©Copyright 2016
Laina Mercer

Space-Time Smoothing Models for Surveillance and Complex Survey Data

Laina Mercer

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

Doctor of Philosophy

University of Washington

2016

Reading Committee:

Jon Wakefield, Chair

Barbara McKnight

Adrian Raftery

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Space-Time Smoothing Models for Surveillance and Complex Survey Data

Laina Mercer

Chair of the Supervisory Committee:
Professor Jon Wakefield
Department of Statistics & Department of Biostatistics

Area and time-specific estimates of disease rates, cause-specific mortality rates and other key health indicators are of great interest for health care and policy purposes. Such estimates provide the information needed to identify areas with increased risk, effectively allocate resources, and target interventions. A wide variety of data, such as vital statistics, complex surveys, demographic surveillance sites, and disease registries, are used for these purposes. Unfortunately, the sample size of data available at a granular space-time scale is often too small to provide reliable estimates and uncertainty intervals. Using data from multiple sources and spatial and temporal smoothing is beneficial to alleviate problems of data scarcity. The purpose of the work described herein is to use Bayesian space-time models, to combine data from multiple sources to provide reliable area-based estimates. This work is motivated by estimating rates of health indicators (e.g. diabetes, smoking) by health reporting areas in King County from the Behavioral Risk Factor Surveillance Survey, child mortality by regions in Tanzania from Demographic and Health Surveys and demographic surveillance sites, and cancer-specific incidence and mortality rates in Europe from government data and local registries.

TABLE OF CONTENTS

LIST OF FIGURES

GLOSSARY

BRFSS: Behavioral Risk Factor Surveillance System

GMRF: Gaussian Markov Random Field

ICAR: Intrinsic Conditional Autoregressive

IGMRF: Intrinsic Gaussian Markov Random Field

INLA: Integrated nested Laplace approximation

MCMC: Markov Chain Monte Carlo

PHSKC: Public Health – Seattle & King County

SAE: Small Area Estimation

U5MR: Under 5 Mortality Rate

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor, Jon Wakefield. He has been patient and encouraging throughout my graduate career and his mentorship in statistics and the Seattle music scene has been invaluable. Jon asked more of me than I thought I was capable of and I am incredibly grateful.

I would like to thank my committee members for their helpful feedback related to my dissertation research as well as their contributions to my education and professional development. I was fortunate enough to take two courses from and serve as a teaching assistant for Barbara McKnight and from her, I learned an incredible amount about statistical communication in a collaborative setting. Through Adrian Raftery's working group I had the wonderful opportunity to share my research, receive feedback from faculty and my peers, and learn about research being conducted at UW and elsewhere. I attribute much of my professional presentation skills to my experiences in his working group. Last, but certainly not least, Stew Tolnay's course on fertility and mortality inspired my fascination with human demography and his clarifying and contextual insights made him my favorite person to talk to after the demography seminar.

From the University of Washington I would like to thank Ken Rice, Robyn McClelland, Thomas Richardson, Patrick Heagerty, Elena Erosheva, David Yanez, Nicole Hamblett, Peter Guttorp, Aimée Dechter, Galen Shorack, Fatema Mookhtiar, Maria Tebb, Ellen Reynolds and Gitana Garofalo for their contributions as instructors, supervisors, mentors and administrators during my time at the University of Washington. I would also like to thank my fellow students from the Biostatistics and Statistics departments, in particular Bailey Fosdick for encouraging me to return to graduate school, Charles Doss for being an inspiring

officemate, and Leigh Fisher and David Benkeser for being friends, teammates, and my go-to statistical resources. Finally I would like to thank the Statistics Department for giving me this opportunity, the Center for Statistics and the Social Sciences for teaching opportunities and funding through a Seed and Travel Grants and the Center for Studies in Demography and Ecology for the support through the National Institute of Child Health and Human Development training grant, T32 HD007543.

I would like to thank my collaborators Athena Pantazis and Sam Clark from UW Sociology; Angelina Lutambi and Honorati Masanja from the Ifakara Health Institute; Lin Song, Amy Laurent and David Solet from Public Health – Seattle & King County; and Jacques Ferlay, Marytn Plummer and Freddie Bray from the International Agency for Research on Cancer for sharing your data, your time, and your enthusiasm for new statistical methods for public health and demographic research.

I would like to thank my friends and family. Devon de Leña, Liza Reines, and Lindsay Jandl I am grateful for your patience and friendship. For their love and encouragement I would like to thank my family Cheryl and Sue Mercer; Rob and Andrew Vander Stoep; Brad, Cindy, and Mitch Hendren; Tim McCall; and Ron, Kathy, Brian, and Claire Delbecq. Finally, I would like to thank my husband Scott Delbecq, who has been supportive of my many academic, professional, and athletic adventures. I am extremely grateful to have such a supportive and loving partner.

DEDICATION

For my family and friends.

Now let's go have some fun!

Chapter 1
INTRODUCTION

To improve the health and well being of a population the most basic need it to understand the current status of health indicators within the population of interest. Ideally, one would want to know how this status has changed over time and how it might vary over space. Unfortunately, in many cases data is not collected with the frequency, consistency, or sample size required to produce reliable estimates of population characteristics of interest, at a granular spatiotemporal scale. The field of Small Area Estimation aims to estimate population characteristics, such as counts or means at a granular scale when at least some of the areas have insufficient samples for traditional methods. This dissertation will draw on the rich literature of space-time disease mapping and survey sampling to generate reliable estimates and associated uncertainty intervals based on data from multiple sources, such as complex surveys, cancer registries, and demographic surveillance sites. This chapter will include a brief introduction to motivating examples and describe the organization of the rest of the dissertation.

1.1 Motivating Examples

1.1.1 Small Area Estimation from Complex Surveys in Washington State

The Behavioral Risk Factor Surveillance System (BRFSS) is the main data source on behavioral risk factors across much of the U.S., providing state and county level estimates, and including ZIP code since 2005. This survey is carried out at the state level in the United States and is the largest telephone-based survey in the world. In the BRFSS survey, interviewees (who are 18 years or older) are asked a series of questions on their health behaviors and provide general demographic information, such as age, race, gender and the zip code in

which they live. The Center for Disease Control (CDC), which runs BRFSS, recommends a sample size of at least 50 respondents for direct estimation (Knutson et al., 2008). For planning interventions and allocating resources it is important for State and County health departments to have sub-state and sub-county level estimates of health indicators, such as smoking rates.

The focus of Chapter 3 is a comparison of various methods that have been proposed for small area estimation with complex surveys. In addition to methods that have been described in the literature we propose a new approach which uses a Bayesian disease mapping approach to spatially smooth the asymptotic distribution of area-level direct estimates and corresponding design-based variances. This chapter includes simulation studies comparing various methods under different non-response scenarios with the aide of post-stratification. These approaches are applied to BRFSS in Washington State to estimate smoking rates at the ZIP code level for the state and, in a collaboration with Public Health Seattle & King County, at the census tracts level within King County, Washington.

1.1.2 Space-Time Smoothing of Child Mortality Rates in Tanzania

In Chapter 4 we consider space-time smoothing of the under 5 mortality rate in resource limited settings. Small area estimates are necessary to understand geographical heterogeneity in health indicators when full-coverage vital statistics are not available. For this endeavor spatio-temporal smoothing is beneficial to alleviate problems of data sparsity. Many people living in low and middle-income countries are not covered by civil registration and vital statistics systems. Consequently, a wide variety of other types of data including many household sample surveys are used to estimate health and population indicators. The use of conventional hierarchical models requires careful thought since the survey weights may need to be considered to alleviate bias due to non-random sampling and non-response.

The application that motivated this work is estimation of child mortality rates in five-year time intervals in regions of Tanzania as part of a collaboration with the Ifakara Health Institute in Dar es Salaam. Data come from Demographic and Health Surveys conducted

over the period 1991–2010 and two demographic surveillance system sites. We derive a variance estimator of under five years child mortality that accounts for the complex survey weighting. For our application, the hierarchical models we consider include random effects for area, time and survey and we compare models using a variety of measures including the conditional predictive ordinate (CPO). The method we propose is implemented via the fast and accurate integrated nested Laplace approximation (INLA).

1.1.3 Modeling of Cancer Incidence and Mortality in Europe

The International Agency for Research on Cancer (IARC), which is the specialized cancer agency of the World Health Organization, provides estimates and predictions of the incidence and mortality of major types of cancer worldwide through the GLOBOCAN project (Parkin et al. 2001, Ferlay et al. 2010, Ferlay et al. 2013). These estimates are widely used by public health organizations globally. Data source and quality varies widely between countries, ranging from complete registry coverage to incomplete national mortality data in countries without vital registration. When data quality is low, the current IARC modeling procedure informally borrows information from neighboring countries and often relies on strong assumptions about incidence-mortality ratios. Their approach requires modeling each country separately and does not provide any measure of uncertainty.

In Chapter 5 we developed an approach to use the data of varying quantity and quality to generate estimates and associated uncertainty intervals for cancer-specific mortality and incidence rates in Europe. This method is applied to breast cancer incidence and mortality data for 40 countries in Europe from 1990-2010. This project is a collaboration with IARC to develop an approach for providing uncertainty estimates along with their incidence and mortality estimates and projections. Resulting estimates are compared with published estimates of age standardized rates from IARC (Ferlay et al., 2013) as well as estimates and uncertainty intervals of cumulative probabilities generated by the Institute for Health Metrics and Evaluation (IHME) (Forouzanfar et al., 2011).

1.2 Methodological Contributions of Dissertation

There are two primary methodological contributions presented in Chapter 3. The first is using the asymptotic distribution of the logit of the true prevalence along with the design-based variance as a working likelihood for small area estimation. The second is embedding the logit prevalence small area estimation model in a Bayesian multiple imputation procedure. The result is a general framework for small area estimation of an outcome that is obtained from a complex survey with missing spatial data.

There are two primary methodological contributions described in chapter 4. The first is the derivation of the design-based standard error for the child mortality rate based on discrete time survival analysis of household survey data. Previous analyses rely on a jackknife estimate. The second is building on the spatial model of the logit rate likelihood discussed in Chapter 3, we develop a space-time smoothing model for combining data from multiple surveys as well as surveillance data by incorporating survey random effects and the design-based variance of the child mortality rate. The result is a framework for smoothing child mortality rates at the sub-national scale over space, time, and data source.

The primary methodological contribution of the work detailed in Chapter 5 is to develop a framework that used different sources and amounts of data between and within countries of Europe. This work requires the development of a joint model of incidence and mortality given incidence as well as an unconditional model of mortality, which shares parameters with the joint model. The result is a coherent method that uses data from all countries, regardless of their available data, and provides reliable uncertainty intervals.

1.3 Organization of Dissertation

The aim of Chapter 2 is to provide a review of the statistical concepts which are heavily relied upon in Chapters 3, 4, and 5. The foundation of the research described in this dissertation is in survey sampling, Gaussian Markov Random Fields, and Bayesian methods. As such, we begin with a review of survey sampling. This is followed by a description of Gaussian

Markov Random Fields and a subset called Intrinsic Gaussian Markov Random Fields, which are often used as multivariate priors for spatial and temporal Bayesian modeling. Lastly, we describe the two methods used for Bayesian computation in this dissertation, Markov Chain Monte Carlo and the Integrated Nested Laplace Approximation.

The core methodological chapters will address the examples described above. A comparison of small area estimation and an application to BRFSS data in Washington state will be described in Chapter 3. Chapter 4 will describe the variance derivation and space-time smoothing of under five child mortality estimates from household surveys and demographic surveillance in Tanzania. A joint estimation approach for cancer incidence and mortality modeling will be described and evaluated in Chapter 5. Finally, Chapter 6 will conclude the dissertation with a summary and a discussion of future work.

Chapter 2

BACKGROUND

2.1 Survey Sampling

2.1.1 Notation

We first establish our notation. We will focus on binary outcomes, and let Y_{ik} represent the binary indicator for the event of interest on the k-th individual, $k = 1, \ldots, N_i$ in the i-th area, $i = 1, \ldots, I$. Common small area characteristics of interest include the true total count, $T_i = \sum_{k=1}^{N_i} Y_{ik}$, or the true proportion, $P_i = \frac{T_i}{N_i}$, in area i, $i = 1, \ldots, I$. We will follow the common convention in the survey sampling literature and denote population values with upper case letters and sampled values with lower case letters. To obtain estimates, a survey is conducted with probabilities of being sampled for the k-th person in area i being denoted π_{ik}. We use s_i to indicate the set of individuals who are sampled from area i with y_{ik} being the observed sample for $k \in s_i$ with $|s_i| = n_i$, so that the latter is the sample size in area i. The design weight for person k in area i we will be represented by w_{ik} and is calculated as the reciprocal of the sampling probability for selection, so that $w_{ik} = \pi_{ik}^{-1}$.

2.1.2 Horvitz–Thompson Estimator

We will focus on estimation of the population proportion P_i. A common and famous estimator was introduced by Horvitz and Thompson (1952). The Horvitz-Thompson estimator is

$$\widehat{P}_i = \frac{1}{N_i} \sum_{k \in s_i} \frac{y_{ik}}{\pi_{ik}}. \tag{2.1}$$

The estimated design variance of the estimator (2.1) of $\widehat{P}_i$, over the randomization distribution (i.e. over the distribution of all samples of fixed size n_i that could have been selected in

area i) is

$$\widehat{\text{var}}(\widehat{P}_i) = \frac{1}{N_i^2} \sum_{k \in s_i, k' \in s_i} \left(\frac{y_{ik} y_{ik'}}{\pi_{ik} \pi_{ik'}} - \frac{y_{ik} y_{ik'}}{\pi_{ikk'}} \right) \tag{2.2}$$

where $\pi_{ikk'}$ is the sampling probability for the pair of individuals k and k' in area i.

2.1.3 Hájek Estimator

Population totals N_i are often unavailable at the analysis stage for surveys with complex designs. For example the Demographic and Health Surveys, which are discussed extensively in Chapter 4 provide 'normalized' weights such that $\sum_{k \in s} w_k^\star = 1$. In these scenarios we can use the Hájek estimator (Hájek, 1971)

$$\widehat{P}_i = \sum_{k \in s_i} w_{ik} y_{ik} / \sum_{k \in s_i} w_{ik}. \tag{2.3}$$

Note, for the Demographic and Health Survey weights, we have $w_k^\star = w_k \times g$, where $g = 1/\sum_{k \in s} w_k$. The Hájek estimator is still a valid estimate of the population proportion because

$$\widehat{P}_i = \sum_{k \in s_i} w_{ik}^\star y_{ik} / \sum_{k \in s_i} w_{ik}^\star = \sum_{k \in s_i} g w_{ik} y_{ik} / \sum_{k \in s_i} g w_{ik} = \sum_{k \in s_i} w_{ik} y_{ik} / \sum_{k \in s_i} w_{ik}.$$

However, it is not possible to estimate population totals, calculated as $\widehat{T}_i = \sum_{k \in s_i} w_{ki} y_{ki}$, with the scaled weights used by the Demographic and Health Surveys.

2.1.4 Post Stratification

In scenarios where the sampling procedure does not allow for stratification along individual demographic characteristics it is often desirable to modify the sampling weights such that the distribution of a characteristic within the population is recovered. For example, if inference is desired on individuals 18 years and above based on a random digit dial phone survey, we may need to modify the weights so that the sum of the weights over age groups $j = 1, ..., J$, recovers a known population total $\sum_{i=1}^{I} \sum_{k \in s_j} w_{ik} = N_j$ is the set. A common strategy is to

produce sample weights as the product of the design weights and the post-stratification adjustment

$$w_{ik}^{\star} = g_j \times w_{ik} \tag{2.4}$$

for person k from area i and stratum j. The post-stratification adjustment is calculated $g_j = N_j / \sum_{i=1}^{I} \sum_{k \in s_j} w_{ik} = N_j / \hat{N}_j$. For a complete description of the construction of post-stratification weights, see Lumley (2010, Section 7.2). When post-stratification is carried out the Horvitz–Thompson estimator is

$$\widehat{P}_i = \frac{1}{N_i} \sum_{k \in s_i} w_{ik}^{\star} y_{ik}. \tag{2.5}$$

2.1.5 Raking

Post-stratification requires population totals for a complete cross-classification of all variables to be used. In many settings this can be difficult for variables beyond age and gender. Raking, which uses iterative proportional fitting, provides an alternative approach which only requires marginal population totals for the variables that are to be used in weight adjustment. The raking procedure is an iterative post-stratification over each variable until the weights stop changing. This procedure has become increasingly common in telephone-based surveys. As of 2011, BRFSS has changed to a raking procedure which includes telephone source, race and ethnicity, regions within state, education level, marital status, age group by gender, gender by race and ethnicity, age group by race and ethnicity, and home ownership status (CDC, 2011). For a further description of the construction of raking weights, see Lumley (2010, Section 7.3).

2.2 Gaussian Markov Random Fields

Assume $\boldsymbol{x} = [x_1, ..., x_n]^T \sim \mathcal{N}(\mu, \Sigma)$ and a labeled graph $G = \{V, E\}$ where $V = \{1, 2, ..., n\}$ and E is such that there is no edge between i and j if and only if $x_i \perp x_j | \boldsymbol{x}_{-ij}$. Then $\boldsymbol{x}$ is called a Gaussian Markov random field (GMRF) with respect to G with mean μ and precision

matrix $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$ if and only if its density is of the form

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

and $Q_{ij} \neq 0 \iff \{i,j\} \in E$ for all $i \neq j$. Furthermore, for any GMRF with respect to a labeled graph $G = \{V, E\}$ with mean μ and precision matrix $\boldsymbol{Q}$ we can express the conditional mean

$$E[x_i|\boldsymbol{x}_{-i}] = \mu_i - \frac{1}{Q_{ii}}\sum_{j:j\sim i} Q_{ij}(x_j - \mu_j)$$

and precision

$$\text{Prec}[x_i|\boldsymbol{x}_{-i}] = Q_{ii}.$$

A random vector $\boldsymbol{x} = [x_1, ..., x_n]^T \in R^n$ is called an intrinsic Gaussian Markov random field (IGMRF) with respect to a labeled graph $G = \{V, E\}$ with 'mean' μ and 'precision' matrix $\boldsymbol{Q}$ if and only if its density is of the form

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

where $\boldsymbol{Q}$ is rank deficient (rank$= n-k$ with $k \geq 1$) and $|\cdot|^*$ is a generalized determinant.

An IGMRF of first order is an improper GMRF of rank $n-1$ where $\boldsymbol{Q}1 = 0$. This implies $\sum_j Q_{ij} = 0$, for all i. The density for an IGMRF is invariant to the addition of any constant and thus the over all level is not specified. We assume $\mu = 0$ and the conditional mean of x_i is the weighted mean of its neighbors

$$E[x_i|\boldsymbol{x}_{-i}] = -\frac{1}{Q_{ii}}\sum_{j:j\sim i} Q_{ij}x_j$$

and the precision is

$$\text{Prec}[x_i|\boldsymbol{x}_{-i}] = Q_{ii}.$$

2.3 Temporal GMRF Models

2.3.1 First Order Random Walk

The first order random walk (RW1) is an IGMRF that is commonly used as a temporally structured prior in Bayesian analyses. If $\boldsymbol{x}$ is assigned a random walk prior, this implies

$$\pi(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x}\right)$$

with $\boldsymbol{Q} = \tau_x K_x$, and where K_x is known as the structure matrix (Clayton, 1996) which reflects the prior assumptions about the relationships between blocks of parameters within $\boldsymbol{x}$ and τ_x is an unknown scalar to be estimated from the data. For the first order random walk the structure matrix takes the form

$$K_x = \begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix}.$$

An alternative representation of the first order random walk relies on the conditional distributions

$$x_t|\boldsymbol{x}_{-t}, \tau_x \sim \mathcal{N}\left(\frac{1}{2}(x_{t-1} + x_{t+1}), 1/(2\tau_x)\right)$$

for $t = 2, ..., T-1$, and $x_1|\boldsymbol{x}_{-1}, \tau_x \sim \mathcal{N}\left(x_2, 1/(\tau_x)\right)$ and $x_T|\boldsymbol{x}_{-T}, \tau_x \sim \mathcal{N}\left(x_{T-1}, 1/(\tau_x)\right)$, for the first and last time points, respectively. Finally, this is also equivalent to penalizing the first order differences, that is, for $t = 1, ..., T-1$

$$(x_{t+1} - x_t)|\tau_x \sim N(0, \tau_x/2).$$

2.3.2 Second Order Random Walk

The second order random walk (RW2) is also an IGMRF that is commonly used as a temporally structured prior in Bayesian analyses. If $\boldsymbol{x}$ is assigned a second order random walk

prior, this implies

$$\pi(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x}\right)$$

where $\boldsymbol{Q} = \tau_x K_x$, τ_x is to be estimated from the data and the structure matrix is of the form

$$K_x = \begin{pmatrix} 1 & -2 & 1 & & & \\ -2 & 5 & -4 & 1 & & \\ 1 & -4 & 6 & -4 & 1 & \\ & \ddots & \ddots & \ddots & \ddots & \\ & 1 & -4 & 6 & -4 & 1 \\ & & 1 & -4 & 5 & -2 \\ & & & 1 & -2 & 1 \end{pmatrix}.$$

The RW2 can also be thought of as penalizing the second differences, that is, for $t = 2, ..., T-1$

$$(x_{t+1} - x_t) - (x_t - x_{t_1})|\tau_x \sim N(0, \tau_x/4)$$

or equivalently,

$$x_{t+1} - 2x_t + x_{t-1}|\tau_x \sim N(0, \tau_x/4).$$

2.4 Spatial GMRF Models

2.4.1 Intrinsic Autoregression

The Intrinsic Conditional Autoregression (ICAR) model is a non-parametric, stochastic smoothing model proposed by by Besag et al. (1991) with

$$U_i|\boldsymbol{U}_{-i}, \tau_u \sim \mathrm{N}\left(\overline{U}_i, \frac{1}{\tau_u m_i}\right), \tag{2.6}$$

where $j \sim i$ signifies that areas j and i are spatial neighbors (*e.g.* share a county boundary), m_i is the number of such neighbors, and $\overline{U}_i = \frac{1}{m_i}\sum_{j\sim i} U_j$ is the mean of the neighbors. The ICAR model is an IGMRF which can be jointly represented as

$$\begin{aligned} \pi(\boldsymbol{U}|\tau_u) &\propto \tau_u^{(k-1)/2}\exp\left(-\tfrac{\tau_u}{2}\sum_{i\sim j}(U_i - U_j)^2\right) \\ &= \tau_u^{(k-1)/2}\exp\left(-\tfrac{\tau_u}{2}\boldsymbol{U}^T K_u \boldsymbol{U}\right) \end{aligned}$$

where the structure matrix has elements

$$K_u = \begin{cases} m_i & i = j \\ -1 & j \sim i \\ 0 & \text{otherwise.} \end{cases}$$

Note, the ICAR is not a proper density as $\boldsymbol{Q}$ is rank deficient, by the number of connected subgraphs (at least one). For the RW1 of Section 2.3.1, RW2 of Section 2.3.2, and ICAR models the variances have *conditional* rather than *marginal* interpretations. The τ_u^{-1} parameter represents the variability of the components U_i conditional on the effects in neighboring areas as shown in (2.6).

2.4.2 The Convolution Model

If the ICAR model is used alone then all unstructured error will be modeled as spatially structured error. This could be misleading in limited data settings and when, say overdispersion is driving the error. To account for this Besag et al. (1991) also suggest including an unstructured random effect $\boldsymbol{V} \sim N(0, \tau_v^{-1} I)$. The total spatial contribution for an area i is $b_i = U_i + V_i$, where U_i is defined in (2.6). This model is often referred to as the 'convolution model' or the 'Besag-York-Mollié' (BYM). The resulting variance is

$$Var(\boldsymbol{b}|\tau_u, \tau_v) = \tau_v^{-1}\boldsymbol{I} + \tau_u^{-1}\boldsymbol{Q}^{-}$$

where $\boldsymbol{Q}^{-}$ is the generalized inverse of $\boldsymbol{Q}$.

2.4.3 The Leroux Model

One potential problem with the BYM approach is that the structured and unstructured terms are not identifiable. The formulation requires $2n$ parameters to summarize n spatial effects. Leroux et al. (2000) proposed a framework to use a weighted sum of the unstructured and structured spatial random effects. The random effects $\boldsymbol{b}$ follow a multivariate normal

distribution with mean zero and

$$Var(\boldsymbol{b}|\tau_b, \phi) = \tau_b^{-1} \left((1-\phi)\boldsymbol{I} + \phi \boldsymbol{Q}\right)^{-1} \tag{2.7}$$

where $\phi \in [0, 1]$ and describes the proportion of effect that is spatially structured. When $\phi = 0$ we have a pure unstructured overdispersion model and when $\phi = 1$ the strictly (structured) spatial ICAR model of 2.4.1. It is also natural to consider the conditional representation of the Leroux model

$$b_i|\boldsymbol{b}_{-i} \sim N\left(\frac{\phi}{1-\phi+\phi m_i}\sum_{j\sim i} b_i, \tau_b^{-1}(1-\phi-\phi m_i)^{-1}\right)$$

where m_i is the number of neighbors of area i as described in Section 2.4.1.

2.4.4 The Dean Model

The Dean Model (Dean et al., 2001) provides an alternative parameterization of the BYM model of Section 2.4.2, with mean

$$\boldsymbol{b} = \frac{1}{\sqrt{\tau_b}}\left(\sqrt{1-\phi}\boldsymbol{V} + \sqrt{\phi}\boldsymbol{U}\right)$$

and covariance

$$\text{Var}\left(\boldsymbol{b}|\tau_b, \phi\right) = \frac{1}{\tau_b}\left[(1-\phi)\,\boldsymbol{I} + \phi \boldsymbol{Q}^{-}\right].$$

The precision parameters of the BYM model can be written in terms of the Dean model parameters with $\tau_u = \tau_b/\phi$ and $\tau_v = \tau_b/(1-\phi)$. This parameterization allows tuning the proportion of variation due to independent or structured effects through the ϕ parameter as well as the overall magnitude of effects through the τ_b parameter. The ϕ and τ_b parameters of the Dean model have a more natural interpretation on the variance scale, compared with the Leroux model. However, the Dean model, like the BYM model, requires $2n$ random effects to describe n distinct areas, which can be a computational burden when n is large.

2.5 Bayesian Analysis and Computation

2.5.1 Hierarchical Bayesian models

At the first stage we employ the rare disease assumption and model lip cancer with a Poisson distribution $Y_i \sim \text{Poisson}(\mu_i)$ with

$$\log(\mu_i) = \log(E_i) + \alpha + U_i + V_i$$

where E_i is the expected case count based on the age distribution and a set of standard rates and U_i and V_i are random effects from the BYM model described in Section 2.4.2. An equivalent model specification is

$$\log(\mu_i) = \log(E_i) + U_i + V_i$$

with $\boldsymbol{V} \sim N(\alpha\mathbf{1}, \tau_v^{-1}\boldsymbol{I})$. This re-parameterization is called hierarchical centering and has been shown to have better mixing properties in an MCMC framework (Gelfand et al., 1995). Sum-to-zero constraints can be removed from samples of $\boldsymbol{U}$ to achieve a similar effect.

At the second stage we assign prior distributions to the parameters $\boldsymbol{\theta} = (\alpha, \boldsymbol{U}, \boldsymbol{V})$

$$\alpha|\alpha_0, \tau_0 \sim \text{flat},$$

$$\pi(\boldsymbol{V}|\tau_v) \propto \exp\left\{-\frac{\tau_v}{2}\left(\boldsymbol{V}^T\boldsymbol{V}\right)\right\},$$

and

$$\pi(\boldsymbol{U}|\tau_u) \propto \exp\left\{-\frac{\tau_u}{2}\left(\boldsymbol{U}^T\boldsymbol{K}\boldsymbol{U}\right)\right\}.$$

Where $\boldsymbol{K}$ is the structure matrix for $\boldsymbol{U}$, *i.e.* has the elements of $\boldsymbol{K}_u$ from Section 2.4.1. The final stage assigns priors to the hyperparameters

$$\tau_u \sim \text{Gamma}(a_u, b_u)$$

and

$$\tau_v \sim \text{Gamma}(a_v, b_v).$$

The a_u, b_u, a_v, b_v parameters are generally selected with a prior belief about the range of area level relative risks (Wakefield, 2009; Fong et al., 2010). Our goal is to learn about the distribution of the parameters of interest given the observed data, $p(\boldsymbol{\theta}|\boldsymbol{y})$. Using Bayes rule we have

$$p(\theta|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\theta) \times \pi(\theta)}{p(\boldsymbol{y})}$$

where $p(\boldsymbol{y}|\theta)$ is the data likelihood, $\pi(\theta)$ are the priors, and $p(\boldsymbol{y})$ is the normalizing constant.

2.5.2 Markov Chain Monte Carlo

Single Site Updates

For the random effects $i = 1, ..., n$ we have the posterior distributions

$$\pi(V_i|\boldsymbol{y}, \boldsymbol{U}, \tau_v) \propto N(\alpha, \tau_v^{-1}) \times \text{Pois}\left(E_i \exp(V_i + U_i)\right)$$

and

$$\pi(U_i|\boldsymbol{y}, \boldsymbol{V}, \tau_u) \propto N\left(\overline{U_i}, \frac{1}{\tau_u m_i}\right) \times \text{Pois}\left(E_i \exp(V_i + U_i)\right).$$

We can use a random walk proposal and the Metropolis Algorithm (Metropolis et al., 1953) with the following steps for a generic parameter θ_i and at iteration k:

1. Simulate a candidate value $\theta^* \sim N(\theta^{(k-1)}, c)$, where c is selected to achieve $\sim$ 20-30% acceptance (Roberts and Rosenthal, 2009).

2. Compute the acceptance probability

$$p_{ik} = \min\left\{\frac{p\left(\boldsymbol{y}|\theta_{-i}^{(k-1)}, \theta_i^*\right)}{p\left(\boldsymbol{y}|\theta^{(k-1)}\right)}, 1\right\}.$$

3. Generate $W \sim \text{Uniform}(0, 1)$.

4. Accept the proposed value $\theta_i^* = \theta_i^{(k)}$ if $W \leq p_{ik}$, otherwise set $\theta_i^{(k-1)} = \theta_i^{(k)}$.

For the precision parameters we can sample directly from the full conditionals

$$\pi(\tau_u|\boldsymbol{U}) \sim \text{Gamma}\left(a_u + (n-1)/2, b_u + \frac{1}{2}\boldsymbol{U}^T\boldsymbol{K}\boldsymbol{U}\right)$$

and

$$\pi(\tau_v|\boldsymbol{V}) \sim \text{Gamma}\left(a_v + n/2, b_v + \frac{1}{2}\boldsymbol{V}^T\boldsymbol{V}\right).$$

For single site updating in WinBUGS (Lunn et al., 2000) the hierarchical centering approach with $\boldsymbol{V}|\alpha, \tau_v \sim N(\alpha\mathbf{1}, \tau_v^{-1}\boldsymbol{I})$ has shown to dramatically improve chain mixing and decrease autocorrelation within posterior samples (Browne, 2004).

Block Updating

It has been shown that updating latent parameters with an IGMRF prior can improve MCMC chain mixing and improve the reliability of results from MCMC analyses (Rue, 2001; Knorr Held and Rue, 2002; Rue and Held, 2005). In this section we will consider a block updating of the ICAR parameters $\boldsymbol{U}$ which have joint posterior distribution

$$\pi(\boldsymbol{U}|\boldsymbol{y}, \boldsymbol{V}, \tau_u) \propto \exp\left\{-\frac{\tau_u}{2}\left(\boldsymbol{U}^T\boldsymbol{K}\boldsymbol{U}\right)\right\} \times \prod_{i=1}^{n} \text{Pois}\left(E_i \exp(V_i + U_i)\right).$$

We would like to propose points from a distribution that looks similar to the posterior of $\boldsymbol{U}|\boldsymbol{y}, \boldsymbol{V}, \tau_u$ so we attempt to find a GMRF that approximates $\pi(\boldsymbol{U}|\boldsymbol{y}, \boldsymbol{V}, \tau_u)$. Unfortunately, $\pi(\boldsymbol{U}|\boldsymbol{y}, \boldsymbol{V}, \tau_u)$ is not a GMRF, however we can use a Taylor approximation to find a quadratic approximation to the Poisson likelihood (see Appendix A.1 for details) and use as our proposal

$$\begin{aligned} q(\boldsymbol{U}|\boldsymbol{y}) &\propto \exp\left\{-\frac{\tau_u}{2}\left(\boldsymbol{U}^T\boldsymbol{K}\boldsymbol{U}\right)\right\} \times \prod_{i=1}^{n} \text{Pois}\left(\mu_i\right) \\ &\propto N_C\left(\boldsymbol{b}(\boldsymbol{u}_0), \boldsymbol{Q}(\boldsymbol{u}_0)\right), \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{b}(\boldsymbol{U}_0) &= \boldsymbol{y} - \exp(\alpha + \boldsymbol{V})\boldsymbol{E}B(\boldsymbol{u}_0) \\ \boldsymbol{Q}(\boldsymbol{U}_0) &= \tau_u \boldsymbol{K} + \exp(\alpha)\text{diag}(\boldsymbol{E}C(\boldsymbol{U}_0)) \\ B(a) &= \exp(a)(1-a) \\ C(a) &= \exp(a) \end{aligned}$$

and $\mathcal{N}_C(\boldsymbol{b}, \boldsymbol{Q})$ represents the Canonical parameterization of a GMRF (see Rue and Held, 2005, Definition 2.2). For a GMRF $\boldsymbol{x}$ with respect to some graph G and canonical parameters $\boldsymbol{b}$ and precision matrix $\boldsymbol{Q} > 0$ the density of $\boldsymbol{x}$ is

$$\pi(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x}\right)$$

written as $\boldsymbol{x} \sim N_C(\boldsymbol{b}, \boldsymbol{Q})$ which is equivalent to $N(\boldsymbol{\mu} = \boldsymbol{Q}^{-1}\boldsymbol{b}, \Sigma = \boldsymbol{Q}^{-1})$.

The $\boldsymbol{U}$'s are updated using the Metropolis-Hastings Algorithm (Metropolis et al., 1953; Hastings, 1970). Note that the approximation depends on $\boldsymbol{U}_0$. I have found that proposing points, $\boldsymbol{U}^*$, using $\boldsymbol{U}_0 = \boldsymbol{U}^{(k-1)}$, where $\boldsymbol{U}^{(k-1)}$ is the point from a previous iteration, generates reasonable proposals. First a point is generated from the proposal

$$\boldsymbol{U}^*|\boldsymbol{y}, \boldsymbol{U}^{(k-1)}, \tau^k, \alpha^{(k-1)} \sim \mathcal{N}_C\left(\boldsymbol{b}\left(\boldsymbol{U}^{(k-1)}\right), \boldsymbol{Q}\left(\boldsymbol{U}^{(k-1)}\right)\right)$$

using Algorithm 2.5 from Rue and Held (2005) or Algorithm 2.6 for sum-to-zero samples.

The Metropolis-Hastings Algorithm requires calculating the transition probabilities $q(\boldsymbol{U}^{(k-1)}|\boldsymbol{U}^*)$ and $q(\boldsymbol{U}^*|\boldsymbol{U}^{(k-1)})$. Evaluating $q(\boldsymbol{U}^{(k-1)}|\boldsymbol{U}^*)$, if we assume $\mathcal{N}_C\left(\boldsymbol{b}\left(\boldsymbol{U}^*\right), \boldsymbol{Q}\left(\boldsymbol{U}^*\right)\right)$, can generate extremely low density values. To improve the approximation we repeat the expansion around the point $\mu^* = \boldsymbol{Q}^{-1}\left(\boldsymbol{U}^*\right)\boldsymbol{b}\left(\boldsymbol{U}^*\right)$ and then evaluate the density of $\boldsymbol{U}^{(k-1)}$ for the distribution $N_C\left(\boldsymbol{b}\left(\mu^*\right), \boldsymbol{Q}\left(\mu^*\right)\right)$ (Knorr Held and Rue, 2002). The steps of the Metropolis-Hastings Algorithm are as follows:

1. Simulate a candidate value $\boldsymbol{U}^*$.

2. Compute the acceptance probability

$$p_{ik} = \min\left\{\frac{\pi(\boldsymbol{U}^*|\boldsymbol{y}, \alpha^{(k-1)}, \tau^{(k)})q(\boldsymbol{U}^{(k-1)}|\boldsymbol{U}^*)}{\pi(\boldsymbol{u}^{(k-1)}|\boldsymbol{y}, \alpha^{(k-1)}, \tau^{(k)})q(\boldsymbol{U}^*|\boldsymbol{U}^{(k-1)})}, 1\right\}.$$

3. Generate $W \sim \text{Uniform}(0, 1)$.

4. Accept the proposed value $\boldsymbol{U}^* = \boldsymbol{U}^{(k)}$ if $W \leq p_{ik}$, otherwise set $\boldsymbol{U}^{(k-1)} = \boldsymbol{U}^{(k)}$.

For convenience I recommend using Algorithm 2.5 from Rue and Held (2005) to model the intercept as the mean of the ICAR portion of the random effects.

2.5.3 Stan

As an alternative to Metropolis-Hastings or Gibbs sampling the Stan Development Team (2015b,a) has built a custom reverse-mode algorithmic differentiation package which allows fast calculation of the gradient of the log posterior which is required for Hamiltonian Monte Carlo (HMC) sampling (Duane et al., 1987), which is another MCMC algorithm. They implement HMC in Stan which can be used via the `RStan` package (Homan and Gelman, 2014; Carpenter et al., 2015) in `R`. Stan was developed to use Bayesian inference for hierarchical models with a large number of parameters that are highly correlated in the posterior. Under conditional sampling schemes these models tend to experience slow convergence. This is a common problem in Gibbs sampling or random walk Metropolis-Hastings algorithms. Slow convergence is often overcome by running very long MCMC chains or reparameterization. However, as the number of parameters increases, in particular the number of parameters that will be correlated in the posterior, these approaches can become prohibitively slow. The HMC reduces the correlation between samples and results in faster convergence.

2.5.4 Integrated Nested Laplace Approximation

MCMC methods for sampling from the posterior distribution can be computationally expensive when there are a large number of parameter or if parameters are correlated in the

posterior distribution. Correlation in the posterior is quite common in spatial and temporal models often requiring a large number of MCMC samples to effectively explore the posterior distribution. The Integrated Nested Laplace Approximation (INLA), introduced by Rue et al. (2009) and implemented in the `INLA` package (Lindgren and Rue, 2015), provides a fast alternative to MCMC for approximating the marginal posterior distribution for models with a latent Gaussian field and a non-Gaussian response. A brief description of INLA is provided below and the full details of INLA can be found in Rue et al. (2009).

Take the vector $\boldsymbol{y}$ as conditionally independent observations with likelihood $\pi(\boldsymbol{y}|\boldsymbol{x}, \theta_1, \theta_2)$. We assume that the latent variable $\boldsymbol{x}$ has the distribution$\mathcal{N}(\mu(\theta_1), \boldsymbol{Q}(\theta_2)) = \pi(\boldsymbol{x}|\theta)$, where θ is a hyperparameters with density $\pi(\boldsymbol{\theta})$. The posterior distribution takes the form

$$p(\boldsymbol{x}, \theta|\boldsymbol{y}) \propto \pi(\theta)|\boldsymbol{Q}(\theta_2)|^{1/2} \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}(\theta_2)\boldsymbol{x} + \sum_i \log\left(p(y_i|x_i, \theta_1)\right)\right).$$

INLA approximates the marginal posterior distributions through a combination of Laplace approximations and numerical integration. In our examples we are typically interested in the marginal distributions $p(\boldsymbol{x}|\boldsymbol{y})$ and $p(\theta|y)$. Adopting the notation of Rue et al. (2009), if we let $\tilde{\pi}(\cdot|\cdot)$ denote an approximate conditional distribution, we can write nested approximations of the marginal posterior distributions as

$$\tilde{\pi}(x_i|\boldsymbol{y}) = \int_\theta \tilde{\pi}(x_i|\boldsymbol{y}, \theta)\tilde{\pi}(\theta|\boldsymbol{y})d\theta, \tag{2.8}$$

$$\tilde{\pi}(\theta_j|\boldsymbol{y}) = \int_{\theta_{-j}} \tilde{\pi}(\theta|\boldsymbol{y})d\theta_{-j}. \tag{2.9}$$

Numerical integration techniques are used to integrate out θ in (2.8) and θ_{-j} in (2.9) and the approximation of the marginal posterior of θ is

$$\tilde{\pi}(\theta|\boldsymbol{y}) \propto \frac{\pi(\boldsymbol{x}, \theta, \boldsymbol{y})}{\tilde{\pi_G}(\boldsymbol{x}|\theta, \boldsymbol{y})}), \text{ evaluated at } \boldsymbol{x} = \boldsymbol{x}^*(\theta)$$

where $\tilde{\pi_G}(\boldsymbol{x}|\theta, \boldsymbol{y})$ is the Gaussian approximation to the full condition of $\boldsymbol{x}$, and $\boldsymbol{x}^*(\theta)$ is the mode of $\pi(\boldsymbol{x}|\theta)$. Additional details can be found in Rue et al. (2009).

2.5.5 Methods Used in this Dissertation

All of the methods for Bayesian analysis and computation mentioned in Section 2.5 were implemented at some stage of research for the dissertation. INLA is very fast and an ideal implementation method if applicable. We were able to use INLA for the methods described in Chapters 3 and 4. However, the methods described in Chapter 5 require a non-linear transformation of parameters, which INLA cannot accommodate. Single site and blocked updates were implemented for Chapter 5, but Stan was selected because it both faster and provided posterior samples with lower autocorrelation than the other MCMC methods.

Chapter 3

SMALL AREA ESTIMATION WITH COMPLEX SURVEYS

3.1 Introduction

Small area estimation (SAE) is used in many fields including education, epidemiology, and global, environmental and public health. Often the surveys carried out to inform SAE are complex in nature, with non-random sampling being carried out for reasons of necessity (i.e., logistical reasons) or to ensure that certain populations of interest are well represented. In addition, post-stratification may be used to re-weight the observations in order to recover known population totals. This approach can account for non-response within the strata used in the post-stratification.

There are two approaches to modeling complex survey data that we shall consider in this chapter. In the first, *design-based* approach weighted estimators are considered, with inference carried out based on the randomization distribution of the samples that could have been collected, i.e., the distribution of the individuals that could appear in the sample. In contrast, a *model-based* approach assumes a hypothetical infinite population from which the responses are drawn. While appealing from a conceptual point of view (since standard statistical modeling machinery can be leaned upon), the modeling approach is difficult to implement since one must model the sampling mechanism, if informative, at least to some extent. For example, if non-random sampling is based on particular inclusion variables (e.g., race or geographical area) then these variables must be included in the model if they are associated with the outcome of interest. Similarly, variables that affect the probabilities of non-response must also be included in the model, again if they are related to the outcome.

The alternative is to assume that variables upon which sampling is based and non-response depends are unrelated to the outcome of interest, which can lead to inaccurate

conclusions. Another impediment to the model-based approach is that the key variables that are required for inclusion may be unavailable in public-use databases, and even if available, the sampling scheme may be highly complex, requiring a model which has a large number of parameters and is therefore difficult to fit. Additionally, in the developing world context, the sub-national population data needed to generate prevalence estimates may be unavailable. Gelman (2007b) describes the issues, and the accompanying discussion (Bell and Cohen, 2007; Breidt and Opsomer, 2007; Little, 2007; Lohr, 2007; Pfefferman, 2007; Gelman, 2007a) gives a range of perspectives on the use of weighted estimators, regression modeling, or a combination of the two.

In this chapter we will consider SAE in the situation in which either the variables upon which sampling was based are unavailable or the scheme is so complex that a simpler approach is desired. SAE has seen a great deal of research interest, with Rao (2003) being a classic text. In the related field of disease mapping, the use of spatial modeling is commonplace (Wakefield et al., 2000), but in this context the data usually consist of a complete enumeration of disease cases in an area, so that no weighting scheme needs to be considered. It is the existence of the weights that causes a major difficulty when one wishes to use spatial smoothing in SAE, and consequently there are relatively few instances of approaches that use spatial smoothing within a model that acknowledges the sampling scheme. In Chen et al. (2014) a new method of incorporating the weights within a spatial hierarchical model was introduced, and various random effects models were compared via simulation. In this chapter we compare the method suggested by Chen et al. (2014) and other suggested methods for accounting for the weighting with a new approach.

As a motivating example, we examine data from the Behavioral Risk Factor Surveillance System (BRFSS). This survey is carried out at the state level in the United States and is the largest telephone-based survey in the world. In the BRFSS survey, interviewees (who are 18 years or older) are asked a series of questions on their health behaviors and provide general demographic information, such as age, race, gender and the ZIP code in which they live. In this chapter we will focus on the surveys conducted in Washington State in 2006 and

2009–2013. In Section 3.6 we will described an analysis of the 2009–2013 BRFSS data to provide census tract estimates of smoking rates in King County, Washington. However, for methodological development and comparisons will focus on the survey conducted in Washington State in 2006, and on the Centers for Disease Control (CDC) calculated variable *Adults who are current smokers*.

