows a RW2 model with precision parameter κ_t, $U(r)$ follows an ICAR model with precision κ_s, and $\epsilon(r) \sim_{iid} N(0, 1/\kappa_\epsilon)$. Both the RW2 and ICAR have a sum-to-zero constraint for identifiability purposes. We fixed $\xi_{c[a]}(p)$ to values similar to those observed in our applications and simulated the structured and unstructured random effects.

We fit the model to FBH data only and considered three different model-fitting approaches for when SBH data was included:

1. $B_{m_s}(a)$ is known and (6.3) is used to include SBH data.
2. The true fertility, $c_{m_s}(a)$, is known and (6.5) is used to include SBH data.
3. The fertility is estimated from FBH data via logistic regression, and $\widehat{c}_{m_s}(a)$ is plugged in for $c_{m_s}(a)$ into (6.5).

Figure 6.1 includes the observed fertility in the SBH data, depicted as blue vertical lines showing the 5th and 95th percentile, and the estimated fertility from FBH data is shown in yellow.

In fitting the mortality model, we use the following as (independent) priors,

$$\beta_{c[a]} \sim N(0, 10^2), \qquad \kappa_t \sim \text{PCprior}(u = 1, \alpha = 0.01),$$
$$\kappa_s \sim \text{PCprior}(u = 1, \alpha = 0.01), \qquad \kappa_\epsilon \sim \Gamma(\text{shape} = 1, \text{scale} = 200),$$

where PCprior is the penalized complexity prior for precision as a hyperprior for κ (Simpson et al., 2017); see Chapter 5 for the exact form.

	True	FBH Only	FBH + SBH: 1	FBH + SBH: 2	FBH + SBH: 3
β_0	-2.25	-2.36 (-2.39, -2.33)	-2.36 (-2.39, -2.33)	-2.36 (-2.39, -2.33)	-2.36 (-2.39, -2.33)
β_1	-3.50	-3.51 (-3.54, -3.48)	-3.52 (-3.55, -3.48)	-3.52 (-3.55, -3.48)	-3.52 (-3.55, -3.48)
β_2	-5.50	-5.50 (-5.58, -5.41)	-5.50 (-5.58, -5.42)	-5.50 (-5.58, -5.42)	-5.50 (-5.58, -5.42)
κ_t	–	556 (159, 1920)	603 (176, 2070)	590 (172, 2070)	592 (173, 2060)
κ_s	125	228 (24, 1430)	237 (32, 1670)	230 (32, 1680)	231 (31, 1750)
κ_ϵ	150	147 (63, 344)	119 (55, 264)	119 (55, 258)	119 (55, 256)

Table 6.1: Comparison of estimates and 95% uncertainty intervals when using FBH only and FBH + SBH data where SBH data is incorporated using one of the 3 approaches.

6.4.2 Results

The three approaches to incorporating SBH yield very similar results (Table 6.1 and Figures 6.2 and 6.3) and are similar to FBH only results. All models underestimate the true intercept for the youngest age group 0–1. This is expected as the Poisson approximation to the Binomial is best when p is small and mortality is highest in the first year of life. Also, there exists some identifiability problems with $U(r)$ and $\epsilon(r)$ and so Figure 6.3 displays results for $U(r) + \epsilon(r)$.

We also derived estimates of U5MR and corresponding measures of uncertainty by region and time period, using a multivariate normal approximation. That is, defining $\widehat{\boldsymbol{\psi}}$ to be the estimates and $\widehat{\boldsymbol{\Sigma}}$ to be the inverse Hessian obtained from using `TMB`, we simulate 1,000 draws, from $\widehat{\boldsymbol{\psi}}^{(i)} \sim N\left(\widehat{\boldsymbol{\psi}}, \widehat{\boldsymbol{\Sigma}}\right)$, where $i = 1, \ldots, 1,000$. These realizations are then combined to give $\widehat{q(5)}^{(i)}$ for each region and period. Figures 6.4 and 6.5 visually depict the estimates with uncertainty (standard deviation), expressed using hatching. For presentation purposes, "SBH + FBH" refers to using the third approach for including SBH data (plugging in fertility estimates from FBH). Both models give similar estimates for U5MR and are close to the truth. Uncertainty is reduced when SBH data is incorporated. For reference, the estimated standard deviation (on the logit U5MR scale) in 1975–1979 was 14%–22% (mean 17%) higher when only FBH was used. In 2005–2009, the estimated standard deviation was

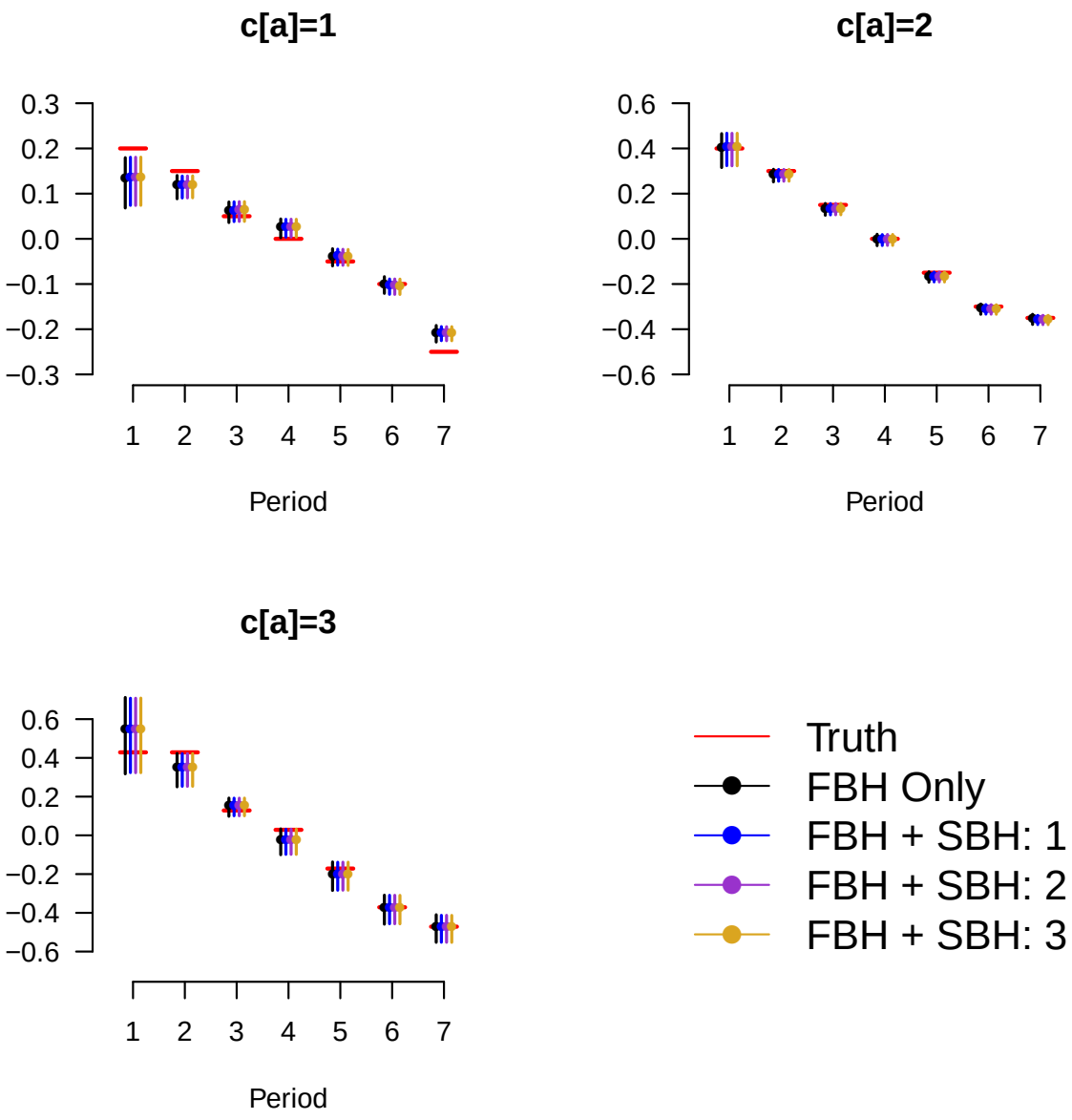

Figure 6.2: RW2 parameters, $\xi_{c[a]}(p)$, in mortality model with 95% uncertainty intervals.

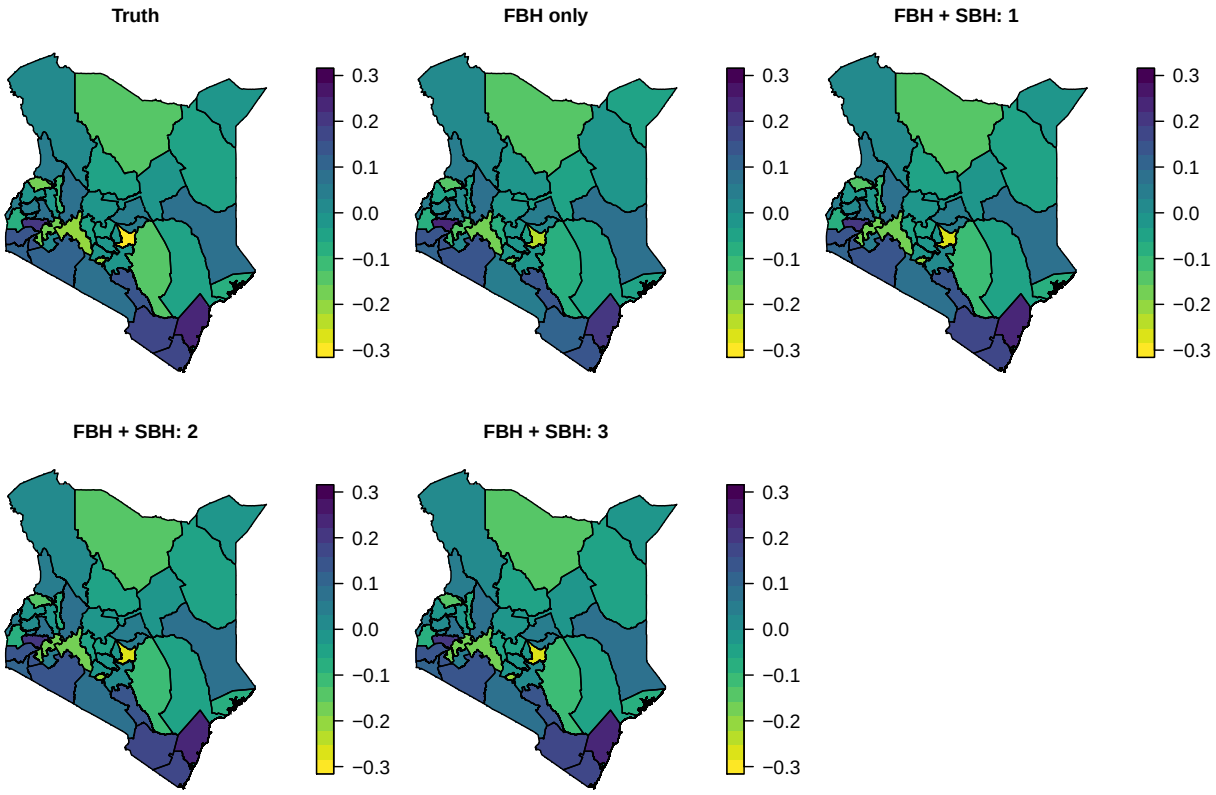

Figure 6.3: Structured and unstructured spatial random effect parameters, $U(r) + \epsilon(r)$, in mortality model. Plotted are estimates.

Survey	# Clusters	# Women	# Births	# Deaths
DHS 2004	521	11,698	35,883	6,534
MICS 2006	1,040	26,211	78,641	11,855
Census 2008		309,851	875,423	150,798
DHS 2010	849	23,020	72,301	11,343
MICS 2013	1,139	24,220	72,568	9,213
DHS 2015	850	24,562	68,074	7,235

Table 6.2: Data summaries, by survey.

96%–131% (mean 115%) higher when only FBH was used.

6.5 Application to Malawi Data

6.5.1 Data

We apply our new approach to data from Malawi. We use the same data sources as Chapter 5. The available data include 3 Demographic and Health Surveys (DHS) taken in 2004, 2010, and 2015; 2 Multiple Indicator Cluster Surveys (MICS) taken in 2006 and 2013; and 1 census taken in 2008 (Table 6.2). All 5 surveys contained FBH information, whereas the census contained only SBH information. Microdata is available on a 10% random sample of the census. In this analysis, we included data from all women who were 15–49 at the time of the survey and from all 26 districts of Malawi.