With respect to the smoking question, $19,502$ respond with "No", $3,733$ with "Yes" and 132 were classified as "don't know/refuse/missing". In the analysis, we remove these latter values. The response variable is therefore a binary indicator and our objective is to estimate the number of individuals who are 18 or older and who are current smokers, in each of 498 ZIP codes in Washington State. We also utilize population estimates from 2006. Table 3.1 summarizes the population and response data. The number of samples per ZIP code shows large variability with a median of 30 and minimum and maximum values of 1 and 384. The spread is apparent in Figure 3.1. Figure 3.11 maps, by ZIP code, the observed number of smokers in the sample (top) and the sample sizes (bottom) and the spatial variability in each map is evident.

Table 3.1: Summary statistics for population data, and the 2006 Washington State BRFSS data on adult current smokers, across ZIP codes.

	Mean	S.D.	Median	Min	Max
Population	12570.0	12931.0	7208.0	11.0	55700.0
Sample Sizes	46.9	54.8	30.0	1.0	384.0
Number of current adult smokers	7.5	9.5	4.0	0.0	67.0

We now describe in greater detail the complex survey scheme that was used by BRFSS in 2006. In this year, the BRFSS survey used land-lines only, and utilized a disproportionate stratified random sample scheme with stratification by county and "phone likelihood". Under this scheme in each county, based on previous surveys, blocks of 100 telephone numbers were classified into strata that are either "likely" or "unlikely" to yield residential numbers.

Telephone numbers in the "likely" strata are then sampled at a higher rate than their "unlikely" counterparts. Once a person is reached at a phone number the number of eligible adults (aged 18 or over) is determined, and one of these is randomly selected for interview. The sample weight, `Sample Wt`, is then calculated as the product of four terms

$$\texttt{Sample Wt} = \texttt{Strat Wt} \times \frac{1}{\texttt{No Telephones}} \times \texttt{No Adults} \times \texttt{Post Strat Wt} \tag{3.1}$$

where `Strat Wt` is the inverse probability of a "likely" or "unlikely" stratum being selected in a particular county, `No Telephones` represents the number of residential telephones in the respondent's household, `No Adults` is the number of adults in the household, and `Post Strat Wt` is the post-stratification correction factor. The latter is based on the number of people in strata defined by gender and age, using the 7 age groups 18–24, 25–34, 35–44, 45–54, 55–64, 65–74, 75+. Estimation will be based on the respondent's outcome, with an accompanying weight, and the population information. Crucially, we will also examine the possibility of leveraging geographic information to smooth rates across ZIP codes.

The structure of the chapter is as follows. In Section 3.2 we describe a number of approaches to formulating hierarchical models that incorporate weighting and in Section 3.4 a number of these methods are compared via a simulation study. In Section 3.5 we return to the 2006 BRFSS data. In Section 3.6 we extend methods with an imputation approach applied to BRFSS data in King County and the paper concludes with a discussion in Section 3.7.

3.2 Review of Small Area Estimation with Sampling Weights

3.2.1 Sampling Models

In this section, in anticipation of the development of a hierarchical smoothing model, we consider various approaches to constructing a likelihood for the observed data.

The simplest approach is to ignore the design and take

$$y_i | P_i \sim \text{Binomial}(m_i, P_i), \tag{3.2}$$

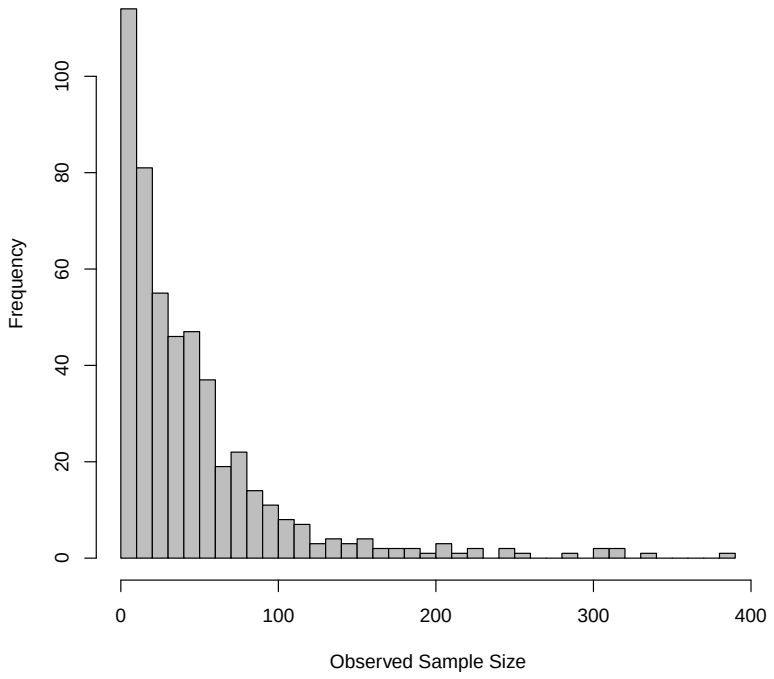

Figure 3.1: For 2006 Washington BRFSS data: histograms of actual sample sizes by ZIP code.

where $y_i = \sum_{k \in s_i} y_{ik}$, where s_i is the indices for the m_i sampled individuals in area i. If the design is informative we would expect bias in the estimator y_i/m_i of P_i. Ignoring the weighting and using this model is often carried out in SAE, for examples see Rao (2003).

As a second method, a number of authors have suggested an approach in which the likelihood is weighted. This approach is often referred to as *pseudo-likelihood*, to acknowledge the sampling design; early references are Binder (1983) and Skinner (1989). In the version we implement the weights are scaled as

$$w_{ik}^{s} = m_i \times \frac{w_{ik}}{\sum_{k=1}^{m_i} w_{ik}},$$

as recommended by Pfeffermann et al. (1998), see also Asparouhov (2006). Defining $y_i^{\mathrm{P}} =$

$\sum_{k \in s_i} w_{ik}^s y_{ik}$, the likelihood is taken as

$$y_i^{\mathrm{P}} | P_i \sim \text{Binomial}(m_i, P_i) \tag{3.3}$$

with the pseudo-log likelihood

$$l(P_i) = y_i^{\mathrm{P}} \log(P_i) + (m_i - y_i^{\mathrm{P}}) \log(1 - P_i).$$

The pseudo-likelihood approach has been used with a spatial smoothing model by Congdon and Lloyd (2010), with the weights being scaled to sum to the sample size m_i. These authors estimate diabetes prevalence for ZIP Code Tabulation Areas (ZCTAs) in the U.S. A drawback with the general approach is that the appropriate standard error is not recovered in the case of clustering. Rabe-Hesketh and Skrondal (2006) utilize a pseudo-likelihood with scaled weights, and use sandwich estimation to provide valid standard error estimates. These authors embed this approach within a multilevel framework but do not consider spatial smoothing. The scaling of the weights with respect to the effective size has been considered by a number of authors including Potthoff et al. (1992) and Longford (1996), with the latter explicitly considering variance components models, though again without considering spatial smoothing.

In a far-reaching paper, Raghunathan et al. (2007) describe an approach for combining data from different surveys. In this chapter we consider their approach as applied to a single survey. They utilize the arcsin square root (also known as the angular) transform: $y_i^{\mathrm{A}} = \arcsin\left(\sqrt{\widehat{P}_i}\right)$ which is the approximate variance stabilizing transformation for binary data and results in $-\pi/2 < y_i^{\mathrm{A}} < \pi/2$. This, and closely related, transforms are discussed in Anscombe (1948). The "effective sample size" m_i^{E} is obtained by solving $\widehat{P}_i(1 - \widehat{P}_i)/m_i^{\mathrm{E}} = \widehat{\mathrm{var}}\left(\widehat{P}_i\right)$ to give

$$m_i^{\mathrm{E}} = \frac{\widehat{P}_i(1 - \widehat{P}_i)}{\widehat{\mathrm{var}}\left(\widehat{P}_i\right)}. \tag{3.4}$$

The asymptotic likelihood used by Raghunathan et al. (2007) is

$$y_i^{\mathrm{A}} | P_i \sim \mathrm{N}\left(\arcsin\left(\sqrt{P_i}\right), \frac{1}{4m_i^{\mathrm{E}}}\right). \tag{3.5}$$

Raghunathan et al. (2007) use (3.5) as the first stage within a hierarchical model, but do not introduce spatial random effects.

Chen et al. (2014) build on this approach by defining the effective number of cases y_i^{E} as the product of the effective sample size m_i^{E} and the weighted proportion $\widehat{P}_i$ to give

$$y_i^{\mathrm{E}} = m_i^{\mathrm{E}} \times \widehat{P}_i. \tag{3.6}$$

The likelihood they assume is

$$y_i^{\mathrm{E}} | P_i \sim \mathrm{Binomial}(m_i^{\mathrm{E}}, P_i).$$

The rationale here is that both numerator and denominator are adjusted for the sampling design, and the use of a binomial likelihood, though not the "true" likelihood, will better reflect the sampling distribution than a normal approximation. In Chen et al. (2014) different random effects models were compared, using the proposed method and direct estimation. A technical but important detail is that the above approach runs into difficulty when $\widehat{P}_i = 0/1$. In these cases we use a smoothing procedure described in Chen et al. (2014) to produce an effective sample size. Briefly, when $\widehat{P}_i = 0/1$ m_i^{E} is undefined. Assuming I areas and J stratum (such as age groups) within area i we have

$$\widetilde{P}_i = \frac{\sum_{j=1}^{J} m_{ij} \widetilde{P_{ij}}}{m_i},$$

where stratum-specific proportions are estimated by assuming P_{ij} follow a beta-binomial model with stratum-specific parameters α_j and β_j. Specifically, we assume

$$P_{ij} | \alpha_j, \beta_j \sim \mathrm{Beta}(\alpha_j, \beta_j), \quad i = 1, ..., I.$$

The stratum-specific parameters, α_j and β_j, are estimated using a methods of moments approach. The full procedure is described in the supplement of Chen et al. (2014).

As an alternative, we propose an approach which summarizes the data in area i via the asymptotic distribution of the Horvitz and Thompson (1952) estimator (or the post-stratified version (2.5)), which we denote $\widehat{P}_i = \widehat{T}_i / N_i$, where $\widehat{P}_i = \frac{1}{N_i} \sum_{k \in s_i} y_k w_k$ and w_k is the sampling

weight. We can also calculate the associated variance estimator $\text{var}(\widehat{P}_i) = \text{var}(\widehat{T}_i)/N_i^2$, where the form of $\text{var}(\widehat{T}_i)$ is given in (2.2). In this way the design is acknowledged in both the estimator and the variance. We could simply take $\widehat{P}_i|P_i \sim \text{N}\left(P_i, \widehat{\text{var}}(\widehat{P}_i)\right)$, but this does not constrain the probability to lie in (0,1) which we might anticipate would cause difficulties, in particular in areas with small m_i. As an alternative we define the area-level data summary as $y_i^{\text{L}} = \log\left[\widehat{P}_i/(1-\widehat{P}_i)\right]$ as the empirical logistic transform of $\widehat{P}_i$. The likelihood is then taken as the asymptotic distribution

$$y_i^{\text{L}}|P_i \sim \text{N}\left(\log\left[\frac{P_i}{1-P_i}\right], \frac{\widehat{\text{var}}(\widehat{P}_i)}{\widehat{P}_i^2(1-\widehat{P}_i)^2}\right). \tag{3.7}$$

3.2.2 Hierarchical models

We examine the use of three-stage models with the first stage given by one the forms, (3.2), (3.7), (3.3), (3.5), (3.6) described in the previous sections. At the second stage of the model we introduce the random effects on the "natural scale" and denote the area-specific parameter on this scale as θ_i. For model (3.5) this is the arcsin square root scale, while for all other models it is the logistic scale. We consider two different second stages for each of the five likelihoods, independent normal random effects only, and independent plus spatial random effects. The non-spatial normal second stage is defined as

$$\theta_i = \beta_0 + V_i, \tag{3.8}$$

with $V_i|\sigma_v^2 \sim_{iid} \text{N}(0, \sigma_v^2)$, so that for four of the models $\exp(\beta_0)$ is the area-level odds of the event of interest in an area with $V_i = 0$. Another interpretation is as the median odds of the event of interest across areas. For model (3.5), β_0 is the the arcsin square root of the frequency of the event of interest in an area with $V_i = 0$.

The second random effects model we consider is the convolution' model (Besag et al., 1991) described in Section 2.4.2:

$$\theta_i = \beta_0 + V_i + U_i, \tag{3.9}$$

with $V_i|\sigma_v^2 \sim_{iid} \text{N}(0, \sigma_v^2)$ and U_i following an ICAR as described in Section 2.4.1. In what

follows we take the conventional approach of spatial epidemiology in which two areas are considered neighbors if they share a common boundary.

In general one may specify proper subjective priors for σ_v^2 and σ_u^2 based on the context. However, when one carries out a simulation study to evaluate Bayesian procedures, one is faced with the thorny question of which priors to use, so that one does not favor one approach over another. In our case this is pertinent since the two scales have quite different ranges on the random effects, for example, the whole of the real line for the logit models, and $[-\pi/2, \pi/2]$ for the arcsin square root model.

Browne and Draper (2006a) carried out an extensive simulation study in which the bias and coverage probabilities were examined as a function of various characteristics, including the priors on the variance components. They found that the uniform prior $U(0, 1/\epsilon)$ (for small ϵ), or an improper uniform prior, on a generic random effects σ^2 produced reasonable behavior. Lambert (2006) prefers uniform priors on the standard deviation, σ, which is further supported by Gelman (2006) who states: "In fitting hierarchical models, we recommend starting with a non-informative uniform prior density on standard deviation parameters σ". This view is also supported by Browne and Draper (2006b). In the simulation study of Section 3.4 we take an improper uniform prior on σ_v and, to aid in stability, a Gamma(0,5,0.008) prior on the spatial conditional precision σ_u^{-2}. The latter prior gives a 95% range on the more interpretable σ_u scale of (0.056,4.04). We use an improper flat prior on β_0.

3.2.3 Inference for Counts

The point estimate of the population count of interest T_i is

$$\widehat{T}_i = \widehat{P}_i \times N_i, \tag{3.10}$$

where $\widehat{P}_i$ is the direct estimator (2.5) and the variance is

$$\widehat{\text{var}}(\widehat{T}_i) = \widehat{\text{var}}(\widehat{P}_i) \times N_i^2. \tag{3.11}$$

Under a Bayesian approach one may summarize the posterior distribution for T_i using quantiles. If a point estimate is required then it is given by (3.10) with $\widehat{P}_i$ being replaced by the

posterior mean or median. The posterior variance $\text{var}(T_i|y)$ is given by (3.11) with $\widehat{\text{var}}(\widehat{P}_i)$ replaced by the posterior variance $\text{var}(P_i|y)$.

3.3 Implementation

To implement the methods compared in this chapter we have utilized the `survey` and `INLA` packages within the `R` computing environment. The `survey` package (Lumley, 2010) is used to obtain the direct estimates and appropriate variances (and hence effective sample sizes). To fit the random effects models we use the integrated nested Laplace approximation (INLA) as implemented by Lindgren and Rue (2015). The INLA approximation is extremely fast and the method has been investigated in many different scenarios with Fong et al. (2010) looking specifically at generalized linear mixed models. For additional details see Section 2.5.4 and Rue et al. (2009). The code to fit the models described in the chapter is available at `http://faculty.washington.edu/jonno/software.html`.

3.4 Simulation Study

We now present a simulation study to compare five of the estimators described in the previous section. The estimators we compare are the naive binomial (3.2), pseudo-likelihood (3.3), the arcsin square root transform (3.5), the numerator and denominator effective sample size adjusted binomial (3.6) and the logit normal (3.7). In each case we consider two random effects models: independent random effects only, and the convolution model with both independent and spatial ICAR random effects. We also include a pair of models with no hierarchical smoothing. One uses the binomial model (3.2) and the other uses the Horvitz–Thompson estimator and its associated variance estimator (and therefore adjusts for design bias). We follow Chen et al. (2014) and report two sets of simulations, one to address non-response bias, and another to address selection bias.

To evaluate the estimates, three statistics will be compared: the squared bias, the variance and the mean squared error (MSE). Let S denote the total number of simulations (which we take as $S = 100$) and T_i the true (but unobserved) count of the event of interest in area i

(which is kept constant across simulations). The summary statistics are

$$\begin{aligned}
\text{Bias} &= \frac{1}{I}\sum_{i=1}^{I}\left(\overline{\widehat{T}}_i - T_i\right), \quad \text{where} \quad \overline{\widehat{T}}_i = \frac{1}{S}\sum_{s=1}^{S}\widehat{T}_i^{(s)}, \\
\text{Variance} &= \frac{1}{I}\sum_{i=1}^{I}\left(\frac{1}{S-1}\sum_{s=1}^{S}(\widehat{T}_i^{(s)} - \overline{\widehat{T}}_i)^2\right), \\
\text{MSE} &= \text{Bias}^2 + \text{Variance}
\end{aligned}$$

and we expect methods with good performance to display low MSE.

3.4.1 Non-Response Bias

In all simulation studies, we take as geography the ZIP codes of Washington State. For direct comparison with the results in Chen et al. (2014) we set the parameters of the simulation based on diabetes. We simulate cases using a probability of diabetes p_{ij} for individuals in area i and post-stratification group j. There are $J = 6$ groups consisting of three age bands and two genders. We examine five scenarios with varying prevalence and response rates. In each of the five scenarios, simple random sampling of individuals is carried out within each area. However, individuals within area i respond to the survey within post-stratification group j with response probabilities q_{ij}, $i = 1, \ldots, I$, $j = 1, \ldots, J$. Thus, we assume that missingness depends on post-stratification group only and so, conditional on the post-stratification group, the missingness does not depend on the observed response. The sample sizes m_i are taken as the actual number of individuals who responded in the Washington State 2006 BRFSS survey. The five scenarios are:

Scenario 1:

In scenario 1, we consider the ideal situation in which every individual selected responds to the survey. The prevalence of diabetes in area i and group j, are the same in each area so that $p_{ij} = p_j$; these values are given in Table 3.2.

Scenario 2:

In scenario 2, we introduce non-response with the response rate being the same in each

Table 3.2: Diabetes prevalence rates p_{ij} in area i, $i = 1, \ldots, I$, and by post-stratification group, $j = 1, ..., 6$, corresponding to age and gender. In scenarios 1, 2, 3 and 5 the rates are fixed across areas and the values listed are based on the National Surveillance Data from the CDC (Chen et al., 2014). In scenario 4 the values vary, with spatial structure, across areas, with the first figure in each cell denoting the median rate, and the figures in parentheses a 95% range.

		Age		
	Scenario	18-44	45-74	75+
Female	1, 2, 3, 5	0.017	0.15	0.17
	4	0.017 (0, 0.034)	0.15 (0.085, 0.21)	0.17 (0, 0.32)
Male	1, 2, 3, 5	0.014	0.16	0.19
	4	0.014 (0, 0.027)	0.16 (0.089, 0.23)	0.19 (0, 0.33)

area but differing by age and gender, i.e. $q_{ij} = q_j$ for $j = 1, ..., 6$. The rates used are given in Table 3.3. The response rates increase with age, and women have higher rates than men.

Scenario 3:

In this scenario the response rates for each group vary between areas via the stochastic relationship:

$$\text{logit}(q_{ij}) = \text{logit}(q_j) + b \times \epsilon_i,$$

where $\epsilon_i \sim_{iid} \text{N}(0, 1)$. The median response rate, $\exp(q_j)/[1 + \exp(q_j)]$ is the same as in Scenario 2. We set $b = 0.35$ to give 95% ranges for the response rates in each of the six groups as in the figures in parentheses in Table 3.3.

Scenario 4:

In scenario 4, we introduce spatial dependence into the underlying prevalence rates. This dependency is induced by adding a spatially correlated area-level covariate x_i:

$$\text{logit}(p_{ij}) = \text{logit}(p_j) + b \times x_i.$$

To simulate spatially correlated covariates x_i, we employ a zero mean, unit variance ICAR

Table 3.3: Response rates q_{ij} in area i, $i = 1, \ldots, I$ and by age and gender groups, $j = 1, \ldots, 6$. In scenarios 1 and 4 there is full response. In scenario 2 the response rates are fixed across areas but vary by group. In scenario 3 the response rates vary, without spatial structure, across areas, with the first figure denoting the median rate, and the figures in parentheses a 95% range. In scenario 5 the response rates vary, with spatial structure, across areas, with the first figure in each cell denoting the median rate, and the figures in parentheses a 95% range.

		Age		
	Scenario	18-44	45-74	75+
Female	1, 4	1	1	1
	2	0.55	0.65	0.8
	3	0.55 (0.38, 0.70)	0.65 (0.48, 0.79)	0.80 (0.67, 0.89)
	5	0.55 (0.46, 0.65)	0.65 (0.57, 0.74)	0.80 (0.74, 0.86)
Male	1, 4	1	1	1
	2	0.50	0.60	0.75
	3	0.50 (0.34, 0.66)	0.60 (0.43, 0.75)	0.75 (0.60, 0.86)
	5	0.50 (0.41, 0.60)	0.60 (0.51, 0.69)	0.75 (0.68, 0.82)

model. Details on how to simulate from ICAR models can be found in Rue and Held (2005). We choose $b = 0.2$ to allow variation in the prevalence rates between area; Table 3.2 gives the marginal (across areas) 95% ranges for the prevalence rates. In this scenario everyone responds.

Scenario 5:

In scenario 5, we allow the response rate for each group to vary between areas by adding a spatial component to the variation:

$$\text{logit}(q_{ij}) = \text{logit}(q_j) + b \times x_i,$$

where x_i is again simulated from a zero mean, unit variance ICAR model. We let $b = 0.3$ to give the 95% ranges in Table 3.3.

The results of this simulation are summarized in Table 3.4 and we make the following observations:

- The non-hierarchical models have the lowest bias since there is no shrinkage. In scenarios 1 and 4 in which there is no non-response the unadjusted estimator has smallest bias while for all other scenarios the Horvitz-Thompson estimator is best. For the hierarchical models, the spatial models have lower bias than the independent models.

- The variance of the non-hierarchical models is very large, clearly showing the benefits of hierarchical modeling, as is well-known. The unadjusted binomial, logit and pseudo-likelihood models have relatively low variance with the independent version of the last of these giving the lowest variance.

- In terms of MSE the effective sample size spatial model gives the best performance in the three scenarios in which there is non-response bias (scenarios 2, 3 and 5), closely followed by the pseudo-likelihood spatial model, with the arcsin square root having the next best performance. In scenarios 1 and 4 (which recall have full response) the spatial unadjusted binomial model slightly out-performs the effective sample size model.

3.4.2 Selection Bias

To investigate the potential for selection bias, we let Z_{ik} denote a binary design variable that dictates whether the k-th individual in area i is selected to be surveyed or not. We use the population simulated from scenario 1 in the simulation study for non-response, and assign the status of the design variable for individual k in area i based on

$$\Pr(Z_{ik} = 1|Y_{ik} = 1) = s, \quad \Pr(Z_{ik} = 1|Y_{ik} = 0) = 0.1.$$

These probabilities provide a correlation between the design variable Z and the outcome variable Y and we examine the extent of the correlation by assigning s values of 0.1, 0.3, 0.5 and 0.8. We still take the ZIP code geography of Washington State and the total sample size m of the 2006 Washington State BRFSS. Let p_i^z denote the proportion of population with $Z = 1$ in area i. We set the sample size $m_i = mp_i^z / \sum p_i^z$, and within each area $m_i/2$ are selected with $Z = 1$ and $m_i/2$ with $Z = 0$. Oversampling populations with certain characteristics is a common technique in surveys. The information on the variable Z is only used when conducting the survey (and in calculating the sample weights), and is considered unavailable at the time of analysis.

Table 3.5 gives the results, and the picture is not as clear cut as with the non-response set of simulations. However, we make the following observations:

- In terms of bias, again the non-hierarchical approaches are best since they do not employ shrinkage. The adjusted estimators have the lowest bias in all cases but the one in which there is no selection bias ($s = 0.1$), in which case the unadjusted estimator performs best.

- With respect to variance, the independent random effects models perform best with the unadjusted binomial having the lowest variance in all cases but the one with the worst selection bias. For this case the independent pseudo-likelihood model has the lowest variance.

- In terms of MSE, if there is no selection bias the unadjusted hierarchical models both perform well. For the second and third levels of selection bias ($s = 0.3, 0.5$) the pseudo-likelihood approaches perform best, with the spatial versions being a little better than the independent version. For the most extreme selection bias case each of the adjusted hierarchical models perform reasonably, with the effective sample size model being best.

In conclusion, no one model is superior in all situations, though adjustment for the weights almost always provides the lowest MSE and hierarchical smoothing is clearly a good idea.

3.5 Washington State 2006 ZIP code BRFSS Example

We apply the sample weighted Bayesian hierarchical models we described in Section 3.2 to the Washington State 2006 BRFSS data introduced in Section 3.1. Sampling weights are taken to be the final weights used in the BRFSS survey, as in (3.1). These weights range between 1.2 and 4675 across ZIP code. The effective sample sizes and number of observations used in the effective sample size approach are calculated using the design-based Horvitz-Thompson variance estimator. Figure 3.3 gives the effective sample sizes, as calculated from (3.4), plotted against the observed sample sizes. In the majority of cases the effective sample size is lower than the observed sample size, so that the design is producing a loss in information, when compared to simple random sampling, in those areas.

We fitted each of the models that were considered in the simulations. With respect to priors, a Ga(0.5,0.0008) prior was initially taken for each of the spatial precision, σ_u^{-2}, and an improper uniform prior on the non-spatial standard deviation, σ_v. Later we examine sensitivity to the prior on σ_u^{-2}.

Figure 3.4 presents the boxplots of logit-transformed estimates of adult smoking prevalence by ZIP code under different approaches. For comparison we also include the design-based estimates, which are denoted as "Direct" in the figure. It is clear that the direct estimates exhibit a large amount of between ZIP code variations, with some extreme values.

The variation is significantly reduced by all of the Bayesian hierarchical models. The pseudo-likelihood approach and the effective sample size approach give very similar estimated adult smoking prevalences both in terms of the location and spread. The boxplots for the unadjusted binomial model have a slightly lower location, reflecting selection and non-response bias.

As we saw in Section 3.4, choosing an appropriate hierarchical model is not straightforward, with one possibility being to report not a single set of estimates. It is also interesting to identify areas with a large number of samples (so that a good idea of the "truth" may be obtained). Within these areas one may repeatedly select small samples and then investigate model can most accurately reproduce the totals. This procedure is carried out for three areas, with the results appearing in Tables 3.6–3.8, with Table 3.9 summarizing over all three areas. Unfortunately the conclusions are far from clear cut for these data, though hierarchical modeling is obviously preferable. In ZIP code 98801 the use of the weights is clearly beneficial and the logit normal model produces the smallest MSE. For ZIP code 98802 the use of the weights is not as beneficial and the unadjusted hierarchical binomial model performs best. Finally, for ZIP code 99347, the logit normal model gives the lowest MSE. Overall (Table 3.9) the logit spatial normal and unadjusted independent binomial models models produce low MSE. In Figure 3.5 we see that, across the study region, the estimated smoking prevalence and total counts predicted by the logit normal and unadjusted binomial model are quite seen to be quite similar. Here we report results under the logit spatial model.

In Figure 3.6 we display a map of the estimated total number of adult smokers by ZIP code. The predicted counts are highest around the Puget Sound area (the channel running north-south with many small, highly populated, ZIP codes) and the central/south area. These areas correspond to King, Snohomish and Spokane counties and the Yakima valley, which are the most populated counties in Washington State. Figure 3.6 provides a map of a measure of the uncertainty, namely the 95% intervals of the predicted smoking counts, using the spatial logit model. We see that, not surprisingly, the greatest uncertainly lies in the areas with the largest estimated counts.

To investigate the sensitivity of our estimates to the prior distribution selected for the spatial precision, σ_u^{-2}, we fit our models on the Washington BRFSS data varying the prior and comparing the posterior medians of σ_u and σ_v as well as the proportion of total variance contributed by the spatial component. In addition to the Ga(0.5,0.008) prior that was used for the simulation study we considered Ga(1.0, 0.026) and Ga(0.35, 0.001), which have a 95% range for the residual odds of (0.5, 2.0) and (0.1, 10), respectively (Wakefield, 2009). In Table 3.10 we report the variance parameter estimates and we see little sensitivity to the prior chosen for σ_u^{-2} and the maps of counts show only small changes under the different priors (within a given class of models). With 498 ZIP codes the insensitivity is not unexpected. In general, it is worthwhile to investigate the sensitivity of results to variance parameter prior specification, since we would not expect the stability seen here to be repeated when the number of areas is small. The values in Table 3.10 do vary by model, however, with the greatest discrepancy being between the arcsin square root scale and the logit model. It is a little surprising that the proportion spatial is so much lower under under the arcsin square root model, when compared to the models on the logit scale.

3.6 Application to King County Census Tracts

Since 1994, ZIP code of the respondents residence has been added to the BRFSS for the King County sample. However, ZIP codes are still relatively large and ZIP code defined areas do not align well with census tracts (CTs) or census block group based areas such as cities and the King County Health Reporting Areas (HRAs). To overcome this limitation, Public Health – Seattle & King County (PHSKC) have included a nearest intersection question in the BRFSS since 2005. The intersections are geocoded to define sub-county areas with more granularity and flexibility while protecting confidentiality of the respondents by not asking their home addresses.

Over the 5-year period from 2009 through 2013, for the 396 census tracts in King County, the sample size of BRFSS respondents ranged from 4 to 108, with a median sample of 28. CT level estimates of smoking rates and other health risk factors (such as obesity) would be

incredibly useful for PHSKC for planning and resource allocation purposes. However, the CDC recommends a sample size of at least 50 for BRFSS direct estimation (Knutson et al., 2008), but unfortunately, only 12% had a sample size of 50 or larger, which makes direct estimation (i.e., generating CT estimates based on the data within each CT), an unreliable option. The aim of this project was to combine small area estimation and multiple imputation techniques to create a procedure for generating census tract level estimates of risk factors based on the risk factor and ZIP code information collected in BRFSS. For illustration, our procedure has been used to generate smoking rates by census tract in King County, Washington.

For this analysis of the BRFSS data, we examined five year combined data from 2009 to 2013 with 16,109 subjects, excluding 440 subjects with non-King County ZIP codes or missing ZIP code. Of the 16,109, 80% answered the nearest intersection questions with two cross street addresses. Exploratory multivariable logistic regression showed that among all respondents, missing CT is associated with younger age, non-white race/ethnicity, and people living in South and East King County. Additionally, Figure 3.7 displays the direct estimates for ZIP code level smoking rates using the full data and the estimates based on observations with known census tract ($n = 12,027$). For the nearly 20% of ZIP codes missing 41–60% of data, the impact on the direct estimate of removing the observations, can be quite large. Given these results, we conclude that excluding subjects with missing CT might mask some of the sub-county disparities at the CT level; therefore we chose to employ a multiple imputation (MI) procedure so that all subjects could be included.

MI is a method that can reduce differential bias because it does not assign a subject into a fixed single CT. Rather, through an iterative process, subjects with missing CT are randomly allocated to CT within a ZIP code multiple times. Each of the multiple allocations is based on the ratio of residential addresses in a CT to the total number of residential addresses in the entire ZIP code. We used the ratio provided by the 2011Q1 HUD ZIP to CT crosswalk table HUD (2012), which will be described further in Section 3.6.2. The allocation process was integrated into the empirical logit normal SAE model described in (3.7) and was repeated

100 times. In addition, MI takes the uncertainty of missing data imputation into account in calculating the standard errors of the estimates (Little and Rubin, 2002).

3.6.1 Notation

The observed data is as follows

- $\boldsymbol{x}$ are the binary responses,
- $\boldsymbol{w}$ are the sampling weights,
- $\boldsymbol{z}$ are the zip codes, and
- $\boldsymbol{u}_O$ are the observed census tracts.

3.6.2 Multiple Imputation Approach

The Department of Housing and Urban Development (HUD) provides the estimated proportion of a ZIP code's total residential addresses, which fall within each census tract (CT). For the M observations with missing census tracts ($k \in s_M$), we assume $p(\boldsymbol{u}_k|z_k) \sim$ Multinomial $(1, \boldsymbol{p}_{z_k})$, where

$$p_{i,z_k} = \frac{\text{Number of residential addresses in } z_k \text{ and } i}{\text{Number of residential addresses in } z_k}.$$

Note, for all census tracts that do not overlap z_k, $p_{i,z_k} = 0$.

Missing census tract for person k are imputed based on a multinomial distribution with the HUD probabilities (p_{i,z_k}). For a single imputation each observation k is assigned to census i creating a set of indices s_i. The imputation is repeated D times, resulting in $d = 1, ..., D$ complete observation data sets (no missing census tracts).

3.6.3 Hierarchical Bayesian Model

For each complete data set d our approach is to summarize the data in census tract i via the asymptotic distribution of the Hájek estimator of P_i from (2.3), with corresponding variance

estimator $\text{var}(\widehat{P}_i)$. In this way the design is acknowledged in both the estimator and the variance. We define the area-level data summary as $y_i = \log\left[\widehat{P}_i/(1-\widehat{P}_i)\right]$ as the empirical logistic transform of $\widehat{P}_i$ given in (3.7). This approach constrains the probability to lie in (0,1). The likelihood is then taken as the asymptotic distribution

$$\eta_i \sim \text{N}\left(\log\left[\frac{P_i}{1-P_i}\right], \frac{\widehat{\text{var}}(\widehat{P}_i)}{\widehat{P}_i^2(1-\widehat{P}_i)^2}\right).$$

We employ a three-stage models with the first stage given by η_i. At the second stage of the model we introduce the spatial random effects terms, corresponding to the convolution model of Besag et al. (1991), and denote the area-specific parameters as

$$\eta_i = \mu + \theta_i + \phi_i$$

where μ is the overall risk level, $\phi_i|\sigma_\phi^2 \sim_{iid} \text{N}(0, \sigma_\phi^2)$ is an independent census tract random effect, and θ_i following an ICAR model (Besag and Kooperberg, 1995; Rue and Held, 2005). For the third stage we use assign Gamma(0,5,0.008) priors on the spatial conditional precision σ_θ^{-2} and the *iid* precision parameter σ_ϕ^{-2}. The Gamma prior gives a 95% range on the interpretable σ_θ and σ_ϕ scale of (0.056,4.04). We use an improper flat prior on α. Unlike the previous sections, the MI procedure generates D estimates for each area which need to be combined to generate our CT estimates.

3.6.4 Combining Estimates

Our goal is to describe the posterior distribution of η given the observed data. In a full Bayesian procedure we would want

$$\begin{aligned} p(\boldsymbol{\eta}|\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{u}_O, \boldsymbol{z}, \psi) &= \sum_{\boldsymbol{u}_M} p(\boldsymbol{\eta}, \boldsymbol{u}_M|\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{z}, \boldsymbol{u}_O, \psi) \\ &= \sum_{\boldsymbol{u}_M} p(\boldsymbol{\eta}|\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{z}, \boldsymbol{u}_O, \boldsymbol{u}_M, \psi) p(\boldsymbol{u}_M|\boldsymbol{z}) \end{aligned}$$

where ψ are the hyperpriors. We can approximate

$$\begin{aligned}\sum_{\boldsymbol{u}_M} p(\boldsymbol{\eta}|\boldsymbol{x},\boldsymbol{w},\boldsymbol{z},\boldsymbol{u}_O,\boldsymbol{u}_M,\psi)p(\boldsymbol{u}_M|\boldsymbol{z}) &\approx \frac{1}{D}\sum_{d=1}^{D} p\left(\boldsymbol{\eta}|\boldsymbol{x},\boldsymbol{w},\boldsymbol{z},\boldsymbol{u}_O,\boldsymbol{u}_M^{(d)},\psi\right)\\ &= \frac{1}{D}\sum_{d=1}^{D} p\left(\boldsymbol{\eta}|\boldsymbol{y}^{(d)},\boldsymbol{z},\boldsymbol{u}_O,\boldsymbol{u}_M^{(d)},\psi\right)\end{aligned}$$

where for each $k \in s_M$, $\boldsymbol{u}_k^{(d)}|\boldsymbol{z} \sim \text{Multinomial}\,(1,\boldsymbol{p}_{z_k})$. For each d we have

$$\widehat{P}_i^{(d)} = \left(\sum_{k:u_{k,i}=1} x_k w_k + \sum_{k:u_{k,i}^{(d)}=1} x_k w_k\right) \Big/ \left(\sum_{k:u_{k,i}=1} w_k + \sum_{k:u_{k,i}^{(d)}=1} w_k\right)$$

and and $y_i^{(d)} = \text{logit}\left(\widehat{P}_i^{(d)}\right)$. It is equivalent to write

$$p\left(\boldsymbol{\eta}|\boldsymbol{x},\boldsymbol{w},\boldsymbol{z},\boldsymbol{u}_O,\boldsymbol{u}_M^{(d)},\psi\right) = p\left(\boldsymbol{\eta}|\boldsymbol{y}^{(d)},\boldsymbol{z},\boldsymbol{u}_O,\boldsymbol{u}_M^{(d)},\psi\right)$$

and our approximation as

$$\sum_{\boldsymbol{u}_M} p(\boldsymbol{\eta}|\boldsymbol{x},\boldsymbol{w},\boldsymbol{z},\boldsymbol{u}_O,\boldsymbol{u}_M,\psi)p(\boldsymbol{u}_M|\boldsymbol{z}) \approx \frac{1}{D}\sum_{d=1}^{D} p\left(\boldsymbol{\eta}|\boldsymbol{y}^{(d)},\boldsymbol{z},\boldsymbol{u}_O,\boldsymbol{u}_M^{(d)},\psi\right).$$

Our hierarchical Bayesian model takes the form

$$\begin{aligned}p\left(\boldsymbol{\eta}|\boldsymbol{y}^{(d)},\boldsymbol{z},\boldsymbol{u}_O,\boldsymbol{u}_M^{(d)},\psi\right) &\propto p\left(\boldsymbol{y}^{(d)}|\boldsymbol{\eta},\boldsymbol{z},\boldsymbol{u}_O,\boldsymbol{u}_M^{(d)}\right) p\left(\boldsymbol{\eta}|\psi\right)\\ &= \prod_{i=1}^{I} \underbrace{p\left(y_i^{(d)}|\eta_i,\boldsymbol{z},\boldsymbol{u}_O,\boldsymbol{u}_M^{(d)}\right)}_{\mathcal{N}(\eta_i,V_{DES,i})} \underbrace{p\left(\boldsymbol{\eta}|\psi\right)}_{\text{Spatial Model}}.\end{aligned}$$

The approximation can be written as

$$\begin{aligned}p(\boldsymbol{\eta}|\boldsymbol{y},\boldsymbol{u}_O,\boldsymbol{z},\psi) &= \sum_{\boldsymbol{u}_M} p(\boldsymbol{\eta}|\boldsymbol{y},\boldsymbol{z},\boldsymbol{u}_O,\boldsymbol{u}_M,\psi)p(\boldsymbol{u}_M|\boldsymbol{z})\\ &\approx \frac{1}{D}\sum_{d=1}^{D}\prod_{i=1}^{I} p(y_i^{(d)}|\eta_i,\boldsymbol{z},\boldsymbol{u}_O,\boldsymbol{u}_M^{(d)})p(\boldsymbol{\eta}|\psi)\end{aligned}$$

where $\boldsymbol{u}_k^{(d)}|\boldsymbol{z} \sim \text{Multinomial}\,(1,\boldsymbol{p}_{z_k})$. Finally, $\boldsymbol{\eta}$ is marginalized via

$$p(\boldsymbol{\eta}|\boldsymbol{y},\boldsymbol{u}_O,\boldsymbol{z}) = \int p(\eta,\psi|\boldsymbol{y},\boldsymbol{u}_O,\boldsymbol{z})d\psi = \int p(\eta|\boldsymbol{y},\boldsymbol{u}_O,\boldsymbol{z},\psi)p(\psi|\boldsymbol{y},\boldsymbol{u}_O,\boldsymbol{z})d\psi.$$

Given D set of smoothed estimates the mean posterior estimate is

$$E(\eta_i|\boldsymbol{y},\boldsymbol{z},\boldsymbol{u}_o) \approx \frac{1}{D}\sum_{d=1}^{D}\hat{\eta}_i^{(d)} = \hat{\eta}_i$$

where $\hat{\eta}_i^{(d)} = E\left(\eta_i|\boldsymbol{y},\boldsymbol{z},\boldsymbol{u}_O,\boldsymbol{u}_M^{(d)}\right)$. Similarly we find a variance

$$V(\eta_i|\boldsymbol{y},\boldsymbol{z},\boldsymbol{u}_O) \approx \frac{1}{D}\sum_{d=1}^{D}V_i^{(d)} + \frac{1}{D-1}\sum_{d=1}^{D}\left(\hat{\eta}_i^{(d)} - \hat{\eta}_i\right)^2 = \hat{V}_i.$$

where $V_{i,d} = V(\eta_i|\boldsymbol{y},\boldsymbol{z},\boldsymbol{u}_O,\boldsymbol{u}_M^{(d)})$ (the posterior variance of η_i based on the dth complete dataset). This variance estimate has contributions from within and between the sets of estimates (Little and Rubin, 2002, Ch. 10). Finally, estimates for census tract i are derived from expit$(\widehat{\eta}_i)$ and $100(1-\alpha)\%$ credible intervals are generated using

$$\left[\text{expit}\left(\widehat{\eta}_i - z_{1-\alpha/2}\times\sqrt{\widehat{V}_i}\right), \text{expit}\left(\widehat{\eta}_i + z_{1-\alpha/2}\times\sqrt{\widehat{V}_i}\right)\right].$$

3.6.5 Results

Figure 3.8 displays the average sample size by CT after CTs are imputed with the complete case CT sample size. We see that the majority of CTs gain approximately 10 samples through the imputation. Figure 3.9 compares the direct estimate (based on data only within each CT) to the smoothed estimate generated by the MI and hierarchical Bayesian model approach. We see that the CTs with fewer than 50 observations experience a greater degree of smoothing. A similar pattern is observed in Figure 3.10, which compares the direct estimates with the estimates resulting from SAE and MI.

Figures 3.11, 3.12, and 3.13 display maps smoking rates for the direct estimates, smoothed census tract estimates based on the complete case, and the multiple imputation analysis, respectively. As expected, based on the exploratory plots of ZIP code level estimates, when we include the observations with the missing census tract the result is often a higher smoking estimate. Figure 3.13 illustrates areas in eastern and southern King county with higher smoking rates based on the multiple imputation approach compared with the complete case

SAE shown in 3.12. The magnitude of difference between the complete case and MI procedures is shown in Figure 3.14 and suggest the majority of large differences are increases in smoking rates estimated by the MI approach.

3.7 Discussion

In this chapter we have considered random effects models that account for the sampling weights that are common in SAE. The simulations of Section 3.4 clearly illustrate the benefits of hierarchical modeling, namely large reductions in the variance of parameter estimation when compared with non-hierarchical approaches. These simulations also show that non-response and selection bias can be reduced via the incorporation of the weights. Various methods have been proposed to achieve this aim and the use of the empirical logit normal model and the effective sample size adjusted numerators and denominators approach (Chen et al., 2014) both perform well in the reported simulations and application to the 2006 BRFSS in Washington State.

We find modeling the empirical logit transformation of the weighted estimator and incorporating the design-based variance preferable to modifying the sample size or outcomes, as it has similar performance and greater flexibility. In Chapter 4 we will use the empirical logit approach to model the under five child mortality rate, which is not modeled as a simple binary response. It is straight forward to extend the empirical logit approach in this setting, but the other methods investigated in this chapter would not extend well in this setting.

In the context of generating census tract level estimates of smoking rates in King County we found a combination of multiple imputation and smoothing of the empirical logit to be a useful approach to account for missing CT. In this scenario SAE is needed to provide estimates given limited sample sizes and the sub-county level. The hierarchical Bayesian model can be useful at the CT or larger sub-county geographic levels and can be applied to the BRFSS indicators for showing place-based disparities with reasonable precision. SAE has increasingly become a useful tool to meet the demand of presenting data at more granular levels and has been utilized for local public health programs (Song et al., 2016).

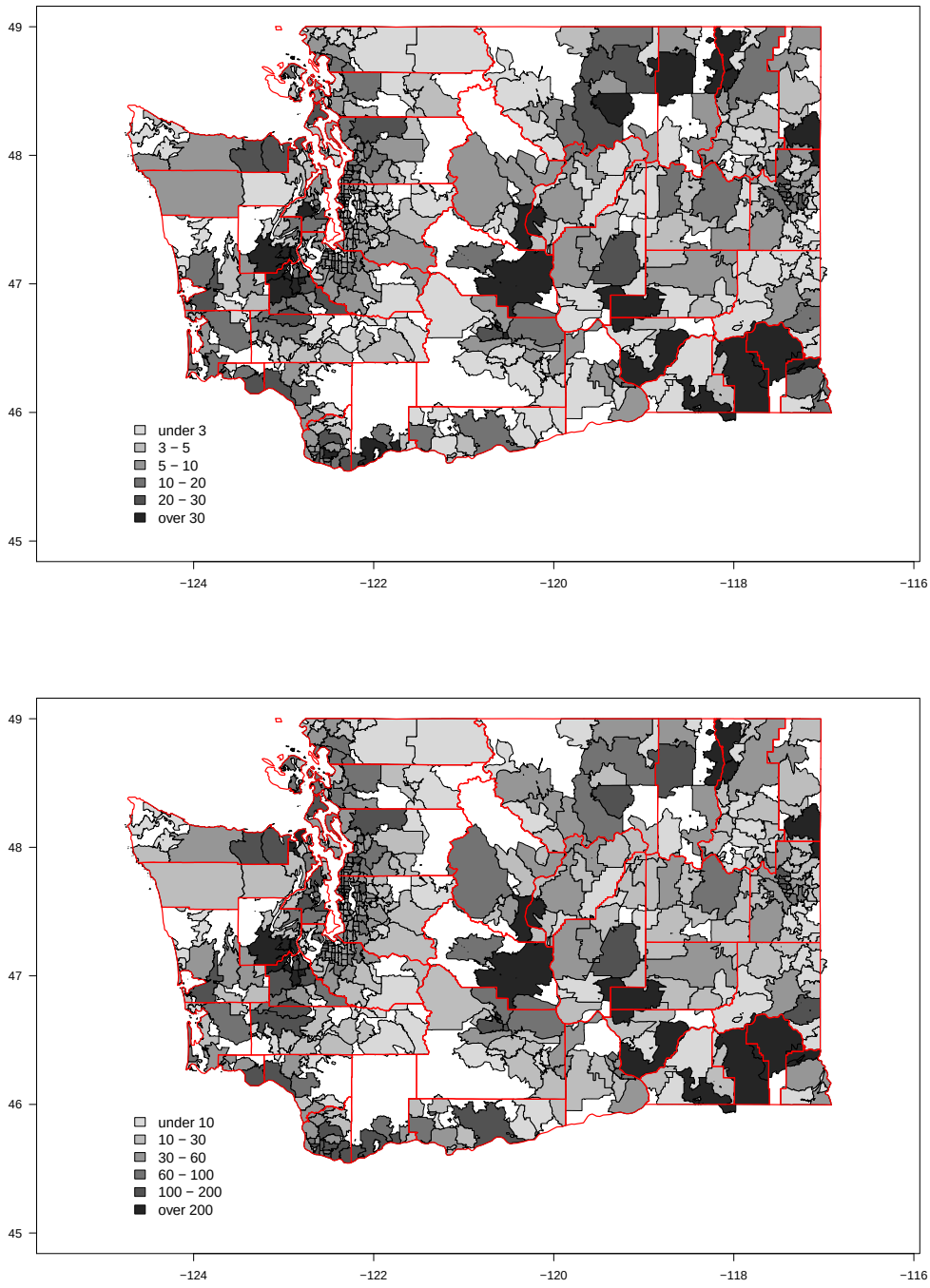

Figure 3.2: Maps of the observed number of adult current smokers (top) and the observed BRFSS sample size (bottom) in Washington State ZIP codes in 2006. County boundaries are indicated.

Table 3.4: Simulation results to examine non-response bias for five different scenarios. Tables 3.2 and 3.3 gives the prevalence and response parameters that change across scenarios. Non-hierarchical unadjusted use the observed y_i and m_i and non-hierarchical adjusted use the Horvitz–Thompson estimator. Numbers in **bold** correspond to the row minimums.

($\times 10^3$)	Non-Hierarchical		Unadjust Binom		Logit Normal		Pseudo-likelihood		Arcsin Sqrt		Effect Samp Size	
Bias^2	Unadjust	Adjust	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial
Scenario 1	**2.2**	2.7	26.0	18.9	54.9	40.7	32.1	22.4	9.3	10.8	15.6	12.2
Scenario 2	14.2	**2.5**	56.0	42.9	54.0	40.0	34.4	24.1	10.0	9.8	17.7	13.4
Scenario 3	15.4	**7.1**	51.3	38.2	52.5	37.9	32.5	22.3	9.0	10.4	15.7	11.9
Scenario 4	**2.2**	2.9	36.7	23.6	64.9	42.8	43.7	27.1	18.8	16.1	25.2	17.8
Scenario 5	12.9	**2.7**	55.2	41.7	53.7	39.1	33.9	23.6	9.9	10.2	17.3	13.2
Variance	Unadjust	Adjust	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial
Scenario 1	233.8	220.8	6.5	7.5	**3.0**	4.8	4.5	6.0	25.8	26.5	15.5	15.4
Scenario 2	253.8	210.3	6.1	7.6	**2.8**	4.7	2.9	4.8	20.9	23.1	12.2	12.6
Scenario 3	252.7	210.4	8.3	9.1	**3.8**	5.3	4.4	5.7	25.3	25.4	15.8	15.3
Scenario 4	236.1	222.8	10.0	10.2	**5.1**	6.8	7.2	8.1	29.6	27.8	19.8	18.6
Scenario 5	249.1	206.1	5.8	7.2	**2.6**	4.4	2.9	4.6	20.9	22.6	12.0	12.2
MSE	Unadjust	Adjust	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial
Scenario 1	236.0	223.4	32.6	**26.4**	57.9	45.6	36.7	28.5	35.1	37.3	31.1	27.6
Scenario 2	268.0	212.8	62.1	50.5	56.7	44.7	37.4	28.9	30.9	32.9	29.9	**26.0**
Scenario 3	268.2	217.5	59.6	47.4	56.3	43.2	36.9	28.0	34.3	35.7	31.4	**27.2**
Scenario 4	238.3	225.7	46.7	**33.8**	70.1	49.6	50.8	35.2	48.3	43.9	45.0	36.3
Scenario 5	262.0	208.8	61.0	48.9	56.3	43.6	36.9	28.2	30.8	32.8	29.3	**25.4**

Table 3.5: Simulation results to examine the effect of selection bias. Non-hierarchical unadjusted use the observed y_i and m_i and non-hierarchical adjusted use the Horvitz–Thompson estimator. Selection is based on $s = \Pr(Z_{ik} = 1|Y_{ik} = 1)$ where Z_{ik} is a binary variable upon which sampling is based. Numbers in **bold** correspond to the row minimums.

$(\times 10^3)$	Non-Hierarchical		Unadjust Binomial		Logit Normal		Pseudo-likelihood		Arcsin Sqrt		Effect Samp size	
Bias^2	Unadjust	Adjust	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial
$s = 0.1$	**4.4**	6.3	23.5	17.6	5.8	7.7	19.4	14.9	20.1	26.5	5.8	5.5
$s = 0.3$	591.7	**5.0**	806.9	712.2	9.7	11.4	22.7	16.2	27.8	39.1	8.7	11.1
$s = 0.5$	1726.0	**3.0**	2141.6	1964.2	10.7	13.6	31.3	21.7	27.2	41.8	11.8	17.3
$s = 0.8$	3490.5	**1.7**	4116.1	3858.8	27.8	19.0	44.9	30.2	17.6	22.6	18.6	19.6
Variance	Unadjust	Adjust	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial
$s = 0.1$	315.7	511.2	**8.2**	8.8	214.8	128.7	100.0	96.5	167.6	161.0	214.8	210.1
$s = 0.3$	473.0	425.5	**15.0**	15.4	116.6	110.9	45.9	43.0	122.3	112.2	147.3	139.0
$s = 0.5$	532.8	322.4	**8.6**	12.0	51.8	46.3	8.6	9.3	58.3	49.7	65.1	57.2
$s = 0.8$	544.5	167.8	1.8	7.5	0.6	2.3	**0.4**	1.4	4.3	6.6	2.3	3.2
MSE	Unadjust	Adjust	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial
$s = 0.1$	320.1	517.4	31.7	**26.5**	220.6	136.4	119.4	111.4	187.7	187.6	220.6	215.7
$s = 0.3$	1064.7	430.6	821.9	727.6	126.3	122.3	68.5	**59.2**	150.2	151.3	156.0	150.1
$s = 0.5$	2258.8	325.4	2150.2	1976.2	62.5	59.9	39.9	**30.9**	85.5	91.5	76.9	74.5
$s = 0.8$	4035.0	169.5	4117.9	3866.3	28.4	21.3	45.3	31.6	21.9	29.1	**20.9**	22.8

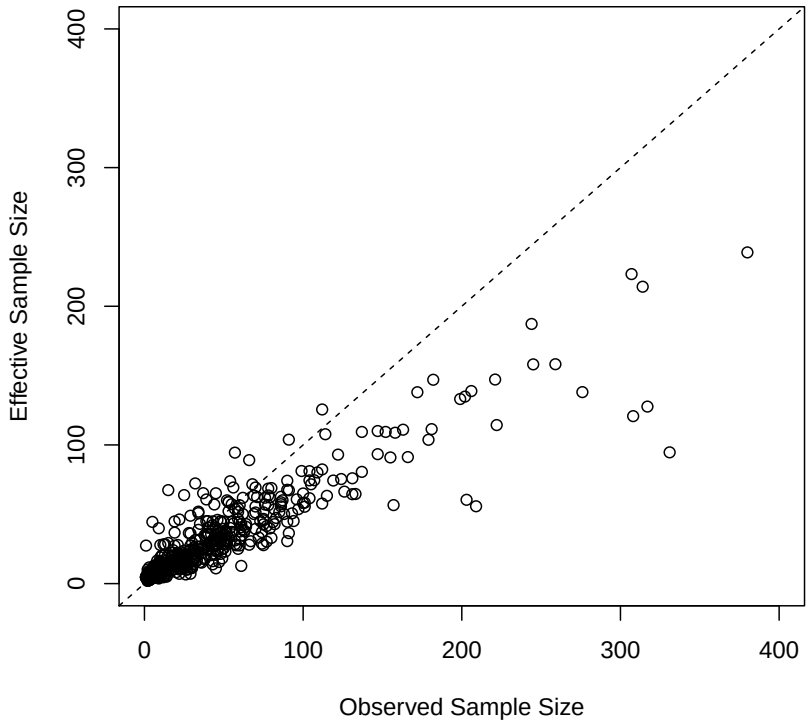

Figure 3.3: For 2006 Washington BRFSS data: effective sample sizes versus observed sample sizes.

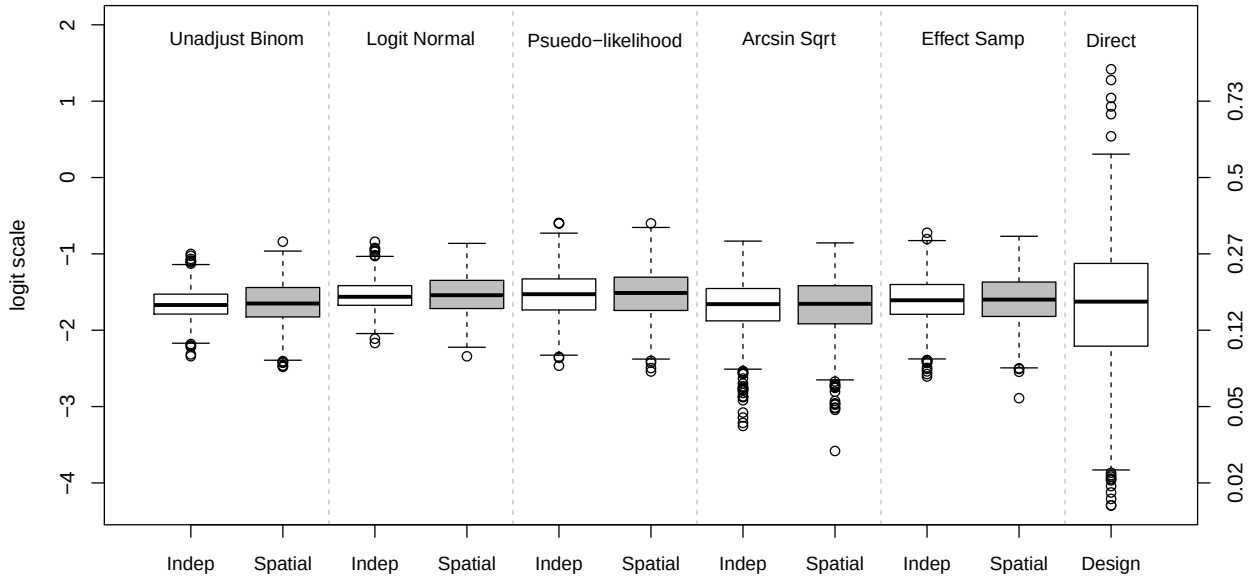

Figure 3.4: Smoking prevalence estimates across Washington State ZIP codes in 2006, using various approaches.

Table 3.6: Model validation on the totals for ZIP code 98801. The true total is estimated to be 5,085. Numbers in **bold** correspond to the row minimums.

($\times 10^3$)	Non-Hierarchical		Unadjust Binomial		Logit Normal		Pseudo-likelihood		Arcsin Sqrt		Effect Samp size	
Bias^2	Unadjust	Adjust	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial
$n = 10$	1459.4	126.6	392.3	763.0	11.0	470.5	**4.1**	268.4	671.5	1671.7	386.1	1178.3
$n = 30$	1539.4	33.7	560.9	839.6	26.2	446.9	**10.2**	171.6	582.1	1405.0	344.1	1000.0
$n = 50$	1433.5	3.8	657.1	878.0	12.2	355.1	**2.4**	90.2	273.9	867.8	162.8	643.2
Variance	Unadjust	Adjust	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial
$n = 10$	9419.8	21804.1	193.1	**53.2**	391.4	143.9	1479.7	772.1	1007.2	549.6	997.7	508.7
$n = 30$	2713.4	7982.3	287.4	**100.9**	619.5	258.0	2180.4	1447.6	1559.1	983.4	1358.8	799.8
$n = 50$	1600.4	4511.8	312.4	**128.5**	619.2	282.6	1864.8	1374.0	1193.5	786.8	1124.4	711.9
MSE	Unadjust	Adjust	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial
$n = 10$	10879.3	21930.6	585.4	816.2	**402.4**	614.4	1483.8	1040.5	1678.8	2221.4	1383.8	1687.0
$n = 30$	4252.8	8016.0	848.3	940.5	**645.7**	704.9	2190.6	1619.3	2141.2	2388.4	1702.8	1799.7
$n = 50$	3033.8	4515.7	969.5	1006.5	**631.5**	637.7	1867.2	1464.1	1467.4	1654.6	1287.3	1355.1

Table 3.7: Model validation on the totals for ZIP code 98802. The true total is estimated to be 2,612. Numbers in **bold** correspond to the row minimums.

($\times 10^3$)	Non-Hierarchical		Unadjust Binomial		Logit Normal		Pseudo-likelihood		Arcsin Sqrt		Effect Samp size	
Bias^2	Unadjust	Adjust	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial
$n = 10$	0.8	**0.0**	110.3	125.8	371.1	289.3	291.7	277.7	14.4	0.1	56.5	18.4
$n = 30$	**0.0**	1.8	73.1	95.5	246.1	218.5	133.2	148.1	3.1	0.9	29.0	9.2
$n = 50$	0.1	**0.0**	48.2	72.2	164.6	162.5	67.8	84.4	3.7	0.0	20.8	7.7
Variance	Unadjust	Adjust	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial
$n = 10$	3938.0	5608.7	80.4	**29.0**	115.8	50.9	330.7	194.2	358.1	239.4	304.4	187.3
$n = 30$	1374.3	2165.4	148.3	**65.0**	203.5	102.3	550.7	388.7	586.4	445.7	475.8	337.6
$n = 50$	729.0	1181.1	142.9	**71.1**	205.7	113.2	453.0	346.4	461.0	364.7	389.3	292.6
MSE	Unadjust	Adjust	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial
$n = 10$	3938.8	5608.7	190.7	**154.8**	486.9	340.2	622.4	471.9	372.5	239.5	360.9	205.7
$n = 30$	1374.4	2167.2	221.4	**160.5**	449.6	320.7	683.9	536.8	589.6	446.5	504.8	346.8
$n = 50$	729.1	1181.1	191.1	**143.3**	370.2	275.6	520.8	430.8	464.8	364.7	410.1	300.3

Table 3.8: Model validation on the totals for ZIP code 99347. The true total is estimated to be 360. Numbers in **bold** correspond to the row minimums.

($\times 10^3$)	Non-Hierarchical		Unadjust Binomial		Logit Normal		Pseudo-likelihood		Arcsin Sqrt		Effect Samp size	
Bias^2	Unadjust	Adjust	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial
$n = 10$	3.1	**0.9**	4.0	3.0	1.1	1.4	0.9	1.2	5.5	6.9	3.9	4.8
$n = 30$	3.6	**0.3**	3.9	3.0	1.2	1.5	0.6	0.7	4.4	5.7	3.3	4.1
$n = 50$	3.4	**0.1**	3.7	3.0	1.2	1.5	0.3	0.5	3.3	4.4	2.5	3.2
Variance	Unadjust	Adjust	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial
$n = 10$	44.6	84.7	0.9	**0.3**	1.6	0.6	6.2	3.5	4.2	2.7	4.3	2.6
$n = 30$	13.8	38.1	1.6	**0.6**	2.4	1.1	11.3	8.0	5.2	3.8	4.9	3.4
$n = 50$	8.0	25.4	1.7	**0.7**	2.7	1.3	11.2	8.6	4.6	3.4	4.6	3.3
MSE	Unadjust	Adjust	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial
$n = 10$	47.7	85.5	5.0	3.2	2.7	**2.1**	7.1	4.7	9.7	9.6	8.2	7.3
$n = 30$	17.3	38.4	5.4	3.6	3.7	**2.6**	11.9	8.7	9.7	9.4	8.2	7.4
$n = 50$	11.4	25.6	5.4	3.7	3.9	**2.8**	11.5	9.0	7.9	7.8	7.1	6.5

Table 3.9: Model validation across the three ZIP codes 98801, 98802 and 99347. Numbers in **bold** correspond to the row minimums.

($\times 10^3$)	Non-Hierarchical		Unadjust Binomial		Logit Normal		Pseudo-likelihood		Arcsin Sqrt		Effect Samp size	
Bias^2	Unadjust	Adjust	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial
$n = 10$	1463.3	**127.4**	506.6	891.7	383.2	761.3	296.8	547.3	691.5	1678.7	446.5	1201.5
$n = 30$	1543.0	**35.9**	637.8	938.1	273.5	666.8	143.9	320.5	589.6	1411.5	376.3	1013.2
$n = 50$	1436.9	**4.0**	709.1	953.2	178.0	519.0	70.6	175.1	280.9	872.3	186.2	654.1
Variance	Unadjust	Adjust	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial
$n = 10$	13402.4	27497.5	274.5	**82.5**	508.8	195.4	1816.6	969.9	1369.5	791.7	1306.3	698.5
$n = 30$	4101.5	10185.7	437.3	**166.5**	825.4	361.4	2742.5	1844.3	2150.8	1432.9	1839.6	1140.7
$n = 50$	2337.4	5718.3	457.0	**200.3**	827.6	397.1	2328.9	1728.9	1659.1	1154.8	1518.3	1007.8
MSE	Unadjust	Adjust	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial	Indep	Spatial
$n = 10$	14865.7	27624.9	**781.1**	974.2	892.0	956.7	2113.4	1517.2	2061.0	2470.4	1752.8	1900.1
$n = 30$	5644.5	10221.6	1075.1	1104.6	1098.9	**1028.2**	2886.4	2164.8	2740.4	2844.4	2215.9	2154.0
$n = 50$	3774.3	5722.3	1166.0	1153.5	1005.6	**916.1**	2399.5	1904.0	1940.0	2027.1	1704.5	1661.9

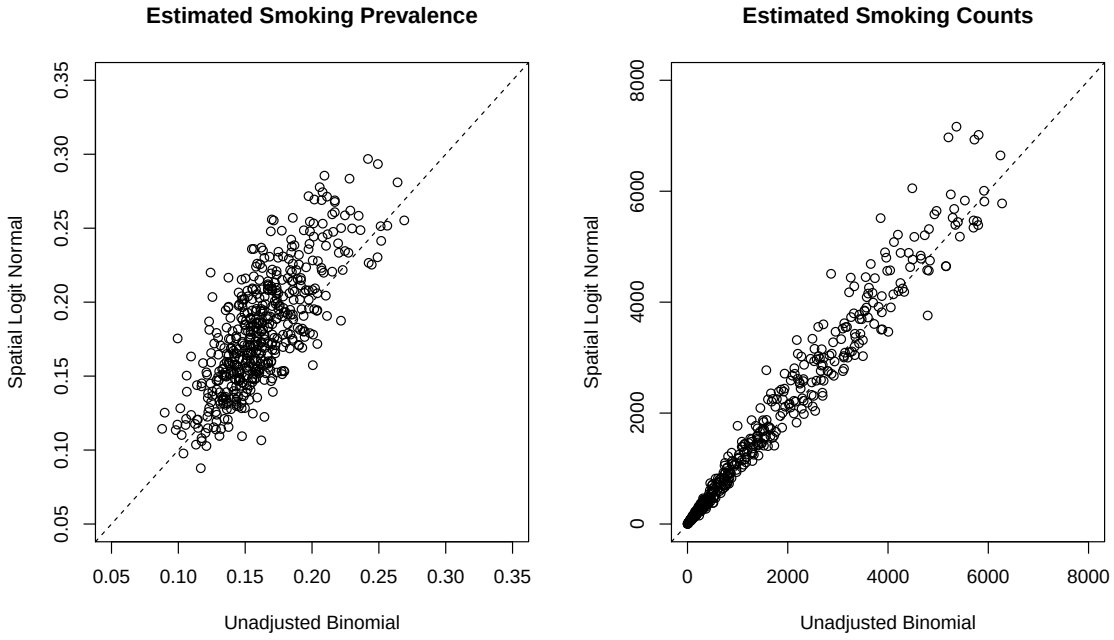

Figure 3.5: Comparison of estimated smoking prevalence (left) and estimated smoking counts (right) across ZIP codes under the spatial logit and unadjusted binomial models.

Table 3.10: Comparison of posterior medians of spatial and non-spatial standard deviations, σ_u and σ_v, and proportion of total variance that is spatial p_s, for three different priors on the spatial precision σ_u^{-2} for each of the hierarchical models described in Section 2.

	Prior for σ_u^{-2}	Unadj Bin	Lgt Normal	Pseudo-lkd	Arcsin Sqrt	Eff Samp
	Ga(0.50,0.008)	0.35	0.30	0.31	0.065	0.29
σ_u	Ga(1.00,0.026)	0.34	0.29	0.30	0.079	0.28
	Ga(0.35,0.001)	0.35	0.31	0.31	0.053	0.29
	Ga(0.50,0.008)	0.17	0.25	0.38	0.078	0.41
σ_v	Ga(1.00,0.026)	0.17	0.25	0.38	0.073	0.41
	Ga(0.35,0.001)	0.16	0.25	0.38	0.081	0.41
	Ga(0.50,0.008)	0.79	0.76	0.78	0.12	0.77
p_s	Ga(1.00,0.026)	0.78	0.75	0.76	0.14	0.75
	Ga(0.35,0.001)	0.79	0.77	0.78	0.099	0.78

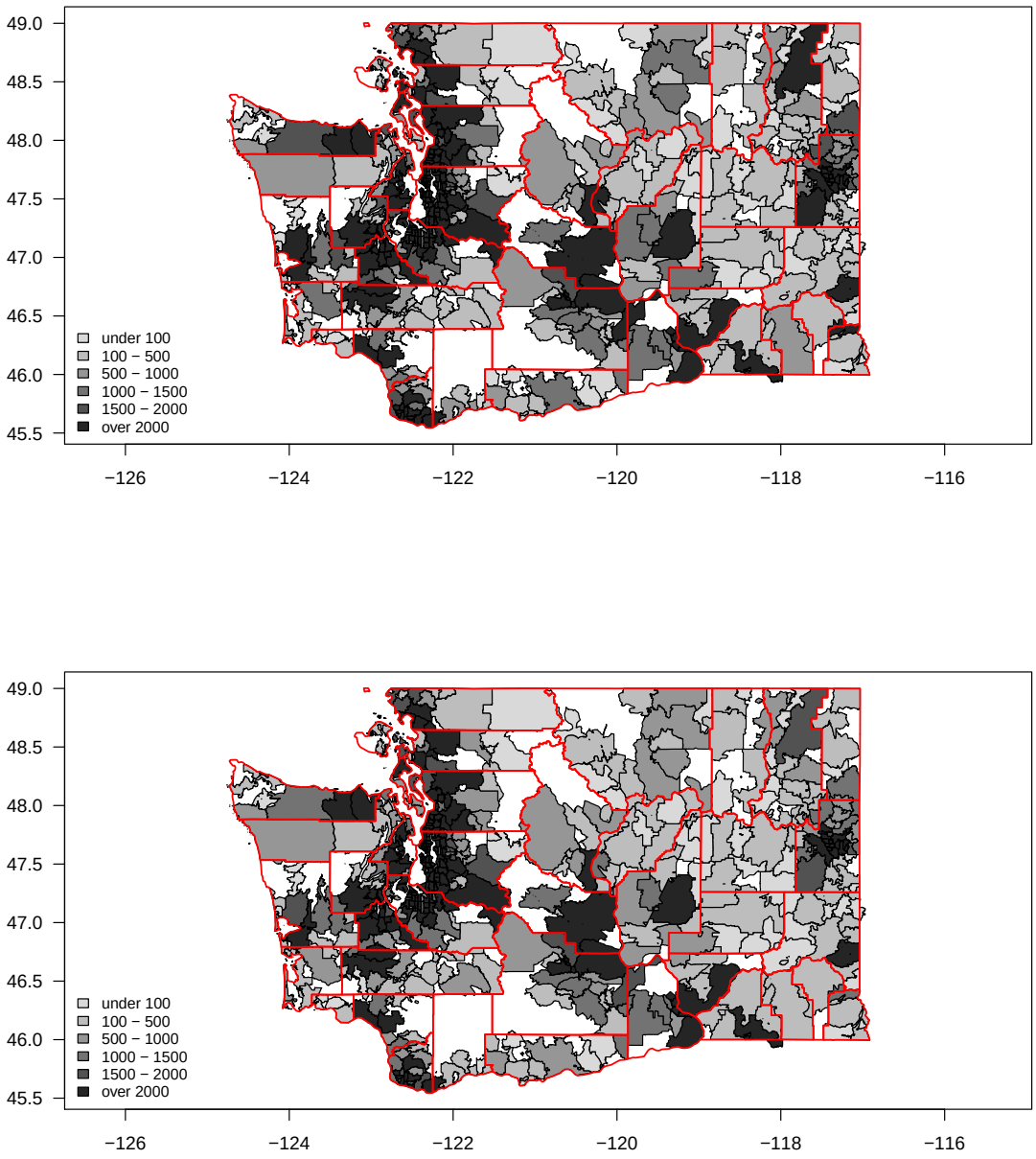

Figure 3.6: Top: Predicted total adult smokers by ZIP code in Washington State in 2006, under the Logit Normal spatial model. County boundaries are indicated. Bottom: The 95% interval of the predicted total adult smokers by ZIP code in Washington State in 2006, under the spatial logit normal model.

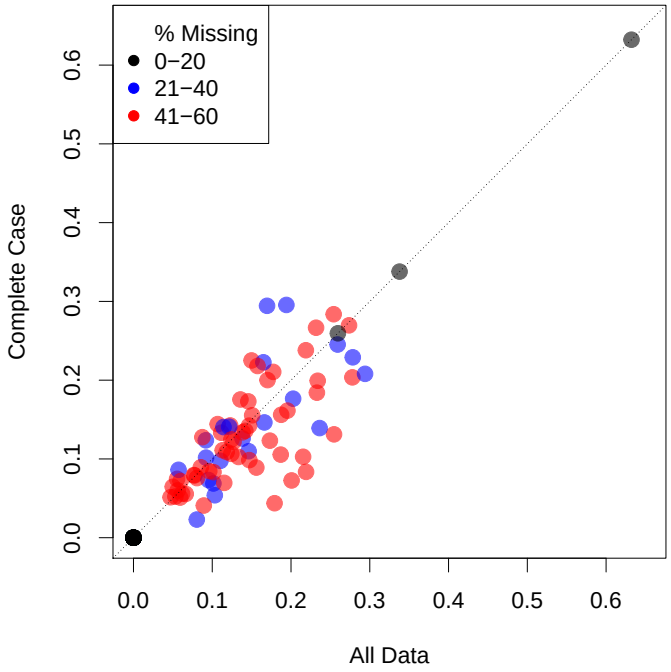

Figure 3.7: Comparison of ZIP code smoking rates with and without respondents with missing census tract.

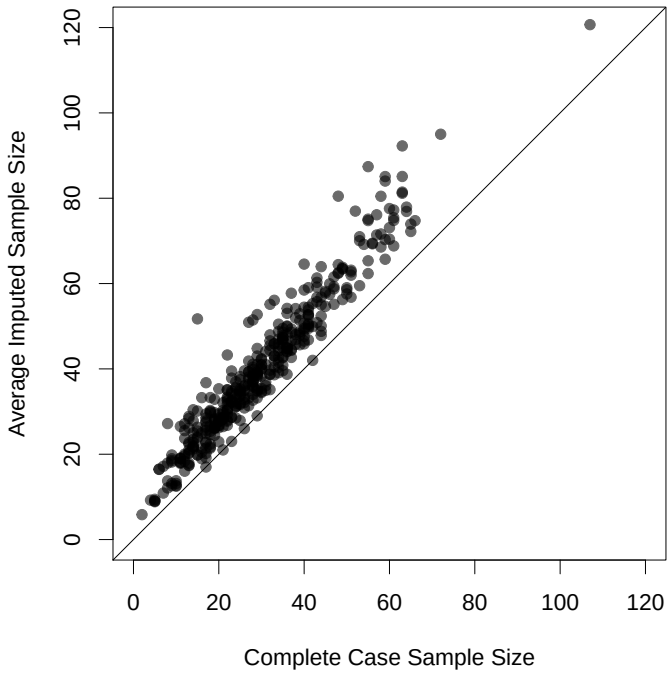

Figure 3.8: The average Census Tract sample size for the imputed data in King County, Washington from 2009-2013.

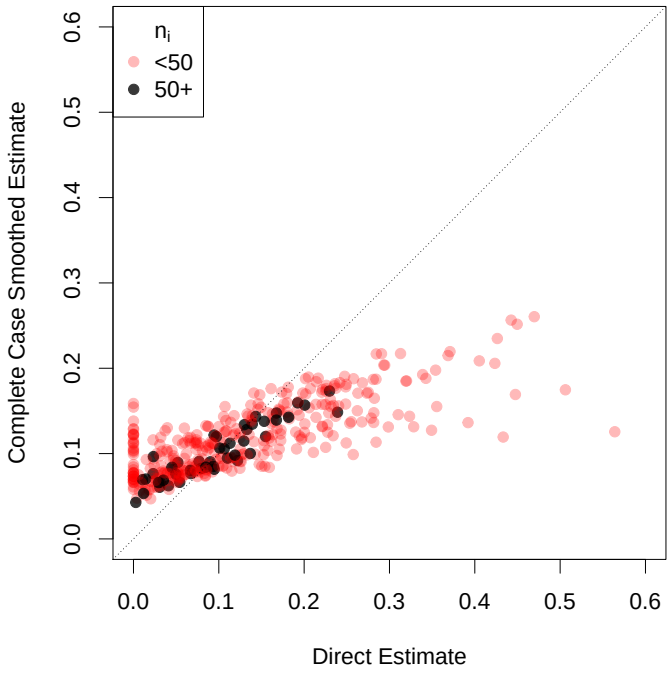

Figure 3.9: The impact of smoothing the complete case data on smoking rates, by complete case sample size in King County, Washington from 2009-2013.

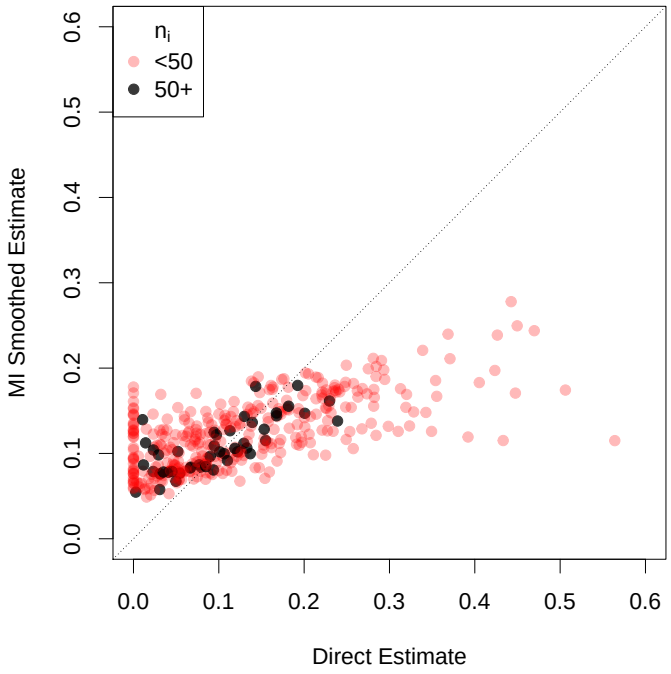

Figure 3.10: The impact of smoothing and the multiple imputation procedure on smoking rates in King County, Washington from 2009-2013.

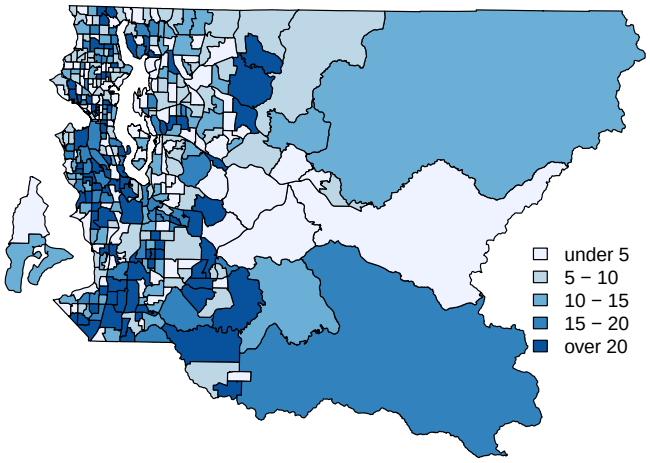

Figure 3.11: Maps of the raw smoking rate (top left), smoking rates based on the complete case analysis (top right), and the smoking rate including the multiple imputation (bottom) in King County, Washington from 2009-2013.

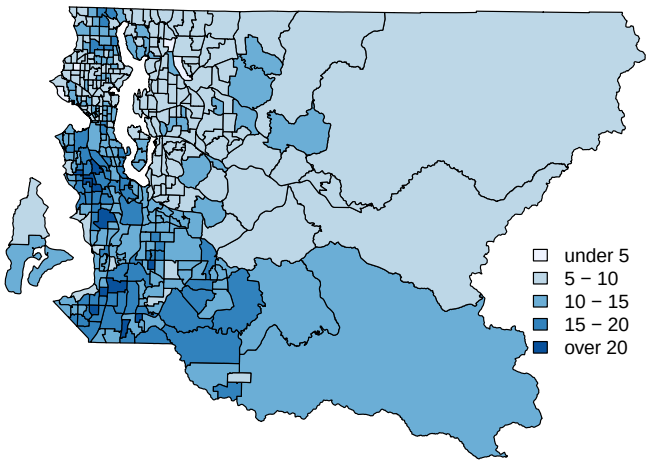

Figure 3.12: Maps of the raw smoking rate (top left), smoking rates based on the complete case analysis (top right), and the smoking rate including the multiple imputation (bottom) in King County, Washington from 2009-2013.

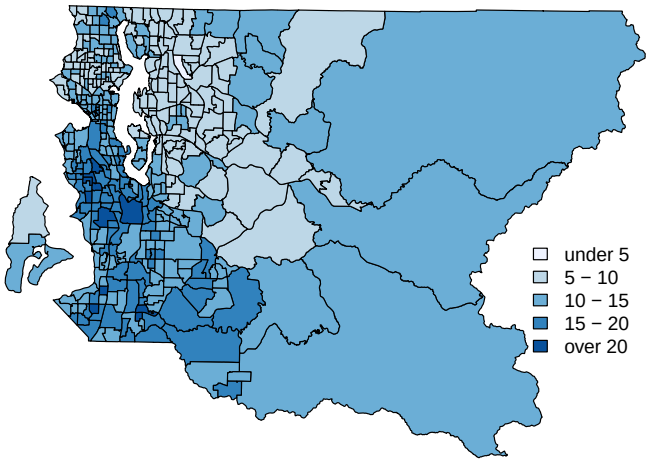

Figure 3.13: Maps of the raw smoking rate (top left), smoking rates based on the complete case analysis (top right), and the smoking rate including the multiple imputation (bottom) in King County, Washington from 2009-2013.

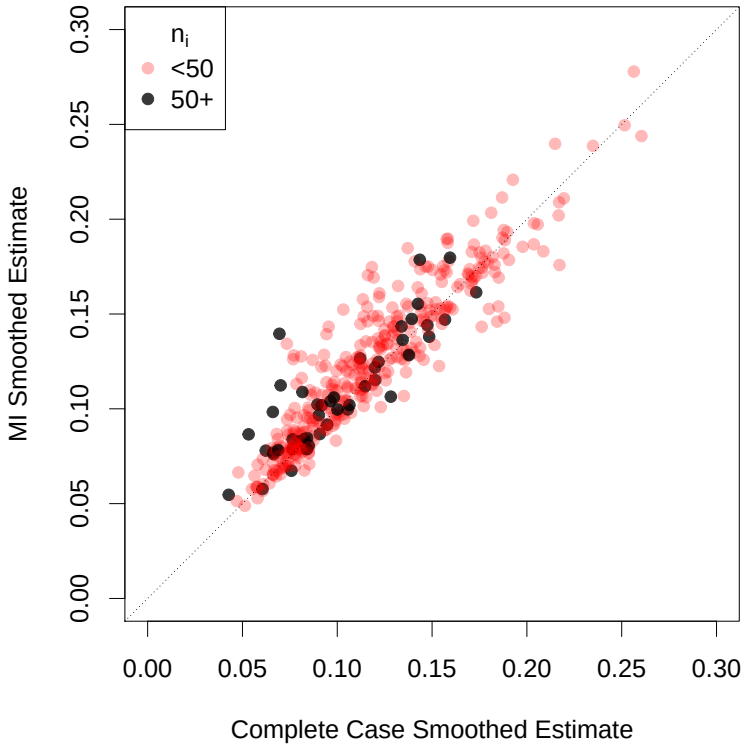

Figure 3.14: Comparison of the complete case SAE and multiple imputation and SAE procedure on smoking rates in King County, Washington from 2009-2013.

Chapter 4

SMOOTHING MODELS FOR ESTIMATION OF CHILD MORTALITY FROM DEMOGRAPHIC SURVEILLANCE SYSTEMS AND COMPLEX SURVEYS DATA

4.1 Introduction

Over the past fifteen years the United Nations' (UN) Millennium Development Goals (MDGs) (UN, 2000) have focused the world's attention on improving key indicators of development, health and wellbeing. The requirement to monitor progress toward the MDGs has revealed a stunning absence of data with which to measure and monitor key indicators related to the MDGs in much of the developing world, and this has led to great interest in improving both the data and our ability to use it. In 2015 the UN and its partners are taking stock of experience with the MDGs and coordinating the establishment of a new set of global goals (UN, 2014d) – the Sustainable Development Goals (SDGs) (UN, 2014e). Even before the SDGs are finalized the UN Secretary General has called for a *Data Revolution for Sustainable Development* and appointed a high level advisory group to define what it should be (UN, 2014b). The aim is clear: to rapidly improve the coverage, quality, availability and timeliness of the data used to measure and monitor progress toward the SDGs. Simultaneously there is sustained, strong interest in improving civil registration, vital statistics (CRVS) and the functioning of statistical offices across the developing world (World Bank and World Health Organization, 2014; Paris21, 2014). The key challenges are improving coverage (UN, 2014a) and timeliness of reporting.

In this context of far-reaching interest in improving data and methods available to monitor indicators of the SDGs and improve CRVS, in this chapter we develop a general approach that combines data from different sources and provides temporal, subnational-specific esti-

mates with uncertainty that accounts for the different designs of the data collection schemes. We demonstrate the method by calculating spatio-temporal estimates of child mortality in Tanzania using data from multiple Demographic and Health Surveys (DHS) (USAID, 2014) and two health and demographic surveillance system (HDSS) sites (INDEPTH Network, 2014).

Reducing child mortality is MDG 4 (UN, 2014c), and over the past fifteen years a great deal of effort and resources have been spent in order to meet MDG 4 targets at the national level in many developing nations. This has driven work to develop better methods to estimate trends in child mortality at the national level, and two groups have produced globally comparable trends in child mortality for all nations. The United Nations Inter-agency Group for Child Mortality Estimation (UN IGME) recently developed a Bayesian B-spline Bias-reduction (B3) method (Alkema et al., 2014), and the Institute for Health Metrics and Evaluation (IHME) uses a Gaussian process regression (Wang et al., 2014). All of these methods produce national estimates through time with measures of uncertainty. None are designed to reveal variation in child mortality within countries. A recent paper by Dwyer-Lindgren et al. (2014) compared many Bayesian space-time smoothing models to produced sub-national estimates of U5MR for Zambia. The major methodological limitation of this approach is that it does not incorporate area-specific sampling variability at the first stage of analysis, which we show can be quite variable for small areas.

In this chapter we combine data from multiple surveys with different sampling designs, and construct subnational estimates through time with uncertainty that reflects the various data collection schemes. Data come from traditional cluster sample surveys (DHS) and two HDSS sites. HDSS sites intensively monitor everyone within a given area, typically to monitor the effects of health intervention trials of various types. Estimates of child mortality from both sources of data are useful but potentially flawed in different ways. National cluster sample surveys are generally not able to produce useful subnational estimates, and HDSS sites are not designed to be nationally representative, and are also thought to fall prey to the Hawthorne effect by which the communities of these sites have improved health

outcomes because they are under observation and, more concretely, because of the trials being conducted.

We construct subnational estimates of Tanzanian child mortality through time with uncertainty intervals. This problem is challenging because in addition to requiring smoothing over space and time, we must also account for the survey design. When sampling is not simple and random and the design variables (upon which sampling was based) are not available, the complex sampling design is accounted for by constructing design weights. Inference is then carried out using design-based inference, e.g. using Horvitz-Thompson estimators (Horvitz and Thompson, 1952). In contrast, a conventional space-time random effects framework, for example, Knorr-Held (2000), is model-based, and requires an explicit likelihood to be specified. In this chapter, we marry these two approaches by constructing a working likelihood based on the asymptotic distribution of a design-based estimator and then smooth using a space-time-survey hierarchical prior.

The organization of this chapter is as follows. In Section 4.2 we describe the two data sources upon which estimation will be based. In Section 4.3 the calculation of child mortality estimates with an appropriate standard error is described using discrete time survival models. The derivation of the standard error for under age 5 child mortality rate is described in Section 4.4. A simulation study to compare the derived standard error to the standard jackknife error used by DHS is explored in Section 4.5. Hierarchical Bayesian space-time models are introduced in Section 4.6. The results of our modeling efforts of under five mortality rates (U5MR) within Tanzania from 1980–2010 are given in Section 4.7 and discussed in Section 4.8.

4.2 Data Sources

We focus on child mortality using data from five Tanzanian Demographic and Health Surveys (TDHS), one Tanzania HIV and Malaria Indicator survey (THMIS), and two health and demographic surveillance system (HDSS) sites in Tanzania, Ifakara and Rufiji. Over the period 1980–2010 estimates of child mortality from the two types of data sources (surveys,

surveillance sites) are generally similar but, as described above, different in useful ways. The HDSS estimates are accurate (low bias) and precise (small variance) measurements for comparatively small, geographically-defined populations, and the household survey estimates are less accurate and much less precise but representative of large populations.

4.2.1 Health and Demographic Surveillance System

The Ifakara Health Institute (IHI), Tanzania runs a number of health and population research projects including two HDSS sites – Ifakara and Rufiji. We collaborated with IHI to estimate child mortality using data from the Ifakara and Rufiji HDSS sites.

The HDSS data are generated through repeated household visits. For the data we use, each household was visited three times per year at regular intervals. During each visit a 'household roster' was updated and all new vital and migration events for all members of the household were recorded. In addition, potentially many other questions were asked as part of both routine and 'add-on' studies. For our purposes we require only the basic core HDSS data that include information on dates of birth, death and migration – the information necessary to accurately identify observed person time, categorize that time by calendar period and age, and identify the outcome of interest, death. The Ifakara and Rufigi HDSS sites contribute data to the Morogoro and Pwani regions of Tanzania, respectively.