Births and time periods were aligned using the same process from Chapter 5, because births and deaths of children can occur in the year a survey was taken. Therefore, the year of survey is taken to be 6 months and every preceding time period is taken to be 1 year. It is convenient to assume that all births occur in the middle of the interval so that aside from births in the most recent time period, the exposure time to death will be in whole years. Adjustment terms are then only needed for the time period corresponding to when

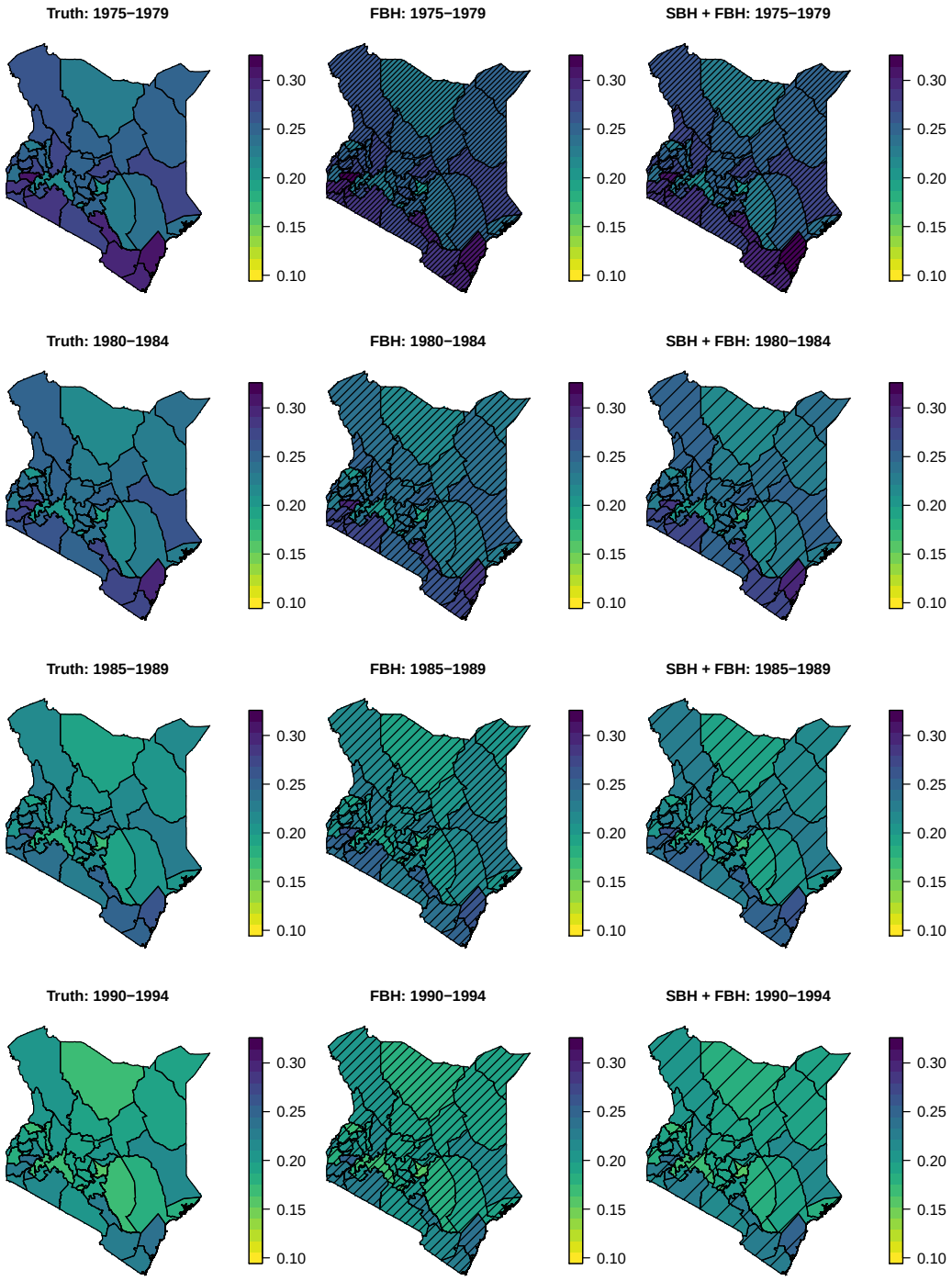

Figure 6.4: Truth and estimates of the U5MR in the first 4 periods by Kenyan counties. Uncertainty (standard deviation of logit U5MR) Denser hatching indicates more uncertainty.

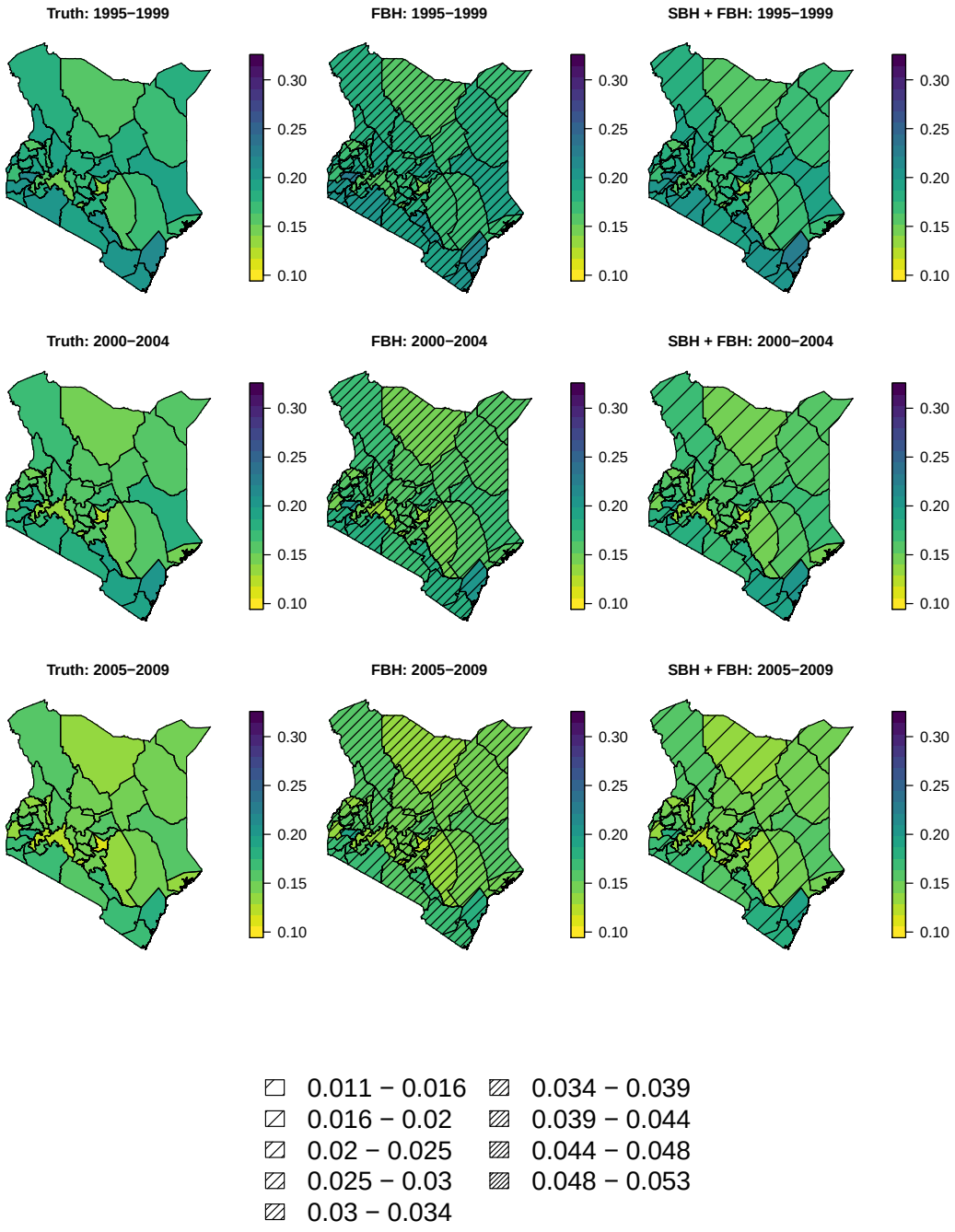

Figure 6.5: True and estimates of the U5MR in the 3 most recent periods by Kenyan counties. Uncertainty (standard deviation of logit U5MR) is represented by hatching. Denser hatching indicates more uncertainty.

the survey occurred.

6.5.2 Fertility Model

The following model is fit to FBH to derive estimates of fertility,

$$\text{logit}(f(m, \boldsymbol{x}(t))) = \beta_m + \beta_{strata} + \xi_{c[m]}(p) + U(r) + \epsilon(r)$$

where β_m are fixed effects for mother's age, β_{strata} is a fixed effect for strata, $\xi_{c[m]}(p)$ is a mother's age group (roughly 5-years, see (6.8)) specific RW2 model (with precision κ_t) in time in roughly 5-year time periods p, $U(r)$ is an ICAR model (with precision κ_s) and $\epsilon(r)$ is unstructured error on regions (i.e., $\epsilon(r) \sim_{iid} N(0, \kappa_\epsilon^{-1})$). This is similar to the fertility model (5.6) used in Chapter 5, but with terms for regions.

We take the following as priors,

$$\begin{aligned}
\kappa_t &\sim \text{PCPrior}(u = 0.5, \alpha = 0.01), \\
\kappa_s &\sim \text{PCPrior}(u = 0.5, \alpha = 0.01), \\
\kappa_\epsilon &\sim \Gamma(1, 0.0025).
\end{aligned}$$

From fitting this model, posterior medians for $\widehat{f}(m)$ are extracted by region, period and woman's age. This is then used to find $\widehat{c}(m)$, which are plugged into (6.5) for fitting the mortality model.

The results are slightly different from the application in Chapter 5. Here, we find the urban effect to be -0.21 in this analysis, compared to -0.36 from the application in Chapter 5 (posterior medians). Figure 6.6 shows the posterior median of the spatial random effect term, $U(r)+\epsilon(r)$. We had not included this term in our application in Chapter 5, and as can be seen from the figure, the central region neither the i.i.d. or ICAR term make a significant contribution in that region.

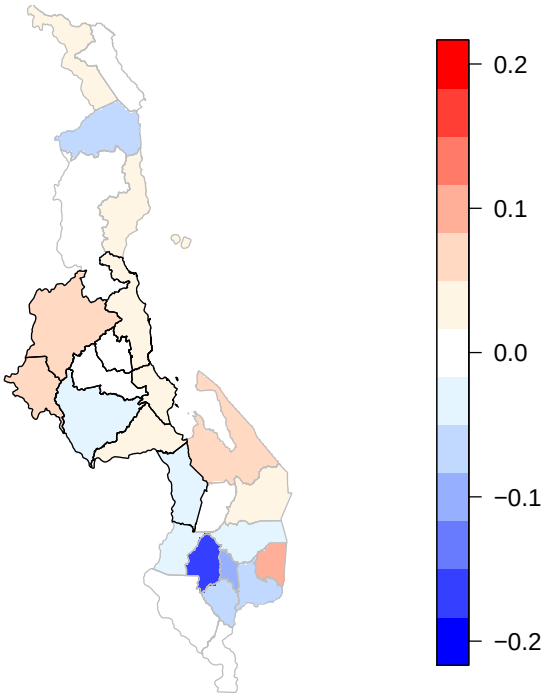

Figure 6.6: Posterior median of $U(r) + \epsilon(r)$ from fertility model. Outlined in black are the districts in the Central region of Malawi.

6.5.3 Mortality Model

Again, we use a model similar to the mortality model (5.7) from Chapter 5,

$$\log({}_1q_a(\boldsymbol{x}(t))) = \log \text{HIV}(p,s) + \beta_{SBH,strata} + \beta_{b[a]} + \xi_{c[a]}(p) + \beta_{strata} + U(r) + \epsilon(r,p) \quad (6.9)$$

where

$$b[a] = \begin{cases} 0 & a = 0, \\ 1 & a = 1, \\ \vdots & \vdots \\ 4 & a = 4 \\ 5 & a = 5, 6, \ldots \end{cases} \qquad c[a] = \begin{cases} 0 & a = 0, \\ 1 & a = 1, \ldots, 4 \\ 2 & a = 5, 6, \ldots \end{cases}$$

An offset term for HIV bias is used, $\log \text{HIV}(p,s)$; see Chapter 5 for a discussion on HIV bias and Figure 5.5 for the terms used. Another term is used to acknowledge bias in the census by urban/rural strata, $\beta_{SBH,strata}$. Both of these terms are not involved in predicting U5MR. Fixed effects are included for age-group of child and strata. Random effects in time and by region are included in the model. Here, $\xi_{c[a]}(p)$ follows a RW2 in period by age-group of child with common precision κ_t, $U(r)$ follows an ICAR by region r with precision κ_s and $\epsilon(r,p)$ is i.i.d. normal error across region and period with common precision κ_ϵ. Note that the probability of dying in the next year given survival to age a is assumed to be the same for children who live to age 5, i.e. for all $a \geq 5$. The temporal trend in the probability of death in the next year is the same given survival to ages 1–4, but the probabilities are different for these ages.