4.2.2 Household Surveys

Full TDHS surveys that collected data necessary for child mortality estimates were conducted in Tanzania in 2010, 2004–05, 1999, 1996, and 1991–92, in addition to the THMIS that included child mortality which was conducted in 2007–08. The 2010 TDHS, 2007–08 THMIS and 2004–05 TDHS surveys used 2-stage cluster samples. First, enumeration areas were sampled from the 2002 Tanzania census and second, a systematic sampling of households within each enumeration area was carried out. The 1999 TDHS, 1996 TDHS and 1991–92 TDHS used a 3-stage cluster design, first selecting wards and branches using the 1988 Tanzania Census as a sampling frame, second using probability proportional to size sampling

to select enumeration areas from each selected ward or branch, and third selecting households from a new list of all households in each selected enumeration area. The same first and second stage units were used for all three of the surveys. For all surveys stratification by urban/rural and region was done at the first stage, with oversampling of Dar es Salaam and other urban areas. The surveys were designed to be nationally representative and to be able to provide estimates of contraceptive prevalence at the regional level. All six household surveys contributed observations to the 21 mainland regions of Tanzania.

All women age 15 to 49 who slept in the household the night before were interviewed in each selected household and response rates were high (above 95% for households in all surveys). TDHS provides sampling (design) weights, assigned to each individual in the dataset. Limited information is provided for each survey concerning the calculation of survey weights, but the general explanation indicates that raw survey weights are the inverse of the product of the 2–3 probabilities of selection from each stage. These raw weights were then adjusted to reflect household response and individual response rates. The 1991–92 Tanzania DHS final report Demographic and Health Surveys (1992) indicates that "final individual weights were calculated by normalizing them for each area so that the total number of weighted cases equals the total number of unweighted cases", but this normalization is not discussed in later reports (Demographic and Health Surveys, 1997, 2000, 2005, 2010) or the DHS statistics manual (Rutstein and Rojas, 2006). For the purposes of our analysis of child mortality, children identified by the women who were interviewed contributed exposure time and deaths.

The data were organized into child-months from birth to either death or date of the mother's interview. The number of clusters, women, and children contributing data to each five year time period for each household survey is shown in Figures 4.1–4.3, respectively. The total number of children contributing person time within each five year time period by region is shown in Figure 4.4.

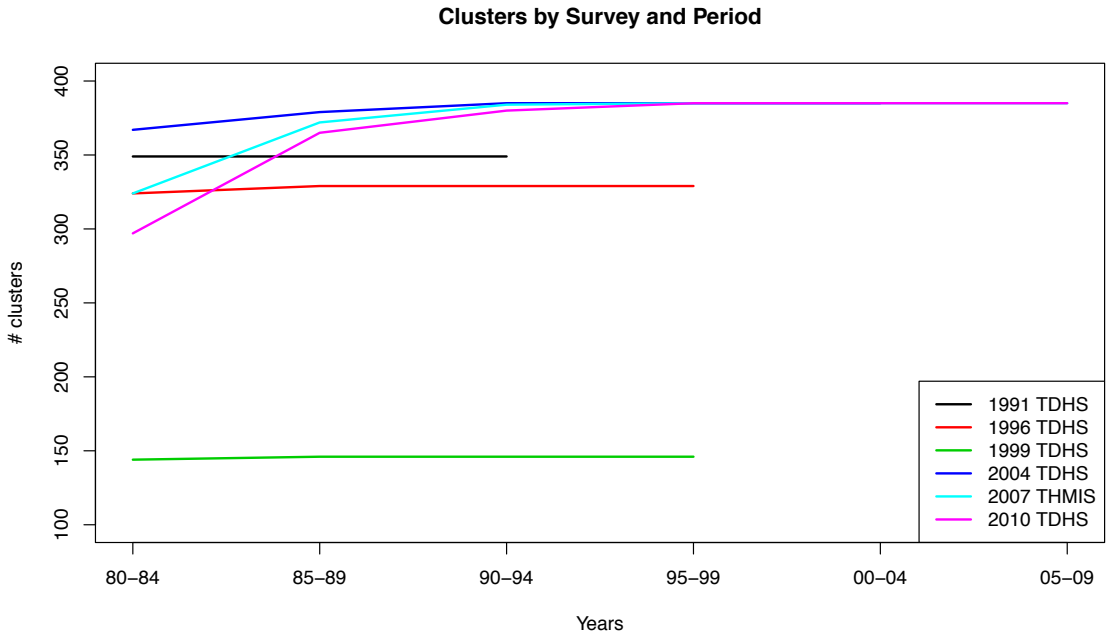

Figure 4.1: Number of sampled enumeration areas with data for each five year time period by survey.

4.3 Calculating child mortality with Discrete Time Survival Models

We modeled child mortality using discrete time survival analysis (DTSA) (Allison, 1984; Jenkins, 1995). Our main aim is to examine the change in risk as a function of age and historical period. DTSA allows us to easily estimate the predicted probabilities which can be used directly in traditional mortality analysis methods such as life tables, in our case to calculate U5MR. We wish to estimate the U5MR and define ${}_nq_x = \Pr(\text{dying before } x + n \mid \text{lived until } x)$ and the discrete hazards model splits the [0,5) period into J intervals $[x_1, x_2), [x_2, x_3), \ldots, [x_J, x_{J+1})$, where $x_{j+1} = x_j + n_j$ so that n_j is the length of the interval beginning at x_j, $j = 1, \ldots, J$. Then U5MR is calculated as

$$ {}_5q_0 = 1 - \prod_{j=1}^{J} (1 - {}_{n_j}q_{x_j}). \tag{4.1} $$

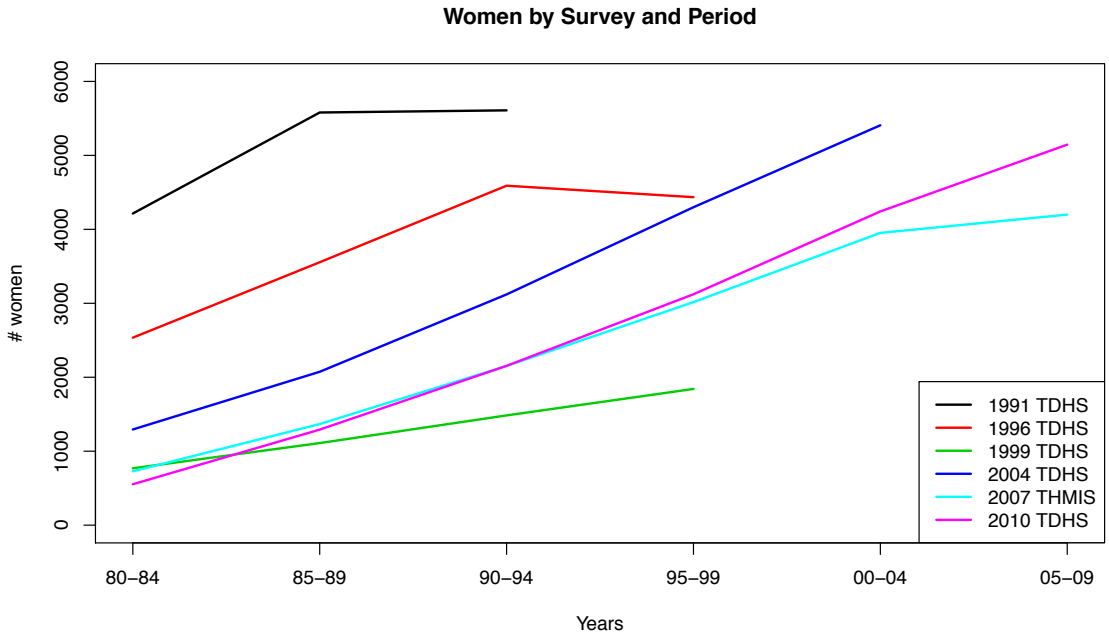

Figure 4.2: Number of sampled women with data for each five year time period by survey.

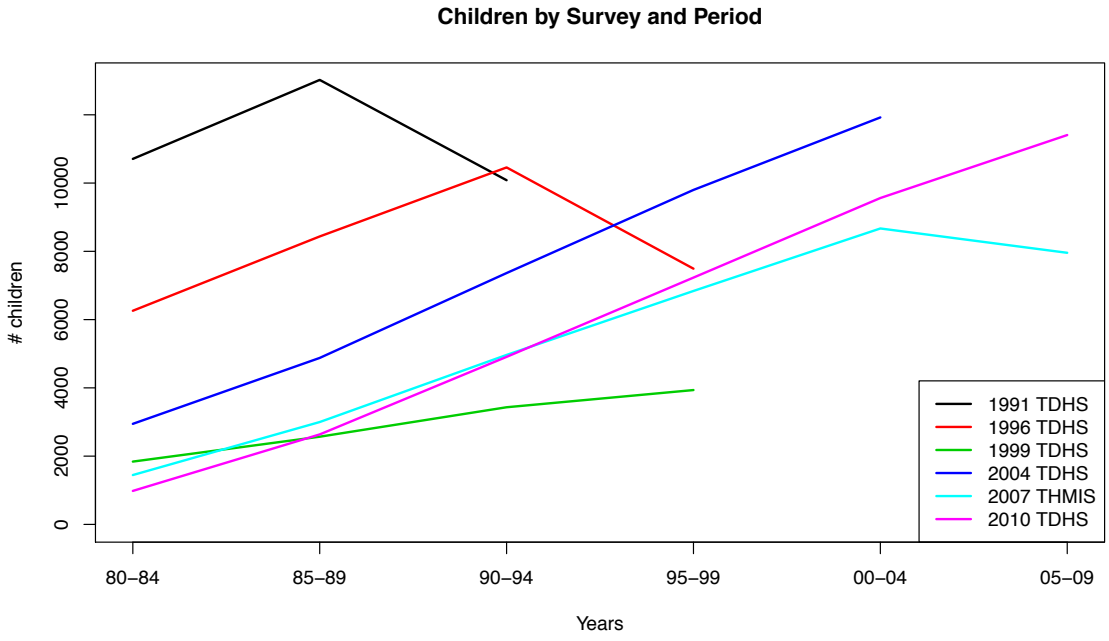

Figure 4.3: Number of children with data for each five year time period by survey.

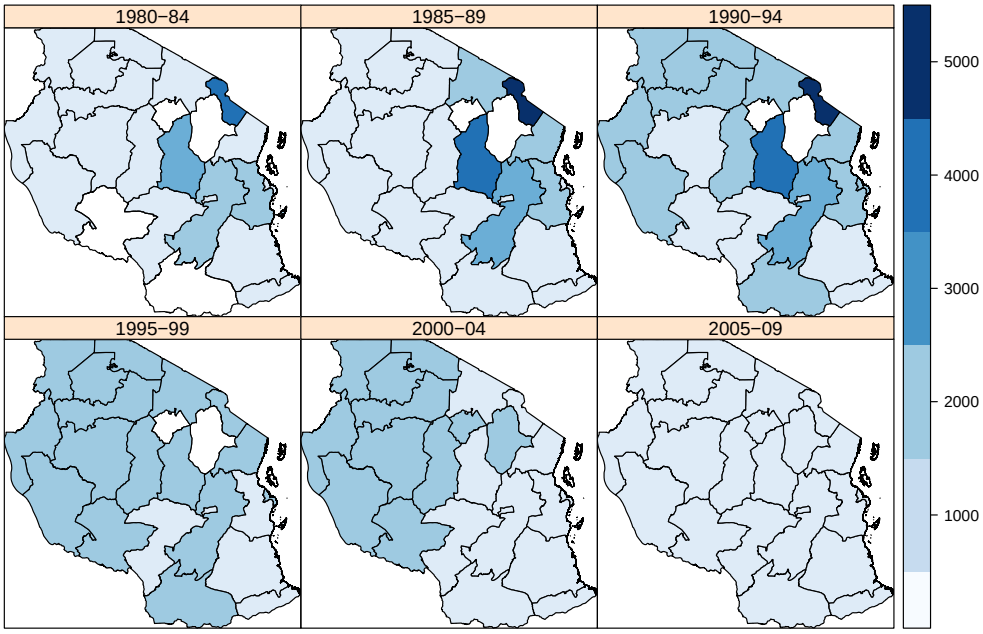

Figure 4.4: Number of children for each five year time period over all surveys by region.

For our purposes, ${}_5q_0$ is calculated by dividing the first 60 months into six intervals ($J = 6$), $[0,1), [1,12), [12,24), [24,36), [36,48), [48,60)$ with $(x_1, ..., x_6) = (0, 1, 12, 24, 36, 48)$ and $(n_1, ..., n_6) = (1, 11, 12, 12, 12, 12)$. Data were organized as child-months where each child was at risk during each month observed from birth up to and including the month of their death. The observed data consist of, for each birth, a binary sequence up to length 60 with 0/1 corresponding to survival/death. For example, a child that died in their fourth month would contribute one child-month to the first age category and three to the second age category. The first three child-months would be assigned a 0 outcome and the final month would be assigned a 1.

We use logistic regression to estimate the monthly probability of dying conditional on the state of the child at the beginning of the month. The monthly probability of death for each interval, ${}_1q_x$, is the probability of dying in $[x, x+1)$ given that $x \in [x_j, x_j + n_j)$ and can be estimated using a logistic generalized linear model (GLM) with J factors for age intervals,

$\text{logit}\,({}_1q_x) = \beta_j$ for $x \in [x_j, x_j + n_j)$.

In the complex survey context that is relevant for the Tanzanian household surveys, an important consideration is that the design weights must be acknowledged. In a finite population of size N we may fit the model $Y_i|\boldsymbol{\beta} \sim \text{Binomial}(1, p_i)$, with $p_i = \exp(\boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}))/[1 + \exp(\boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}})]$ and $\boldsymbol{\beta}$ a $J \times 1$ vector. The finite population parameter $\boldsymbol{B}$ is the solution to the score equations:

$$\sum_{i=1}^{N} x_{ij}\left[y_i - \frac{\exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{B})}{1 + \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{B})}\right] = 0 \text{ for } j = 1, \ldots, J.$$

When a survey is taken, following Binder (1983), a design-based estimate of $\boldsymbol{B}$ is given by the solution to

$$\sum_{i \in s} w_i x_{ij}\left[y_i - \frac{\exp(\boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{B}})}{1 + \exp(\boldsymbol{x}_i^{\mathrm{T}}\widehat{\boldsymbol{B}})}\right] = 0 \text{ for } j = 1, \ldots, J$$

where s represents the units included in the sample.

Once $\widehat{B}_j$ are estimated we can calculate $\widehat{{}_1q_x} = \exp(\widehat{B}_j)/[1+\exp(\widehat{B}_j)]$ for $x_j \in [x_j, x_j+n_j)$. The complement of surviving each month of the interval $[x_j, x_j + n_j)$ is used to calculate ${}_{n_j}\widehat{q}_{x_j} = 1 - \left(1 - \text{expit}(\widehat{B}_j)\right)^{n_j}$ where $B_j = \text{logit}({}_1q_x)$ for $x \in [x_j, x_j + n_j)$. Since

$$\begin{aligned} 1 - {}_{n_j}\widehat{q}_{x_j} &= \left(1 - \text{expit}(\widehat{B}_j)\right)^{n_j} = \left(\frac{e^{\widehat{B}_j} + 1}{e^{\widehat{B}_j} + 1} - \frac{e^{\widehat{B}_j}}{e^{\widehat{B}_j} + 1}\right)^{n_j} \\ &= \left(\frac{1}{e^{\widehat{B}_j} + 1}\right)^{n_j} = \left(e^{\widehat{B}_j} + 1\right)^{-n_j}. \end{aligned}$$

In terms of $\widehat{\mathbf{B}}$, (4.1) can be written as

$${}_5\widehat{q}_0 = 1 - \prod_{j=1}^{J}(1 - {}_{n_j}\,\widehat{q}_{x_j}) = 1 - \prod_{j=1}^{J}\left(e^{\widehat{B}_j} + 1\right)^{-n_j}.$$

4.4 Derivation of Standard Error for U5MR

In Section 4.6 we will construct, for a generic U5MR, a working likelihood based on the asymptotic distribution

$$y = \text{logit}\,({}_5\widehat{q}_0) \sim \text{N}(\eta, \widehat{V}_{\text{DES}})$$

where $\eta = \log[{}_5q_0/(1 - {}_5\, q_0)]$ and $\widehat{V}_{\text{DES}}$ is the asymptotic (design-based) variance estimate of $\text{logit}\,({}_5\widehat{q}_0)$, which is obtained via the delta method, and requires the variance of $\boldsymbol{B}$.

The design-based variance of $\boldsymbol{B}$ is more difficult to derive than the estimator $\widehat{\mathbf{B}}$, and the following is based on Roberts et al. (1987), using the notation of that paper. Suppose we have a saturated logistic model (as in our example) with I groups (factor levels). Let N_i be the true number of individuals who fall in group i and $N = \sum_{i=1}^{I} N_i$ the total number of individuals in the population. Also let N_{i1} be the number of individuals responding in group i. The ratio estimator of the proportion responding is

$$p_i = \frac{\widehat{N}_{i1}}{\widehat{N}_i} = \frac{\sum_{k \in s_i} w_k y_k}{\sum_{k \in s_i} w_k}$$

where s_i is the random set of sampled units that fall in group i and w_k is the design weight associated with surveyed response y_k, $k = 1, \ldots, n$ (so that n is the size of the survey). Also let

$$w_i = \frac{\widehat{N}_i}{\widehat{N}} = \frac{\sum_{k \in s_i} w_k}{\sum_{k \in s} w_k}$$

where S is the random set of all the samples. The design-based variance estimator of $\boldsymbol{p} = [p_1, \ldots, p_I]^{\text{T}}$ will be denoted $\widehat{\boldsymbol{V}}$, and obviously depends on the design chosen. The pseudo-MLEs of the fractions responding in group i will be denoted $\widehat{f}_i$. Then the variance-covariance of the estimator of $\boldsymbol{B}$ is given by equation (2.4) of Roberts et al. (1987):

$$\widehat{\text{var}}(\widehat{\boldsymbol{B}}) = n^{-1}\widehat{\Delta}^{-1}\boldsymbol{D}(w)\widehat{\boldsymbol{V}}\boldsymbol{D}(w)\widehat{\Delta}^{-1} \tag{4.2}$$

where $\Delta = \text{diag}(w_1\widehat{f}_1(1 - \widehat{f}_1), \ldots, w_I\widehat{f}_I(1 - \widehat{f}_I))$ and $\boldsymbol{D}(w) = \text{diag}(w_1, \ldots, w_I)$.

We choose to model the parameter $\eta = \text{logit}({}_5q_0)$. Ultimately we will use a Gaussian distribution for the first stage of our hierarchical model, so we would like a parameter that

can take values along the whole real line. We have

$$\begin{aligned}\eta = \text{logit}({}_5q_0) &= \log\left(\frac{{}_5q_0}{1-{}_5q_0}\right) \\ &= \log\left(\frac{1-\prod_{j=1}^{J}\left(e^{B_j}+1\right)^{-n_j}}{\prod_{j=1}^{J}\left(e^{B_j}+1\right)^{-n_j}}\right) \\ &= \log\left(\prod_{j=1}^{J}\left(e^{B_j}+1\right)^{n_j}-1\right)\end{aligned} \tag{4.3}$$

The asymptotic distribution of the MLE is $\widehat{\boldsymbol{B}} \sim N(\boldsymbol{B}, \boldsymbol{\Sigma})$ and from the delta method we can find the asymptotic distribution of $\hat{\eta}$:

$$\widehat{\eta} \sim N\left(\text{logit}({}_5q_0), \widehat{V}_{\text{DES}}\right) \tag{4.4}$$

where

$$\widehat{V}_{\text{DES}} = \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{B}}^T \widehat{\boldsymbol{\Sigma}} \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{B}}. \tag{4.5}$$

where $\widehat{\boldsymbol{\Sigma}}$ is $\widehat{\text{var}}(\widehat{\boldsymbol{B}})$ from (4.2). The weights depend on the design and $\widehat{\boldsymbol{\Sigma}}$ can be extracted from the `svyglm()` function within the `survey` package.

Then, if we define:

$$\gamma = \prod_{j=1}^{J}\left(e^{B_j}+1\right)^{n_j}$$

and set $\eta = \log(\gamma - 1)$ we find, after some algebra,

$$\frac{\partial \eta}{\partial B_j} \;=\; \frac{\gamma}{\gamma-1} \times \left[n_j \times \text{expit}(B_j)\right].$$

The value of $\frac{\partial \eta}{\partial \boldsymbol{B}}$ is used in (4.5) for the asymptotic distribution, (4.4) which is used to derive $100(1-\alpha)\%$ confidence intervals for ${}_5q_0$ as

$$\left[\text{expit}\left(\hat{\eta} + \sqrt{\widehat{V}_{\text{DES}}} \times z_{\alpha/2}\right), \text{expit}\left(\hat{\eta} + \sqrt{\widehat{V}_{\text{DES}}} \times z_{1-\alpha/2}\right)\right]. \tag{4.6}$$

4.5 Simulation to test coverage performance of derived SE

Pedersen and Liu (2012) discuss the difficulties associated with deriving a variance estimate in the context of child mortality estimates. DHS typically uses a jackknife estimator, $V_{\text{JACK}}({}_5\widehat{q}_0)$

of ${}_5\widehat{q}_0$, which for cluster sampling is

$$\widehat{V}_{\text{JACK}} = \frac{N_c - 1}{N_c} \sum_{c=1}^{N_c} ({}_5\widehat{q}_{0(c)} - {}_5\widehat{q}_0)^2 \tag{4.7}$$

where N_c is the number of clusters and ${}_5\widehat{q}_{0(c)}$ is the estimate based on all of the data while holding out the c-th cluster (Lohr, 2009, ch. 9). A $100(1-\alpha)\%$ confidence interval for ${}_5q_0$ is based on

$$\left[\quad {}_5\widehat{q}_0 - z_{1-\alpha/2} \times \sqrt{\widehat{V}_{\text{JACK}}} \quad , \quad {}_5\widehat{q}_0 + z_{1-\alpha/2} \times \sqrt{\widehat{V}_{\text{JACK}}} \quad \right]. \tag{4.8}$$

4.5.1 Data Generation

A simulated dataset was created to assess the coverage properties of an interval based on V_{DES} and V_{JACK}. To simulate the estimation process within one region 100,000 women were assigned to 500 clusters. The number of births for each woman were generated from a Poisson distribution with rate of 3. Each birth was assigned a calendar month between 0 and 119 (a 10 year period).

We considered three scenarios to generate the deaths within the first 60 months of life. Deaths were assigned based on a multinomial distribution with the following discrete hazards p_0 for the first month, p_1 for months 2–12 and p_2 for months 13–60. As we are only interested in death within the first 5 years, the remaining probability was assigned to a 61st category for death after the age of 5 years.

In the first scenario we assume that monthly probabilities of death are constant between clusters and assumed

$$\begin{aligned} \text{logit}(p_0) &= \alpha_0 \\ \text{logit}(p_1) &= \alpha_0 + \alpha_1 \\ \text{logit}(p_2) &= \alpha_0 + \alpha_2 \end{aligned}$$

where α_0, α_1, and α_2 were chosen to correspond to probabilities p_1, p_2, p_3 of 0.04, 0.004, and 0.001, respectively as shown in shown in Figure 4.5. This distribution of probability

represents the highest risk in the first month of life, an elevated risk for the remainder of the first year, and the lowest risk for the following 4 years. These probabilities result in an expected U5MR of 124.5 deaths per 1,000, which is similar to the late 1990s U5MR in Tanzania.

In the second and third scenario we generated data with cluster-specific probabilities

$$\begin{aligned} \text{logit}(p_{0,c}) &= \alpha_0 + \epsilon_c \\ \text{logit}(p_{1,c}) &= \alpha_0 + \alpha_1 + \epsilon_c \\ \text{logit}(p_{2,c}) &= \alpha_0 + \alpha_2 + \epsilon_c \end{aligned}$$

for clusters $c = 1, ..., 500$. Scenario 2 had low between cluster variability with $\epsilon_c \sim N(0, \sigma = 0.1)$ resulting in the U5MR varying from 81.7 to 156.2 deaths per 1,000 births. Scenario 3 had higher between cluster variability with $\epsilon_c \sim N(0, \sigma = 0.3)$ resulting in the U5MR varying from 54.0 to 265.9 deaths per 1,000 births. The distribution of cluster-specific U5MRs are shown in Figure 4.6. As we are only interested in death within the first 5 years, the remaining probability was assigned to a 61st category for death after the age of 5.

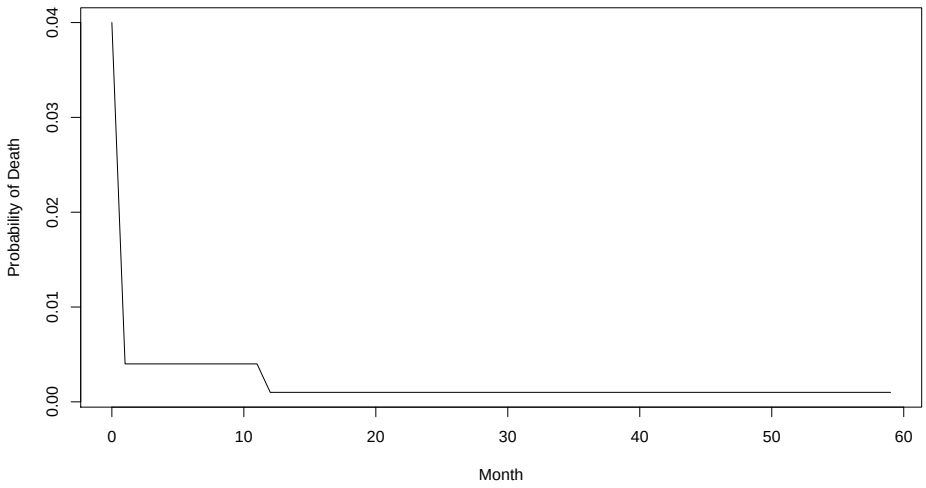

Figure 4.5: Monthly probability of death for the first 60 months.

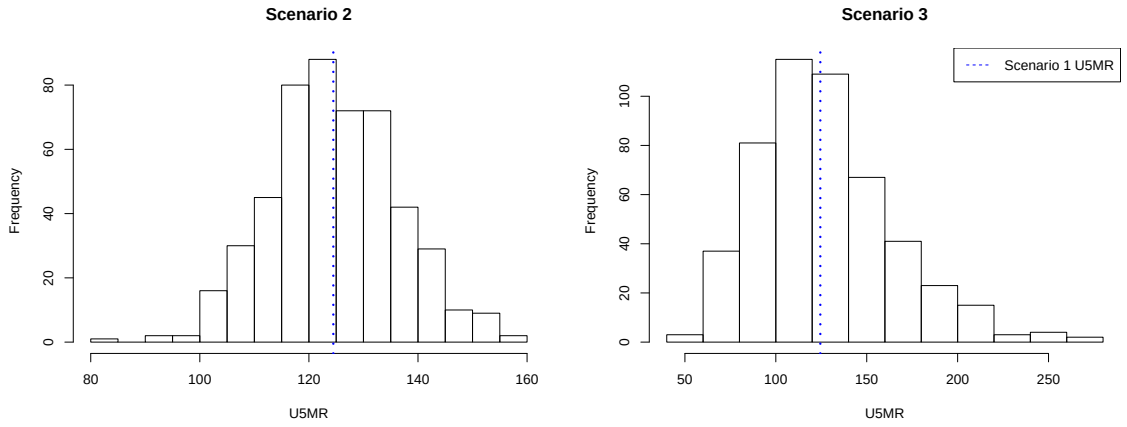

Figure 4.6: Simulated cluster-specific under 5 child mortality rates (U5MR).

When creating estimates for a particular 5 year calendar period from the household survey and HDSS data, births before or during this period contribute person-time from children who are born or die or are censored in other 5 year time periods. To mimic this aspect of the real data, births were simulated from a wider time period and then subsetted to include only observations within the relevant time period. Children who survive past 60 months only contribute person time for months 0–59. This subset included person-time from approximately 240,000 unique births from 91,000 mothers and the U5MR in these populations were 131.2, 133.3, and 136.4 per 1,000 for scenarios 1, 2, and 3, respectively.

4.5.2 Simulation Procedure

We employed a two stage cluster sampling design. At the first stage n_{c} clusters (analogous to enumeration areas in the TDHS designs) were randomly selected from the N_{c} available. At the second stage, suppose cluster c is selected, then n_{W} women were randomly selected from the $N_{\mathrm{W}c}$ total women within the selected cluster. The resulting sampling weights for a mother selected in cluster c is

$$w_{\mathrm{E}c} = \frac{N_{\mathrm{c}}}{n_{\mathrm{c}}} \times \frac{N_{\mathrm{W}c}}{n_{\mathrm{W}}}.$$

The number of clusters was set at $N_c = 500$ for all simulations and the number selected at the first stage was one of $n_c = 15, 25$. Sample sizes within clusters (n_W) varied by 5 within 10–30. For each combination of n_c and n_W we draw 1,000 samples and the corresponding delta method and jackknife confidence intervals were created based on $\widehat{V}_{DES}$ and $\widehat{V}_{JACK}$, respectively. The sample coverage of each interval type was calculated as the average number of intervals that contained the true population U5MR.

4.5.3 Results

The coverage of the delta method and jackknife intervals by number of clusters and within sample size cluster are shown in Figure 4.7. Results are much as one would expect from clustered sampling, coverage improves when there are more clusters and within a given number of clusters there is little gain in precision when increasing the sample size. Generally the performance of the delta method and jackknife intervals is very similar. Figure 4.8 displays the average 95% confidence interval width by number of clusters and sample size within cluster. As expected, the intervals narrow as either the number of first stage clusters or the number of samples within each cluster increases and scenario 3, which has the most between cluster variability, has the widest intervals.

The TDHS sampling scheme is generally around 25 clusters with a within sample size of approximately 20 women, which corresponds to the blue lines at a sample size of 20. This suggest that TDHS intervals may be slightly anti-conservative. We prefer the delta method as it is generally applicable (i.e., to a variety of designs) and has a far smaller computational burden. We conclude that the asymptotic normal sampling distribution and the delta method variance result in sufficiently accurate confidence interval coverage for the cluster and sample sizes considered in our application. Consequently, we will use the asymptotic distribution with the delta method variance as a working likelihood for our hierarchical modeling.

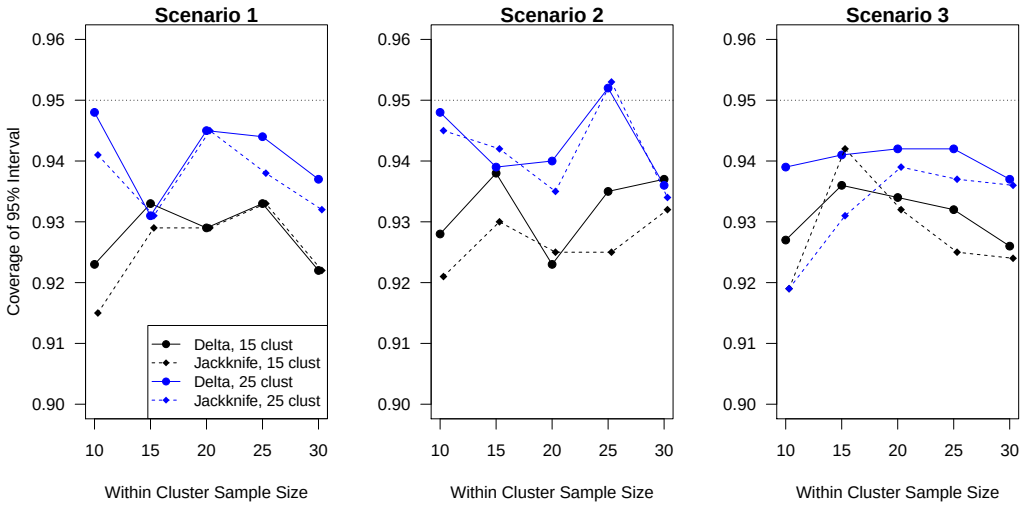

Figure 4.7: Coverage of jackknife and delta method 95% confidence intervals by number of clusters and within cluster sample size. Sample size values are offset horizontally to improve the visibility.

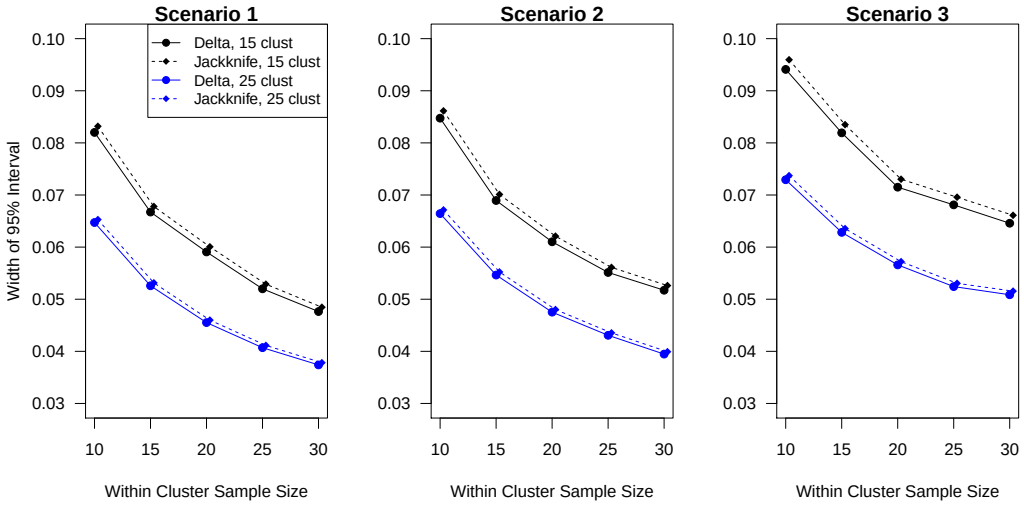

Figure 4.8: Mean width of jackknife and delta method 95% confidence intervals by number of clusters and within cluster sample size. Sample size values are offset horizontally to improve the visibility.

4.6 Combining Data Sources in Hierarchical Bayesian Space-Time Model

4.6.1 The First Stage

Let ${}_5\widehat{q}_{0its}$ represent the estimate of U5MR from survey s in region i and in period t. A model-based approach to inference with survey data may be carried out if the design variables upon which sampling were based, and associated population totals, are available (Gelman, 2007b). Unfortunately, not all of these variables are available for the Tanzanian household surveys. As an alternative we have selected to summarize the data in area i at time point t from survey s via the asymptotic distribution of the logit transformed estimator:

$$y_{its} = \log\left[\frac{{}_5\widehat{q}_{0its}}{1 - {}_5\widehat{q}_{0its}}\right].$$

We define the area, period and survey summary as $\eta_{its} = \log\left[{}_5q_{0its}/(1 - {}_5q_{0its})\right]$. Motivated by the performance and flexibility of the empirical logit approach proposed in Section 3.7, we take as working likelihood the asymptotic distribution

$$y_{its} \mid \eta_{its} \sim \mathrm{N}\left(\eta_{its}, \widehat{V}_{\mathrm{DES},its}\right). \tag{4.9}$$

Dwyer-Lindgren et al. (2014) also used the logit transformed estimator, but did not incorporate design effects and failed to incorporate the precision of the direct estimates by assuming a common variance across all observations.

4.6.2 Second Stage Smoothing Models

We wish to smooth over time period, region and surveys, but would like as parsimonious a model as possible, to avoid overfitting. At the second stage of our model we adopt a model similar to the 'Type I' inseparable space-time model of Knorr-Held (2000). However, unlike Knorr-Held (2000) our data provides multiple observations for each area i and time point t through the THMIS, five TDHS and two HDSS, denoted as surveys s. Thus we consider models that allow the option of survey-specific effects. The survey effects could be constant over time and space, could vary with time, vary with space, or vary by time and space.

The six candidate models we consider are given in Table 4.1, with the caption containing the random effects specification. There are two temporal terms with α_t being independent and identically distributed random effects that pick up short-term fluctuations with no structure, and γ_t being given an (intrinsic) random walk prior of order 1 or 2 (models type 'a' or 'b'), to pick up local temporal smooth fluctuations, for $t = 1, \ldots, T = 6$ time periods. Five-year time periods were chosen because survey-specific regional sample sizes can be quite small. The UN IGME has only recently moved to annual estimates at the national-level because the sample size of recent DHS has increased (Pedersen and Liu, 2012). We are combining recent and older DHS at a regional-level and thus sample sizes are not sufficiently large to produce reliable annual estimates.

There are also two spatial terms, corresponding to the convolution model of Besag et al. (1991) as described in Section 2.4.2. The independent random effects are denoted θ_i and the intrinsic conditional autoregressive (ICAR) terms are ϕ_i for $i = 1, \ldots, I = 21$ regions of Tanzania (Section 2.4.1). The latter perform local geographical smoothing. The space-time interaction terms δ_{it} are taken to be independent, which corresponds to the Type I interaction model of Knorr-Held (2000). Type II-IV interaction models were considered, which include spatial and/or temporal structure on the prior for δ_{it}, but these models didn't not substantially modify estimates, so Type I was selected for parsimony.

There are $S = 8$ different surveys that are carried out over the various time periods (since mothers are surveyed on their complete birth history and so report on births from previous time periods), the five TDHS and THMIS surveys cover all 21 regions over the different time periods they were administered and the HDSS sites contribute data for one region each in the last three time periods. The independent random effects ν_s allow for these surveys to have a systematic displacement from the true logit of U5MR. The interactions ν_{ts} and ν_{is} allow these displacements to vary with period and space, respectively, while ν_{its} allow the complete interaction between survey, period and area. Model I contains crossed random effects only, since each area is represented in each of the time periods. Models II–VI contain a combination of nested and crossed random effects. The random walk and ICAR models

are described briefly in Chapter 2 and fully in Rue and Held (2005).

Table 4.1: Random effects models for time period t, region i and survey s. In all models μ is the intercept and $\alpha_t \sim_{iid} \text{N}(0, \sigma_\alpha^2)$, $\theta_i \sim_{iid} \text{N}(0, \sigma_\theta^2)$, $\phi_i \sim \text{ICAR}(\sigma_\phi^2)$, $\delta_{it} \sim_{iid} \text{N}(0, \sigma_\delta^2)$. Specific models contain random effects with distributions $\nu_s \sim_{iid} \text{N}(0, \sigma_{\nu 1}^2)$, $\nu_{is} \sim_{iid} \text{N}(0, \sigma_{\nu 2}^2)$, $\nu_{ts} \sim_{iid} \text{N}(0, \sigma_{\nu 3}^2)$, $\nu_{its} \sim_{iid} \text{N}(0, \sigma_{\nu 4}^2)$. In the 'a' models $\gamma_t \sim \text{RW1}(\sigma_\gamma^2)$ and in the 'b' models $\gamma_t \sim \text{RW2}(\sigma_\gamma^2)$.

Model	Linear Predictor η_{its}
I	$\mu + \alpha_t + \gamma_t + \theta_i + \phi_i + \delta_{it}$
II	$\mu + \alpha_t + \gamma_t + \theta_i + \phi_i + \delta_{it} + \nu_s$
III	$\mu + \alpha_t + \gamma_t + \theta_i + \phi_i + \delta_{it} + \nu_s + \nu_{is}$
IV	$\mu + \alpha_t + \gamma_t + \theta_i + \phi_i + \delta_{it} + \nu_s + \nu_{ts}$
V	$\mu + \alpha_t + \gamma_t + \theta_i + \phi_i + \delta_{it} + \nu_s + \nu_{ts} + \nu_{is}$
VI	$\mu + \alpha_t + \gamma_t + \theta_i + \phi_i + \delta_{it} + \nu_s + \nu_{ts} + \nu_{is} + \nu_{its}$

4.6.3 Hyperpriors Specification

For a generic set of independent random effects we specify priors on the precision τ such that a 95% prior interval for the residual odds ratios lies in the interval [0.5,2] which leads to Gamma$(a_{\text{MARG}}, b_{\text{MARG}})$ priors for precisions (Wakefield, 2009) with $a_{\text{MARG}} = 0.5$, $b_{\text{MARG}} = 0.001488$. For the RW1, RW2 and ICAR models the precisions have *conditional* rather than *marginal* interpretations. Let $\boldsymbol{z}$ represent a random effect from a improper GMRF with "mean" $\mathbf{0}$ and "precision" $\tau^\star \boldsymbol{Q}$. Following the supplementary materials of Fong et al. (2010), we gain compatibility by calculating an approximate measure of the average marginal "variance" of $\boldsymbol{z}$ in the situation with $\tau^\star = 1$; call this average c. Then to put on the same scale we take $a_{\text{COND}} = a_{\text{MARG}}$ and $b_{\text{COND}} = b_{\text{MARG}}/c$. In the above description, the words mean, precision, variance are written in italics to acknowledge that strictly speaking these quantities do not exist since the distribution is improper. However, one may calculate a

generalized inverse using the equation given at the end of Section 4.4 of Fong et al. (2010). This method is closely related to that later described by Sørbye and Rue (2014). For the Tanzania data this leads to gamma priors for the RW1 of $\tau_\gamma \sim \text{Gamma}(0.5, 0.00153)$, for the RW2 of $\tau_\gamma \sim \text{Gamma}(0.5, 0.00286)$, and for the ICAR of $\tau_\phi \sim \text{Gamma}(0.5, 0.00360)$.

Our model contains a relatively complex combination of nested and crossed random effects and we described a particular approach to hyperprior selection. As with any such suggestion, it is beneficial to examine prior sensitivity. For our application we considered three different prior distributions corresponding to 95% ranges for the residuals odds ratios of [0.5,2], [0.2,5], [0.1,10]. Figure 4.9 illustrates the sensitivity of point and interval estimates for the hyperparameters. Ultimately, [0.5,2.0] was selected for the results presented in this chapter. We see sensitivity for the spatial random effects, though the total spatial random effects contributions remain relatively constant since as the structured random effects increase, the unstructured random effects decrease. The structured temporal random effects are robust, which is reassuring since these provide the largest contribution to the overall variability; these are well-estimated, however, since the trend is strong. Similarly, the unstructured survey-area random effects are robust, but all of the standard deviations of the remaining independent random effects show modest increases as the prior moves further from zero.

4.6.4 Computation

Model fitting was carried out within the `R` computing environment. Weighted logistic regressions were fit using the `svyglm()` function from the `survey` package (Lumley, 2004) from which the design-based variance was extracted. The hierarchical Bayesian space-time models were fitted using the Integrated Nested Laplace Approximation (INLA) (Rue et al., 2009) as implemented in the `INLA` package. INLA provides a fast alternative to MCMC for approximating the marginal posterior distributions of Markov random field (MRF) models. There is now extensive evidence that the approximations are accurate for space-time modeling, see for example Fong et al. (2010), Schrödle and Held (2011), and Blangiardo and Cameletti

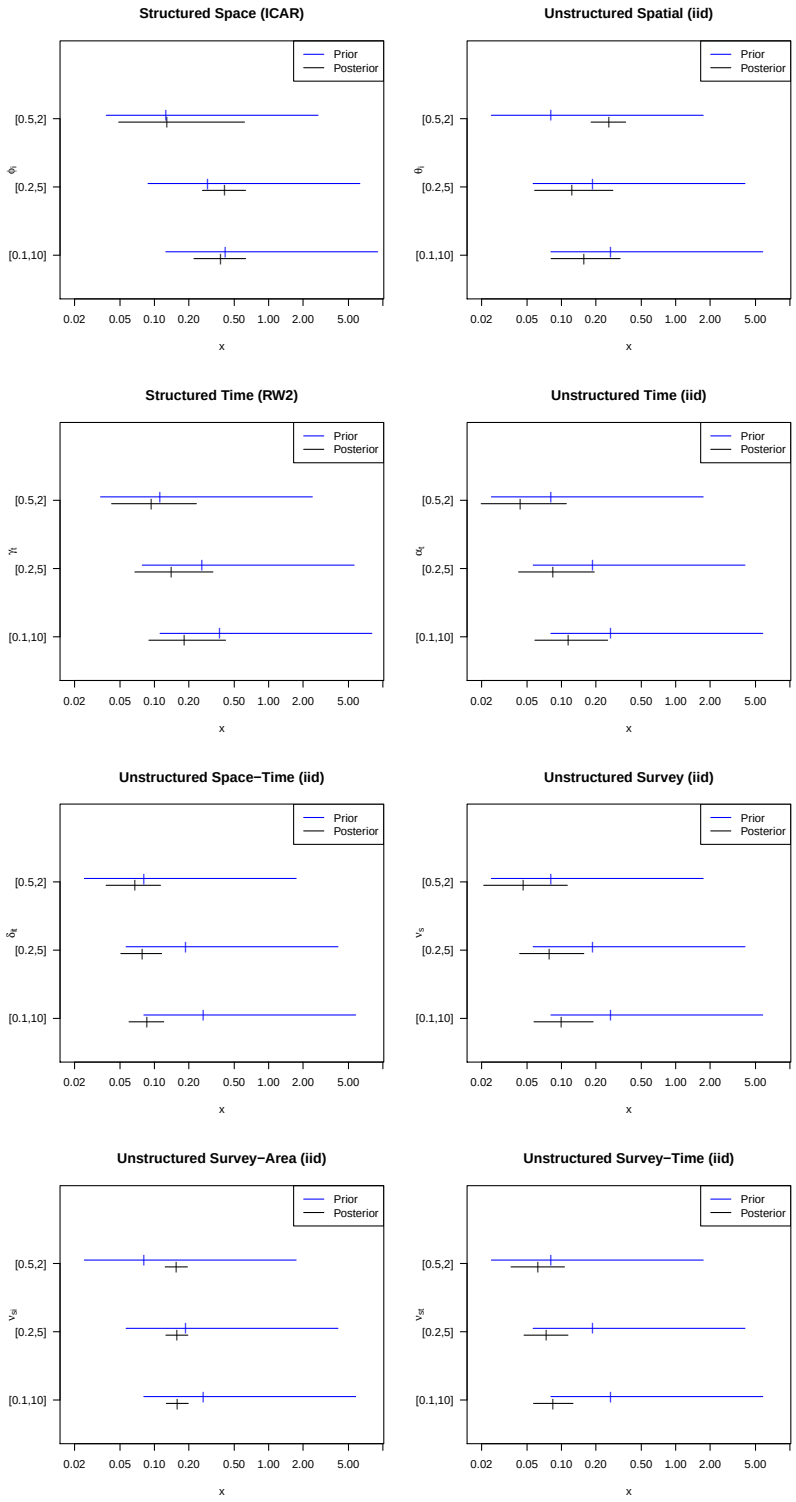

Figure 4.9: Prior sensitivity of the standard deviations of the eight random effects in the model. The three priors are based on 95% prior intervals on the residual odds ratios of [0.5,2], [0.2,5], [0.1,10].

(2015).

4.6.5 Model Selection

In Table 4.1 we describe twelve plausible random effects specifications (allowing for RW1 or RW2 models). A number of approaches have been described for comparing models, including the conditional predictive ordinate (CPO), the deviance information criteria (DIC) as introduced by Spiegelhalter et al. (2002) and the normalizing constants $p(\boldsymbol{y}|M)$ for the twelve models indexed by M. Let $\boldsymbol{y}_{-its}$ represent the vector of data with the observation from region i, time period t and survey s removed. The idea behind the CPO is to predict the density ordinate of the left-out observation, based on those that remain. Specifically, the CPO for observation i, t, s is defined as:

$$\text{CPO}_{its} = p(y_{its}|\boldsymbol{y}_{-its}) = \int p(y_{its}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{y}_{-its})\ d\boldsymbol{\theta} = E_{\theta|y_{-its}}\left[p(y_{its}|\boldsymbol{\theta})\right]$$

where $\boldsymbol{\theta}$ represents the totality of parameters and in the U5MR setting the distribution of $y_{its}|\boldsymbol{\theta}$ is $\text{N}(\eta_{its}, \widehat{V}_{\text{DES},its})$. The CPOs can be used to look at local fit, or one can define an overall score for each model:

$$\text{LCPO} = \log\left(\text{CPO}\right) = \sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{s=1}^{S}\log \text{CPO}_{its},$$

and good models will have relatively high values of LCPO. Held et al. (2010) discuss shortcuts for computation (i.e. avoidance of fitting the model $I \times T \times S$ times) using INLA.

We also calculate another widely-used model comparison measure, the deviance information criteria, or DIC (Spiegelhalter et al., 2002). To define the DIC with respect to a generic set of parameters $\boldsymbol{\theta}$, first define an 'effective number of parameters' as

$$p_D = E_{\theta|y}\{-2\log[p(\boldsymbol{y}|\boldsymbol{\theta})]\} + 2\log[p(\boldsymbol{y}|\overline{\boldsymbol{\theta}})] = \overline{D} + D(\overline{\boldsymbol{\theta}})$$

where D is the deviance, $\overline{\boldsymbol{\theta}} = E[\boldsymbol{\theta}|\boldsymbol{y}]$ is the posterior mean, $D(\overline{\boldsymbol{\theta}})$ is the deviance evaluated at the posterior mean and $\overline{D} = E[D|\boldsymbol{y}]$.

The DIC is given by

$$\text{DIC} = D(\overline{\boldsymbol{\theta}}) + 2p_D = \overline{D} + p_D,$$

so that we have the sum of a measure of goodness of fit and model complexity. We are wary of interpretation of DIC in our setting, since Plummer (2008) has shown that DIC is prone to inappropriately under-penalize large models such as the ones we are fitting, see also Spiegelhalter et al. (2014).

4.7 Applying Methods to household surveys and HDSS sites in Tanzania

We fit Models Ia–VIb (as summarized in Table 4.1) to the Tanzania survey and surveillance data and Table 4.2 provides the summaries of various model comparison summaries. Model Vb is the favored model according to both the DIC, LCPO, and log of the normalizing constant criterion. Results for Models Vb and VIb are very similar, but we see from the effective number of parameters that even though the number of 3-way interaction random effects is 573, there are only 13 effective parameters due to the closeness of the interactions to zero. Hence, from this point onwards we shall report summaries with respect to Model Vb. We begin by summarizing the posterior distribution, and then describe regional trends.

4.7.1 Summarizing the Posterior Distribution

Table 4.3 provides numerical summaries and the proportion of total variation explained by each random effect. The total variance is

$$\sigma_\alpha^2 + s_\gamma^2 + \sigma_\theta^2 + s_\phi^2 + \sigma_\delta^2 + \sigma_{\nu_s}^2 + \sigma_{\nu_{si}}^2 + \sigma_{\nu_{st}}^2$$

where s_γ^2 and s_ϕ^2 are empirical estimates of the marginal variances in the RW2 and ICAR models. The structured temporal and unstructured spatial random effects explain 77% of the total variation. Hence, there is strong temporal structure and large spatial heterogeneity, which we shall discuss subsequently. The third largest contribution to the variation is 11% from the survey-space interaction. Different survey teams are sent to different regions which explains to some extent this relatively large contribution.

Table 4.2: Model comparison: p_D is the effective degrees of freedom, as defined for the calculation of the deviance information criteria (DIC), which also uses the deviance evaluated at the posterior mean, $\overline{D}$; LCPO is defined as $\sum_{its} \log(CPO_{its})$. In the 'a' models $\gamma_t \sim$ RW1(σ^2_γ) and in the 'b' models $\gamma_t \sim$ RW2(σ^2_γ).

Model	No Pars	$\log p(\boldsymbol{y})$	p_D	$\overline{D}$	DIC	LCPO
Ia	181	-310.9	74.5	409.3	483.8	-294.5
IIa	189	-304.6	80.1	384.2	464.3	-287.3
IIIa	313	-257.6	118.9	221.8	340.7	-193.5
IVa	223	-302.2	88.6	367.5	456.2	-283.4
Va	347	-254.6	121.8	210.1	332.0	-183.1
VIa	920	-255.0	134.5	199.4	334.0	-183.9
Ib	181	-308.2	74.2	409.1	483.3	-293.7
IIb	189	-301.9	79.8	383.9	463.7	-286.4
IIIb	313	-255.0	118.6	221.7	340.3	-192.9
IVb	223	-299.5	88.2	367.4	455.6	-282.5
Vb	347	**-252.1**	121.6	209.9	**331.5**	**-183.1**
VIb	920	-252.5	133.3	200.2	333.4	-183.4

Table 4.3: Summaries of variance components. The proportion of variation is calculated as the contribution the relevant set of random effects makes to the total variation. In the case of the RW2 and ICAR models, the relevant contribution is evaluated empirically, since the variance parameter is conditional rather than marginal.

Variance	Interpretation	Median (95% Interval)	Percentage Variation
σ^2_α	Indept Time	0.002 (0.001, 0.012)	1.3
σ^2_γ	RW2 Time	0.009 (0.002, 0.054)	46.0
σ^2_θ	Indept Space	0.068 (0.033, 0.133)	31.3
σ^2_ϕ	ICAR Space	0.017 (0.002, 0.378)	4.9
σ^2_δ	Indept Space-Time Interaction	0.005 (0.001, 0.013)	2.3
$\sigma^2_{\nu_s}$	Indept Survey	0.002 (0.001, 0.013)	1.4
$\sigma^2_{\nu_{st}}$	Indept Survey-Time Interaction	0.004 (0.001, 0.011)	2.0
$\sigma^2_{\nu_{si}}$	Indept Survey-Space Interaction	0.024 (0.015, 0.038)	10.9

4.7.2 Model Validation

To validate the model we removed all of the observations in area i for time point t and then generated 95% intervals around the posterior mean ${}_5\tilde{q}_{0,it}$ using the variance of the observed response, defined as $\tilde{S}^2_{its} = \tilde{\sigma}^2_{it} + \widehat{V}_{\text{DES,its}}$, where $\tilde{\sigma}^2_{it}$ is the variance of the posterior distribution of logit$({}_5\tilde{q}_{0,it})$ and $\widehat{V}_{\text{DES,its}}$ is the design-based variance described in Section 4.3. This was completed for the 21 regions and 6 time points. Intervals contained the design-based estimates 92.5% of the time overall. Figure 4.10 displays the intervals created with the variance of the observed logit response around the posterior mean with the observed point estimates from each survey, by region for each time interval. Time/area-specific coverages range from 89.9–96.9% and the coverage for the final time point is 93.2%.

4.7.3 Summary of Random Effects

In this section we present graphical summaries of the various random effects present in Model Vb of the Tanzania U5MR model. Figure 4.11 presents the posterior medians of the spatial (ICAR) and unstructured random effects (note that the scales on the different plots differ). All temporally structured random effects are from a RW2.

Figure 4.12 plots the unstructured temporal random effects versus period, along with the survey by period random effects. The unstructured random effects are relatively small in magnitude compared with the survey by period random effects.

Figure 4.13 displays the unstructured time random effects (α_t) compared with the structured time random effects (γ_t). The structured random effects have a much larger range than the unstructured effects. Figure 4.14 displays structured time random effects (γ_t) by time. There is a noticeable negative trend in the random effects.

Figure 4.15 provides maps of the unstructured space-time random effects (δ_{it}). There is no clear spatial pattern to the high-valued and low-valued random effects by time period. So, based on this plot, there is little evidence of spatially structured space-time interaction (Type III interaction), which is consistent with the model comparison statistics given in

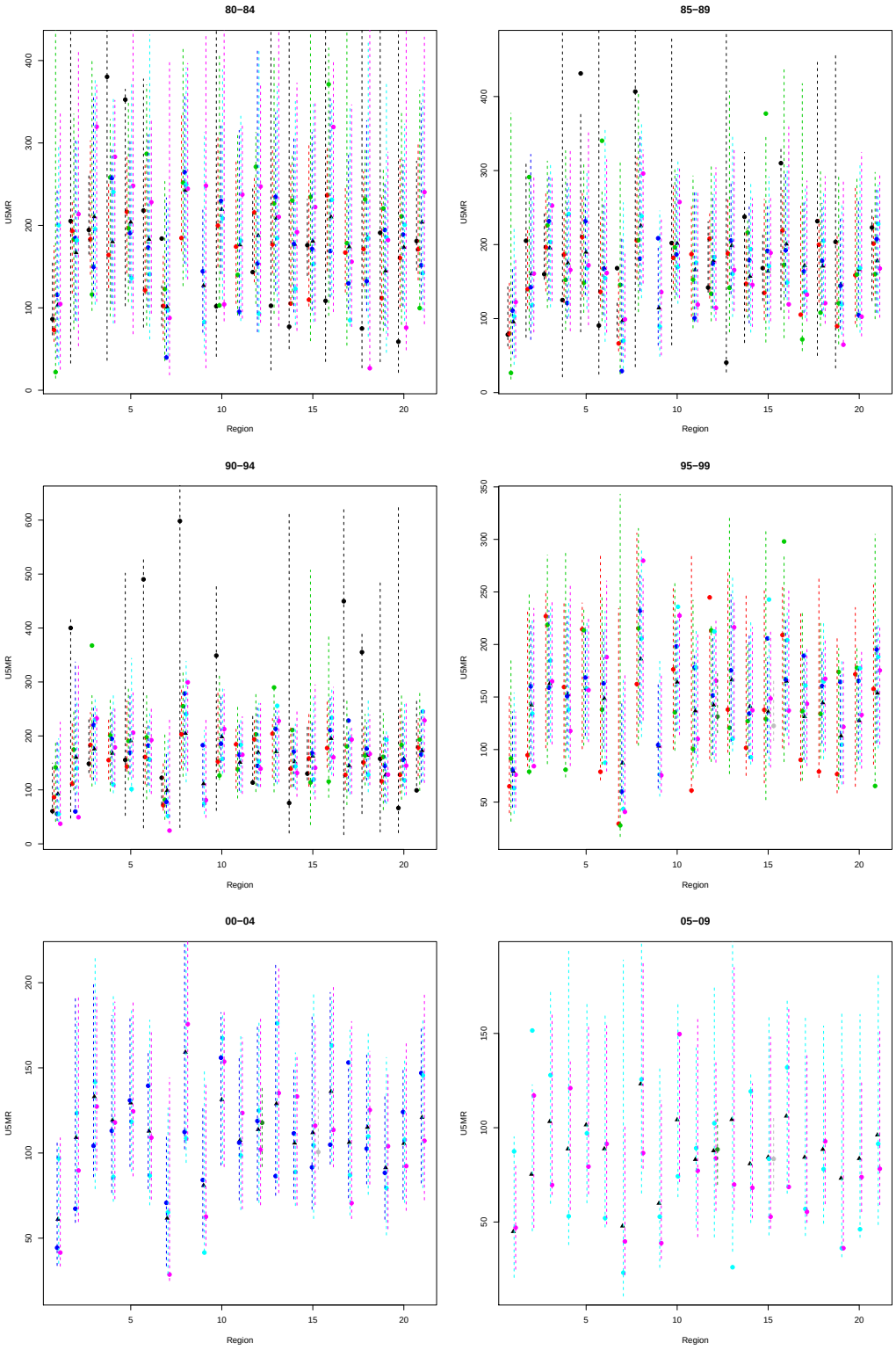

Figure 4.10: Intervals based on the variance of the observed logit response and region and time-specific direct estimates.

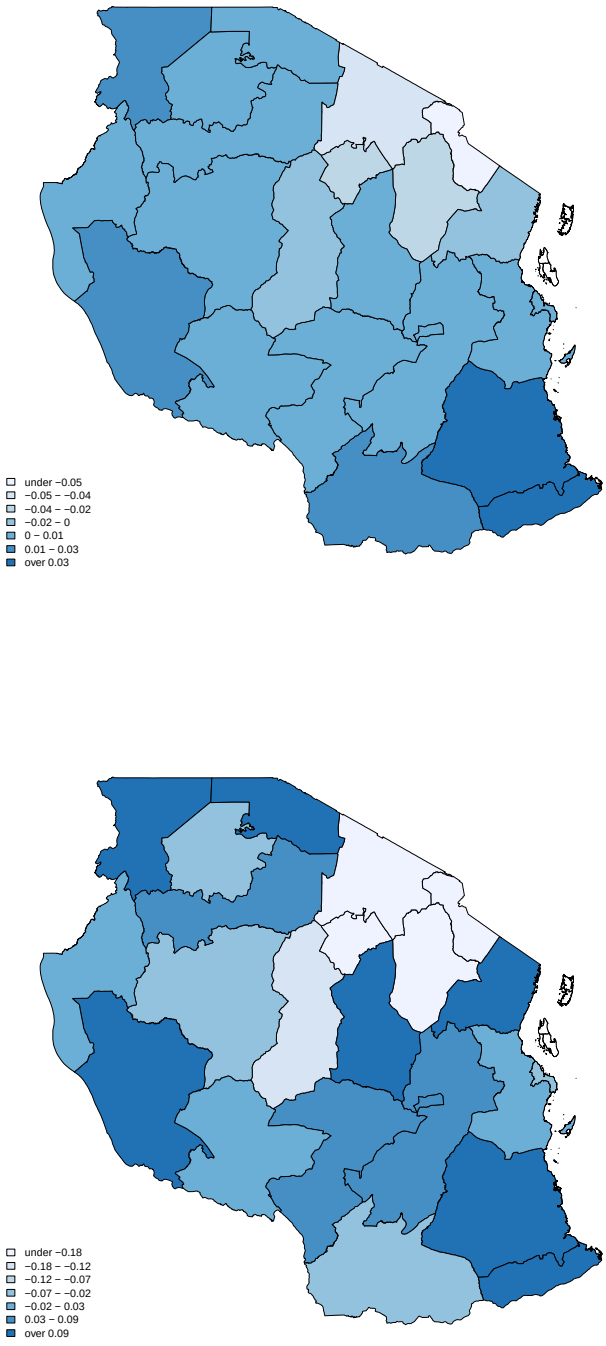

Figure 4.11: ICAR random effects, ϕ_i (top) and unstructured spatial random effects, θ_i (bottom).

Table 4.2. Similarly, the plots in Figure 4.16 which display the survey-area random effects (ν_{is}) along with the magnitude of the survey-specific (ν_s) random effect, do not show similar spatial patterns over the different time periods, suggesting there is not strong evidence for a temporally structured space-time interaction (Type II interaction).

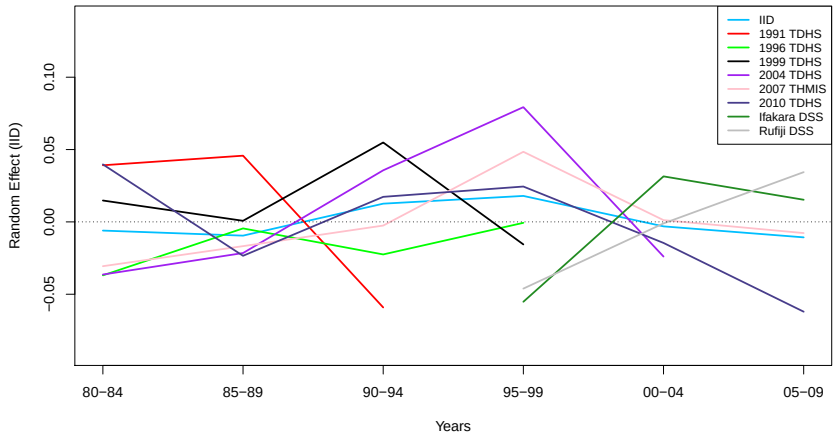

Figure 4.12: Unstructured time (α_t) and survey-time (ν_{st}) random effects.

4.7.4 Comparison of weighted and unweighted estimates

Figure 4.17 provides maps of the inverse-variance weighted Horvitz-Thompson regional estimates of child mortality. Unlike the smoothed maps provided in the main text these maps display some unlikely temporal trends. Figure 4.18 displays the differences in weighted and unweighted regional estimates of U5MR and associated variances for all surveys. We see that the weighting makes a significant impact on many of the region U5MR estimates and variances, with weighted estimates often being 25% higher or lower than unweighted estimates. Similarly, variances calculations which ignore the design range from 50–200% of the design-based variance.

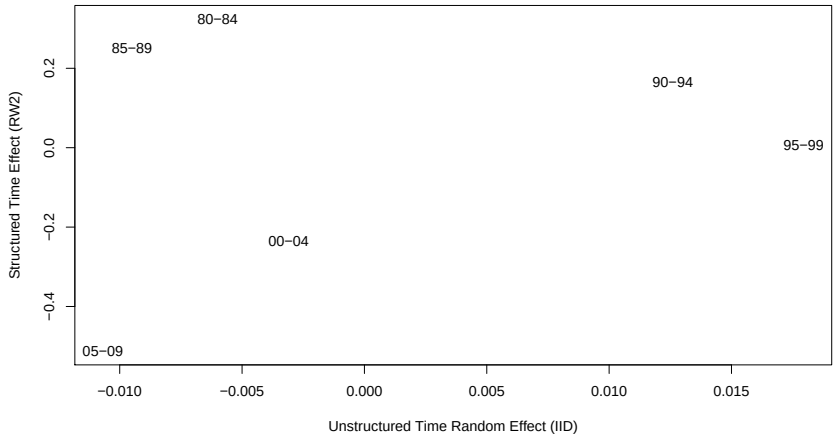

Figure 4.13: Unstructured time (α_t) and structured time (γ_t) random effects.

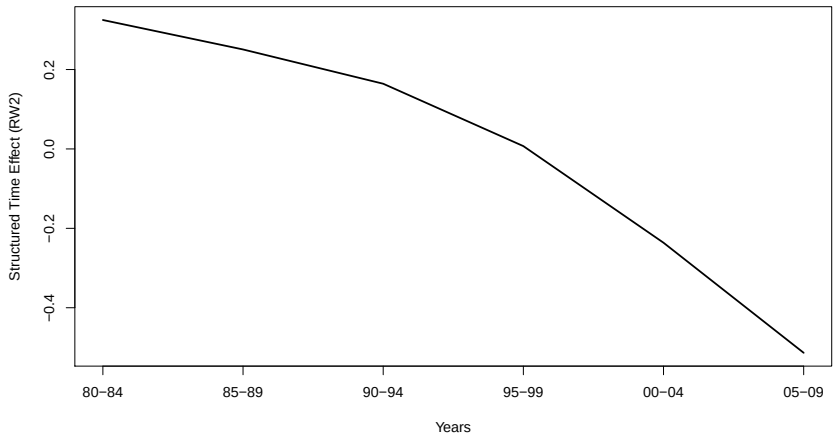

Figure 4.14: Structured time (γ_t) random effects.

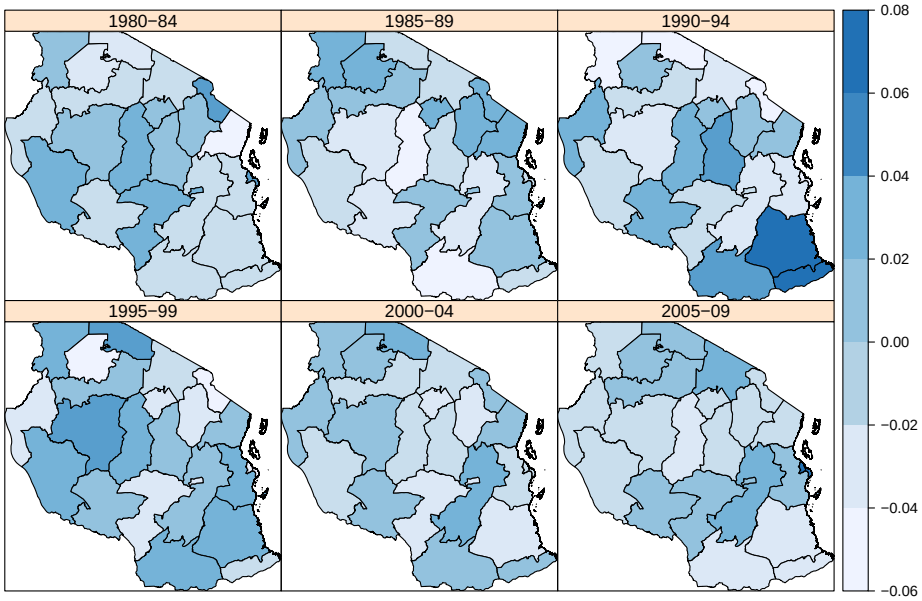

Figure 4.15: Unstructured space-time random effects (δ_{it}).

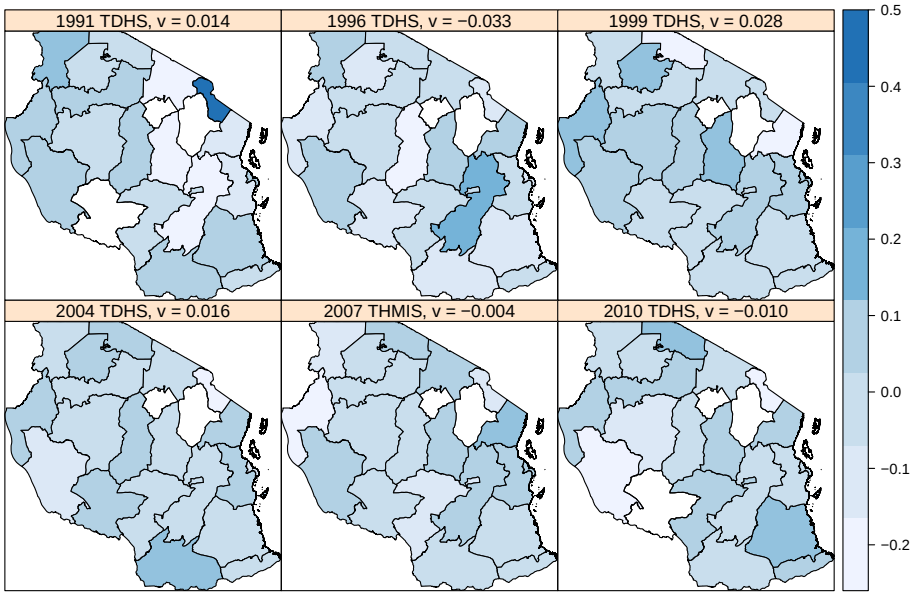

Figure 4.16: Survey (ν_s) and survey-area (ν_{si}) random effects. The median random effect (ν_s) is given in the heading of each plot. There are five Demographic and Health Surveys (DHS) and one Tanzania HIV and Malaria Indicator survey (THMIS).

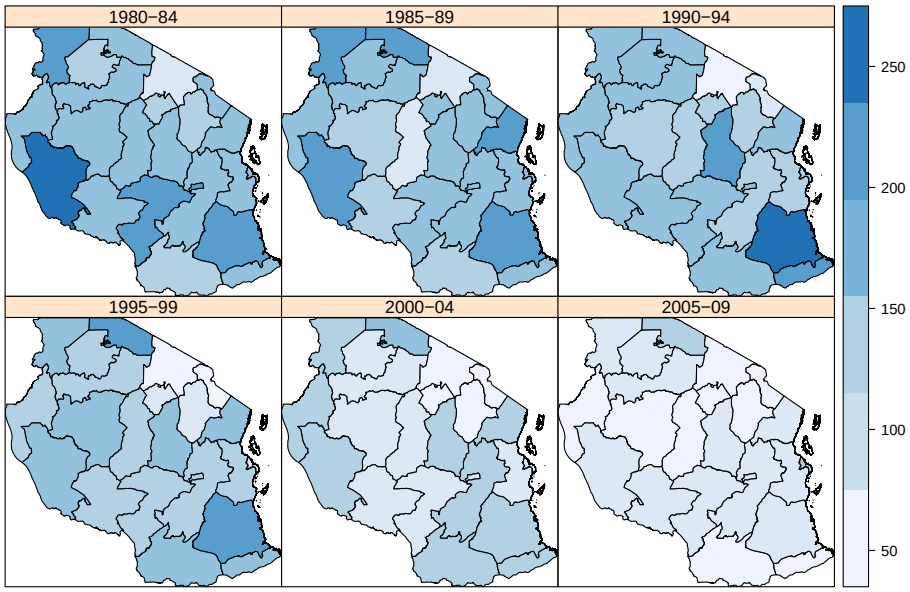

Figure 4.17: Inverse-variance weighted Horvitz-Thompson regional estimates of child mortality (per 1000 births).

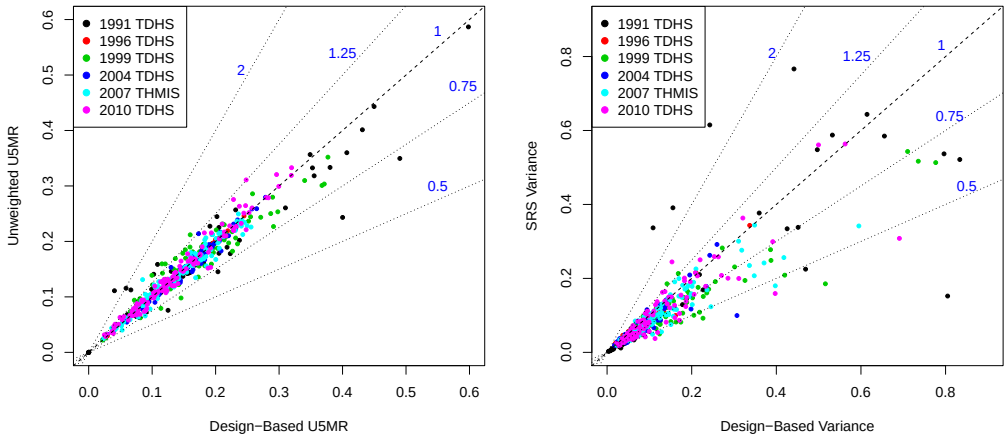

Figure 4.18: Comparison of design-based and unweighted estimates of U5MR and variances. Values in blue indicate the slope of the lines.

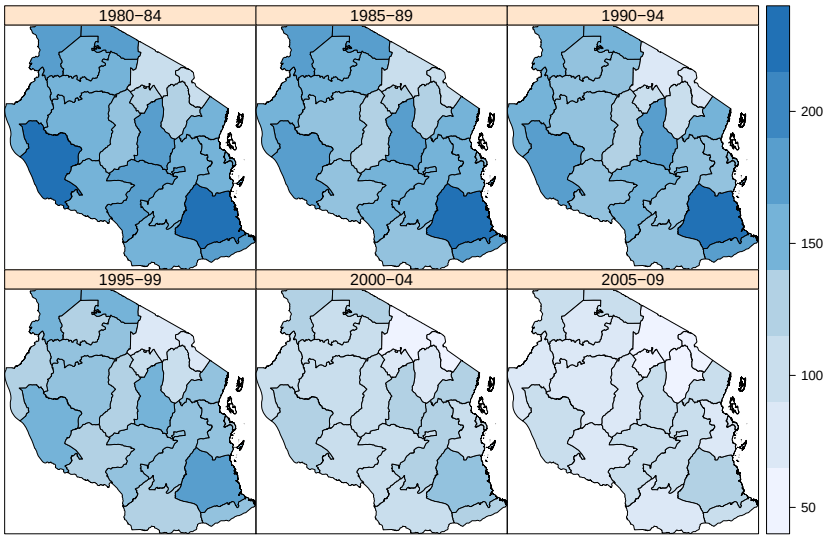

Figure 4.19: Smoothed regional estimates of child mortality (per 1000 births).

4.7.5 Regional Estimates and Projections

For region i and 5 year period t, estimates, projections, and credible intervals of U5MR taken from posterior draws of

$$ {}_5q_{0,it} = \text{expit}\left(\mu + \alpha_t + \gamma_t + \theta_i + \phi_i + \delta_{it}\right). $$

Figure 4.19 shows maps of the posterior median estimates of child mortality (per 1000 births) by region for the six observed 5 year time periods. Child mortality has decreased markedly over the 30 year period considered but overall more than 5% of infants still die before they turn 5, and there are strong regional differences. Figures 4.20 and 4.21 display the observed direct estimates and smoothed results for the Morgoro and Pwani regions, respectively. Additionally, each plot shows the projected U5MR for the 2010–2014 time period. The direct estimates have a great deal of variability between surveys, especially for the first four time points and design-based intervals are very wide.

Figure 4.22 shows the smoothed values for each region, by each time point and a projection into the 2010–2014 period. Figure 4.23 shows the percent decrease in each region compared

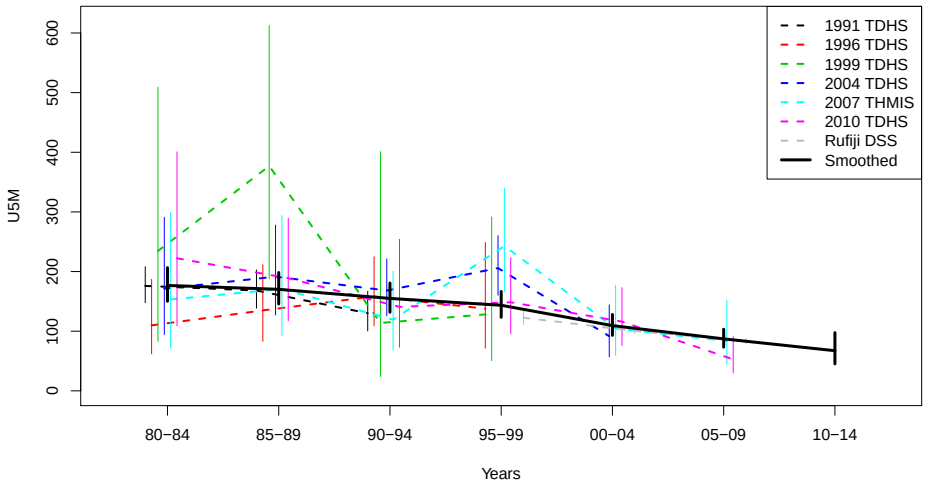

Figure 4.20: Regional five-year direct and model-based smoothed of ${}_5q_0$ in Pwani, TZA with 95% confidence intervals.

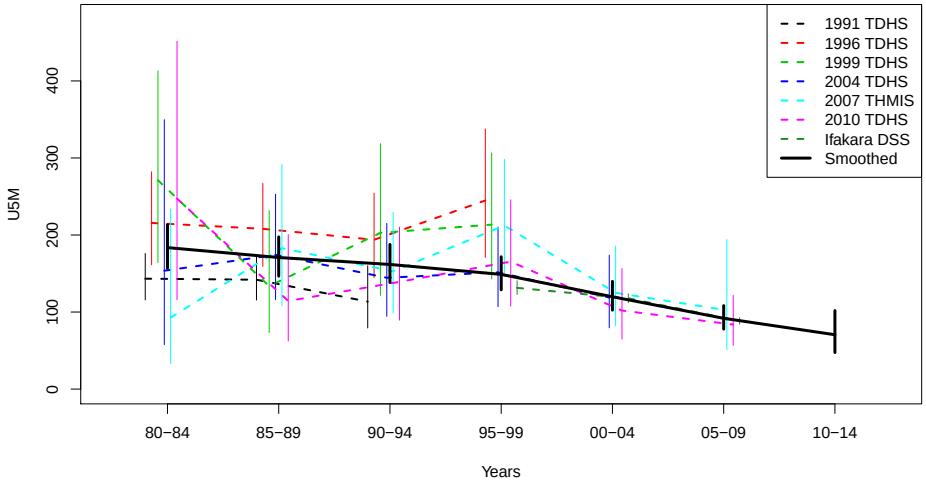

Figure 4.21: Regional five-year direct and model-based smoothed estimates of ${}_5q_0$ in Morogoro, TZA with 95% confidence intervals.

to the 1985–1989. The line at -66% corresponds to the the fourth millennium development goal of a two thirds reduction in child mortality by 2015.

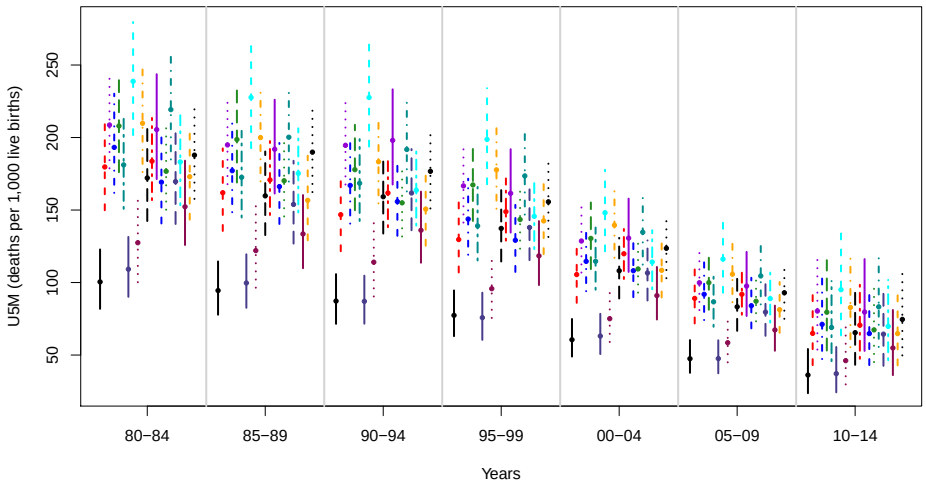

Figure 4.22: Posterior medians and 95% intervals for the 21 regions of Tanzania and a projection for 2010–2014.

4.8 Discussion

We have described a general method for spatiotemporal smoothing of a health outcome, with the data arising from complex surveys and surveillance. The method was illustrated with child mortality in regions of Tanzania over 1980–2009 using data from household surveys and surveillance sites. A great advantage of the model is that there is a fast implementation within the R computing environment using the existing `survey` and `INLA` packages. As an example, fitting the most complex model for the Tanzania data took just 18.7 seconds on a Macbook Pro[1].

In our hierarchical modeling approach we explicitly acknowledge the weights by taking as likelihood the design-based sampling distribution of the estimator. As we have shown

[1]processor: 2.9 GHz Intel Core i7; memory: 8 GB 1,600 MHz DDR3.

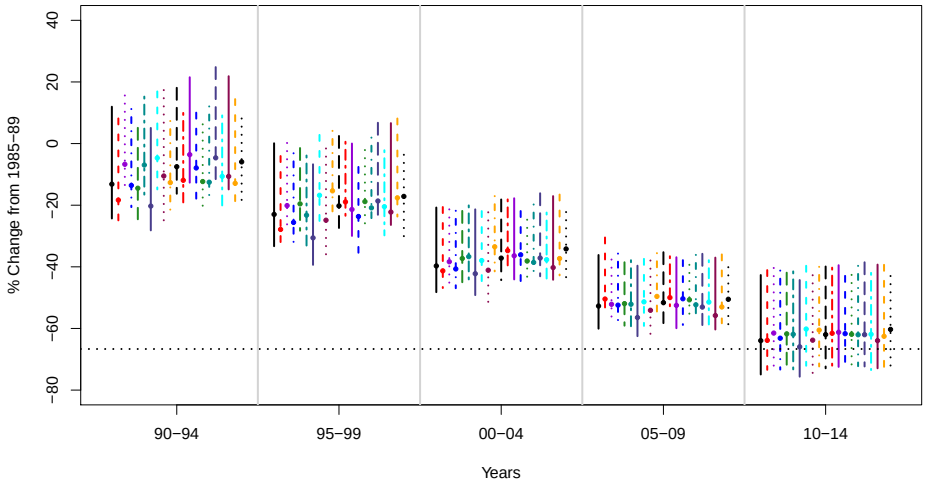

Figure 4.23: Percent reduction in region-specific child mortality since 1985–1989 with projections for 2010–2014.

the impact of the weights on both the estimates and the standard errors can be quite large. Another use of our model is for prediction, with the RW2 terms drawn from the relevant conditional distribution as shown with our projected decreases in child mortality for the 2010-2014.

An integral part of our method involves calculating and pooling estimates of child mortality from household surveys and demographic surveillance sites and allowing both to inform our overall estimates by region and for the country as a whole. A byproduct of this procedure is an ability to carefully compare the DHS-based and demographic surveillance-based estimates of child mortality in the regions that include HDSS sites. As Figures 4.20 and 4.21 makes clear, the central estimates from the two different data collection schemes are very similar. This adds more weight to similar findings by others (Byass et al., 2007; Fottrell et al., 2009; Hammer et al., 2006) and reduces concerns about the Hawthorne effect preventing measures of child mortality from HDSS sites from being more widely relevant, i.e. similar to surrounding populations.

Although we have demonstrated our method with a single country and outcome, it is sufficiently general to be applied to produce spatiotemporal estimates of a variety of indicators. Because this approach provides consistent, precise estimates across both time and space utilizing data from a variety of sources, including complex sample surveys, accounting for study designs, it should be considered as an approach for producing subnational estimates of child mortality and other key health, demographic and development indicators. However, countries with a substantial HIV/AIDs burden may suffer from underreporting biases. The UN IGME pre-processes data in a number of countries, including Tanzania, to take account of underreporting biases because of HIV/AIDS. We base our analysis on direct national estimates of U5MR, and so do not adjust for this bias, but our smoothed results do not differ substantially from the UN results at the national level and so we believe that any bias from this source will be small.

The world's rapidly growing appetite for timely, subnational estimates of key development indicators will continue to motivate innovative new developments in both data collection and analysis. In addition to providing a means to improve indicator estimates using different sources of data, our results also hint at the possibility of eventually creating integrated data collection and analysis schemes that build on existing infrastructure to yield some of the functionality of full-coverage CRVS. Clark et al. (2012) and Ye et al. (2012) begin to discuss ideas in this vein, e.g. how one might utilize both sample surveys and demographic surveillance to continuously provide indicators equivalent to what is normally produced by vital registration. The method and results we present in this chapter encourage future development of those ideas.

Chapter 5

JOINT MODELING OF EUROPEAN BREAST CANCER INCIDENCE AND MORTALITY

5.1 Introduction

Accurate estimation of cancer incidence and mortality are important for research and planning. Estimates provided by the World Health Organization (WHO) Global Burden of Disease 2004 update (Mathers et al., 2009) suggests that cancer is currently one of the leading causes of death worldwide and that the proportion of deaths attributed to cancer is expected to increase. Thus, cancer rates are of particular interest in societies with aging populations or in the developing world where life expectancy is extending and mortality rates due to communicable diseases are decreasing.

The International Agency for Research on Cancer (IARC), which is the specialized cancer agency of the WHO, provides estimates and predictions of cancer incidence and mortality of major types of cancer worldwide through the GLOBOCAN project (Parkin et al., 2001; Ferlay et al., 2010, 2013). These estimates are used by public health organizations globally. Data quality and sources vary widely between countries from complete registry coverage to incomplete national mortality data in countries without vital registration. When data quality is low, the current IARC modeling procedure informally borrows information from neighboring countries. The IARC approach leverages a number of different models, depending on data availability and quality within each country and, acknowledging the difficulties associated with such an endeavor, but does not provide any measures of uncertainty to accompany estimates.

In recent years the Institute for Health Metrics and Evaluation (IHME) has done extensive work to compile large amounts of data and complete descriptive analyses on many public

health and demographic indicators. Efforts include estimates of child mortality (Rajaratnam et al., 2010a; Wang et al., 2014), maternal mortality (Hogan et al., 2010), all cause mortality by gender (Rajaratnam et al., 2010b), worldwide education levels (Gakidou et al., 2010), breast and cervical cancer (Forouzanfar et al., 2011), and the global burden of all cancers (Fitzmaurice et al., 2015). In all of these analyses a large quantity of data, from diverse sources, are processed and pooled to generate the desired country-specific rates. Unlike the estimates produced by IARC, the IHME estimates include an uncertainty interval. However, the usual method used to generate the intervals can be quite complex and statistical properties, such as frequentist coverage, of the intervals are unknown. Further, the same basic method is used generically for a variety of endpoints, which one would expect to be suboptimal.

Missing and incomplete data is a frequent concern when cause-specific mortality is examined. One advantage when studying cancer rates is that cancer registry data are quite common. In the best cases (for example, in Scandinavia), countries have complete cancer registries and vital registration. However, in over half of the countries worldwide only local incidence data is available and mortality data ranges from complete vital registration to no available mortality data. When registry and/or mortality data is incomplete, relationships between incidence and mortality are used to estimate unmeasured incidence and/or mortality. Often this relationship is estimated via the mortality-incidence (MI) ratio (Forouzanfar et al., 2011; Uhry et al., 2013) or a mixture of survival analyses, MI ratios, and incidence-mortality (IM) ratios (Ferlay et al. 2010a,b, 2013).

We use data provided by IARC to generate estimates of country- and age-specific rates of cancer cases and deaths that would appear in the health care system for 40 countries in Europe from 1990-2010. In their best year thirty-nine of the 40 countries in the four United Nations-defined areas of Europe have some national mortality data. Among these 39 countries, 23 also have national incidence data, 10 countries have mortality and incidence data from local registries, and 6 countries have no incidence. Montenegro has no data for incidence or mortality.

In an effort to standardize the estimation approach IARC has created an alphanumeric scoring system which independently describes the availability of incidence and mortality data in each country. These scores are available for all countries that are included in the GLOBOCAN estimates. The methods described in this chapter condense these scores into four data categories based on national or sub-national source and the presence or absence of incidence or mortality data in Europe. The data is discussed further in Section 5.3.

Our aim is to use Bayesian space-time smoothing models to borrow strength between countries to provide estimates of national incidence and mortality, along with measures of uncertainty. Cancer is a complex set of diseases and incidence varies across space, time and different populations (with an obvious difference being age structure). There are common underlying biological processes that lead to cancer, and so we would expect similarities across space, and temporal changes in cultural and environmental risk factors are likely to be small suggesting that temporal smoothing is merited. But we would like to avoid constancy of rates in time and space due to the differences in cultural practices and environmental conditions. There are also differences in the probability of incident cases appearing in surveillance which can vary greatly by country.

Our overall approach to modeling incidence and mortality is to directly model mortality (which is more universally available) and the MI ratio. The latter relationship can be estimated from countries with good quality incidence and mortality data. Nearly all the countries in Europe have national mortality data, but have only sub-national incidence data available from local registries. For countries with sub-national registry data the MI ratio is estimated based on the local incidence and mortality data and then the unobserved incidence (in areas not covered by sub-national registries) is estimated based on the difference between the national and sub-national mortality. The smoothed average MI ratio is used in countries with only national mortality data to estimate the national incidence.