The following are used as priors:

$$\kappa_t \sim \text{PCPrior}(u = 0.5, \alpha = 0.01),$$
$$\kappa_s \sim \text{PCPrior}(u = 0.5, \alpha = 0.01),$$
$$\kappa_\epsilon \sim \Gamma(1, 0.0005),$$

	FBH Only	FBH + SBH
β_0	-2.374 (-2.432, -2.317)	-2.372 (-2.432, -2.312)
β_1	-3.450 (-3.525, -3.375)	-3.459 (-3.541, -3.377)
β_2	-3.725 (-3.802, -3.647)	-3.740 (-3.823, -3.656)
β_3	-4.080 (-4.161, -3.999)	-4.101 (-4.187, -4.014)
β_4	-4.705 (-4.793, -4.615)	-4.730 (-4.824, -4.635)
β_5	-5.286 (-5.477, -5.096)	-5.318 (-5.530, -5.106)
β_U	-0.313 (-0.347, -0.279)	-0.299 (-0.332, -0.267)
β_{SBH}		0.160 (0.148, 0.171)
$\beta_{SBH,U}$		-0.165 (-0.203, -0.127)
κ_t	24.1 (9.22, 63.3)	16.0 (6.28, 40.1)
κ_s	91.7 (46.6, 181)	108 (55.6, 208)
κ_ϵ	245 (149, 402)	178 (121, 260)

Table 6.3: Comparison of estimates and 95% uncertainty intervals when using FBH only and FBH + SBH data in Malawi application.

and for all fixed effects, the parameters are given independent $N(0, 100)$ priors except for the SBH bias terms which are given independent $N(0, 10)$ priors.

6.5.4 Mortality Model Results

The model combining SBH and FBH takes about 8 minutes to fit using `TMB`. Fixed effect estimates and 95% uncertainty intervals in the FBH only model are similar to the results from the FBH and SBH only model (Table 6.3). They are also fairly consistent with the results from Chapter 5. We again have nonzero SBH bias terms, and these are similar to what was observed in that chapter. In this application we find the interpretation to be that the SBH data result in an increase of 17.4% in the period hazard ratio in rural areas (vs 18.6% in the yearly hazard odds found previously) and a decrease of 0.50% in urban areas (vs a decrease of 5.5%).

Figures 6.7 and 6.8 display the estimates and uncertainty for the RW2 and ICAR pa-

rameters. Estimates for the RW2 remain mostly the same, though with some difference in the 2005-2009 time period (period 9). When adding SBH, the estimate is pulled up, which is what we also noticed in the application from Chapter 5. In general, there appears to be a decrease over time, with some leveling off in the middle time periods, in part due to the HIV epidemic.

Figures 6.9–6.15 compare the estimated U5MR and 95% uncertainty for the 26 districts by urban/rural strata for the 5-year time periods between 1980 and 2015. Results were also aggregated across strata, using proportion of reported births from the census, for each district and are displayed in the right panel. Also depicted are weighted estimates derived from FBH data and a combined approach where the weighted estimates are combined with using the Brass method with jackknife variance on SBH data. See Chapter 5 for details on both of these approaches. Estimates tend to be fairly similar over time. For the 9 districts in the central region, posterior medians and 95% credible intervals obtained from using the data augmentation procedure as described in Chapter 5 are plotted in grey. For these districts where the data augmentation method was applied, results are fairly consistent with the Poisson approximation approach we used here. Uncertainty tends to be larger in this particular application and the estimates shifted downwards in earlier time periods, which could be attributed to differences in the mortality model, namely the addition of $\epsilon(r, p)$ in (6.9).

6.6 Discussion

In this chapter, we propose a statistical formulation of the popular Brass method to accommodate SBH data into U5M modeling. Our approach is based on a Poisson approximation to the data generative model. For computation, the `R` package `TMB` is used to obtain empirical Bayes type estimates with associated uncertainty. This is very fast, with results from our application to Malawi being obtained in under 10 minutes.

Our method involves two stages where in the first stage a fertility model is fit to FBH data. An estimate of the fertility probability derived from this model is then plugged into

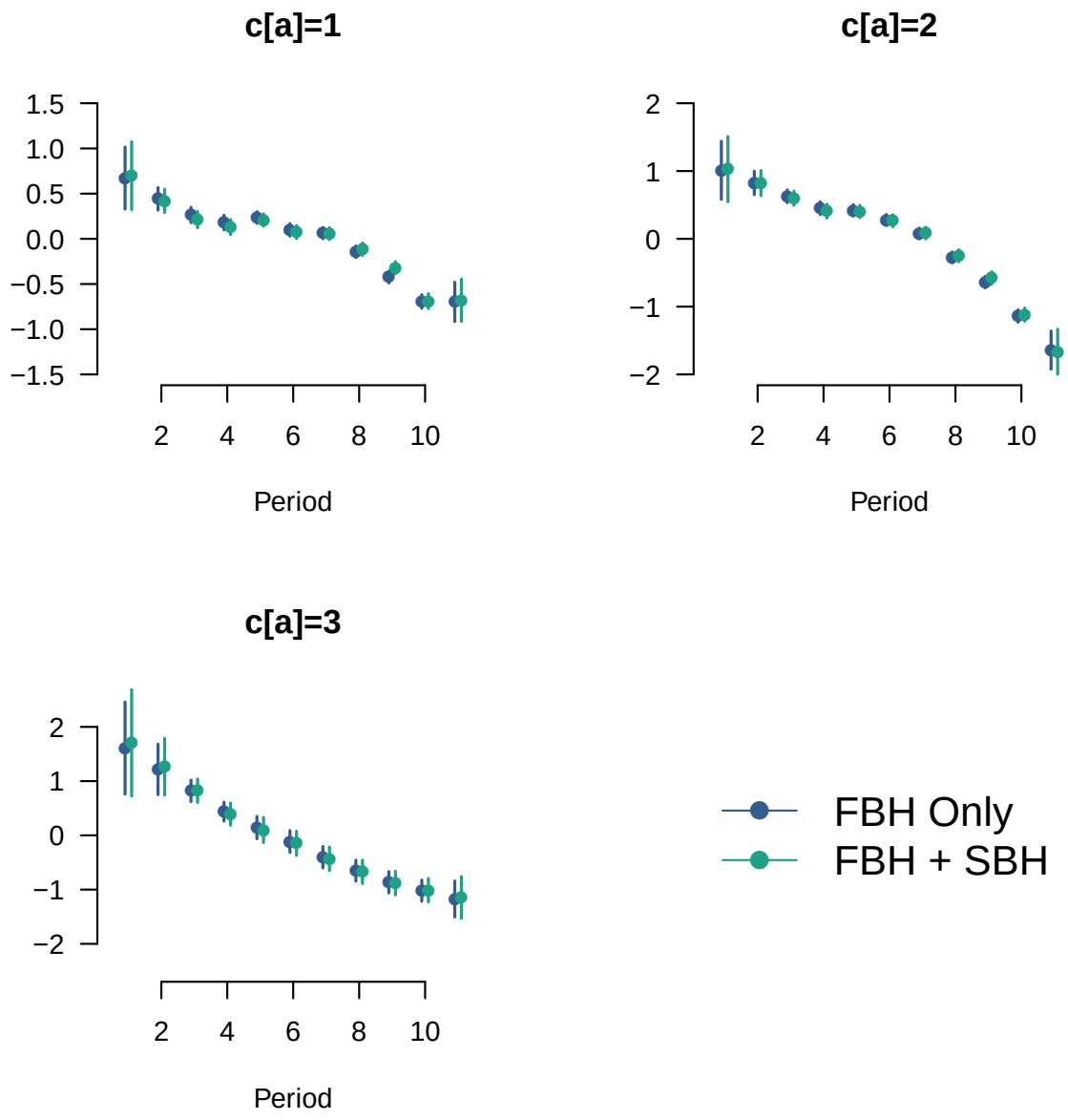

Figure 6.7: Estimates and 95% uncertainty intervals for the RW2 parameters, $\xi_{c[a]}(p)$

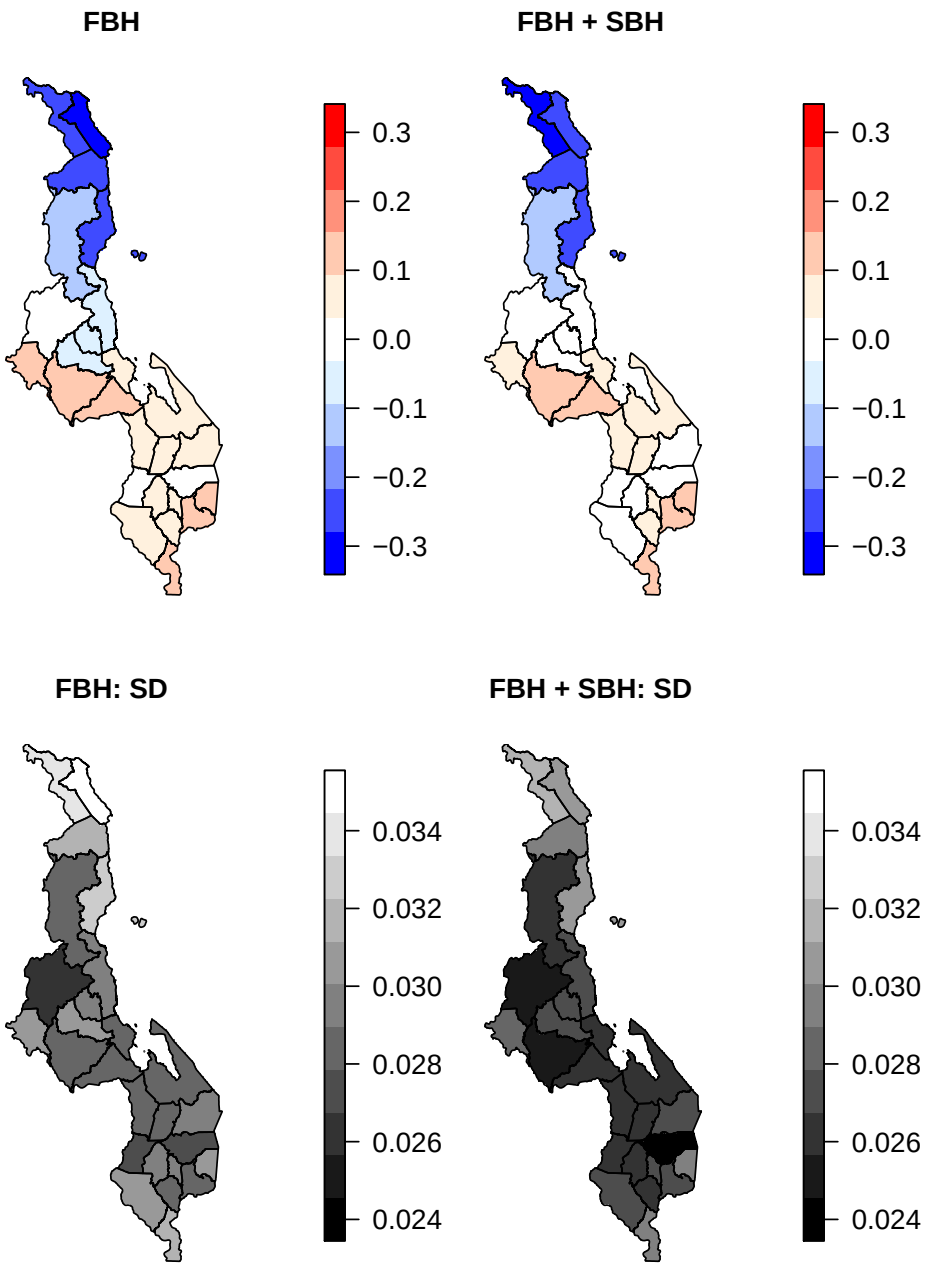

Figure 6.8: Estimates and standard deviation for the ICAR parameters, $U(r)$

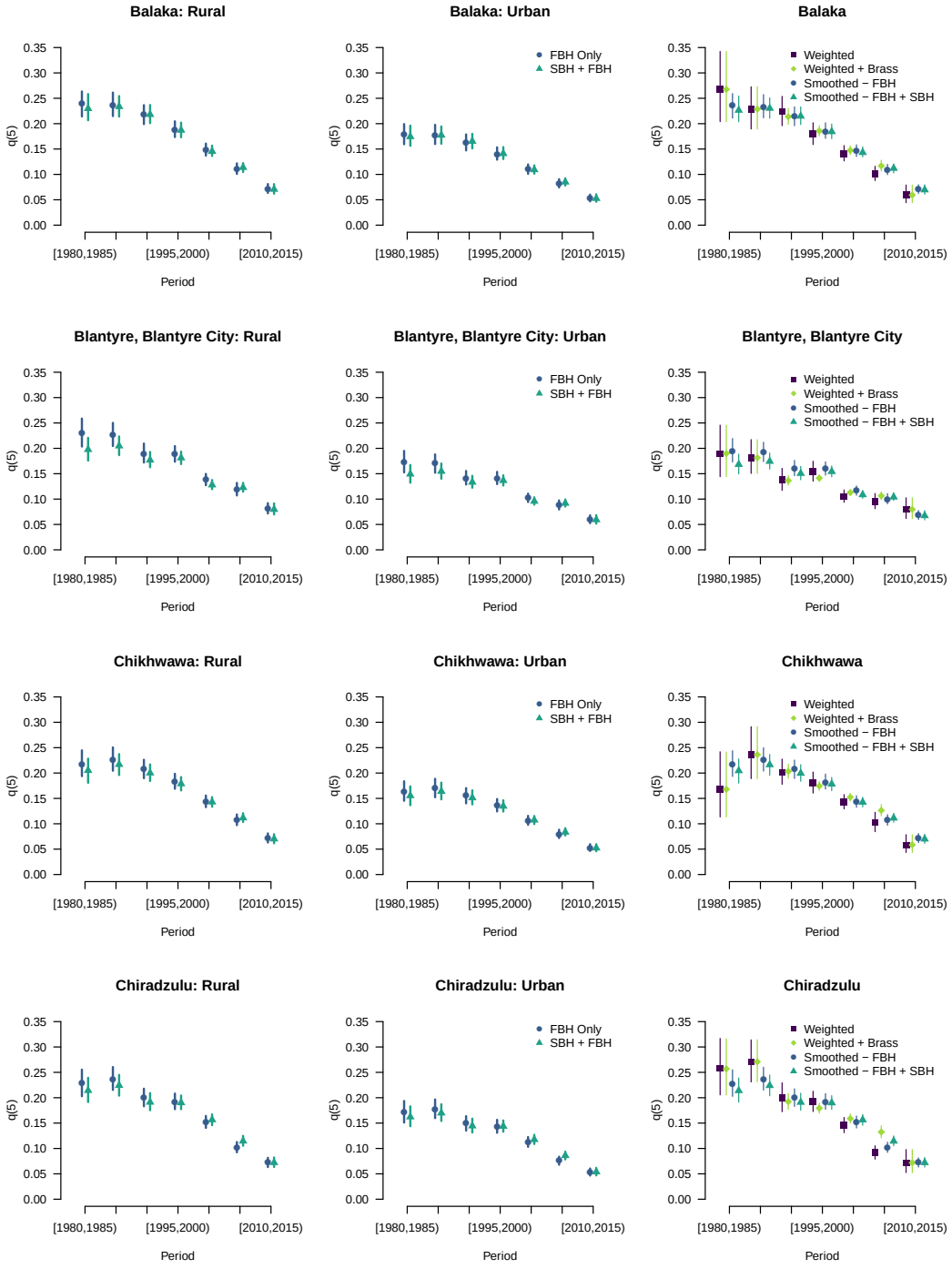

Figure 6.9: Comparison of estimated $q(5)$ for 4 districts. Vertical lines represent 95% uncertainty intervals.

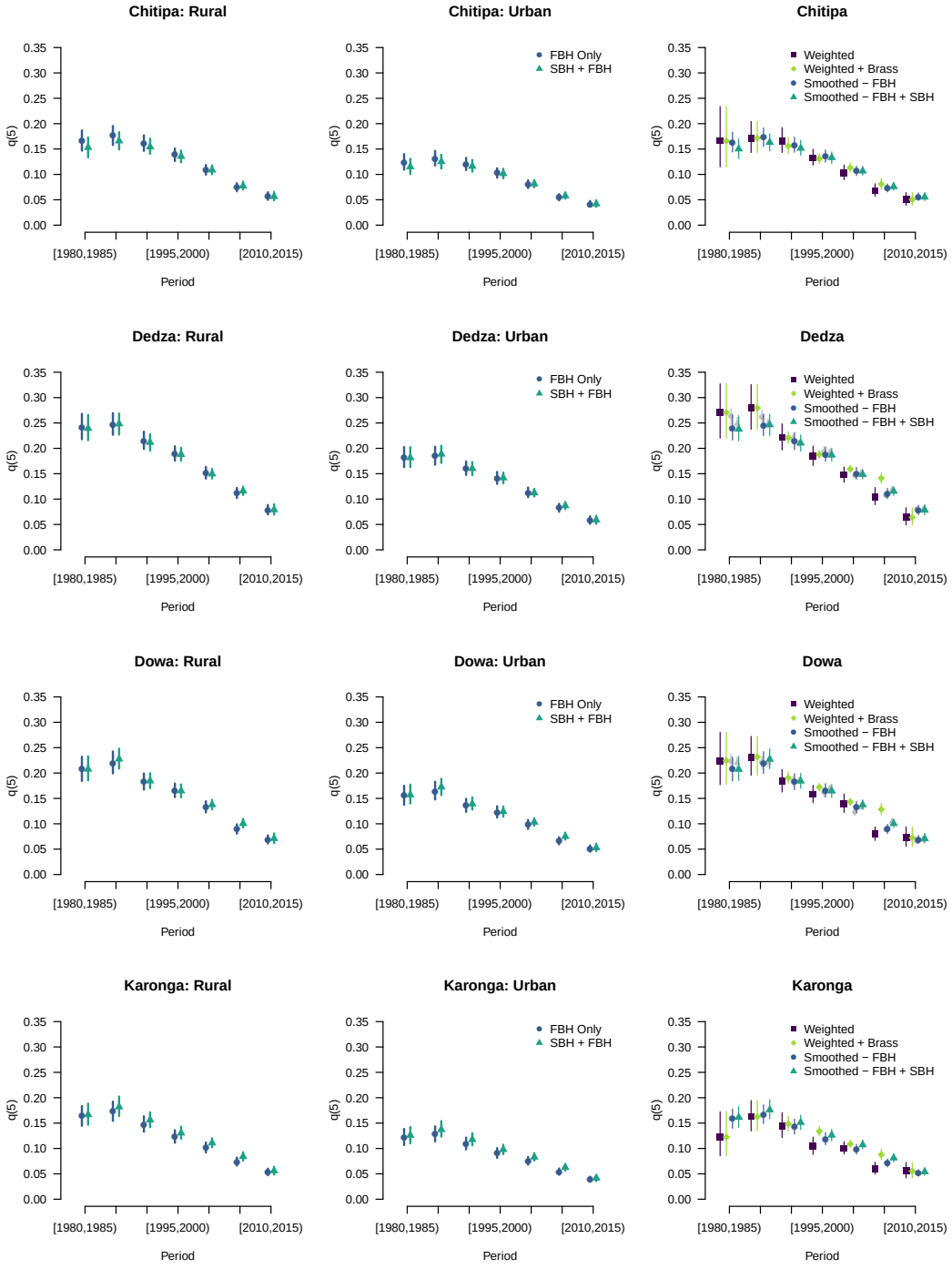

Figure 6.10: Comparison of estimated $q(5)$ for 4 districts. Vertical lines represent 95% uncertainty intervals.

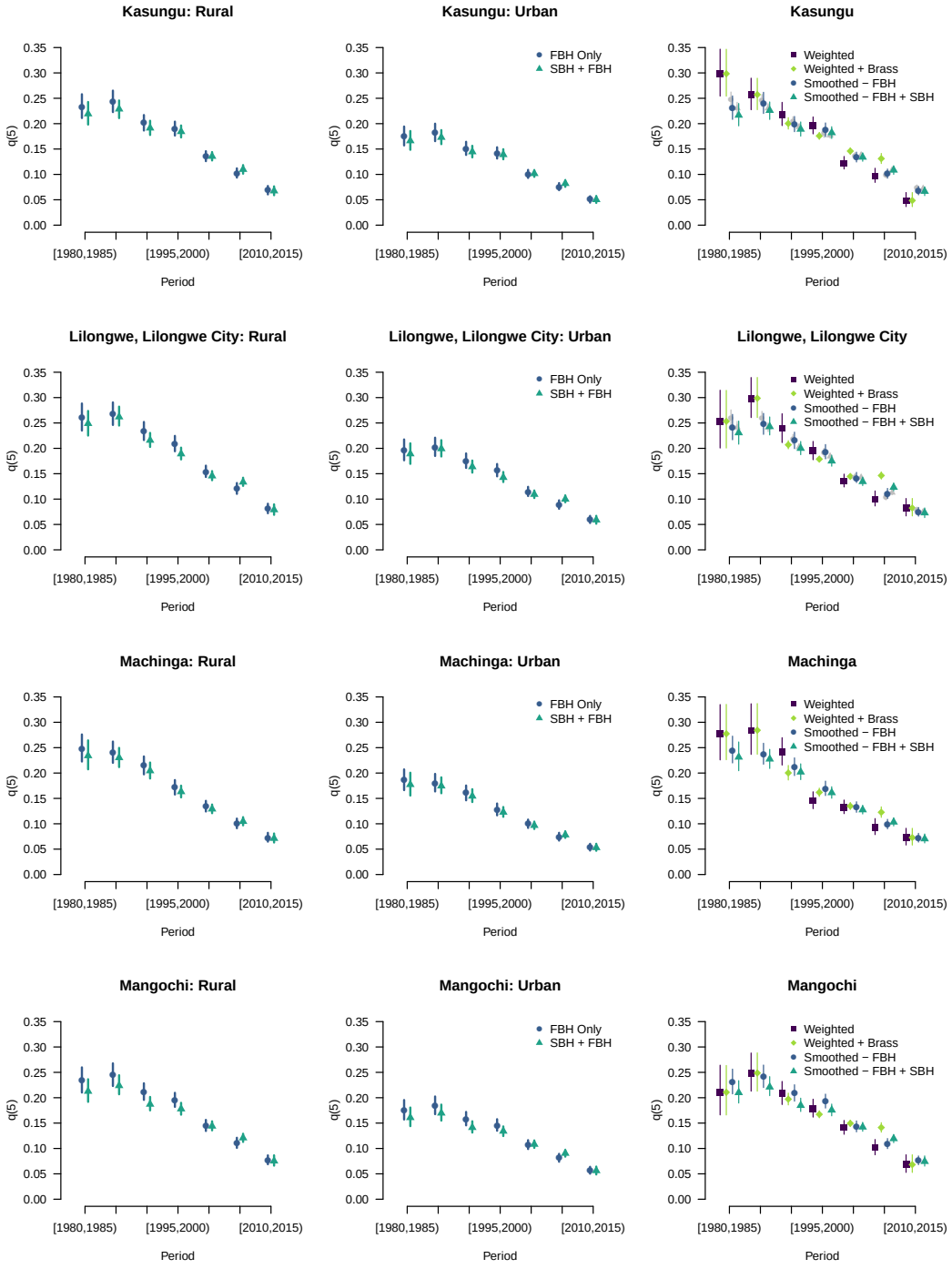

Figure 6.11: Comparison of estimated $q(5)$ for 4 districts. Vertical lines represent 95% uncertainty intervals.