The rest of this chapter is organized as follows. In Section 5.2 our notation is established. The national and sub-national registry data provided by IARC is described in Section 5.3. In Section 5.4 we briefly discuss the factors associated with breast cancer incidence and

mortality rates as well as potential sources of bias. Section 5.5 describes the estimation methods employed by IARC and IHME. The joint modeling of incidence rates and the MI ratio is developed in Section 5.6. A spatial simulation study is detailed in Section 5.7. Section 5.8 presents models for age and time to be considered in our full analysis of European breast cancer data, which is conducted in Section 5.9 and discussed in Section 5.10.

5.2 Notation

We begin by establishing notation with a representing 5 year age group, c representing country, t denoting time in years, L being local (registry) data and R the remainder data (not covered by registries). For a generic country:

- N_{act}^L = Population in time t, age group a, local area in country c (population covered by all registries)

- Y_{act}^L = Incident reported cases in time t, age group a, local area in country c (total of all registries)

- Z_{act}^L = Reported deaths (Mortality) in time t, age group a, local area in country c (total of all registries)

- N_{act}^R = Population in time t, age group a, local area in country c (population not covered by registries)

- Y_{act}^R = Incident reported cases in time t, age group a, local area in country c in the remainder population (not covered by registries)

- Z_{act}^R = Reported deaths (Mortality) in time t, age group a, local area in country c in the remainder population (not covered by registries)

- $Y_{act} = Y_{act}^L + Y_{act}^R$, so that Y_{act} is all reported cases in time t, age group a, country c

- $Z_{act} = Z^L_{act} + Z^R_{act}$, so that Z_{act} is all reported deaths in time t, age group a, country c
- $N_{act} = N^L_{act} + N^R_{act}$ where N_{act} is total population in time t, age group a, country c
- $p_{act} = P(\text{Reported incidence } |t, a, c)$
- $q_{act} = P(\text{Reported mortality } |t, a, c)$
- $r_{act} = \Pr(\text{Reported mortality} \mid \text{Reported incidence}, a, c, t)$ (MI ratio)

5.3 European Breast Cancer Data

The quality scores for cancer incidence and mortality data established by IARC are shown in Tables 5.1 and 5.2. Incidence scores are based on the availability and coverage of registries within the country. The number of registries in each country and the coverage of any given registry varies quite a bit within and between countries. Additionally, some countries only provide national or regional rates and it is quite difficult to asses the reliability of these rates. Figure 5.1 displays the incidence scores by country. Twenty-nine of the countries have high quality incidence data, two have low coverage registries, 4 have regional or national rates, and six have no incidence data. Mortality scores are based on the availability and quality of vital registration. Figure 5.2 displays the mortality scores by country. Thirty-three of the countries have have high or medium quality data, seven have low quality, and one country has no data.

The current implementation of the methods described in this chapter rely on an aggregated version of the IARC scores and are limited to countries within Europe. We will refer to countries as having one of four types of data; (I) national incidence and mortality, (II) sub-national incidence and mortality (from registries) and national mortality, (III) only national mortality, and (IV) no available data. The relationship between the IARC scores and our categories are shown in Table 5.3. The most complete data type by country is shown in the top of Figure 5.3. The bottom of Figure 5.3 shows the available data by country

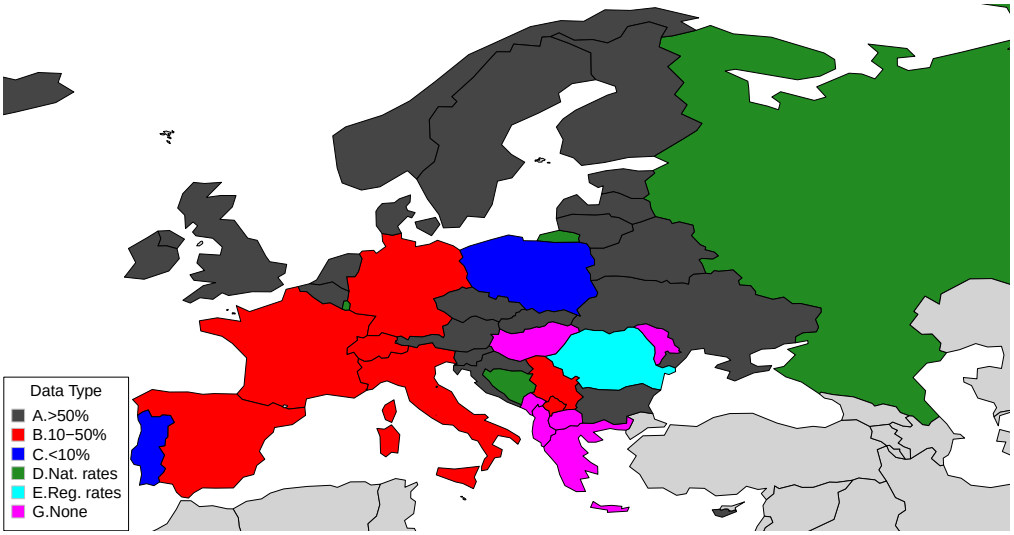

Figure 5.1: Quality of available incidence data as described in Table 5.1.

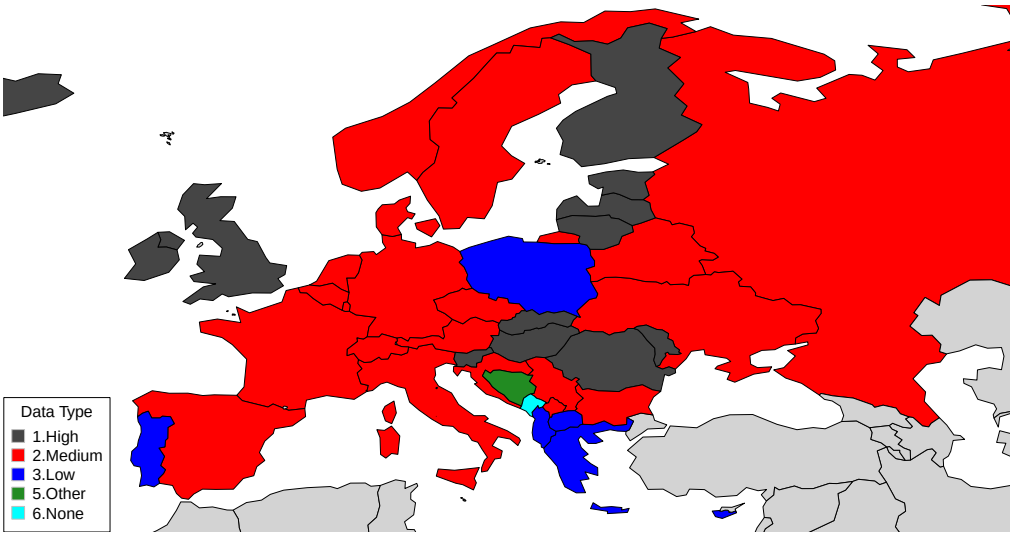

Figure 5.2: Quality of available mortality data as described in Table 5.2.

Table 5.1: Source and quality of incidence data from IARC.

Score	Data Quality & Source
A	High quality national or regional (coverage greater than 50%) data
B	High quality regional data (coverage between 10% and 50%)
C	High quality regional data (coverage lower than 10%)
D	National data (rates)
E	Regional Data (rates)
F	Frequency data
G	No data

Table 5.2: Source and quality of mortality data from IARC.

Score	Data Quality & Source
1	High quality complete vital registration
2	Medium quality complete vital registration
3	Low quality complete vital registration
4	Incomplete or sample vital registration
5	Other sources (cancer registries, verbal autopsy surveys, etc...)
6	No data

from 1990-2010. Complete details about the number of registries, years of incidence data, years of mortality data, and scores for the 40 European countries are in Table B.1 located in Appendix B.

To protect patient confidentiality we do not have information about individual registries beyond cases, deaths, and catchment population. For the purposes of analyses discussed in this chapter we have collapsed registry data by summing over all registries by year and age

group within each country. This is a slight deviation from the approach employed by IARC which uses a weighted average of registry rates scaled by the square root of the catchment population.

Table 5.3: Categorization for our methodology based on source and quality of mortality data from IARC.

	A	B	C	D	E	F	G
1	I	–	–	–	II	–	III
2	I	II	–	I/III	–	–	–
3	I	–	II	–	–	–	III
4	–	–	–	–	–	–	–
5	–	–	–	II	–	–	–
6	–	–	–	–	–	–	IV

5.4 Factors associated with breast cancer incidence and mortality rates

In our setting we are trying to estimate the number of cases that would appear in the health care system. This is not etiology, but a descriptive and surveillance-related effort. There are many factors that may influence the reported rates of breast cancer in each country. As shown in Section 5.3 the data quality and sources vary between countries. The location of registries and/or oncological centers may not be representative of the country as a whole which will introduce bias. Additionally we would expect screening (which is related to a better health care system and development) to increase incidence rates. Lastly, risk factors vary between countries which affects the true underlying rates. With country-level data and limited data on risk factors, it is difficult to disaggregate the impact of the many factors that may effect reported incidence and mortality rates (*e.g.* registry bias and differences in risk factors), but in this section we will briefly discuss some of the factors and their likely effects.

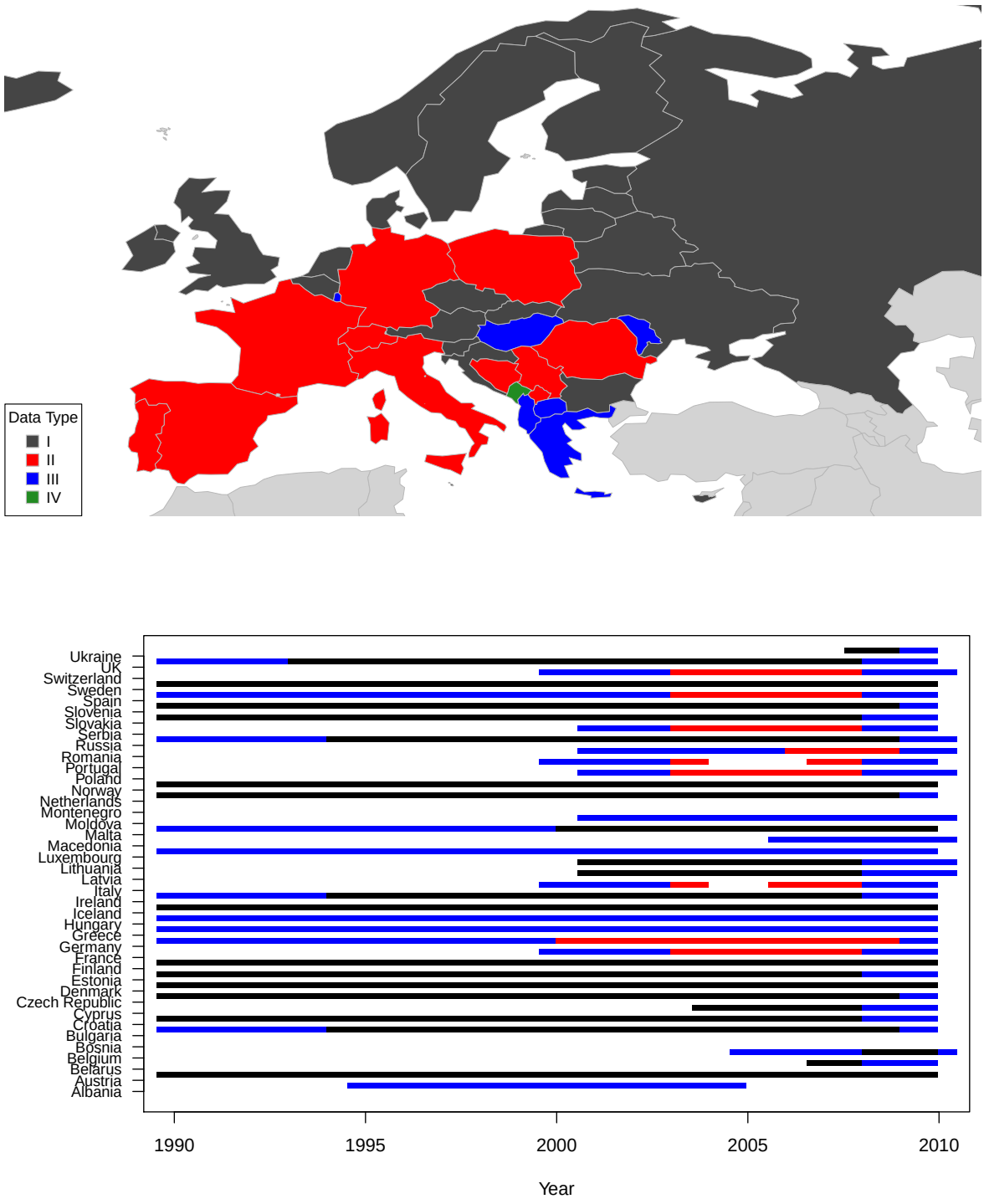

Figure 5.3: **Top**: Data type for most complete year of incidence and mortality data available in each country. Type I includes national incidence and mortality data. Type II has local incidence and mortality from registries as well as national mortality. Type III includes only national mortality. Type IV has no available data. **Bottom**: Data type by country from 1990-2010. Type IV (Montenegro) is blank.

There are many potential sources of bias in both the registry and national cancer data. Systematic underreporting of incident cases is likely if screening is difficult to access. Poor access to screening can be associated with higher than expected mortality, because cases are identified at more advanced stages, and often access to treatment options are likely limited in low-screening settings. This could lead to higher than expected mortality rates given the reported incidence. Registries may also suffer from the fact that they do not represent a random sample from the population of interest. One could imagine that registries are established primarily in urban centers or in areas with known risk factors. In either setting, registries may not be representative of the country as a whole.

Another important consideration are the risk factors in each country, as these could increase or decrease the true rates. CDC (2016) reports the primary risk factors for breast cancer as:

- age (getting older),
- genetic mutations (such as BRCA1 and BRCA2),
- early menstrual period,
- later or no pregnancy, in particular first pregnancy after age 30 or no full-term pregnancies,
- menopause after age 55,
- not being physically active,
- overweight/obese after menopause,
- breast density,
- use of combination hormone therapy,

- oral contraceptives,
- personal history of breast cancer or non-cancer breast diseases,
- family history of breast cancer,
- previous radiation therapy treatment,
- exposure to diethylstilbestrol (DES), and
- drinking alcohol.

Age is straight forward to incorporate in a modeling strategy and will be discussed in later sections, but most of the other risk factors are difficult to find reliable data on at a country level over a long time period. The notable exception being childbearing age and the total fertility rate (TFR), which is the expected number of children for a woman experiencing the age-specific fertility rates over a given time period. Estimates of the mean age of childbearing and TFR from 1950-2015, by five-year time periods is available via the World Population Prospects 2015 Revision (United Nations and Social Affairs, 2015). Figure 5.4 shows the trends by country and United Nations (UN) region (Figure 5.6) of mean childbearing age in Europe. A visually striking pattern is that in general the mean childbearing ages in Eastern Europe tend to be among the lowest ages over all time periods. Figure 5.5 shows the TFR over the same time period. In the time periods before 1985 we see the Western and Northern Europe TFRs tend to be among the lowest in Europe.

Earlier births and at least one full-term pregnancy are reported to be associated with lower rates of breast cancer (CDC, 2016), so these trends suggests that we might expect lower rates of breast cancer in Eastern Europe. To investigate the impact of TFR and mean childbearing age in Europe we fit a log-linear Quasipoisson regression to the incidence of women aged 55-59 years among countries with some national or sub-national incidence data from 1990-2010 (Type I and II from Figure 5.3) as predicted by TFR and mean childbearing

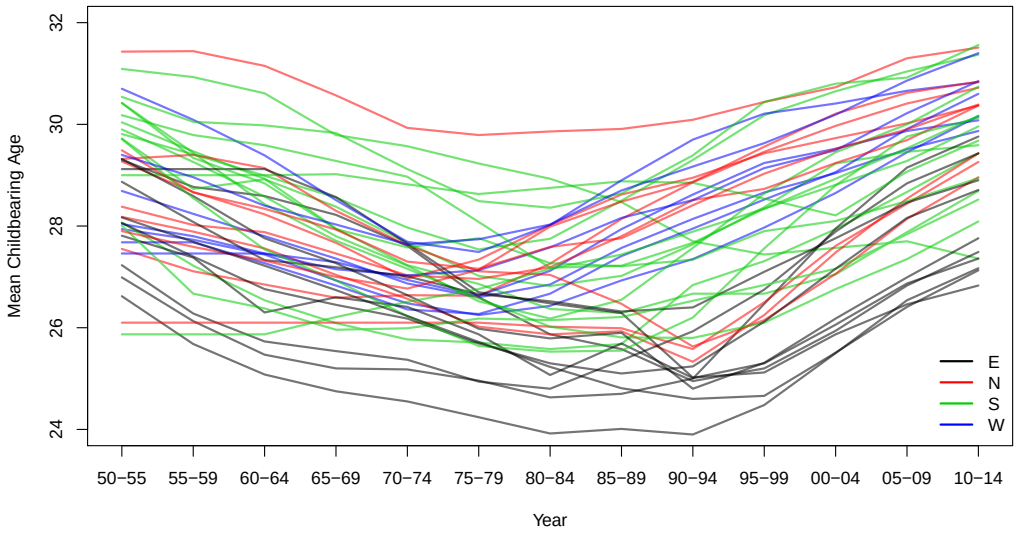

Figure 5.4: Country-specific trends in mean childbearing age by Eastern (E), Northern (N), Southern (S), and Western (W) regions of Europe as shown in Figure 5.6.

age when they were aged 30-34 years. Figure 5.7 shows scatter plots of the relevant data, with colors and symbols representing countries and color intensity increasing over time. We find a one year difference in mean childbearing age to be associated with 1.20 (1.15,1.24) times higher rate of breast cancer and a one child difference in TFR to be associated with 0.74 (0.65, 0.84) times lower rate of breast cancer adjusting for year and TFR or mean childbearing age, respectively. These are ecological associations, *i.e.* they are estimating the association between country-wide reported incidence rate and country-wide TFR and mean childbearing age, and so are not individual-level, but the associations are in the expected direction.

Russia is of particular interest as it has the largest population in Europe and some of the lowest breast cancer rates, with relatively high mortality rates. We know that Russia is likely to suffer from underreporting relative to the true incidence rate due to low screening rates, but

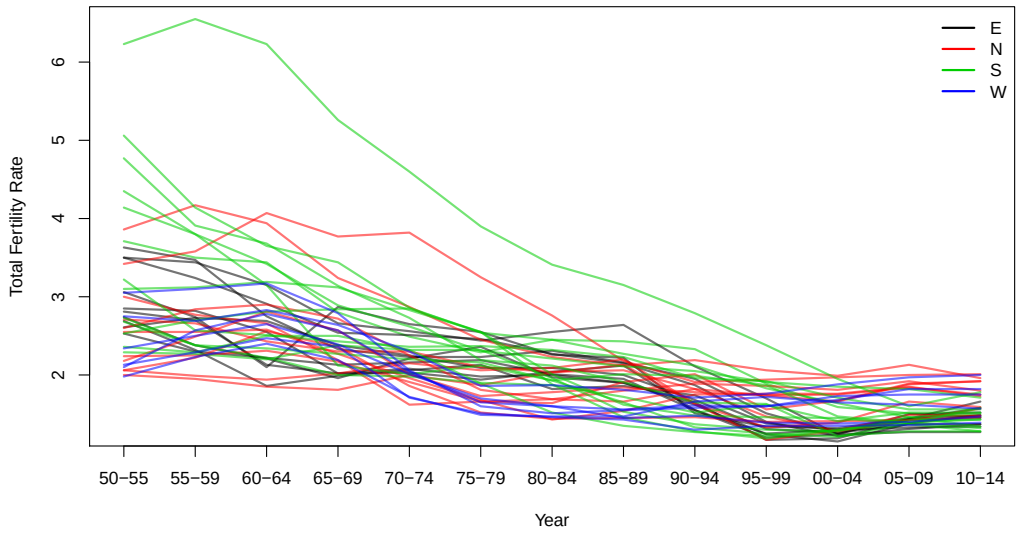

Figure 5.5: Country-specific trends in total fertility rate (TFR) by Eastern (E), Northern (N), Southern (S), and Western (W) regions of Europe as shown in Figure 5.6.

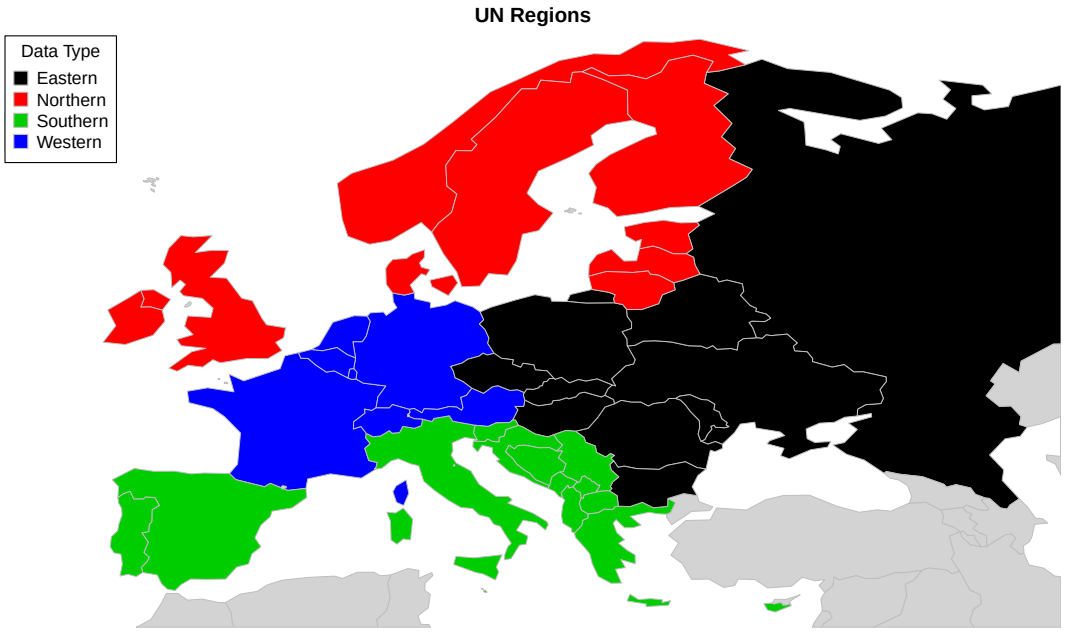

Figure 5.6: United Nations regions of Europe.

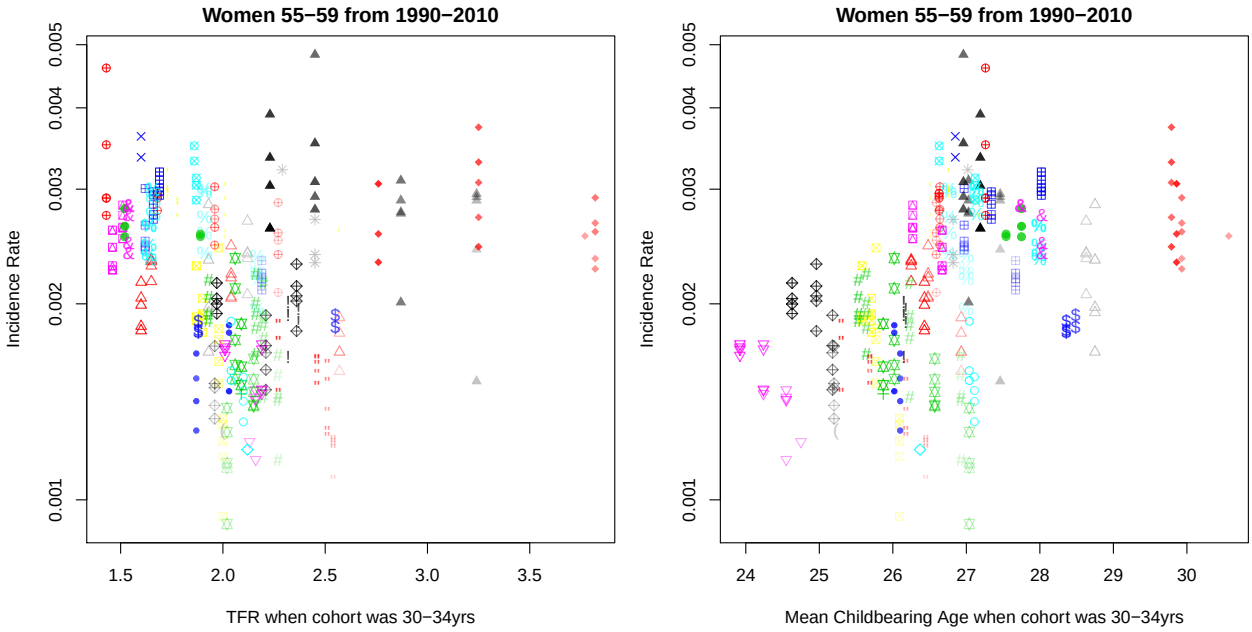

Figure 5.7: Among women aged 55-59 years from 1990-2010 a comparison of TFR (left) and mean childbearing age (right) when the cohort was age 30-34 years to incidence rates. Symbols and colors represent countries and color intensity increases over time.

may also have a legitimately lower breast cancer incidence rate. Figure 5.8 displays the mean childbearing age and total fertility rates from 1980-84, a relevant time period for the current high risk age groups. Russia, and the rest of Eastern Europe clearly have the lowest ages at first birth, suggesting lower breast cancer risk, everything else being equal. Additionally Russia has relatively high TFR when compared to Northern and Western Europe, which again would be associated with lower breast cancer risk.

Figure 5.9 shows the age-specific incidence rates (age 40 years and up) from the St. Petersburg and Munich cancer registries from 2003–2007. We can see that Munich consistently has higher age-specific incidence rates over this period. Munich and St. Petersburg are considered to be quality registries, thus this suggests a lower incidence rate in urban Russia compared with urban Germany. However, Figure 5.10 compares the reported national incidence rates to the St. Petersburg registry rates. In this figure we see that the national rates are consistently lower than the registry rates. This suggests that either the rest of Russia has lower rates of breast cancer incidence, suffers from significant underreporting, or quite likely both.

The true denominator for mortality given incidence would be true incidence, not reported incidence and since we will be using the relationship between reported incidence and mortality in our models (through the mortality-incidence ratio) it is also worthwhile to look at the associations between reported incidence and mortality rates both within and between countries. The over all correlation between observed log reported incidence and log mortality rates is 0.81. However we see some interesting patterns between ages as shown in Figure 5.11, where the strongest associations are see in the very young and older age groups. The strength of correlation also varies between countries as shown in Figure 5.12. The correlations tend to be fairly high in all of Europe, but highest in the East.

An example comparison between East and Northern Europe is shown in Figure 5.13 which shows much higher incidence rates in Norway than Russia. Mortality rates are more similar than incidence rates between the two countries in the younger age groups, but mortality appears to be slightly lower than Norway for some of the older age groups. Correlations are

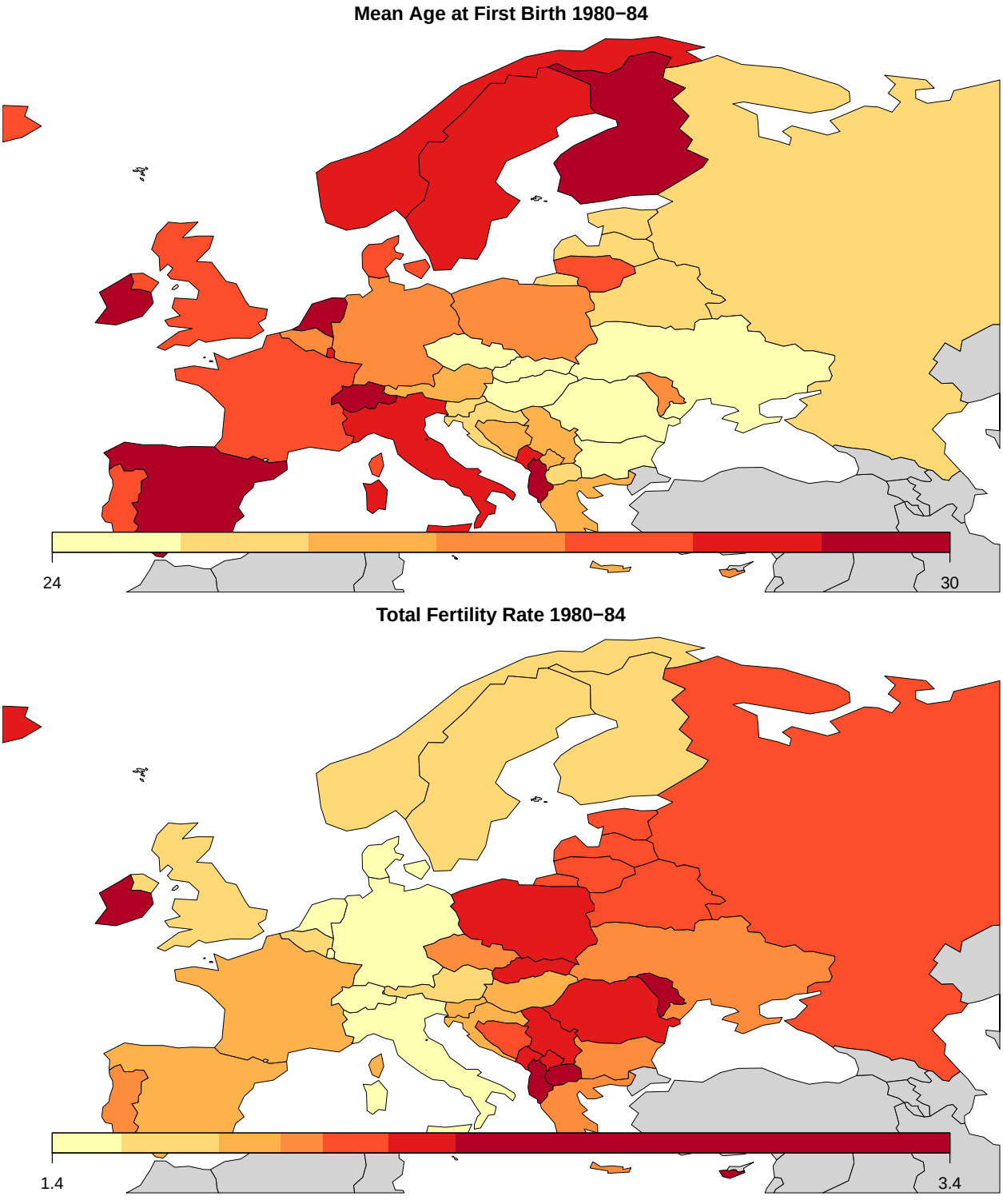

Figure 5.8: Mean childbearing age (top) and total fertility rate (bottom) for countries in Europe from 1980-84.

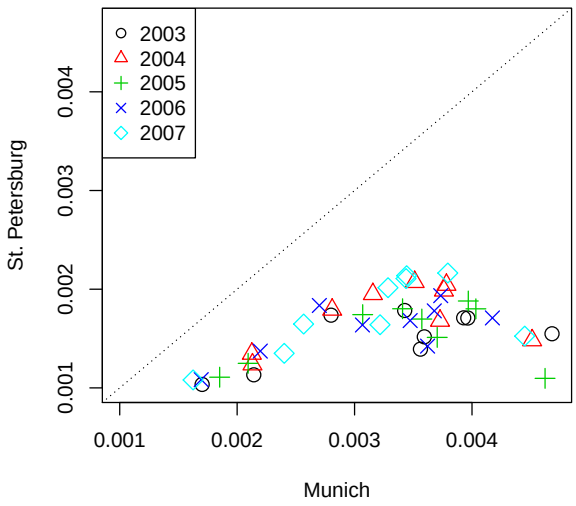

Figure 5.9: Age-specific breast cancer incidence rates from the St. Petersburg and Munich cancer registries from 2003–2007.

0.96 and 0.99 for Norway and Russia, respectively.

We can imagine at least two reasonable narratives to describe differences in incidence and similarities in mortality rates in Europe. The first is that in Eastern Europe incidence rates are lower due to differences in risk factors, but access to treatment is poor, so we see higher mortality given incidence rates than in Northern and Western Europe. A second narrative is that due to poor access to screening, the true incidence rates are much higher than reported incidence rates in Eastern Europe and true mortality given incidence rates could be similar to other parts of Europe, but appear high due to to the underreporting bias in the incidence rates. Based on these descriptive analyses it would seem that the reported rates are the result of a combination of these factors.

In summary, there are many factors related to the reported rates of breast cancer. Based on the exploratory analyses discussed in this section we conclude that reported incidence rates are likely impacted by both true differences in behavioral risk factors as well as underreport-

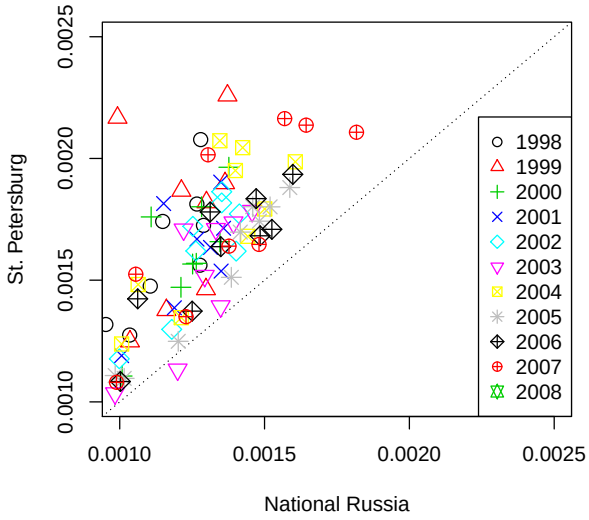

Figure 5.10: Age-specific breast cancer incidence rates reported national for Russia and from the St. Petersburg cancer registries from 1998–2007.

ing of cases due to limited screening access. Our goal is to provide short term projections of the reported number of cases, as these are directly related to health care utilization, relevant for planning and resource allocation, and easier to obtain given the reported data. Thus our focus in the remainder of this chapter will be to estimate the expected number of reported cases and deaths given the current reporting and utilization setting in each country.

5.5 IARC and IHME Methods

5.5.1 IARC Methods

The IARC methods of estimation are country- and cancer-specific and the estimation approach depends upon the amount and quality of the data available for each country by cancer. Generally, European estimates are generated and then followed by the larger effort of generating worldwide estimates (GLOBOCAN). The European 2012 estimates have

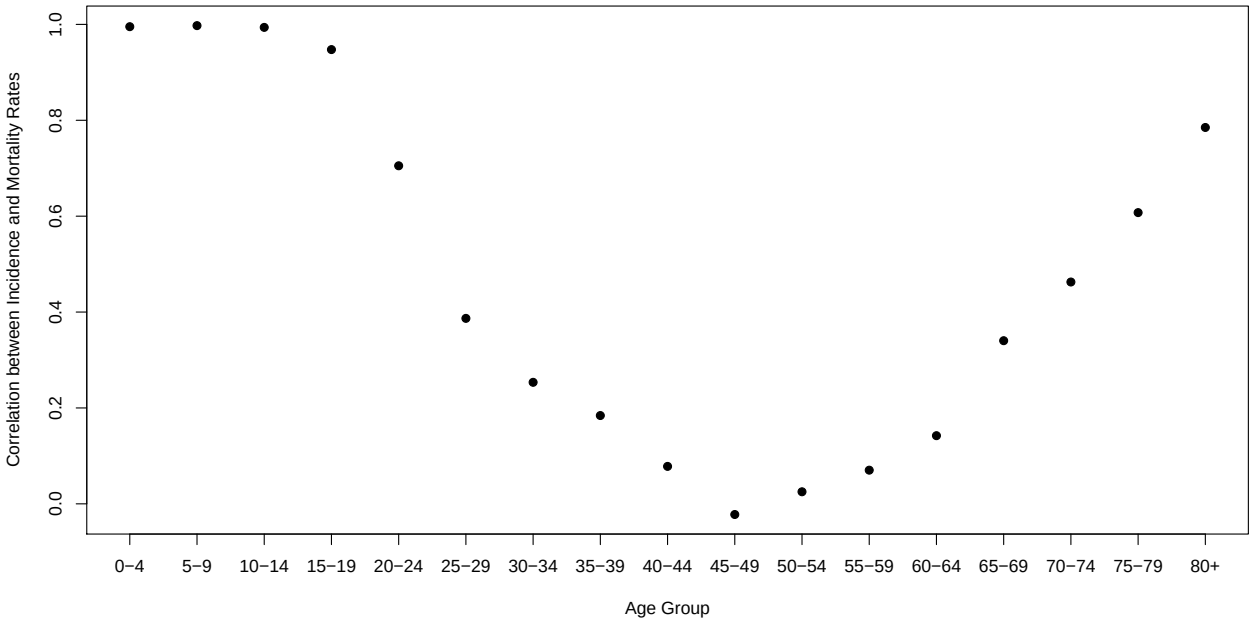

Figure 5.11: Correlation between log reported incidence and log mortality rate by age over all reported years.

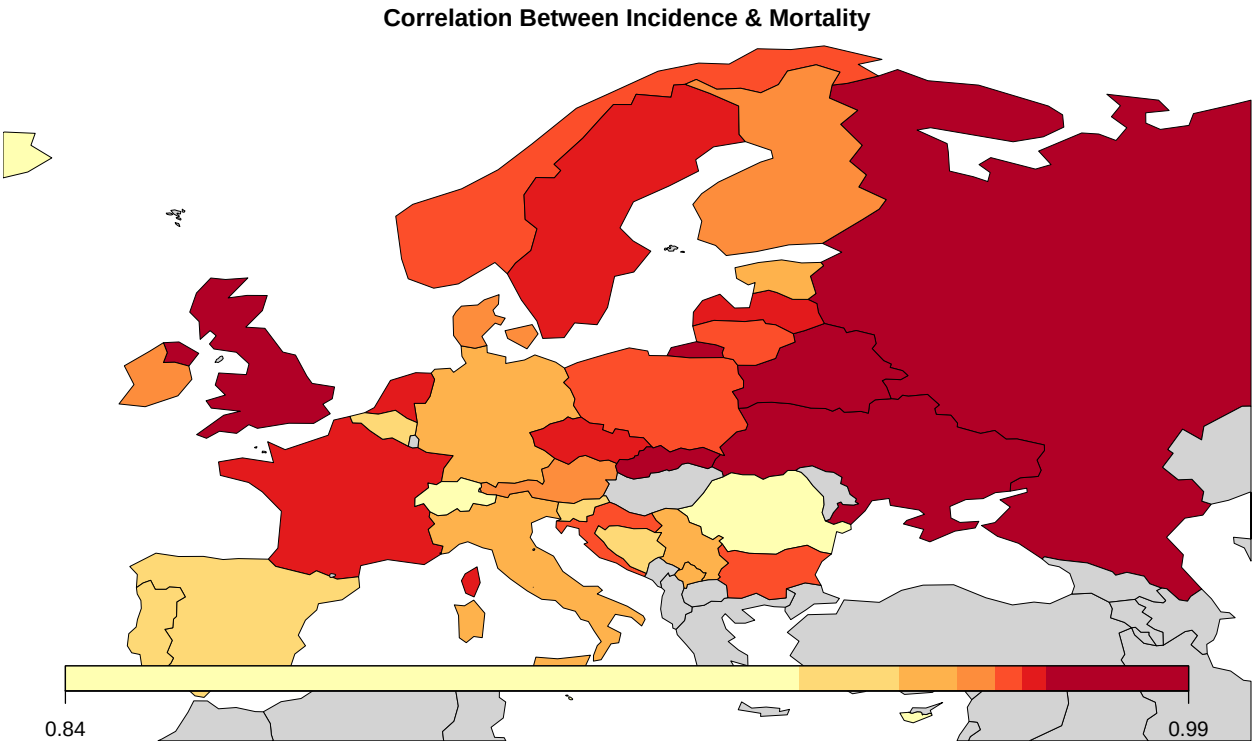

Figure 5.12: Correlation between log reported incidence and log mortality rate over all reported years. Areas in grey did not have available incidence data (Type III and IV countries).

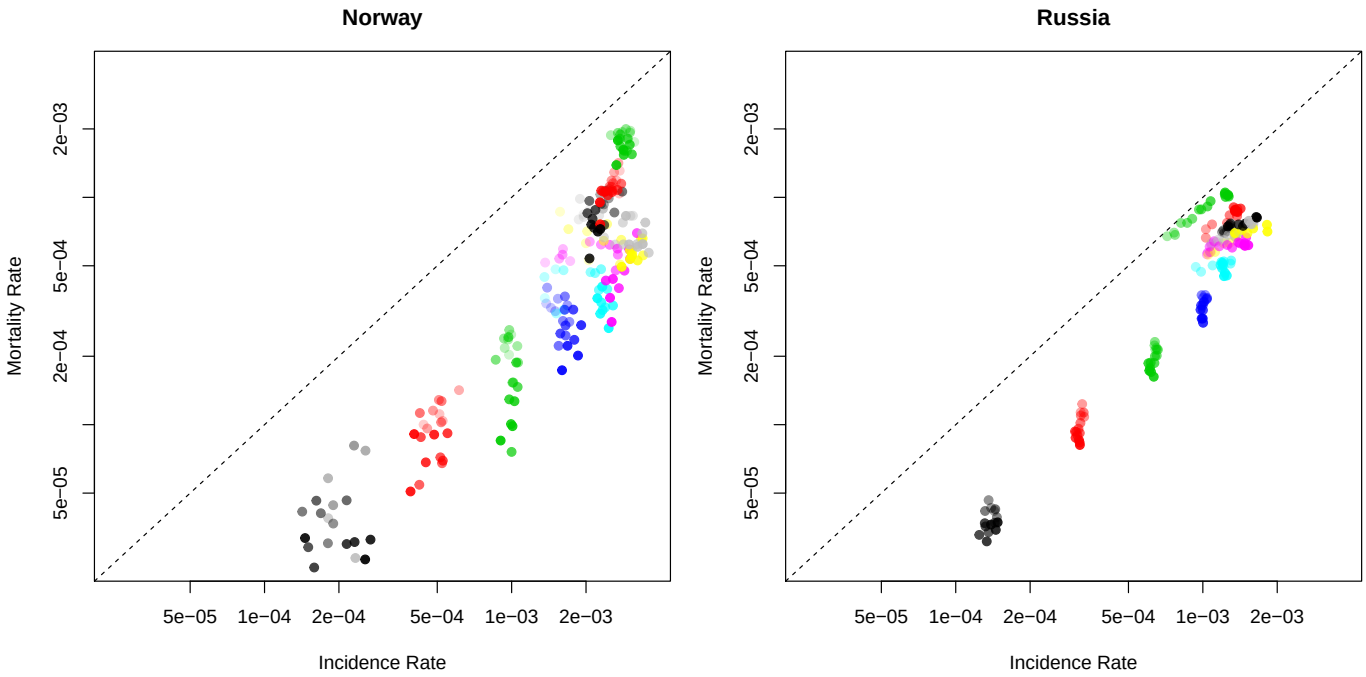

Figure 5.13: Age-specific (colors) rates of incidence and mortality over all reported years (intensity) in Norway (left) and Russia (right).

been published (Ferlay et al., 2013) and the statistical methods were discussed thoroughly in GLOBOCAN 2008 (Ferlay et al., 2010).

The estimation and projection of country-specific mortality rates described by Ferlay et al. (2010) and applied by Ferlay et al. (2013) can be divided into three categories based on available data. The NORDPRED method, an age-period-cohort model described in Møller et al. (2003), was used to generate mortality estimates in the 21 countries with complete national data and at least 15 years of historical data. In the 13 countries with less than 15 years of historical data the DEPPRED method, a linear time prediction model developed by IARC and based on Dyba and Hakulinen (2000), was used to generate mortality estimates. In the 6 remaining countries with less than 5 years of historical data, all of the data from the most recent 5-year period were used as a proxy for disease rates in 2012. Estimates for Montenegro, which had no available data, were based on averages of Bosnia Herzegovina and Serbia.

The approach for estimating country-specific incidence rates for countries with national incidence data are estimated in a similar fashion, with 19 countries using the NORDPRED method, two countries using the DEPPRED method, and 7 countries used available national incidence data from the previous 5-year period as a proxy for disease rates in 2012. The national incidence estimates for the remaining 11 countries rely on national mortality data as well as sub-national incidence and mortality from local registries within country or from a neighboring country. In these 11 countries the age-specific incidence–mortality (IM) ratio (Y^L_{ac}/Z^L_{ac}) was modeled from sub-national registry data and applied to the national mortality data (Z_{ac})

$$\widehat{Y_{ac}} = Z_{ac} \times Y^L_{ac}/Z^L_{ac}$$

to generate estimates of the national incidence (Y_{ac}) in 2012. Ferlay et al. (2013) provided summarization via total estimated cases and deaths as well as the European age standardized rates. The incidence ASR for country c in year t would be

$$ASR_{ct} = \sum_{a=A}^{A} p_{act} \times N^s_a \tag{5.1}$$

where $A = 17$ for the seventeen 5-year age groups (0-5,...,75-79, 80+) and N_a^s are the European age standard populations are shown in Figure 5.14.