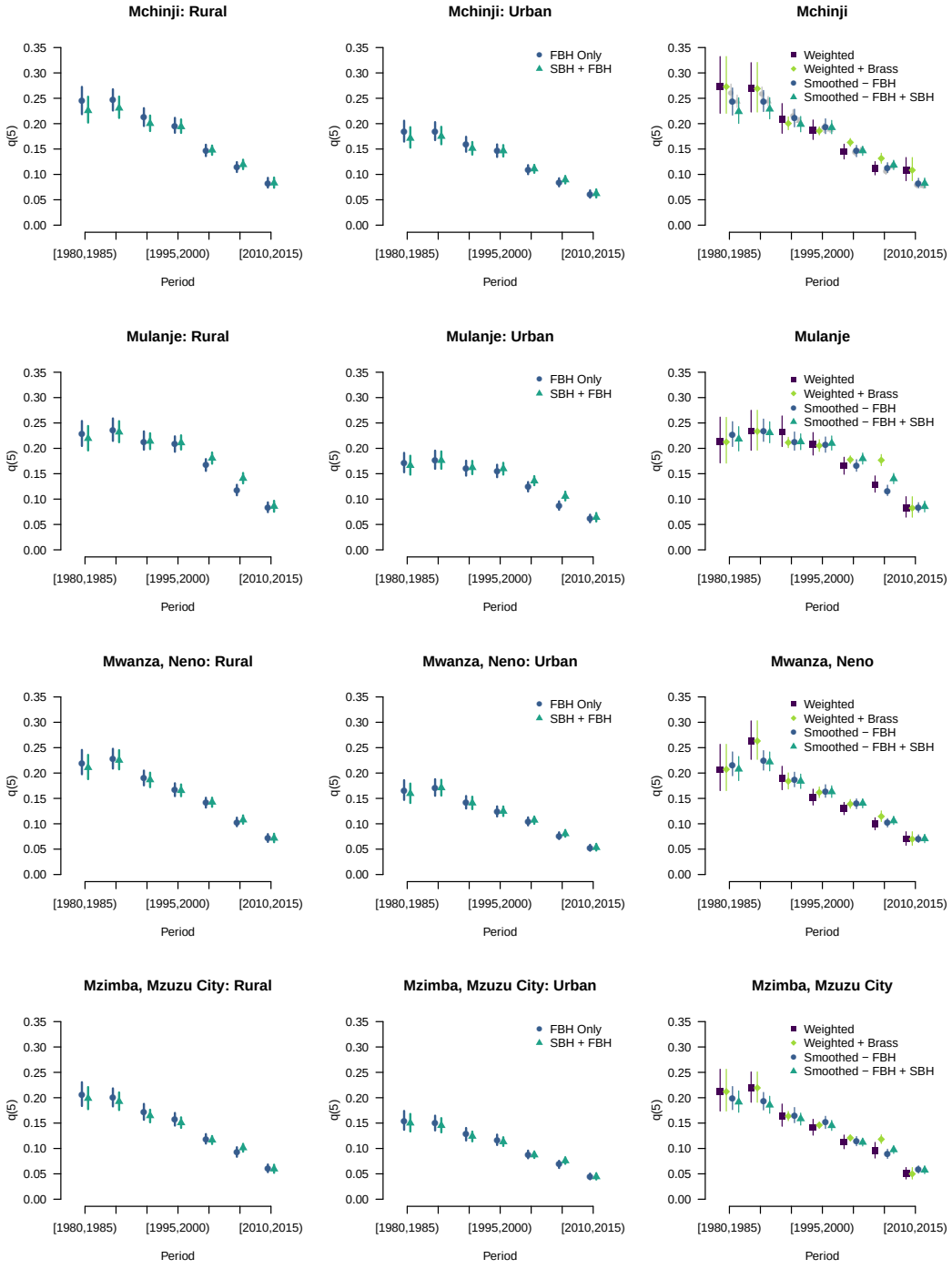

Figure 6.12: Comparison of estimated $q(5)$ for 4 districts. Vertical lines represent 95% uncertainty intervals.

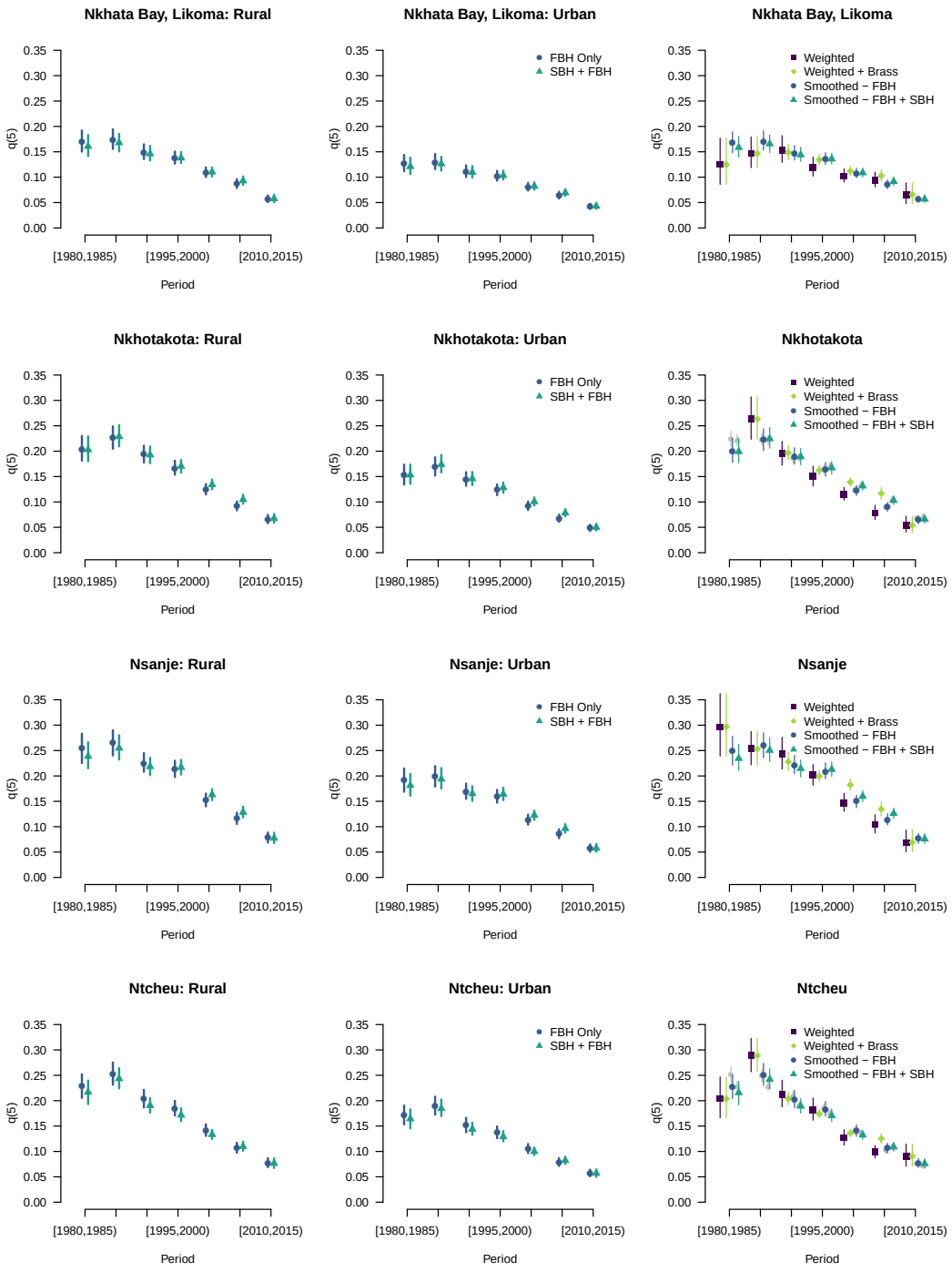

Figure 6.13: Comparison of estimated $q(5)$ for 4 districts. Vertical lines represent 95% uncertainty intervals.

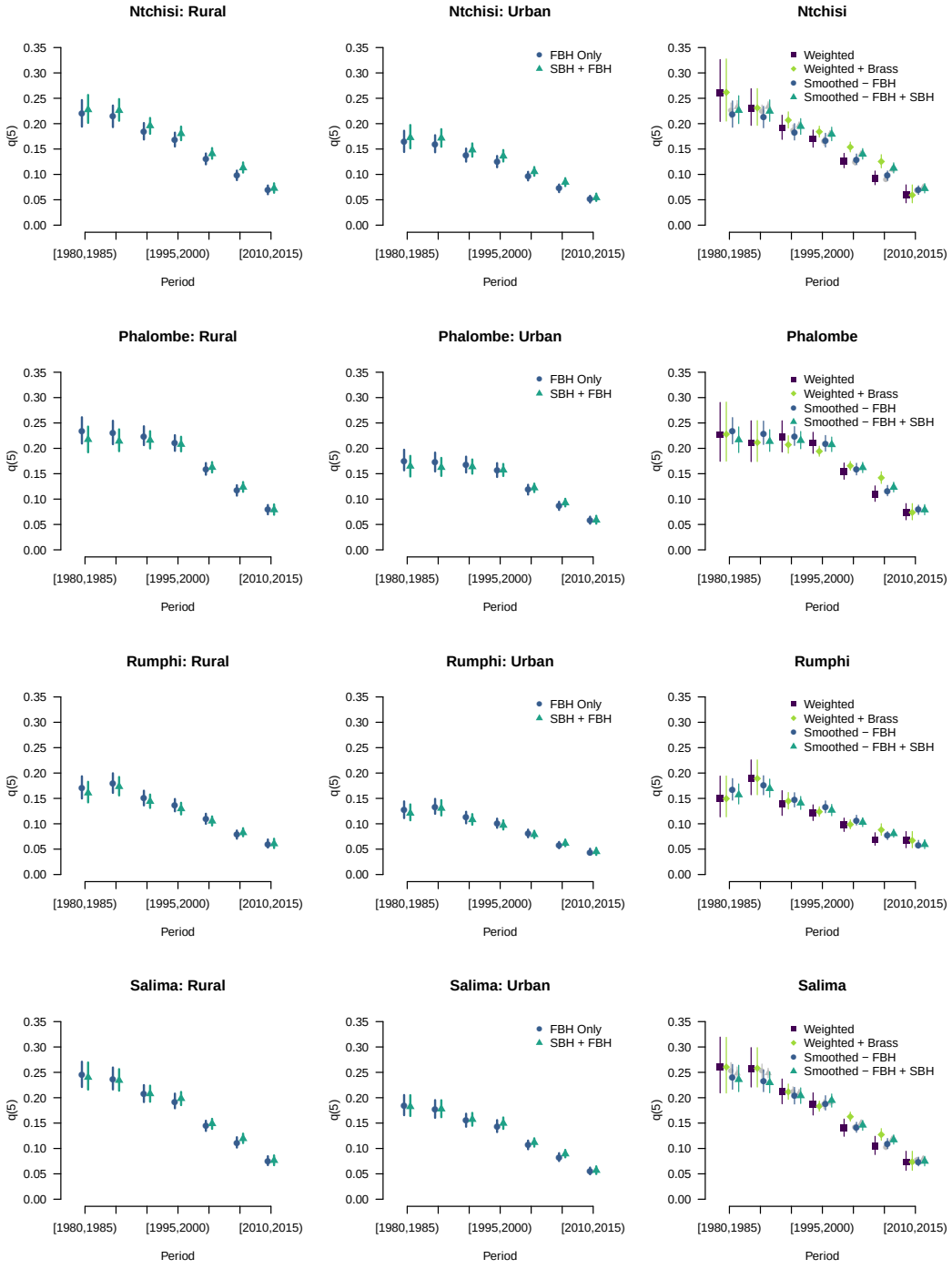

Figure 6.14: Comparison of estimated $q(5)$ for 4 districts. Vertical lines represent 95% uncertainty intervals.