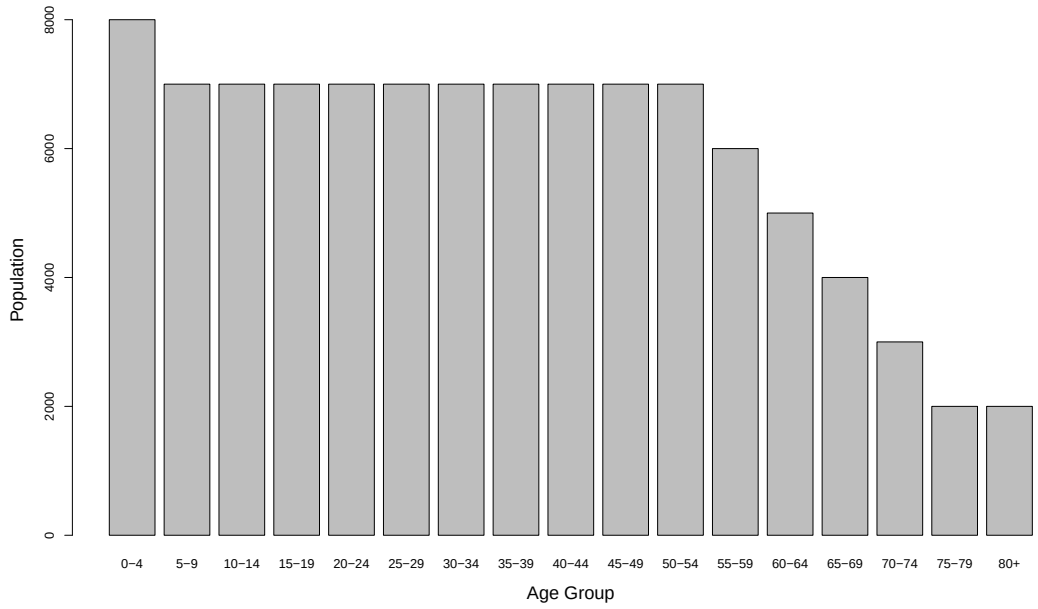

Figure 5.14: European age standardized population.

5.5.2 IHME Methods

The article 'Breast and cervical cancer in 187 countries between 1980 and 2010: a systematic analysis', (Forouzanfar et al., 2011) and the associated 73 page web appendix detail the methods and results of IHME's efforts to generate reliable estimates of worldwide breast and cervical cancer incidence and mortality. The following section provides an overview of the estimation procedure as it is described in the main text and web appendix.

MI Ratio Modeling

The initial step is to model the MI ratio. This step only uses regions which have high quality incidence and mortality data, though some of the MI ratios may be unavailable by age.

For such regions, estimates of the MI ratio may be imputed when mortality data are not available by age. The method assumes that the population proportions by age in countries with available data is the same as those without the available mortality data. These ratios are obtained by averaging over all countries and registry years, using only reliable data. At the end of this stage the MI ratio is available by year, age, and country for each of breast and cervical cancer, in those countries with both incidence and mortality deemed reliable.

For the data discussed in the previous step, the linear mixed model

$$\log\left(\frac{\widehat{r}_{act}}{1-\widehat{r}_{act}}\right) = \text{logit}\ \widehat{r}_{act} = \beta_0 + \beta_1 \text{GDP}_{c,t} + \beta_a + \gamma_c + \theta_t + \epsilon_{act} \tag{5.2}$$

is fit, where $\widehat{r}_{act}$ is the estimated MI ratio from the previous step in country c, for age group a, and year t. The age effects β_a as well as the coefficient of the Gross Domestic Product (GDP) β_1 are treated as fixed effects and γ_c, θ_t are independent random effects for registry and time respectively, and ϵ_{act} represents measurement error.

The rest of the analysis follows fairly closely with the methods previously used by IHME for different outcomes: childhood mortality in 187 countries in 1970 to 2010 (Rajaratnam et al., 2010a; Wang et al., 2014); worldwide mortality in men and women from 1970 to 2010 (Rajaratnam et al., 2010b); maternal mortality for 181 countries in 1980–2008 Hogan et al. (2010). A Gaussian Process Regression (GPR) is used to estimate a nonparametric unknown function in time. The description of the approach suggests the data consist of the logit of the MI ratio by age, country and time, as shown in (5.2), in countries where incidence and mortality are both available, but the mean of the GPR is taken as the estimated version of (5.2), so that effectively a GPR model is being applied to the residuals of the model. This model requires the selection of a covariance function, for which the Matérn is selected. Inference for the GPR is carried out via MCMC using the `pyMC` package (Patil et al., 2010).

Estimation of Trends in Mortality by Age

The next step develops a mortality model. The data are observed mortality (from vital registration, verbal autopsy, registration data) and mortality imputed from incidence data

using the MI model described above. This approach is based on the approach used by Hogan et al. (2010).

The first step was to identify covariates that might be predictors of mortality. For example, alcohol, smoking, education, national income, and total fertility rate, might be considered for breast cancer. These variables are used at the aggregate country-level. Many forwards selection procedures are then carried out. In all, 4,095 models were compared for breast cancer and 225 models for cervical cancer. Subsequently, these models were reduced down to 185 and 28 models for breast and cervical cancer, respectively.

As with the use of GPR in the MI modeling, the mean function is taken as the fitted model, in this case, each of the covariate models. In addition to the comparison of the single models, a weighted combination of 'ensemble' models is also constructed. This is a standard approach with the most common approach being Bayesian model averaging in which case the weights are posterior probabilities on each of the models being true (Raftery et al., 1997). However, the weight function used in Forouzanfar et al. (2011) is non-standard and was described in Foreman et al. (2012).

Based on the ensemble mortality models described above and the MI ratio model shown in (5.2), mortality rates are produced by age and year for all countries. To obtain uncertainty estimates, 1000 draws from the posterior distribution of the predictions of the MI ratios and 1000 draws from the posterior distribution of the mortality distribution were used. Each draw can be converted to an incidence ratio, and intervals can be derived from the distribution of these ratios. The incidence and mortality results are summarized by country as cumulative rates. The cumulative rate ϕ_c, was calculated for women ages 15-80 as

$$\phi_c = 1 - \exp\left(-\sum_{a=1}^{13} 5 \times p_{ac}\right) \tag{5.3}$$

for the thirteen 5-year age groups from 15 to 80 years.

5.5.3 Discussion

Ferlay et al. (2013) relies on an intimate knowledge of the data sources and quality that are

used in the estimates. They have developed a systematic estimation procedure based on the amount and perceived quality of the data. Ferlay et al. (2013) provides details of the modeling strategy for each country. These methods are individually quite straightforward, suggesting it would be possible to reproduce results with access to the same data. A shortcoming of their approach is that no uncertainty intervals are generated for estimates or projections.

The methods described in Forouzanfar et al. (2011) contain many complex steps. Forouzanfar et al. (2011) uses many data sources for the incidence and mortality data and incorporates many country-level covariates, which may add precision to the estimates. Uncertainty intervals are generated based on simulated draws from a posterior distribution. An extensive appendix is provided with details related to the modeling procedure, however there are many complex steps and it is unlikely that the results would be externally reproducible even with access to the same data sources.

5.6 Joint Model of Incidence and MI Ratio

The approach we describe relies on probabilistic models for incidence, mortality, and mortality given incidence. For most countries an alternative would be to rely only on unconditional models for just incidence and mortality. The MI modeling approach facilitates estimating national incidence in countries without out national or without local incidence data by providing an explicit link between mortality and incidence.

5.6.1 Countries with national mortality and incidence

We first consider models for those 23 countries with both national mortality and incidence data. For simplicity, we start by ignoring time and age. Our approach assumes the following two conceptual models for incidence and mortality given incidence, respectively:

$$Y_c|N_c, p_c \sim \text{Poisson}(N_c p_c), \qquad p_c = \exp(\alpha_c^I) \tag{5.4}$$

$$Z_c|Y_c, r_c \sim \text{Binomial}(Y_c, r_c), \qquad r_c = \frac{\exp(\alpha_c^{MI})}{1+\exp(\alpha_c^{MI})}. \tag{5.5}$$

The log linear model in (5.4) acknowledges the rare disease assumption. This pair of models imply the (unconditional) mortality model

$$Z_c|N_c, q_c \sim \text{Poisson}(N_c q_c) \tag{5.6}$$

where $q_c = p_c \times r_c$. We model the parameters as,

$$\alpha_c^I = \alpha^I + b_c^I \tag{5.7}$$

$$\alpha_c^{MI} = \alpha^{MI} + b_c^{MI} \tag{5.8}$$

where the BYM model of Section 2.4.2, the Leroux model described in Section 2.4.3, and a bivariate normal prior were all considered for each of the b_c terms. Exploratory work based on the Leroux prior for incidence and MI model country random effects suggest strong correlation between b_c^I and b_c^{MI}. We account for this by assuming a bivariate model for country random effects

$$\begin{bmatrix} b_c^I \\ b_c^{MI} \end{bmatrix} \sim \mathcal{N}_2 \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \boldsymbol{\Sigma} \right)$$

with an Inverse Wishart prior assigned to the covariance matrix

$$\boldsymbol{\Sigma} \sim \text{Inverse Wishart}\,(df = 4, \boldsymbol{S})$$

where $\boldsymbol{S}$ was selected for an average correlation of 0 and 0.2 standard deviation.

The parameters α_I, α_{MI} and the hyperparameters associated with b_c^I, b_c^{MI} are shared across all countries, regardless of the available data. Let ω_1 denote the set of indices for countries with both national incidence and mortality data, and then let

- $\boldsymbol{y}^{(1)} = \{y_c : c \in \omega_1\}$ national incidence data in type I countries
- $\boldsymbol{z}^{(1)} = \{z_c : c \in \omega_1\}$ national mortality data in type I countries
- $\boldsymbol{b}^{(1)I} = \{V_c^I : c \in \omega_1\}$ country-specific random effects for incidence in type I countries
- $\boldsymbol{b}^{(1)MI} = \{V_c^{MI} : c \in \omega_1\}$ country-specific random effects for MI ratio in type I countries

The likelihood of the data from type I countries is

$$p(\boldsymbol{y}^{(1)}, \boldsymbol{z}^{(1)} | \alpha^I, \alpha^{MI}, \boldsymbol{b}^{(1)I}) = \underbrace{p(\boldsymbol{y}^{(1)} | \alpha^I, \boldsymbol{b}^{(1)I})}_{\prod_{c \in \omega_1} \text{Poisson}(N_c p_c)} \times \underbrace{p(\boldsymbol{z}^{(1)} | \boldsymbol{y}^{(1)}, \alpha^{MI}, \boldsymbol{b}^{(1)MI})}_{\prod_{c \in \omega_1} \text{Binomial}(Y_c, r_c)}.$$

5.6.2 Countries with sub-national mortality and incidence and national mortality

Next we consider a modeling approach for the 10 countries with local incidence and mortality data from registries and national mortality data. In the current implementation we collapse over all registries within each country. For the local data we assume

$$Y_c^L | N_c^L, p_c^L \sim \text{Poisson}(N_c^L p_c), \qquad p_c = \exp(\alpha_c^I) \tag{5.9}$$

$$Z_c^L | Y_c^L, r_c^L \sim \text{Binomial}(Y_c^L, r_c), \qquad r_c = \frac{\exp(\alpha_c^{MI})}{1 + \exp(\alpha_c^{MI})} \tag{5.10}$$

where the intercept parameters are assumed to be of the form (5.8) and (5.8).

We are interested in estimating national incidence in these countries with only local incidence data. Our approach relies on modeling the local incidence and mortality as well as the mortality that is not covered by registries. Let $Z_c^R = Z_c - Z_c^L$ and $N_c^R = N_c - N_c^L$. Our MI ratio assumption states that the registry and national MI ratios are equivalent ($r_c^L = r_c^R = r_c$). Thus we can consider the following model for the remainder mortality:

$$Z_c^R | N_c^R, q_c \sim \text{Poisson}(N_c^R q_c), \quad q_c = \underbrace{\exp(\alpha_c^I)}_{p_c} \times \underbrace{\frac{\exp(\alpha_c^{MI})}{1 + \exp(\alpha_c^{MI})}}_{r_c}. \tag{5.11}$$

We are effectively modeling a single count (the remainder mortality) using two parameters (p_c and r_c) with both of these parameters being influenced by the observed local incidence and mortality data. Naturally, when N^L/N is close to one the parameters will be more heavily influenced by the registry data and when N^L/N is small they will be more heavily influenced by the national mortality data. Let ω_2 denote the set of indices for countries with local incidence and mortality data from registries and national mortality data, and then let

- $\boldsymbol{y}^{(2)} = \{y_c : c \in \omega_2\}$ local incidence data in type II countries

- $\boldsymbol{z}^{(2)L} = \{z_c^L : c \in \omega_2\}$ local mortality data in type II countries
- $\boldsymbol{z}^{(2)R} = \{z_c^R : c \in \omega_2\}$ local mortality data subtracted from national mortality data
- $\boldsymbol{b}^{(2)I} = \{b_c^I : c \in \omega_2\}$ country-specific random effects for incidence in type II countries
- $\boldsymbol{b}^{(2)MI} = \{b_c^{MI} : c \in \omega_2\}$ country-specific random effects for MI in type II countries.

The likelihood of the data from type II countries is

$$p(\boldsymbol{y}^{(2)}, \boldsymbol{z}^{(2)} | \alpha^I, \alpha^{MI}, \boldsymbol{b}^{(2)I}, \boldsymbol{b}^{(2)MI}) = \underbrace{p(\boldsymbol{y}^{(2)} | \alpha^I, \boldsymbol{b}^I)}_{\prod_{c\in\omega_2} \text{Poisson}(N_c^L p_c)} \times \underbrace{p(\boldsymbol{z}^{(2)L} | \boldsymbol{y}^{(2)}, \alpha^{MI}, \boldsymbol{b}^{MI})}_{\prod_{c\in\omega_2} \text{Binomial}(Y_c^L, r_c)} \times \underbrace{p(\boldsymbol{z}^{(2)R} | \alpha^I, \boldsymbol{b}^{RI}, \alpha^{MI})}_{\prod_{c\in\omega_2} \text{Poisson}(N_c^R q_c)}.$$

5.6.3 Countries with national mortality only

There are 6 countries with only national mortality data. In this case we fit the national mortality model (5.6) and then obtain national incidence estimates using imputed r_c estimates. We learn about the national incidence rate p_c by modeling the national mortality rate q_c via the model

$$q_c = p_c \times r_c = \exp(\alpha_c^I) \times \frac{\exp(\alpha_c^{MI})}{1 + \exp(\alpha_c^{MI})} \tag{5.12}$$

Let ω_3 denote the set of indices for countries with national mortality data and no incidence data, and then let

- $\boldsymbol{z}^{(3)} = \{z_c : c \in \omega_3\}$ national mortality data in type III countries
- $\boldsymbol{b}^{(3)I} = \{V_c^I : c \in \omega_3\}$ country-specific random effects for incidence in type III countries
- $\boldsymbol{b}^{(3)MI} = \{V_c^{MI} : c \in \omega_3\}$ country-specific random effects for MI in type III countries

The likelihood of the data for these models is

$$p(\boldsymbol{z}^{(3)} | \alpha^I, \alpha^{MI}, \boldsymbol{b}^{(3)I}, \boldsymbol{b}^{(3)MI}) = \prod_{c\in\omega_3} \text{Poisson}(N_c q_c)$$

from (5.12).

5.6.4 Countries with No Data

During time periods where countries have no data, shown as the blank line in Figure 5.3, spatial effects will be generated from the prior. In these countries we will use

$$p_c = \exp\left(\alpha^I + b_c^{I\star}\right)$$

where $b_c^{I\star}$ will be influenced by neighboring regions and the rest of Europe through the spatial priors. Similarly, we will impute

$$r_c = \frac{\exp\left(\alpha^{MI} + b_c^{MI\star}\right)}{1 + \exp\left(\alpha^{MI} + b_c^{MI\star}\right)}$$

where $b_c^{MI\star}$ is simulated from the prior. Let ω_4 denote the set of indices for countries with no available data, and then let

- $\boldsymbol{b}^{(4)I} = \{V_c^{I\star} : c \in \omega_4\}$ country-specific random effect for incidence in type IV countries
- $\boldsymbol{b}^{(4)MI} = \{V_c^{MI\star} : c \in \omega_4\}$ country-specific random effect for MI in type IV countries

5.7 Spatial Simulation

This section describes a brief simulation study designed to investigate our ability to recover true incidence and mortality rates with the spatial models for incidence and the MI ratio described in Section 5.6.

5.7.1 Generation of Spatially Correlated Data

Five hundred data sets were generated in such a way as to be similar to the observed data for women age 50–54 in Europe for a single year. Population sizes (N_c) were based on the average population over the years with available data from each country. Populations ranged from 7,595 (Iceland) to 4,932,000 (Russia). Parameters were selected to generate incidence rates (p_c) and MI ratios (r_c) with distributions similar to the average observed rates for each country. Incidence rates were generated with $p_c = \exp\left(\alpha^I + V_c^I + U_c^I\right)$ where

$\alpha^I = -6.5, V_c^I \sim N(0, \sigma = 0.5)$, and U_c^I were spatially correlated as described by Besag et al. (1991). Similarly, the MI ratios $r_c = \text{expit}\left(\alpha^{MI} + V_c^{MI} + U_c^{MI}\right)$ where $\alpha^I = -1, V_c^I \sim N(0, \sigma = 0.5)$, and U_c^{MI} were spatially correlated. Among the type II countries, registry coverage (N_c^L/N_c) varied between 5%–75%. Cases were generated from Poisson($N_c p_c$) and Poisson($N_c^L p_c$) for national and registry rates, respectively and deaths were generated from Binomial(Y_c, r_c) and Binomial(Y_c^L, r_c).

5.7.2 Coverage of Credible Intervals

The spatial methods described in Section 5.6 were implemented on the simulated data using a Bayesian approach with computation via MCMC. Figure 5.15 displays the posterior distributions of the predicted incidence and mortality counts (on the log scale) from a sinle simulated data set. As we would expect, type III and IV have the widest intervals among similar size counts and incidence is recovered particularly well in the type I and II countries. Over the 500 simulations, the true incidence was contained in 95.1% of 95% credible intervals. Among type II, III, and IV countries, which we would expect to be the most difficult to estimate, the true incidence values were contained in 89% of 95% credible intervals. The true mortality was recovered in all credible intervals. These results suggest that the joint incidence and MI modeling method is performing adequately.

5.8 Modeling Age and Time

For notational brevity the models described in Section 5.6.1-5.6.4 were purely spatial. However, the full model specification also includes components for time, age, space–age interactions, and space–time interactions. The primary models to be considered are shown in Table 5.4. Five-year age bands, denoted by $\gamma_a^I, \gamma_a^{MI}$, are given RW2 priors as described in Section 2.3.2. Country-specific age effects, are assigned multivariate Normal priors, $\delta_{ac}^I | \tau_{\delta^I} \sim N\left(\mathbf{0}, \tau_{\delta^I}^{-1} \boldsymbol{I}\right)$ and $\delta_{ac}^{MI} | \tau_{\delta^{MI}} \sim N\left(\mathbf{0}, \tau_{\delta^{MI}}^{-1} \boldsymbol{I}\right)$, where $\boldsymbol{I}$ is an appropriately sized identity matrix.

Rates are modeled on an annual time scale. The temporal model takes the same form

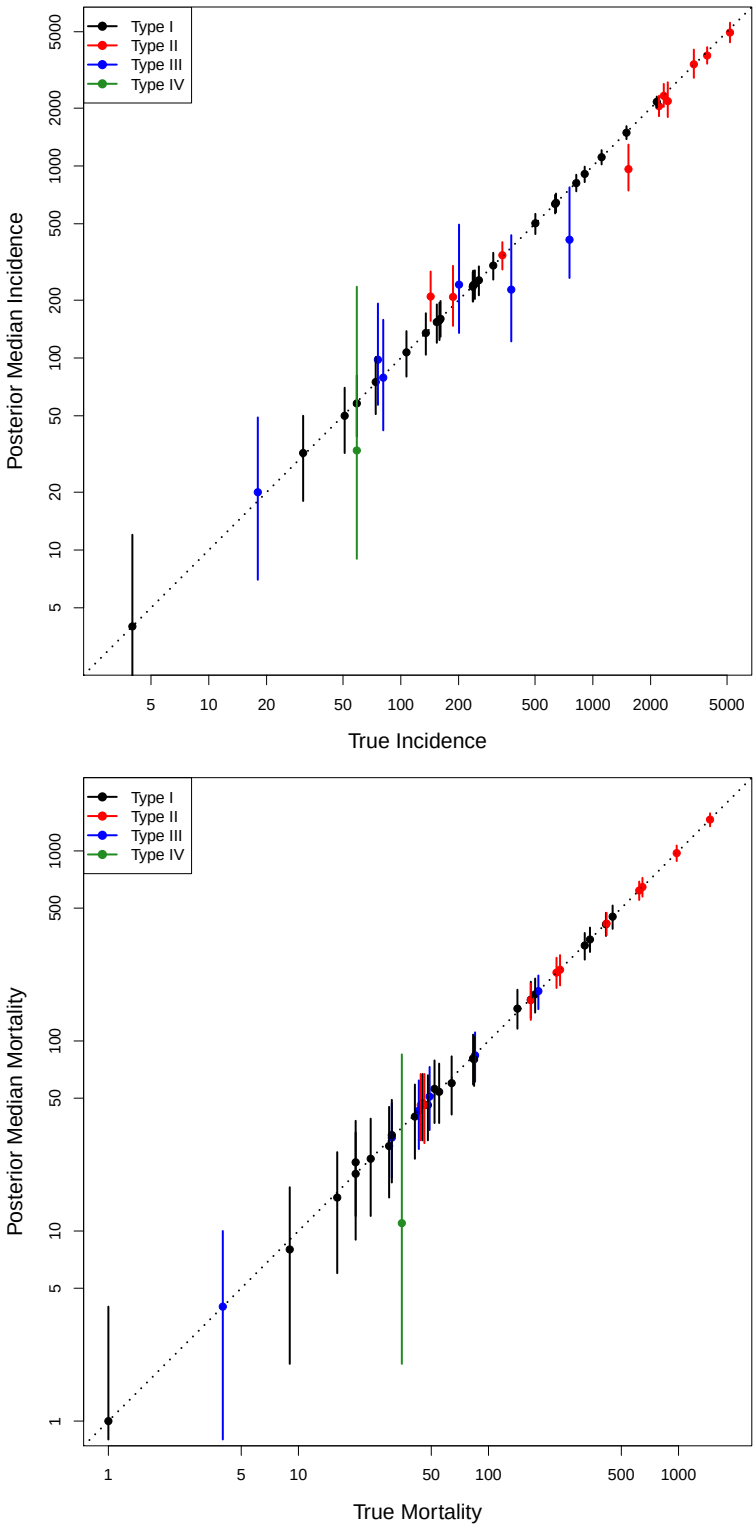

Figure 5.15: Posterior median and 95% credible intervals and observed incidence (top) and mortality (bottom) on the log scale for a single realization of simulated data.

in both the incidence and MI models, aside from Model I, which has no time trend. Model II from Table 5.4, considers an overall linear trend θ, which is assigned a diffuse Normal prior. Model III assumes country-specific linear trends, $\theta_c \sim N\left(\theta, \tau_\theta^{-1}\boldsymbol{I}\right)$, where θ is the overall trend and is assigned a diffuse Normal prior. Model IV considers country-specific linear splines. For knots $k = 1, ..., K$ at times $t_1, ..., t_K$

$$S_c(t) = \theta_{o,c}t + \sum_{k=1}^{K} \theta_{k,c}(t - t_k)_+$$

where

$$(t - t_k)_+ = \begin{cases} t - t_k & t > t_k \\ 0 & t \leq t_k \end{cases}$$

and $\theta_{k,c} \sim N\left(\mathbf{0}, \tau_{\theta_k}^{-1}\boldsymbol{I}\right)$ for $k = 0, ..., K$. For the European breast cancer application, data was available from 1990-2010, but many countries had no data until 2000. Up to 4 knots were considered, but we found that additional knots did not meaningfully change the results, so a single knot was used and placed at the year 2000.

Table 5.4: Random effects models for year t, country c and age a considered for the linear predictors of $\log(p_{act})$ and $\text{logit}(r_{act})$. In all models the intercept has been incorporated into the age effects γ. The remaining parameters are defined as follows, $[\boldsymbol{b}^I, \boldsymbol{b}^{MI}]^T \sim \mathcal{N}\left([0,0]^T, \boldsymbol{\Sigma}\right)$, $\gamma_a \sim RW2$, $\theta \sim$ flat, $\delta_{ac} \sim N(0, \tau_\delta^{-1}\boldsymbol{I})$, and $S_c(t)$ country-specific linear splines.

Model	Linear Predictor
I	$\gamma_a + b_c + \delta_{ac}$
II	$\gamma_a + b_c + \delta_{ac} + \theta t$
III	$\gamma_a + b_c + \delta_{ac} + \theta_c t$
IV	$\gamma_a + b_c + \delta_{ac} + S_c(t)$

5.9 Application to European Breast Cancer

The four models described in Table 5.4 were fit to the breast cancer incidence and mortality data from 40 European countries shown in Figure 5.3, from 1990–2009, with seventeen age groups, (0-4, 5-9,..., 75-80, 80+). Posterior samples were drawn using Stan via the `RStan` package (Homan and Gelman, 2014; Carpenter et al., 2015) in `R` as described in Section 2.5.3. For models I–IV, 2,500 draws from the posterior distribution, after the first 500 iterations were discarded as burn-in, which required approximately 10 hours of computing time. We considered summarization via European age standardized rate (ASR) for both incidence and mortality, which is described in Eq. (5.1) and is used by IARC in the European-specific publications (Ferlay et al., 2013). The remainder of this section describes the model selection procedure, summaries of the posterior distribution, estimated breast cancer incidence and mortality rates and a comparison with breast cancer results published by IARC (Ferlay et al., 2013) and IHME (Forouzanfar et al., 2011).

5.9.1 Model Selection

Models were evaluated using the mean square error (MSE) of the posterior mean and the observed incidence and mortality ASRs in 2010. The bias is calculated as the sum of differences between the observed ASR, ASR_c and the estimated ASR, $\widehat{ASR}_c$

$$\text{Bias} = \frac{1}{C}\sum_{c=1}^{C}\left(\widehat{ASR}_c - ASR_c\right)$$

where the $\widehat{ASR}_c = \frac{1}{S}\sum_{s=1}^{S}\widehat{ASR}_c^{(s)}$ and $s = 1, ..., S$ are the draws from the posterior distribution. The variance is calculated as

$$\text{Variance} = \frac{1}{C}\sum_{c=1}^{C}\left(\frac{1}{S-1}\sum_{s=1}^{S}\left(\widehat{ASR}_c^{(s)} - \widehat{ASR}_c\right)^2\right)$$

and the MSE=Bias^2 + Variance. In general, we would expect more complex models to have reduced bias, with potentially increased variance, if additional parameters do not contribute to the fit. Bias is more complex here, however, since we have two interacting outcomes.

Table 5.5 displays the bias, variance, and MSE of the ASRs for incidence and mortality. Model IV, which has greatest flexibility, results in the the lowest MSE for incidence and near the lowest MSE for mortality and was selected for analysis. The rest of this chapter focuses on the fit of model IV to the European breast cancer data from 1990–2010.

Table 5.5: MSE of posterior means from models described in Table 5.4.

	Incidence			Mortality		
Model	Bias^2	Variance	MSE	Bias^2	Variance	MSE
I	433.7	106.6	540.3	13.5	26.1	39.6
II	218.3	124.2	342.6	4.6	24.0	28.6
III	191.6	118.5	310.2	5.6	24.1	29.7
IV	190.6	116.6	307.2	6.8	22.9	29.7

5.9.2 Summarizing the Posterior Distribution

Figure 5.16 shows the RW2 age effects for the incidence model along with pointwise 95% credible intervals on the incidence rate scale. Incidence rates are extremely low before the age 25 years and increase steadily until ages 65–69 years at which point the rates essentially level off. Figure 5.17 shows the RW2 age effects and 95% credible interval for the MI model on the probability scale. Mortality given incidence decreases among the youngest ages and then increases after age 20–24. Finally, Figure 5.18 provides the mortality rates that result from the incidence and MI age random effects. Mortality rates are very low until age 30 years and increase steadily with age.

The posterior median of the spatial random effects $\boldsymbol{b}^I, \boldsymbol{b}^{MI}$ are shown in Figure 5.19. In general we see higher incidence rates in Western and Northern Europe and higher MI ratios in Eastern Europe. A similar pattern is observed when looking at the slopes $\theta^I, \boldsymbol{\theta}^{MI}$ in Figure 5.20. Though, the pattern looks somewhat different for the spline coefficients $\theta_1^I, \theta_1^{MI}$

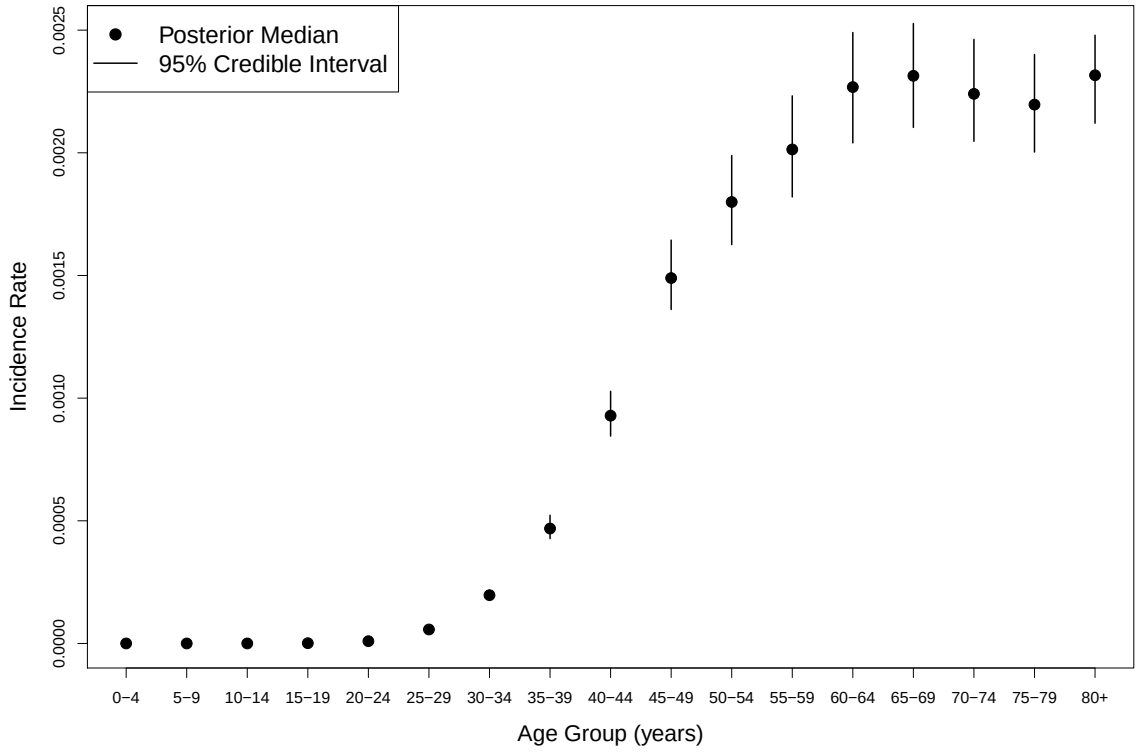

Figure 5.16: Posterior distribution of the random effects for age on the rate scale, $p_a = \exp\left(\gamma^I\right)$ from the Incidence model.

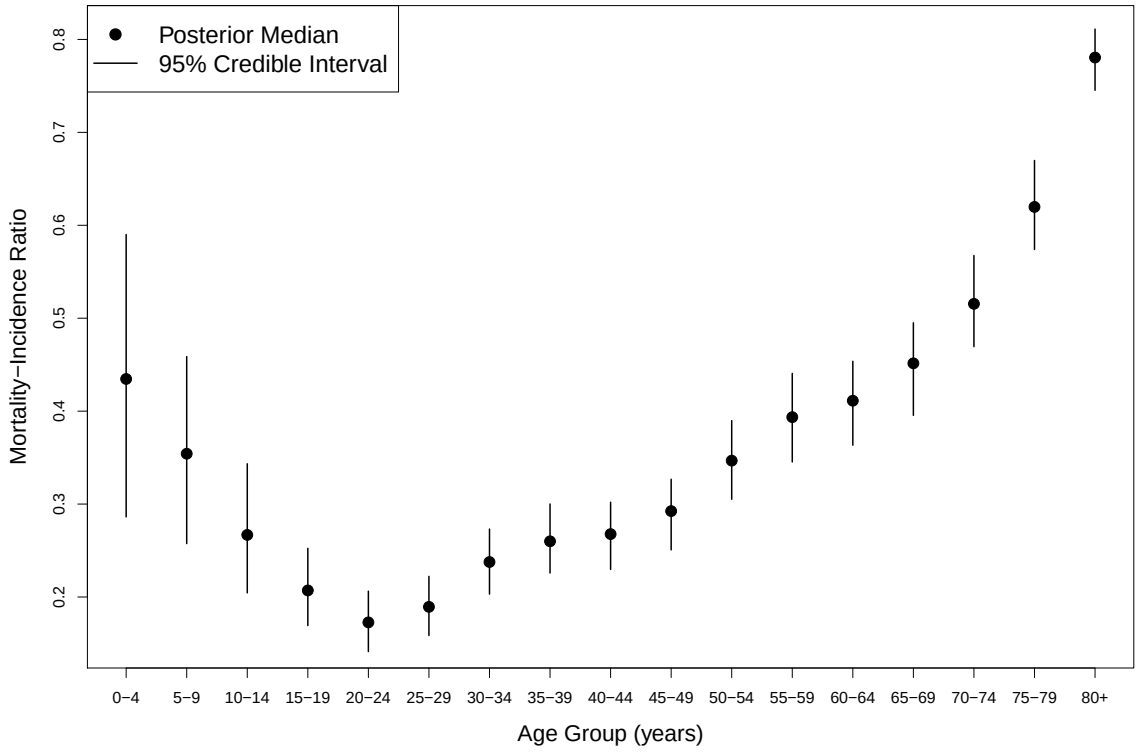

Figure 5.17: Posterior distribution of the random effects for age on the probability scale $r_a = \text{expit}\left(\gamma^{MI}\right)$ from the MI ratio model.

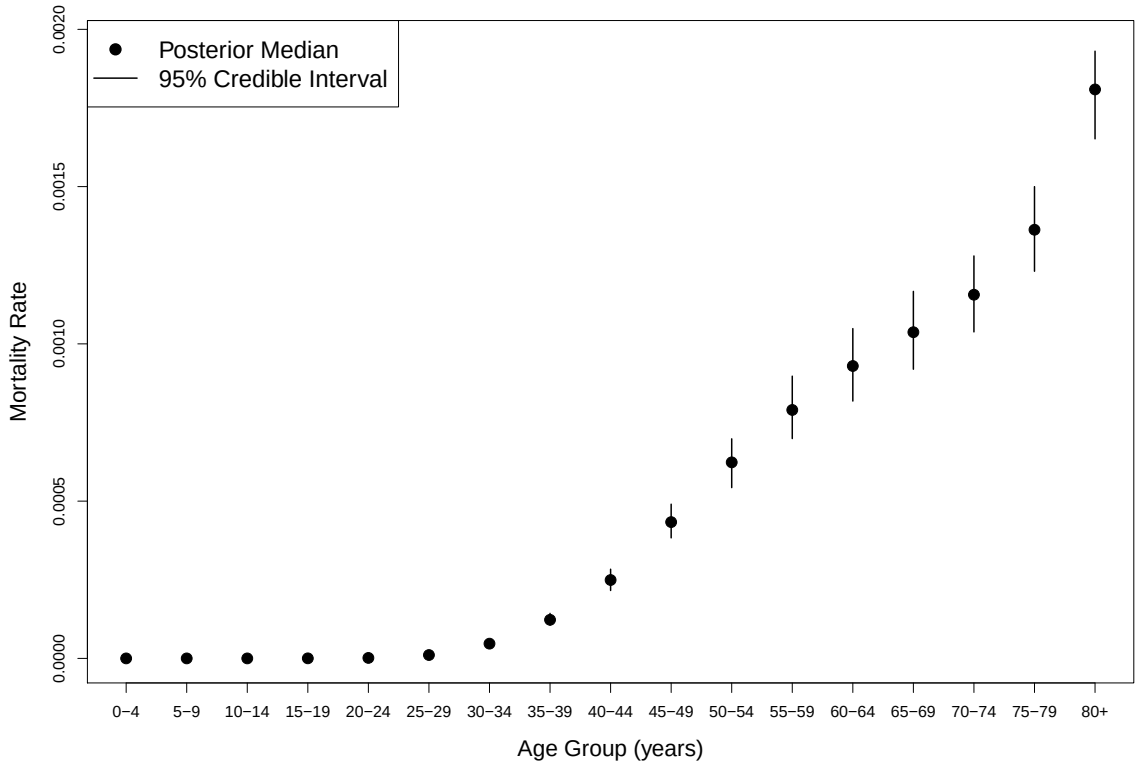

Figure 5.18: Posterior distribution of the age specific effects for mortality $q_a = \exp\left(\gamma^I\right) \times \text{expit}\left(\gamma^{MI}\right)$ from the incidence and MI ratio models.

shown in Figure 5.21, with higher incidence in Central Europe and higher MI ratio effects in Western Europe. Figure 5.22 shows the posterior median of the country random effects which have the Bivariate Normal prior. The strong negative correlation is clear from the figure is clear in the figure.

5.9.3 Validation of Projected European Breast Cancer

As an out of sample validation exercise the models were fit on the European breast cancer data from 1990–2008 and the coverage of observed ASRs in 2009 by the 95% posterior predictive intervals was assessed. The year 2009 was selected because there are fewer countries with national incidence and mortality data available in 2010. There are a total of 37 countries with national mortality data in 2009 and only 8 countries with national incidence. Sixty-three percent of incidence ASRs (5/8) and 89% (33/37) of observed mortality ASRs were contained in the 95% posterior predictive intervals.

Figure 5.23 shows the estimates and the posterior 95% predictive intervals for the observed rates (accounting for Poisson sampling) for the ASRs for each country when compared to the observed value in 2009. For both incidence and mortality the estimates for one country of each type are quite far from the observed values; these are Serbia for incidence and Denmark for mortality. In Figure 5.24 we can see that the surprising result in Denmark is due to higher than normal age specific mortality rates in 2009. The story is less clear in Serbia, which does not appear to be having an unusual year. The results of the validation suggest our predictive intervals should have good coverage properties, though may be a bit too narrow (anti-conservative).

5.9.4 Estimates and Projections of European Breast Cancer

Figures 5.25 and 5.26 display the fitted age-specific incidence and mortality rates with posterior predictive intervals for the Netherlands and Italy, respectively. The Netherlands provides an example of the estimated age-specific annual rates of breast cancer incidence and mortality of a Type I country and Italy provides and example of a Type II country. In both sets

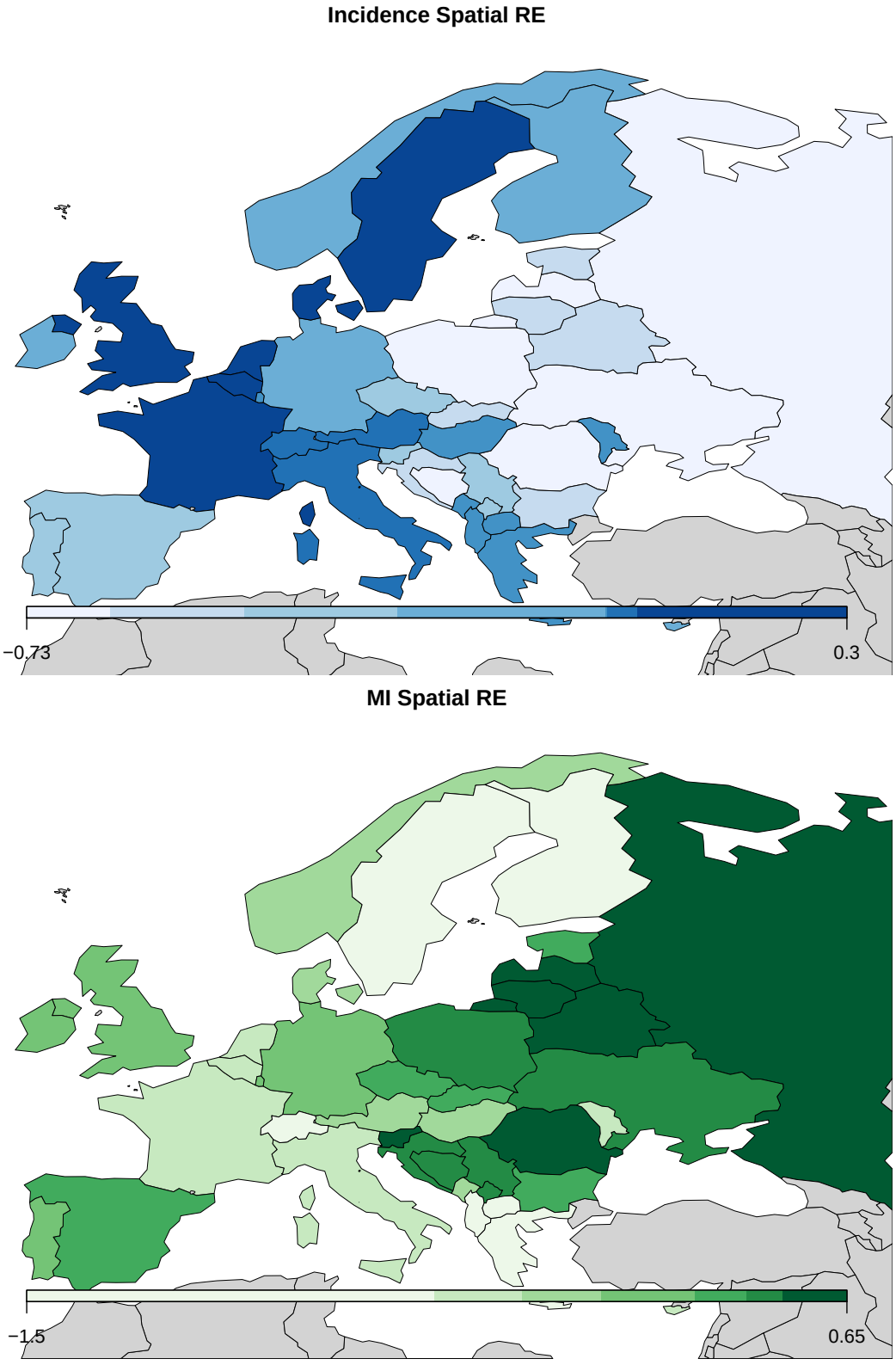

Figure 5.19: Posterior median of the spatial random effects $\boldsymbol{b}^I$ (top) and $\boldsymbol{b}^{MI}$ (bottom).

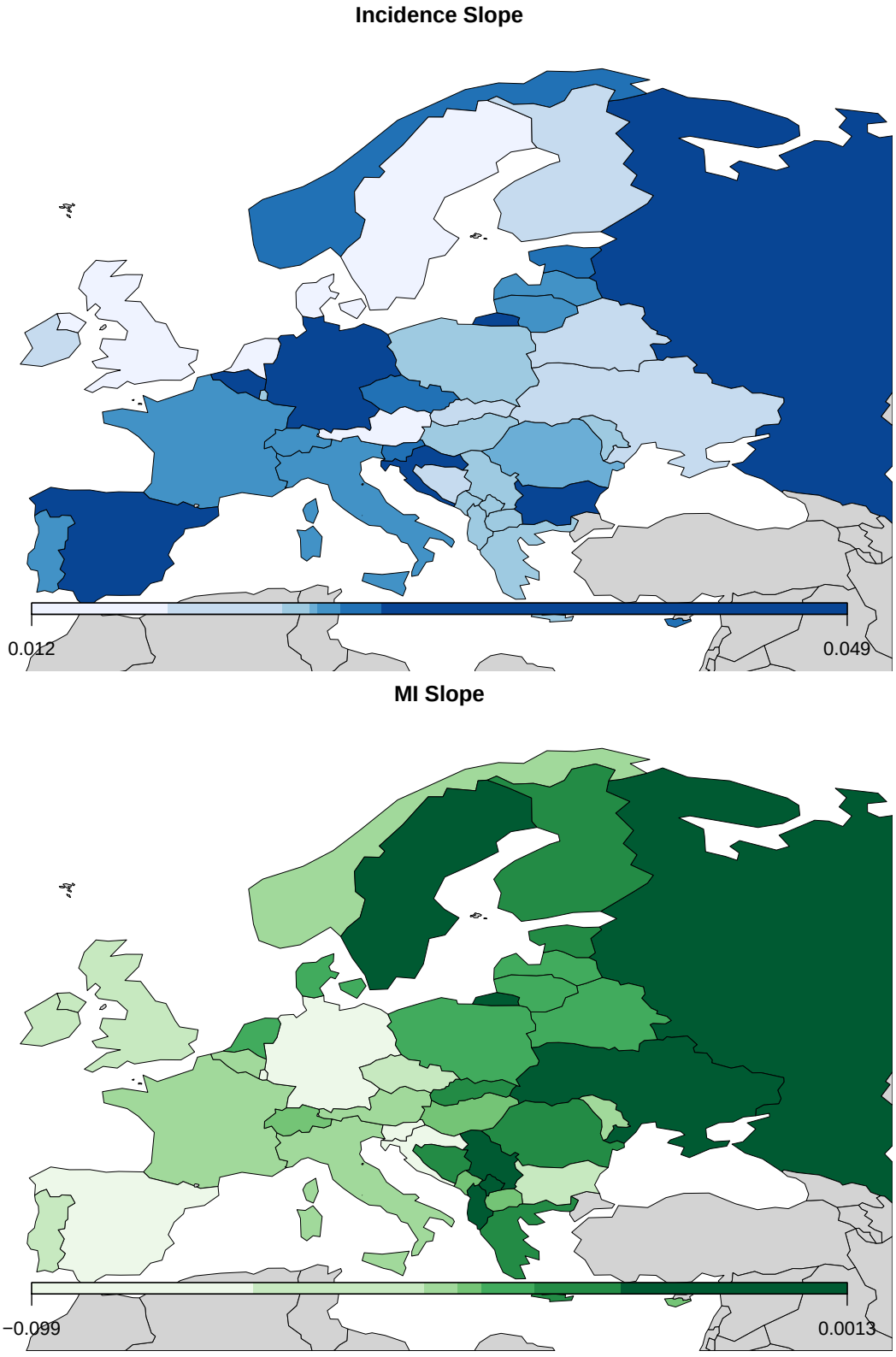

Figure 5.20: Posterior median of the linear time coefficients from Model IV, θ_o^I (top) and θ_o^{MI} (bottom).