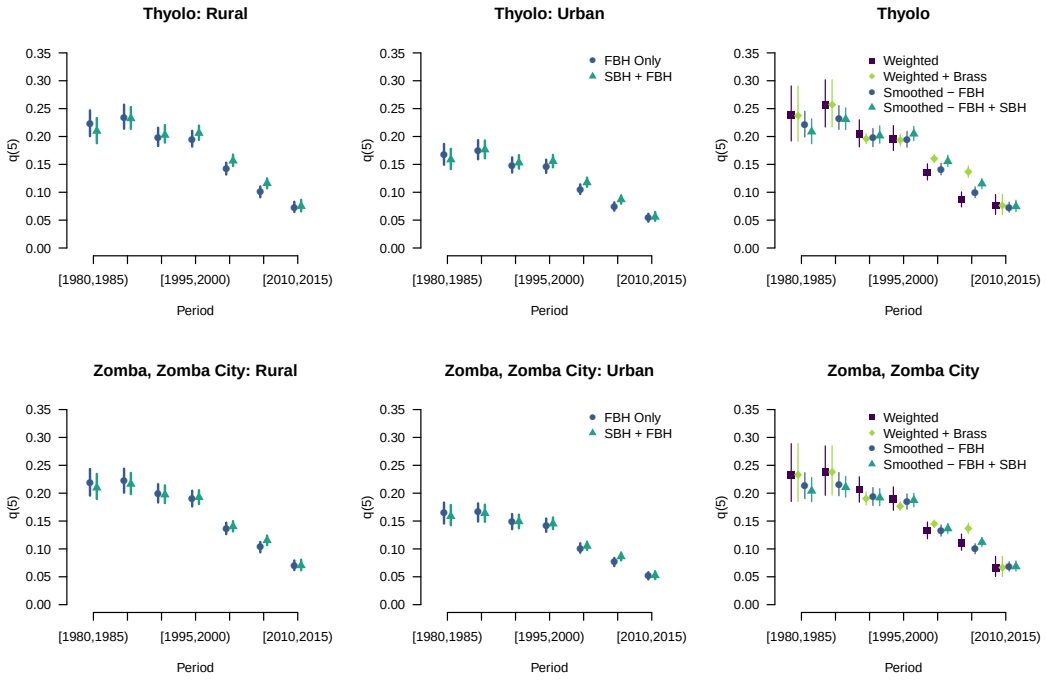

Figure 6.15: Comparison of estimated $q(5)$ for 2 districts. Vertical lines represent 95% uncertainty intervals.

the second stage, where the mortality model is fit to FBH and SBH data. It is important to note that the variance of the fertility estimates is not propagated through into the second stage and the estimates are taken as the truth. This has the potential for undercoverage; however, we did not observe this in our simulation. On a related note, work should be done to investigate the implications of an incorrectly specified fertility model.

As in Chapter 5, the discussions on bias of the surveys and census and what constitutes the "truth" remain relevant. A benefit of this method is that because it runs quickly, it is easier to accommodate biases through the fertility and mortality models.

Future work involves using holdout data to assess performance of our approach, similar to the one described in Chapter 5. Additionally, we wish to compare our method to other existing approaches, namely those proposed by Rajaratnam et al. (2010) and Hill et al. (2015).

Chapter 7

DISCUSSION AND FUTURE WORK

Obtaining estimates and associated measures of uncertainty for health outcomes can be challenging when data are pulled from multiple sources. This is often the case in resource-limited settings where there does not exist high quality vital registration systems, and instead information is obtained via surveys or periodic censuses. Using these types of data can pose a challenge since the information on health outcomes can vary. In this dissertation, we considered three major issues and developed four statistical methods to deal with them.

In Chapter 3, we proposed fitting a spatially continuous model to data that may have been observed at points or aggregated over areas. Having data that have been aggregated over areas, such as the total number of individuals with a particular outcome, are a common feature of census data. Having data associated with points are a common feature of complex household survey data, where in the first stage of sampling, clusters are selected. Clusters, a collection of households, for all intents and purposes are represented by a single point (typically the centroid of the cluster). Using the SPDE approach of Lindgren et al. (2011) to approximate the latent continuous surface and gridded population data, we derived a model that relates the observed data to the latent process. Via a simulation with normally distributed outcome data, we found that including census-type data improved reconstruction of the underlying continuous risk surface over simply using only point data. For this type of outcome data, `R-INLA` (Lindgren and Rue, 2015), can be used. We then applied our method to total counts of lip cancer in Scotland by county, where a Poisson model was used. Here, we developed two main computation strategies. The first was a fully Bayesian approach that used a Hamiltonian Monte Carlo (HMC) algorithm (Neal, 2011). The second was an empirical Bayes (EB) approach implemented using `TMB`, and was based on using a Laplace

approximation to integrate out the spatial random effect. This work has been published (Wilson and Wakefield, 2018b).

In Chapter 4, we derived a data augmentation (DA) algorithm that can handle two scenarios common with survey data. Here, data are associated with points, but there is missing information on where the points are in space, usually a result of confidentiality concerns. In one scenario, called "masking", only the administrative area in which the cluster resides is available. In the second, called "jittering", the point is randomly displaced. The mechanism that results in this missing information can be encoded in a prior for the distribution of the true cluster location. For inference, we suggest an "INLA within MCMC" scheme, where Markov chain Monte Carlo (MCMC) is used for the data augmentation step where the true cluster locations are proposed and INLA is used to update the other parameters. From simulation, we find that inference is improved when data augmentation is used in the "masking" case (over not including the masked data at all). However, in the "jittering" case, there seemed to be no discernible impact on inference when the displaced coordinates were used. Future work is needed to expand the simulation study. Additionally, this method will be applied to household survey data in Kenya.

In Chapter 5, we constructed a DA procedure for incorporating birth history data with non-specific temporal information into modeling the under-five mortality rate (U5MR). Specifically, we considered summary birth histories (SBH), where only the total number of children and number of deaths by woman are obtained. This approach requires a model for fertility and mortality rates, which can incorporate individual level covariates, such as urban/rural, administrative area, and age. Bias in SBH can also be incorporated. Inference was obtained using an MCMC algorithm where the true birth and death times were imputed for SBH in the data augmentation step and the parameters in the fertility and mortality models were updated using HMC. We found, in simulation, incorporating birth records only available as SBH inference is improved as compared to using only full birth histories (FBH), which contain reported birth and death times. The method was applied to data from central Malawi. Using some FBH as holdout data, the DA approach tended to perform best

as measured by mean squared error. This work (Wilson and Wakefield, 2018a) is currently under revision.

In Chapter 6, we proposed a computationally efficient method to combine SBH and FBH data. Notably, this method, based on a Poisson approximation, eliminates the need to simulate the "true" birth and death times for SBH. EB is used for inference and model fitting is done using `TMB`. Via simulation, we investigated the accuracy of the approximation. Finally, this approach was applied to data from Malawi. Work is ongoing to compare this method to other existing approaches, namely Rajaratnam et al. (2010) and Hill et al. (2012).

Finally, we aim to apply the methods developed in this dissertation to modeling the U5MR in sub-Saharan Africa. Many of the methods described here seek to address common problems that are encountered when using birth history data in resource-limited settings.

Future work in modeling survey and census data in these countries include understanding and addressing sources of bias in regards to modeling and model assessment. For example when modeling the U5MR, recall bias is an issue as women may misreport dates of births and deaths or omit children altogether. Additionally, computational efficiency and ease of use have made certain methods more attractive even with statistical limitations. In several of the chapters, we have proposed methods that circumvent some of the more computationally demanding aspects of Bayesian inference by using modeling approximations or through empirical Bayes (Chapters 3 and 6). Even so, these approximations require the user to implement the methods themselves. Ultimately, having software available (e.g., an `R` package) that can take in data at different resolutions, model the underlying process at different resolutions, and output predictions at different resolutions would be desirable.

BIBLIOGRAPHY

Alkema, L. and New, J. R. (2012). Progress toward global reduction in under-five mortality: a bootstrap analysis of uncertainty in millennium development goal 4 estimates. *PLoS Medicine*, 9(12):e1001355.

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*, 8:2122–2149.

Alkema, L., New, J. R., Pedersen, J., You, D., et al. (2014). Child mortality estimation 2013: an overview of updates in estimation methods by the United Nations Inter-agency Group for Child Mortality Estimation. *PLoS ONE*, 9(7):e101112.

Banerjee, S., Carlin, B., and Gelfand, A. (2014). *Hierarchical Modeling and Analysis for Spatial Data, Second Edition*. CRC Press.

Berrocal, V., Gelfand, A., and Holland, D. (2010). A spatio-temporal downscaler for output from numerical models. *Journal of Agricultural, Biological, and Environmental Statistics*, 15:176–197.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistics and Mathematics*, 43:1–59.

Bivand, R., Gómez-Rubio, V., and Rue, H. (2015). Spatial data analysis with `R-INLA` with some extensions. *Journal of Statistical Software*, 63:1–31.

Bivand, R. S., Gómez-Rubio, V., and Rue, H. (2014). Approximate bayesian inference for spatial econometrics models. *Spatial Statistics*, 9:146–165.

Bradley, J., Wikle, C., and Holan, S. (2016). Bayesian spatial change of support for count-valued survey data with application to the American Community Survey. *Journal of the American Statistical Association*, 111:472–487.

Brady, E. and Hill, K. (2017). Testing survey-based methods for rapid monitoring of child mortality, with implications for summary birth history data. *PLoS ONE*, 12:e0176366.

Brass, W. (1964). *Uses of census or survey data for the estimation of vital rates*. United Nations. Paper prepared for the African Seminar on Vital Statistics, Addis Ababa, 14–19 December, 1964.

Burgert, C. R., Colston, J., Roy, T., and Zachary, B. (2013). Geographic displacement procedure and georeferenced data release policy for the demographic and health surveys. Technical report, ICF International. DHS Spatial Analysis Reports No. 7.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76:1–32.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models*. Chapman and Hall/CRC.

Center for International Earth Science Information Network - CIESIN - Columbia University (2016). Gridded population of the world, version 4 (GPWv4): Population count. Accessed 22 November 2016.

Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43:671–682.

Coale, A. J. and Demeny, P. with Vaughan, B. (1983). *Regional Model Life Tables and Stable Populations*. New York: Academic Press, 2nd edition.

Coale, A. J. and Trussell, J. (1977). Annex I: estimating the time to which Brass estimates apply. *Population Bulletin of the United Nations*, 10:87–89.

Corsi, D., Neuman, M., Finlay, J., and Subramanian, S. (2012). Demographic and health surveys: a profile. *International Journal of Epidemiology*, 41:1602–1613.

Cressie, N. and Kornak, J. (2003). Spatial statistics in the presence of location error with an application to remote sensing of the environment. *Statistical science*, 18:436–456.

Cressie, N. and Wikle, C. (2011). *Statistics for Spatio-Temporal Data*. John Wiley and Sons.

Diggle, P., Moraga, P., Rowlingson, B., and Taylor, B. (2013). Spatial and spatio-temporal log-Gaussian Cox processes: Extending the geostatistical paradigm. *Statistical Science*, 28:542–563.

Diggle, P. and Ribeiro, P. (2007). *Model Based Geostatistics*. Springer, New York.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid monte carlo. *Physics letters B*, 195:216–222.

Fan, C., Stafford, J., and Brown, P. (2011). Local-EM and the EMS algorithm. *Journal of Computational and Graphical Statistics*, 20:750–766.

Fanshawe, T. and Diggle, P. (2011). Spatial prediction in the presence of positional error. *Environmetrics*, 22:109–122.

Feeney, G. (1980). Estimating infant mortality trends from child survivorship data. *Population Studies*, 34:109–128.

Filippone, M., Zhong, M., and Girolami, M. (2013). A comparative evaluation of stochastic-based inference methods for Gaussian process models. *Machine Learning*, 93:93–114.

Fong, Y., Rue, H., and Wakefield, J. (2010). Bayesian inference for generalized linear mixed models. *Biostatistics*, 11:397–412.

Fronterrè, C., Giorgi, E., and Diggle, P. (2018). Geostatistical inference in the presence of geomasking: a composite-likelihood approach. *Spatial Statistics*, 28:319–330.

Fuentes, M. and Raftery, A. (2005). Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. *Biometrics*, 61:36–45.

Gabrosek, J. and Cressie, N. (2002). The effect on attribute prediction of location uncertainty in spatial data. *Geographical Analysis*, 34:262–285.

Gelfand, A. (2010). Misaligned spatial data. In Gelfand, A., Diggle, P., Fuentes, M., and Guttorp, P., editors, *Handbook of Spatial Statistics*, pages 517–539. CRC Press.

Gelman, A. and Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–511.

Gething, P., Tatem, A., Bird, T., and Burgert-Brucker, C. (2015). Creating spatial interpolation surfaces with DHS data. Technical report, ICF International. DHS Spatial Analysis Reports No. 11.

Golding, N., Burstein, R., Longbottom, J., Browne, A. J., Fullman, N., Osgood-Zimmerman, A., et al. (2017). Mapping under-5 and neonatal mortality in Africa, 2000–15: a baseline analysis for the Sustainable Development Goals. *The Lancet*, 390:2171–2182.

Gómez-Rubio, V. and Palmí-Perales, F. (2017). Spatial models with the integrated nested laplace approximation within markov chain monte carlo. *arXiv preprint arXiv:1702.03891*.

Gómez-Rubio, V. and Rue, H. (2017). Markov chain monte carlo with the integrated nested laplace approximation. *arXiv preprint arXiv:1701.07844*.

Gotway, C. and Young, L. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association*, 97:632–648.

Hájek, J. (1971). Discussion of Basu. In Godambe, V. and Sprott, D., editors, *Foundations of Statistical Inference*. Holt, Rinehart & Winston.

Hill, K. (2013). Indirect estimation of child mortality. In Moultrie, T. A., Dorrington, R., Hill, A., Hill, K., Timæus, I., and Zaba, B., editors, *Tools for Demographic Estimation*, pages 148–164. International Union for the Scientific Study of Population Paris.

Hill, K., Brady, E., Zimmerman, L., Montana, L., Silva, R., and Amouzou, A. (2015). Monitoring change in child mortality through household surveys. *PLoS ONE*, 10:e0137713.

Hill, K. and Figueroa, M.-E. (1999). Child mortality estimation by time since first birth. In Zaba, B. and Blacker, J., editors, *Brass Tacks: Essays in Medical Demography*, pages 9–19. London: Athlone.

Hill, K., You, D., Inoue, M., and Oestergaard, M. Z. (2012). Child mortality estimation: accelerated progress in reducing global child mortality, 1990–2010. *PLoS Medicine*, 9:e1001303.

Hoeting, J., Madigan, D., Raftery, A., and Volinsky, C. (1998). Bayesian model averaging. Technical Report 9814, Department of Statistics, Colorado State University.

Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685.

Image and Data processing by NOAA's National Geophysical Data Center. DMSP data collected by the US Air Force Weather Agency (2008). Version 4 DMSP-OLS Nighttime Lights Time Series. `https://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html`. Accessed 27 July 2018.

Kenya National Bureau of Statistics (2015). Kenya Demographic and Health Survey 2014. Technical report, Kenya National Bureau of Statistics.

Kristensen, K. (2014). `TMB`: General random effect model builder tool inspired by ADMB. `R` *package version*.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., and Bell, B. M. (2016). TMB: Automatic Differentiation and Laplace approximation. *Journal of Statistical Software*, 70:1–21.

Lee, J., Nguyen, P., Brown, P., Stafford, J., and Saint-Jacques, N. (2017). A local-EM algorithm for spatio-temporal disease mapping with aggregated data. *Spatial Statistics*, 21:75–95.

Li, Y., Brown, P., Gesink, D., and Rue, H. (2012). Log Gaussian Cox processes and spatially aggregated disease incidence data. *Statistical Methods in Medical Research*, 21:479–507.

Lindgren, F. and Rue, H. (2015). Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*, 63:1–25.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic differential equation approach (with discussion). *Journal of the Royal Statistical Society, Series B*, 73:423–498.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9:1–19.

Lumley, T. (2018). survey: analysis of complex survey samples. R package version 3.35.

Marin, J.-M., Mengersen, K., and Robert, C. P. (2005). Bayesian modelling and inference on mixtures of distributions. In Dey, D. and Rao, C., editors, *Handbook of Statistics*, pages 459–507. Elsevier, Amsterdam.

Mercer, L. D., Wakefield, J., Pantazis, A., Lutambi, A. M., Masanja, H., and Clark, S. (2015). Space–time smoothing of complex survey data: small area estimation for child mortality. *The Annals of Applied Statistics*, 9:1889–1905.

Minot, N. and Baulch, B. (2005). Spatial patterns of poverty in vietnam and their implications for policy. *Food Policy*, 30:461–475.

Moraga, P., Cramb, S., Mengersen, K., and Pagano, M. (2017). A geostatistical model for combined analysis of point-level and area-level data using INLA and SPDE. *Spatial Statistics*, 21:27–41.

National Statistical Office - NSO/Malawi and ICF Macro (2011). Malawi Demographic and Health Survey 2010. Final report, NSO/Malawi and ICF Macro, Zomba, Malawi. Available at http://dhsprogram.com/pubs/pdf/FR247/FR247.pdf.

Neal, R. (2011). MCMC using Hamiltonian dynamics. In Brooks, S., Gelman, A., Jones, G., and Meng, X., editors, *Handbook of Markov Chain Monte Carlo*, volume 2, pages 113–162. Chapman and Hall/CRC Press.

Nguyen, P., Brown, P., and Stafford, J. (2012). Mapping cancer risk in Southwestern Ontario with changing census boundaries. *Biometrics*, 68:1228–1237.

Okwi, P., Ndeng'e, G., Kristjanson, P., Arunga, M., Notenbaert, A., Omolo, A., Henninger, N., Benson, T., Kariuki, P., and Owuor, J. (2007). Spatial determinants of poverty in rural kenya. *Proceedings of the National Academy of Sciences*, 104:16769–16774.

Pedersen, J. and Liu, J. (2012). Child mortality estimation: appropriate time periods for child mortality estimates from full birth histories. *PLoS Medicine*, 9:e1001289.

Perez-Heydrick, C., Warren, J., Burgert, C., and Emch, M. (2013). Guidelines on the use of DHS GPS data. Technical report, ICF International. DHS Spatial Analysis Reports No. 8.

Pezzulo, C., T.Bird, Edson, C., Utazi, C., Sorichetta, A., Tatem, A., Yourkavitch, J., and Burgert-Brucker, C. (2016). Geospatial modeling of child mortality across 27 countries in sub-Saharan Africa. Technical report, ICF International. DHS Spatial Analysis Reports No. 13.

Preston, S. H., Heuveline, P., and Guillot, M. (2000). *Demography: Measuring and Modeling Population Processes*. Blackwell Malden, MA.

Rajaratnam, J. K., Tran, L. N., Lopez, A. D., and Murray, C. J. (2010). Measuring under-five mortality: validation of new low-cost methods. *PLoS Medicine*, 7:e1000253.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Application*. Chapman and Hall/CRC Press, Boca Raton.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B*, 71:319–392.

Rutstein, S. and Johnson, K. (2004). The DHS wealth index. Dhs comparative reports no. 6, Calverton, Maryland, USA. Available at http://dhsprogram.com/pubs/pdf/CR6/CR6.pdf.

Rutstein, S. O. and Rojas, G. (2006). *Guide to DHS statistics*. Calverton, MD: ORC Macro.

Silva, R. (2012). Child mortality estimation: consistency of under-five mortality rate estimates using full birth histories and summary birth histories. *PLoS Medicine*, 9:e1001296.

Simpson, D., Lindgren, F., and Rue, H. (2012a). In order to make spatial statistics computationally feasible, we need to forget about the covariance function. *Environmetrics*, 23:65–74.

Simpson, D., Lindgren, F., and Rue, H. (2012b). Think continuous: Markovian Gaussian models in spatial statistics. *Spatial Statistics*, 1:16–29.

Simpson, D., Rue, H., Riebler, A., Martins, T., and Sørbye, S. (2017). Penalising model component complexity: a principled, practical approach to constructing priors (with discussion). *Statistical Science*, 32:1–28.

Sørbye, S. and Rue, H. (2014). Scaling intrinsic Gaussian Markov random field priors in spatial modelling. *Spatial Statistics*, 8:39–51.

Stein, M. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82:528–540.

Taylor, B., Andrade-Pacheco, R., and Sturrock, H. (2018). Continuous inference for aggregated point process data. *Journal of the Royal Statistical Society: Series A*, 181:1125–1150.

Taylor, B., Davies, T., Rowlingson, B., and Diggle, P. (2015). `lgcp`: an R package for inference with spatial and spatio-temporal log-Gaussian Cox processes. *Journal of Statistical Software*, 52:1–40.

The Demographic and Health Surveys Program (2018). Conditions of Use for The DHS Program datasets . `https://dhsprogram.com/data/terms-of-use.cfm`. Accessed: 2019-01-30.

Thorson, J., Skaug, H., Kristensen, K., Shelton, A., Ward, E., Harms, J., and Benante, J. (2015). The importance of spatial models for estimating the strength of density dependence. *Ecology*, 96:1202–1212.

Trussell, J. (1975). A re-estimation of the multiplying factors for the Brass technique for determining childhood survivorship rates. *Population Studies*, 29:97–107.

United Nations (1983). *Manual X: Indirect Techniques for Demographic Estimation*. United Nations, New York.

United Nations (2017). *Sustainable development knowledge platform: transforming our world: the 2030 agenda for sustainable development*. `https://goo.gl/ETnctb`.

United Nations General Assembly (2015). Transforming our world: the 2030 agenda for sustainable development. Resolution Adopted by the General Assembly on 25 September 2015: 70/1. `https://goo.gl/UBddEC`.

Wakefield, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics*, 8:158–183.

Wakefield, J., Fuglstad, G.-A., Riebler, A., Godwin, J., Wilson, K., and Clark, S. J. (2018). Estimating under five mortality in space and time in a developing world context. *Statistical Methods in Medical Research*. To Appear.

Walker, N., Hill, K., and Zhao, F. (2012). Child mortality estimation: methods used to adjust for bias due to AIDS in estimating trends in under-five mortality. *PLoS Medicine*, 9:e1001298.

Wilson, K. and Wakefield, J. (2018a). Child mortality estimation incorporating summary birth history data. *Submitted for Publication*.

Wilson, K. and Wakefield, J. (2018b). Pointless spatial modeling. *Biostatistics*. Published online 6 September, 2018.

WorldPop (2016). Version 2.0 estimates of total number people per grid square for five timepoints between 2000 and 2020 at five year intervals; national totals have been adjusted to match UN Population Division estimates for each time point. DOI: 10.5258/SOTON/WP00004.

Appendix A

APPENDIX

A.1 Appendix for Chapter 3

Trace plots for the fully Bayesian approach are shown in Figures A.1 and A.2, respectively. The trace plot and histogram for the hybrid approach is displayed in Figure A.3.

A.2 Appendix for Chapter 4

Trace plots shown in Figures A.4–A.7 for various simulation scenarios that were considered.

A.3 Appendix for Chapter 5

Figure A.8 contains the trace plots for the the mortality model in the central Malawi application.

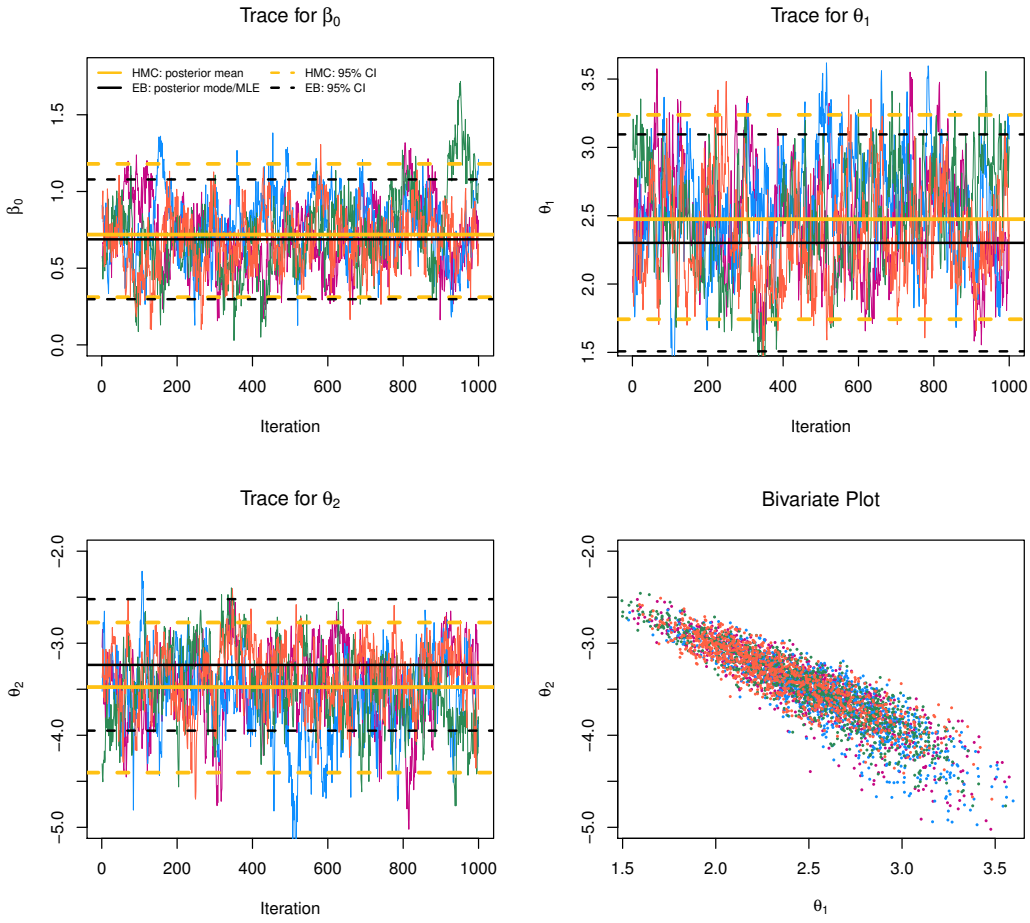

Figure A.1: Trace plots for β_0, ϕ_1, and ϕ_2 in the Scotland example using the fully Bayesian approach. Solid gold lines and dashed gold lines are the posterior means and 95% CI, respectively using HMC. Solid black lines are the EB estimates and the dashed black lines are the corresponding 95% CI using the strictly EB approach.

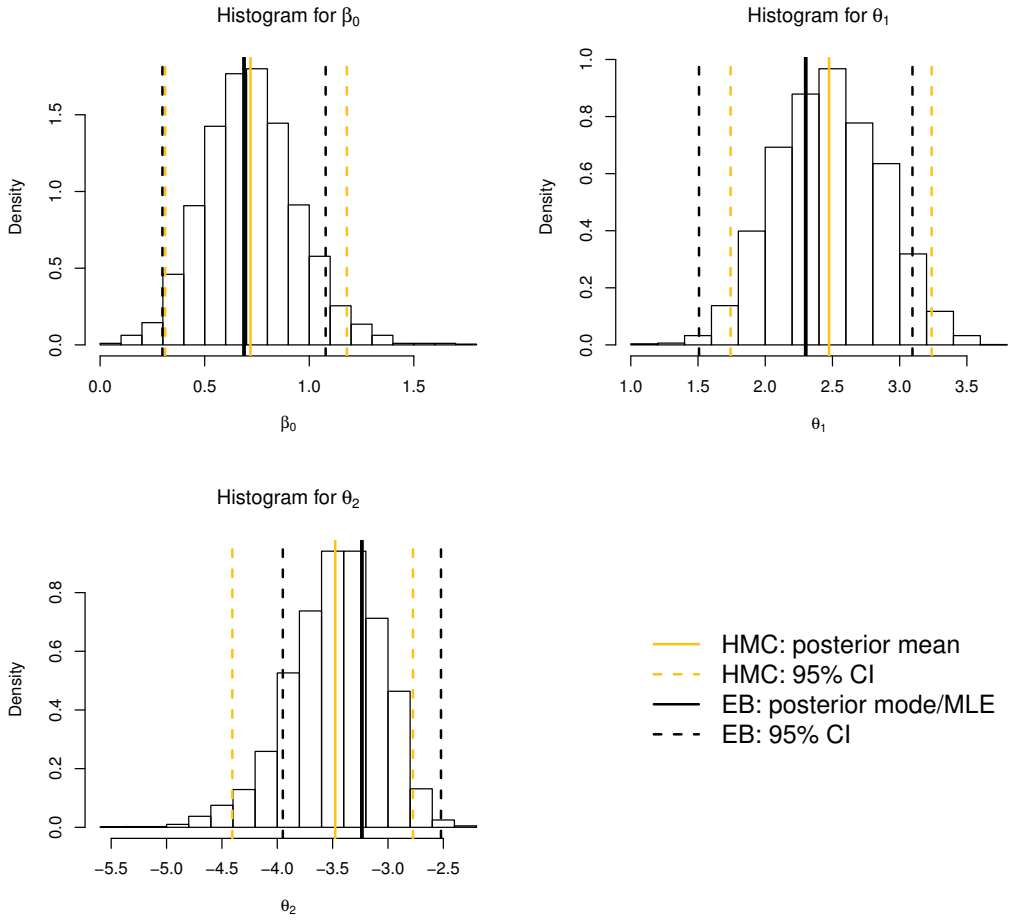

Figure A.2: Univariate posterior distributions for β_0, ϕ_1, and ϕ_2 using the fully Bayesian approach. Solid gold lines and dashed gold lines are the posterior means and 95% CI, respectively using HMC. Solid black lines are the EB estimates and the dashed black lines are the corresponding 95% CI using the strictly EB approach.

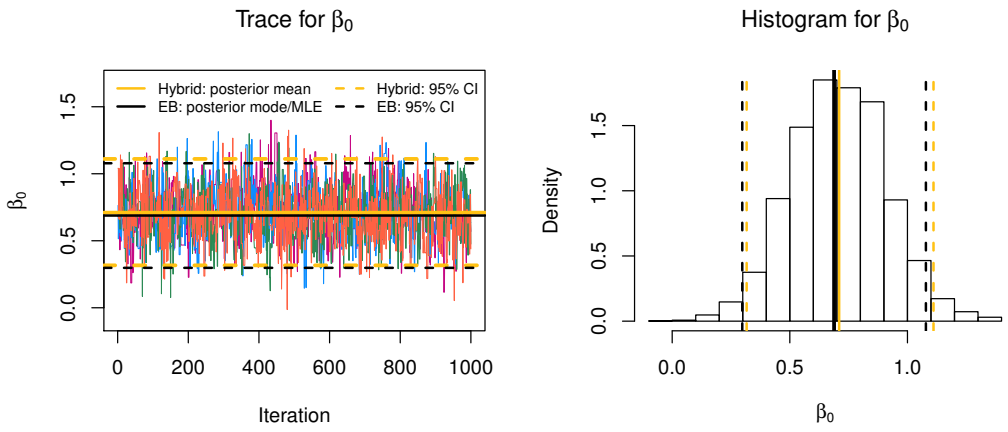

Figure A.3: Trace plot and univariate posterior distribution for β_0 in the Scotland example using the hybrid approach. Solid gold lines and dashed gold lines are the posterior means and 95% CI, respectively. Solid black lines and dashed black lines are the EB estimates and 95% CI, respectively using the strictly EB approach.

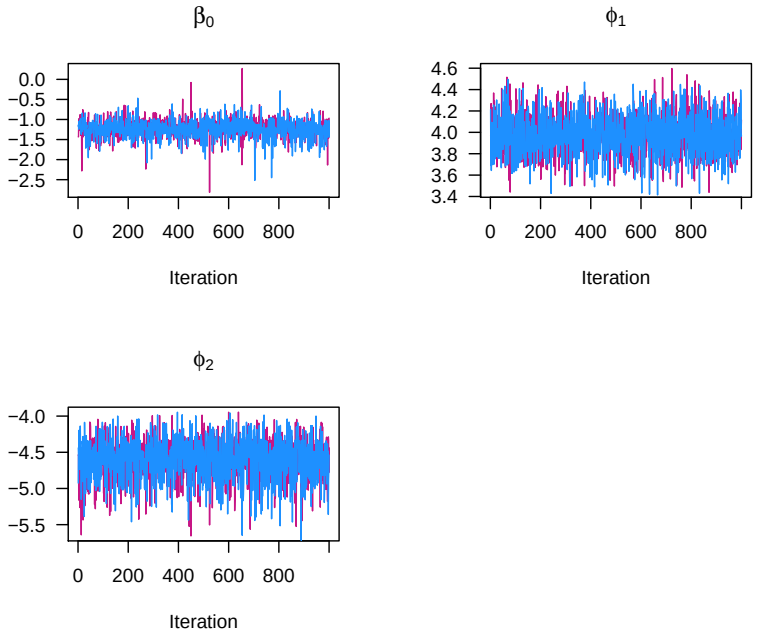

Figure A.4: Trace plots for jittering scenario with no covariate.

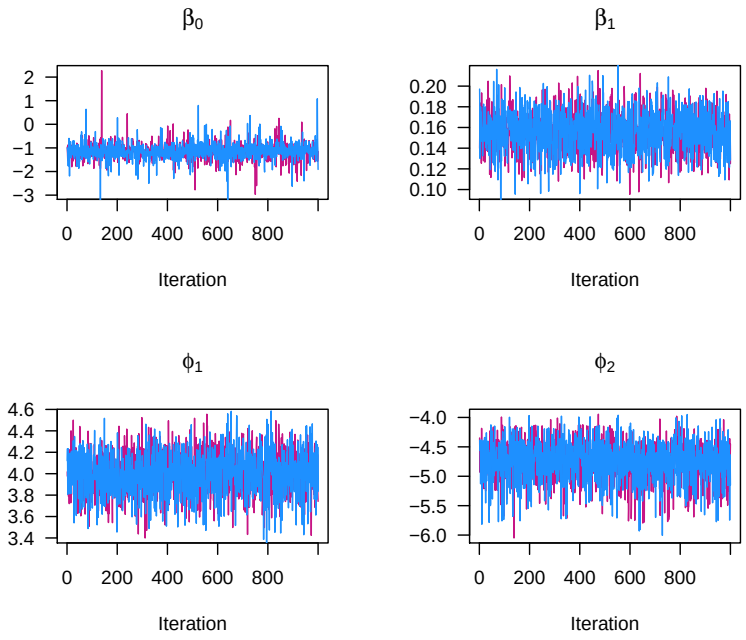

Figure A.5: Trace plots for jittering scenario with covariate.

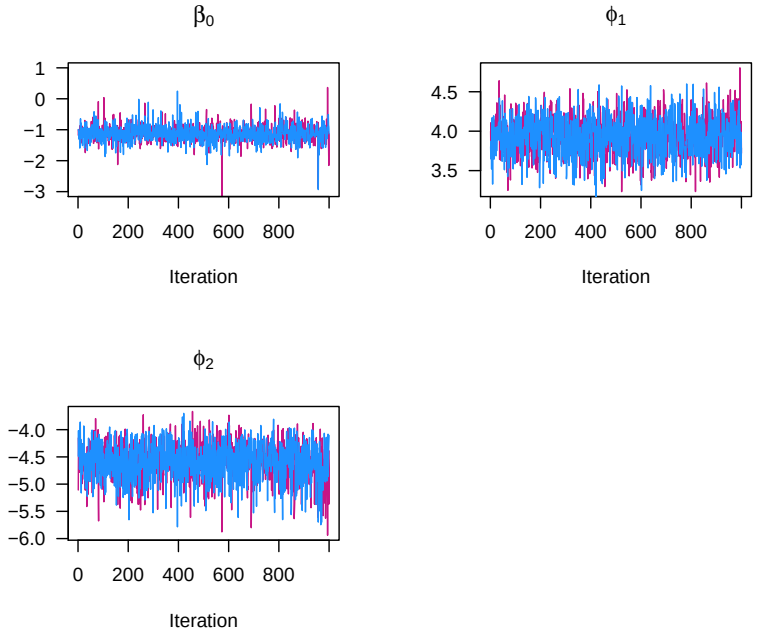

Figure A.6: Trace plots for masking scenario with no covariate.

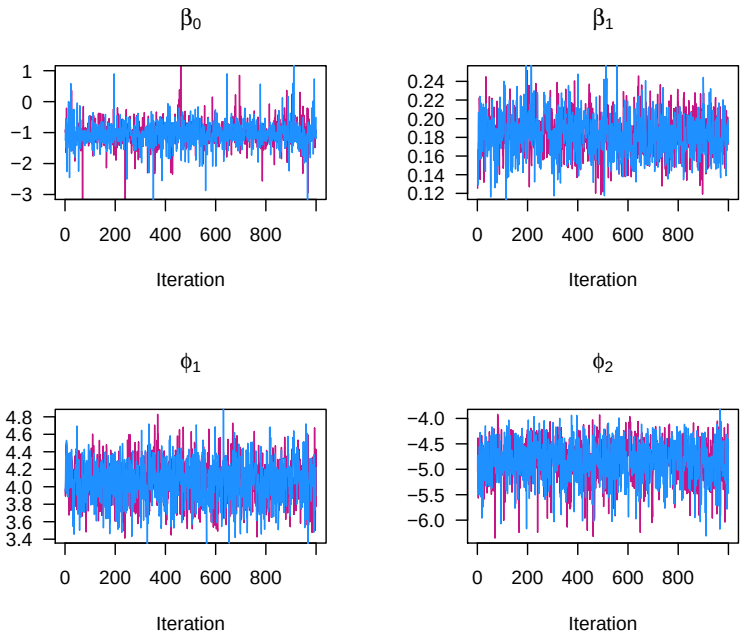

Figure A.7: Trace plots for masking scenario with covariate.

Figure A.8: Trace plots for parameters in the mortality model, $\beta_{c[a]}$, $\beta_{district}$, β_{strata}, $\beta_{SBH,strata}$, and precision parameter κ for the RW2. Red are results from the FBH + SBH model. Blue are results from the FBH only model. Solid lines: posterior medians. Dashed lines: 95% CI.