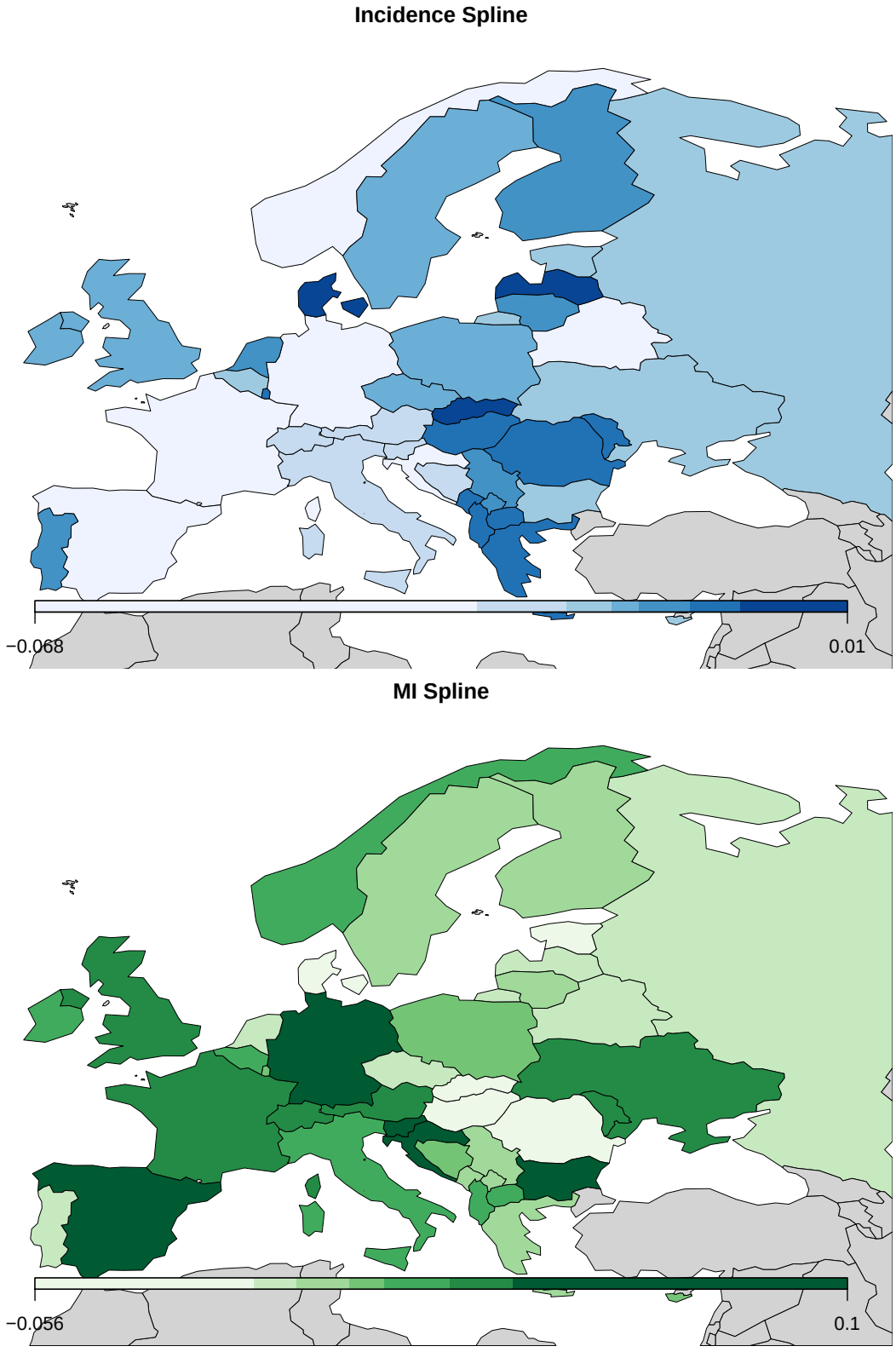

Figure 5.21: Posterior median of the spline coefficient for the knot at the year 2000 from Model IV, θ_1^I (top) and θ_1^{MI} (bottom).

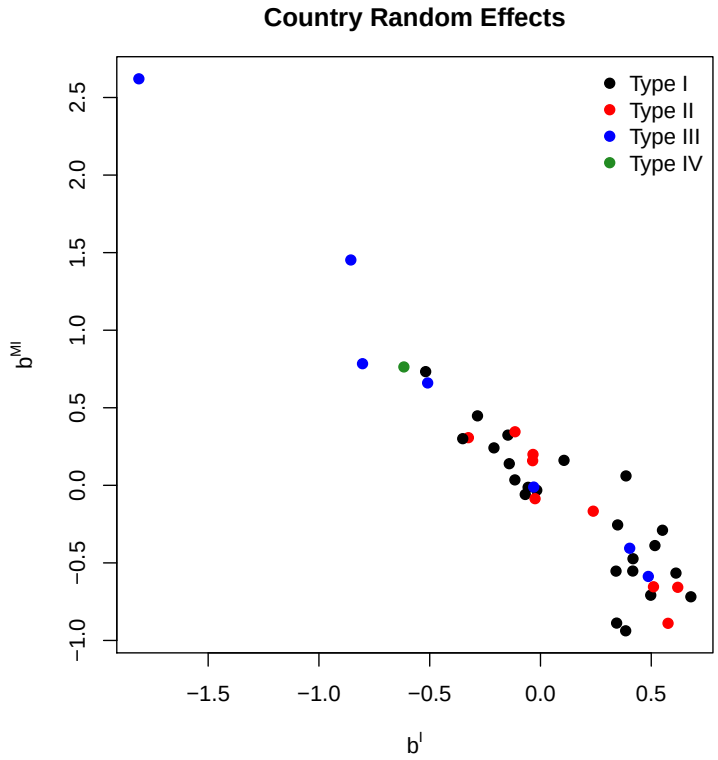

Figure 5.22: The country random effects $\boldsymbol{b}^I$ and $\boldsymbol{b}^{MI}$ with the Bivariate Normal prior.

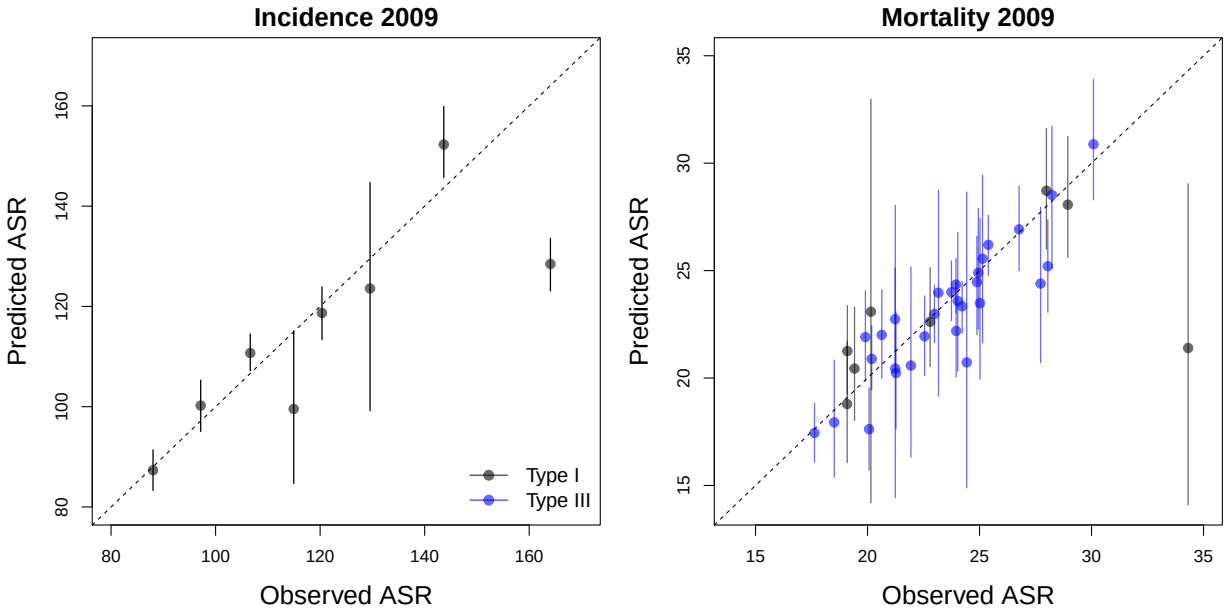

Figure 5.23: Predicted posterior distribution for the ASRs in 2009 based on model fit to data from 1990-2008 compared with observed values.

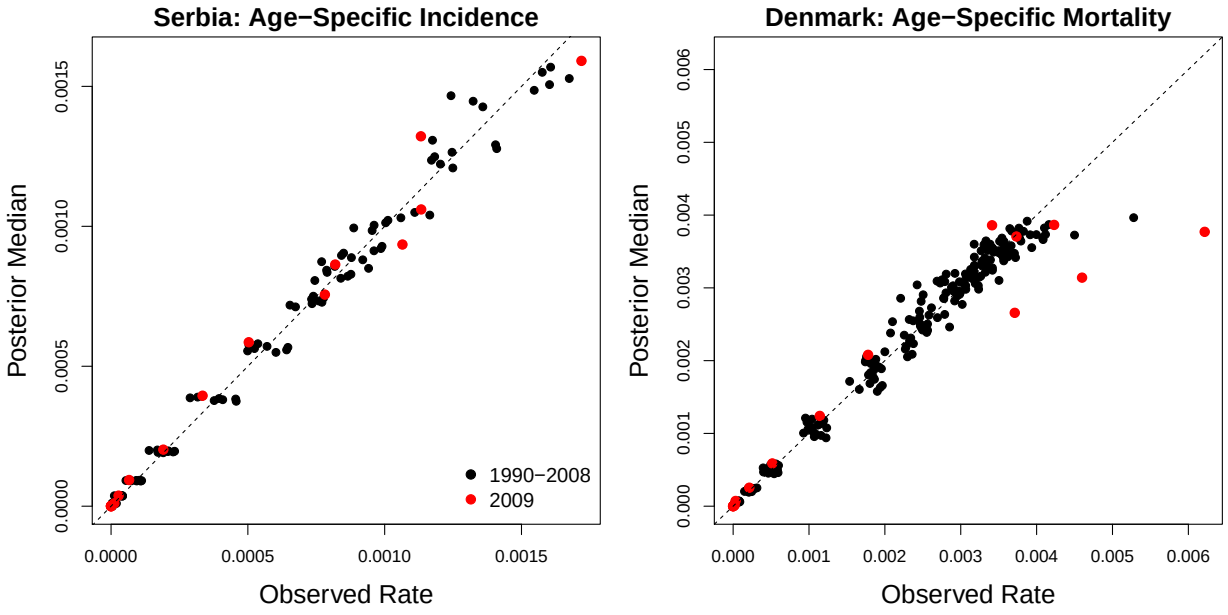

Figure 5.24: Age specific rates for incidence in Serbia (left) and mortality in Denmark (right) for 1990-2009.

of plots solid circles represent observed national data, hollow circles represent sub-national registry data, solid lines signify the posterior median, dashed lines the posterior 95% predictive interval, and age groups are shown by color. Rates are presented on the natural scale to emphasize differences in older age groups.

The Netherlands (Figure 5.25) provides a long time series of national data from 1990–2010. There is clear year-to-year variation in observed incidence and mortality rates, however most observations are contained in the predictive intervals. Italy (Figure 5.26) is limited to sub-national registry data from 2003–2007 and national mortality data from 2000–2009. Year-to-year variability is quite a bit higher in Italy compared with he Netherlands due to variability in rates observed in the sub-national registry data, however most observations from Italy also fall within the predictive intervals.

Figure 5.27 shows the estimated ASRs as of 2010 for incidence and mortality. In the top figure we see that the incidence rates in Western Europe are quite a bit higher than those of Eastern Europe. The lower figure shows that the mortality rates do not vary as much as the incidence rates and do not display the same spatial pattern. The large differences in Eastern and Western breast cancer incidence is likely due to both differences in risk factors as well as differences in the quality of reporting. However, the balance of these factors in each country is unknown and not the focus of this analysis, which aims to estimate reported rates of breast cancer.

Our analysis relies on the same data as the 2012 IARC projections of European cancer rates (Ferlay et al., 2013) which makes a direct comparison possible. Our model also allows us to compare our estimates with breast cancer rates from IHME published by Forouzanfar et al. (2011), which summarized cancer rates via the cumulative incidence rate in 2010 by country shown in (5.3).

Figure 5.28 provides a comparison of breast cancer incidence estimates by IHME on the left and IARC on the right. The incidence rates projected by IARC are contained within our 95% interval in many countries, but we see major differences in, Albania, Montenegro, and Greece. Interestingly, our results are quite similar to IHME for both Albania and Greece.

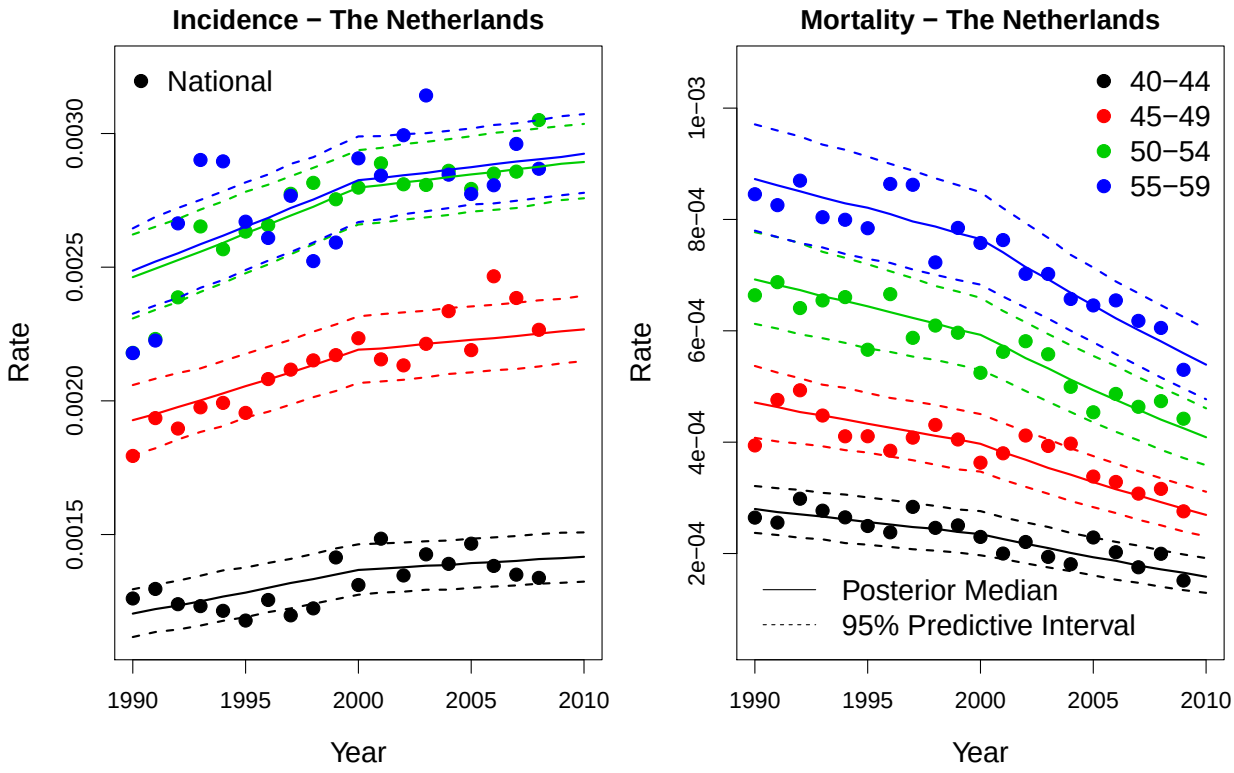

Figure 5.25: Posterior predictive distribution for age specific breast cancer rates in the Netherlands. Colors represent age groups and dashed lines represent the 95% predictive interval. Observed national data are shown in solid circles and registry data is shown with open circles.

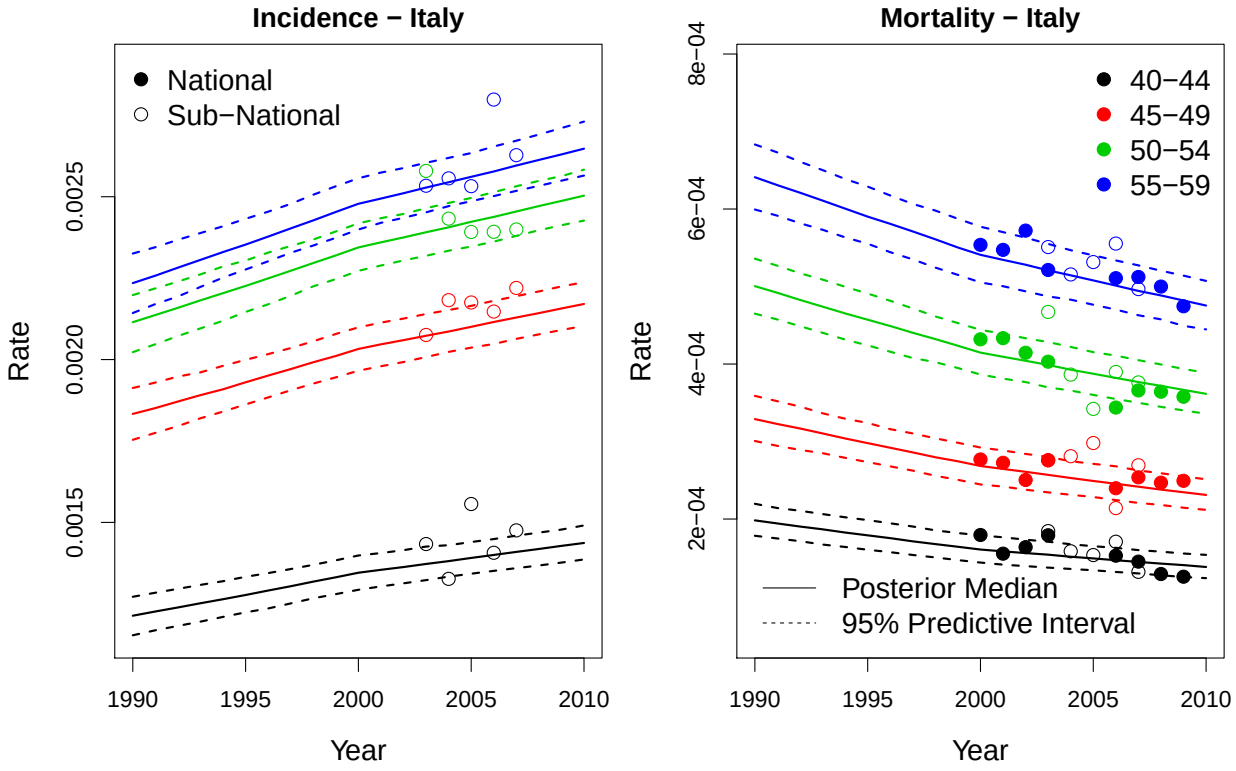

Figure 5.26: Posterior predictive distribution for age specific breast cancer rates in Italy. Colors represent age groups and dashed lines represent the 95% predictive interval. Observed national data are shown in solid circles and registry data is shown with open circles.

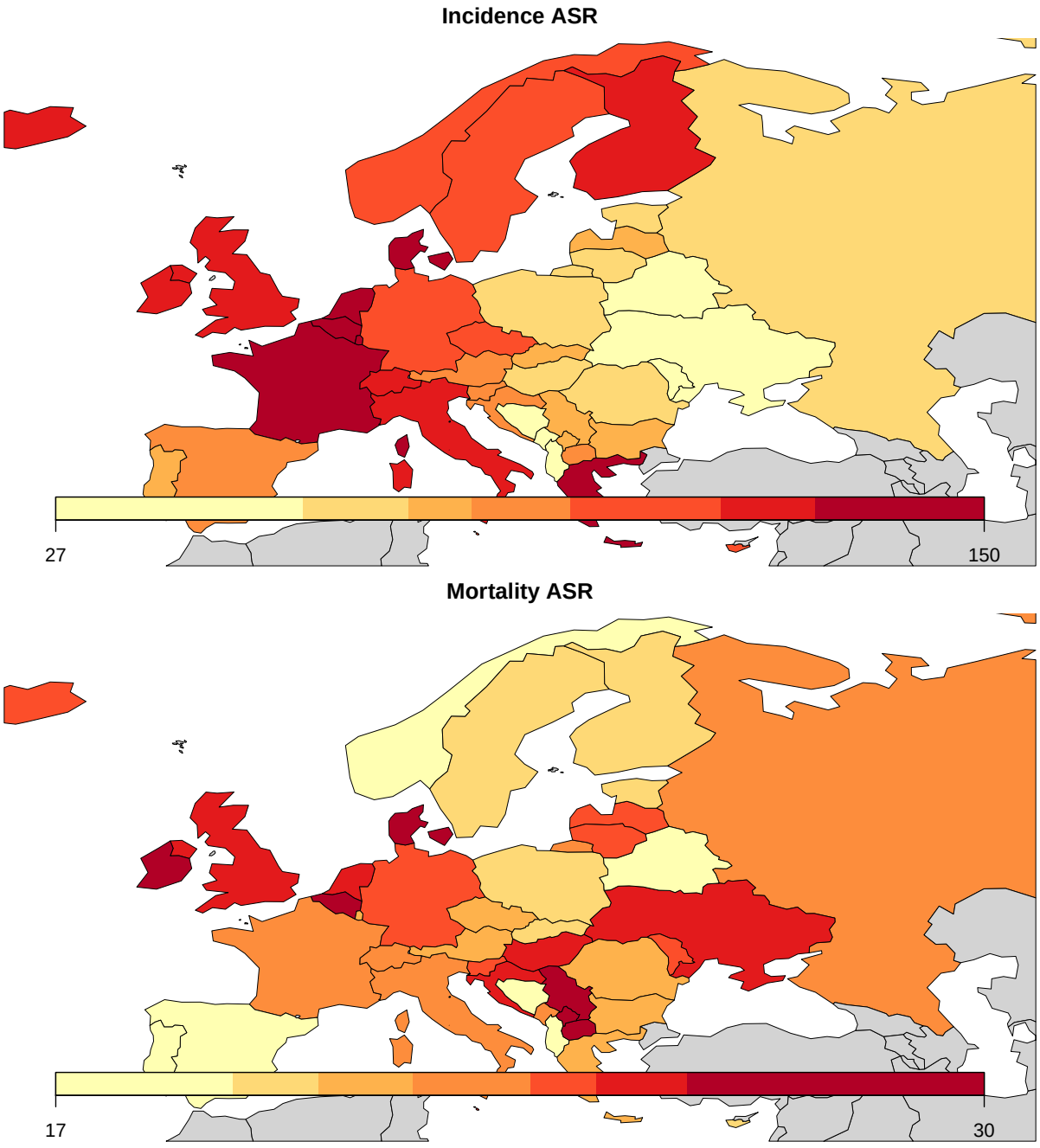

Figure 5.27: Age standardized rates for breast cancer incidence (top) and mortality (bottom) in Europe.

Another interesting comparison is in the width of the 95% intervals shown in the left panel of Figure 5.28 and ordered by width for our method (left) and IHME (right) in Figure 5.29. On average our intervals are narrower than those of IHME, but have a greater range of interval widths overall. In general we would expect the intervals to be narrower for countries with larger populations and better quality data. In the left panel of Figure 5.29 which shows our results we see that to be the case. However, in the right panel (IHME results) we see some very small countries, and many with no incidence data, such as Albania (in blue or green) that have much narrower intervals than larger countries, such as the United Kingdom which has quality national data and a larger population.

Figure 5.30 provides a comparison between our method and the IHME and IARC mortality estimates. The only substantial disagreement between our results and IARC are in Macedonia, which IARC estimates to be quite high. Our mortality estimates differ with the IHME estimates for many countries, without a clear pattern. Again in Figure 5.31 we see the odd pattern of large countries with quality national data having larger uncertainty intervals than smaller countries with poor quality data.

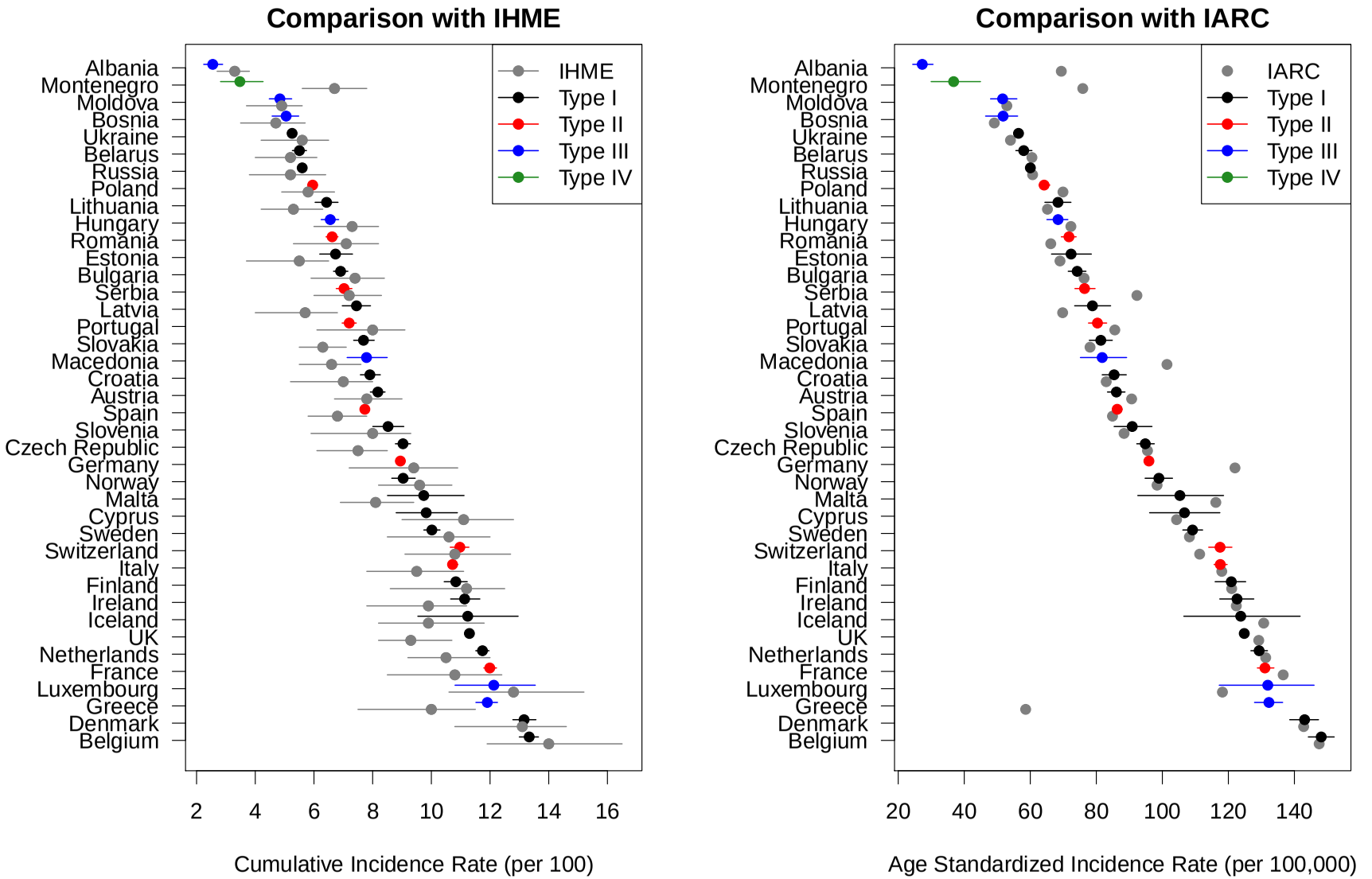

Figure 5.28: Comparison to published breast cancer incidence rates from IHME (left) and IARC (right). Points and lines in grey represent published estimates and intervals by IHME and IARC. Points and lines in black, red, blue, and green represent our estimates with colors corresponding to country types shown in Figure 5.3.

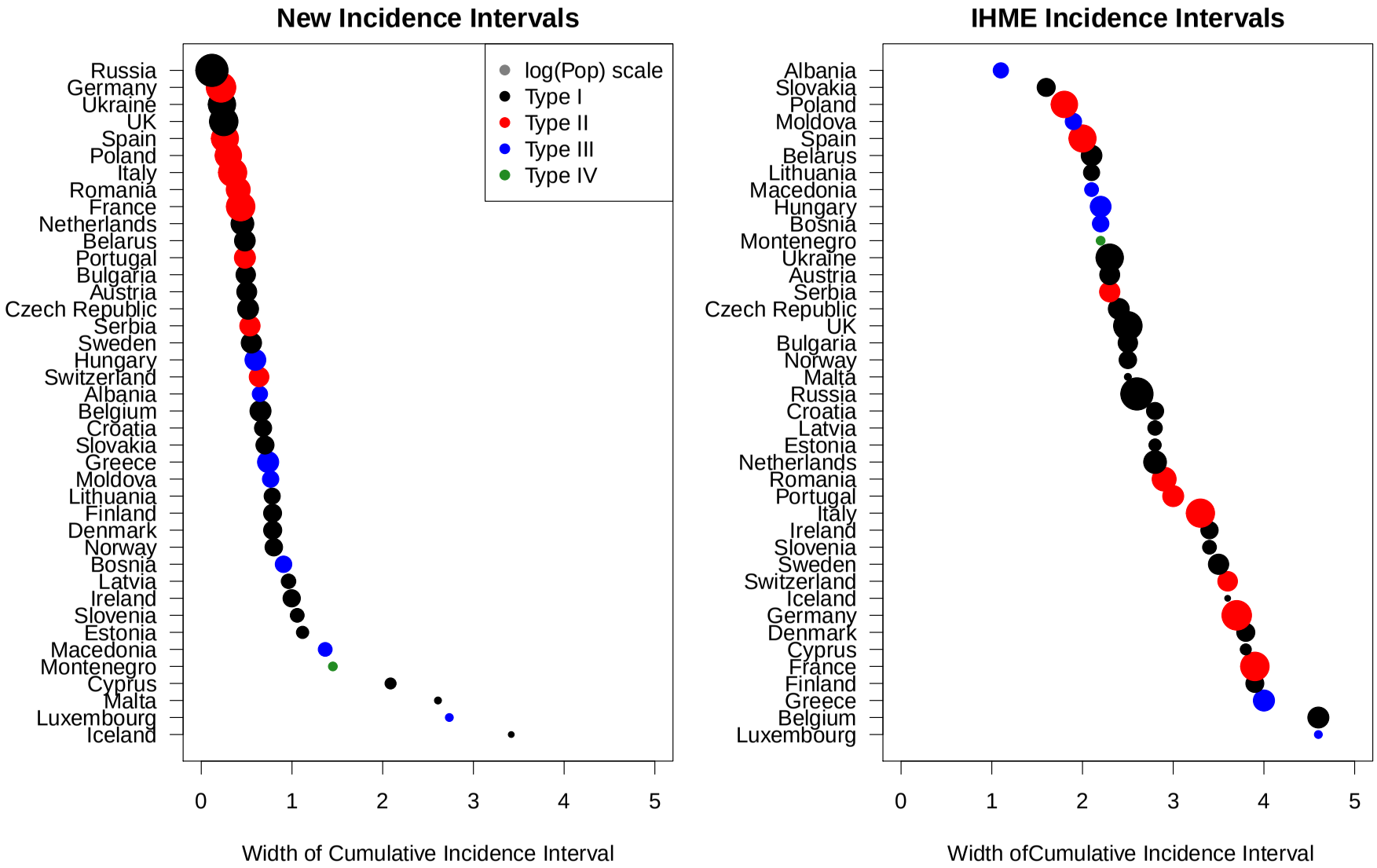

Figure 5.29: Confidence interval widths for incidence rates compared with published confidence intervals from IHME. Size of points represents population size and black, red, blue, and green colors represent our estimates with colors corresponding to country types shown in Figure 5.3.

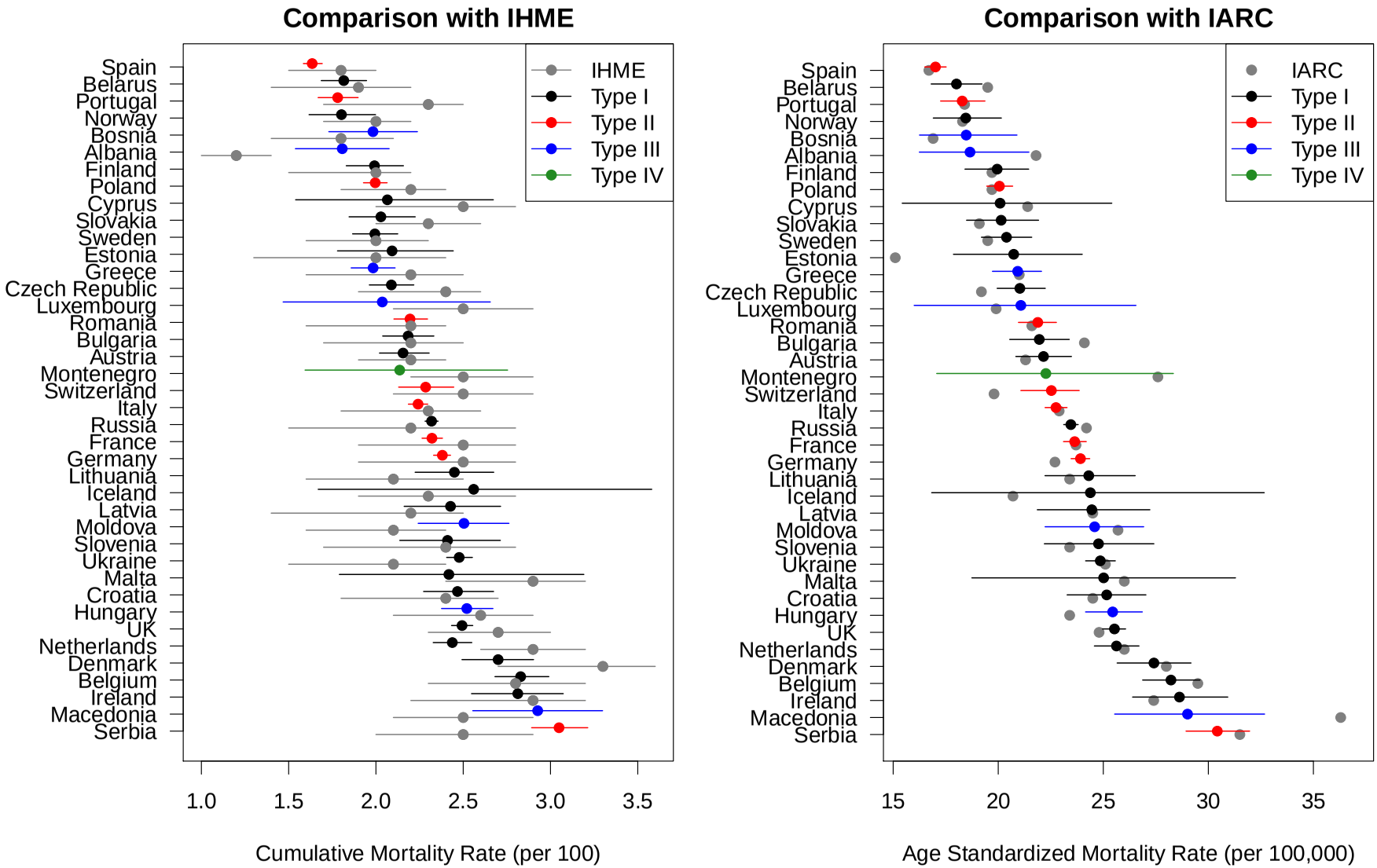

Figure 5.30: Comparison to published breast cancer mortality rates from IHME (left) and IARC (right). Points and lines in grey represent published estimates and intervals by IHME and IARC. Points and lines in black, red, blue, and green represent our estimates with colors corresponding to country types shown in Figure 5.3.

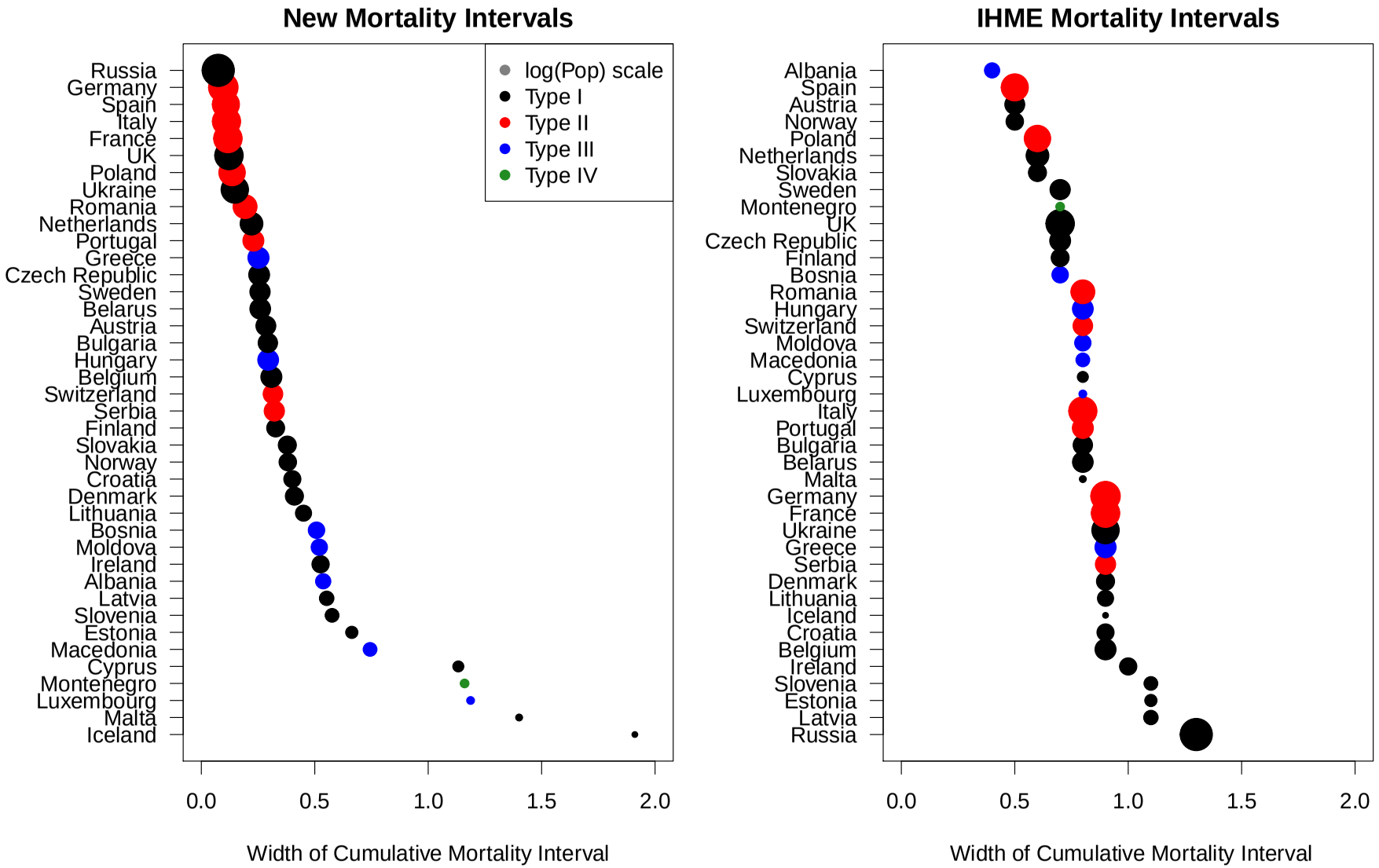

Figure 5.31: Confidence interval widths for mortality rates compared with published confidence intervals from IHME. Size of points represents population size and black, red, blue, and green colors represent our estimates with colors corresponding to country types shown in Figure 5.3.

5.10 Discussion

Simulation and validation results suggest reasonable coverage properties of our credible intervals and our point estimates are very similar to previously published breast cancer estimates from IARC. The width of our uncertainty intervals seems logical when in terms of population size and data quality, which does not appear to be the case for IHME intervals. Additionally, our method has the practical benefit of modeling all of Europe jointly. The current IARC approach requires cancer, gender, and country specific models. With this effort we have at least decreased the modeling burden by a factor of 40.

The work presented in this chapter would benefit from a few additional investigations. First, it may be interesting to consider alternative specifications for the random effects priors, such as RW2 by country for the space-age effects and the addition of a spatially structured random effect like in the BYM model. Second, we would like to apply our methods to generate estimates for other cancers in Europe. Lastly, future work will be focused on extending the method to be used in the developing world context where incidence data is more prevalent than cause-specific mortality data (Ferlay et al., 2010; Forouzanfar et al., 2011).

In summary we developed a principled approach for generating national estimates, projections, and uncertainty intervals for cancer incidence and mortality. Our approach relies on jointly modeling incidence, mortality given incidence (through the MI ratio), and mortality. This approach is flexible and can be simultaneously used for countries with national incidence data, registry incidence data, or no incidence data.

Chapter 6

DISCUSSION AND FUTURE WORK

Quantifying the health and wellbeing of a populations often suffers from a lack of adequate data. In resource-limited settings we find great demand for combining data from multiple sources to generate estimates and meaningful uncertainty intervals to aid in evaluation and resource allocation. In this dissertation we addressed three distinct methodological challenges in an effort to combine data from multiple sources and provide estimates and uncertainty intervals for meaningful population indicators.

In Chapter 3 we compared approaches for incorporating sampling weights in Bayesian hierarchical models for small area estimation. We found that incorporating the weights reduces bias due to non-response and variance is reduced through spatial smoothing and that the empirical logit normal model provides a flexible way to incorporate the design into a working likelihood. This approach was then applied to the 2006 Washington BRFSS data at the ZIP code level (Mercer et al., 2014). Through a collaboration with Public Health – Seattle & King County (PHSKC), these methods were applied to BRFSS data from 2009-2013 in King County at the census tract level. However, approximately 25% of respondents were missing census tract so we devised a multiple imputation procedure to assign census tracts based on crosswalk data from the Department of Housing and Urban Development. This work resulted in census tract level small area estimates of smoking prevalence and an approach that is currently being used by our collaborators at PHSKC for other health indicators (Song et al., 2016).

In Chapter 4 we derived a variance estimator for the under five child mortality rate (U5MR) as estimated by discrete time survival analysis from either household surveys or demographic surveillance sites. We used a Bayesian space-time smoothing approach that

accounted for survey effects and the design-based variances to generate sub-national estimates of U5MR in 21 regions of Tanzania from 1980 to 2014. This work highlighted substantial subnational variability in U5MR and evaluated sub-national progress towards the fourth Millennium Development Goal of reducing children mortality by two-thirds by 2015. This methodological work has been completed (Mercer et al., 2015), but work is ongoing through a group of students at the University of Washington who are working to apply this approach to other low and middle-income countries.

In Chapter 5 we developed an approach for joint modeling of cancer incidence and mortality given incidence in Europe. Our approach facilities incorporating both national and registry based data to generate estimates of incidence and mortality in countries with little or no data. Work is ongoing to investigate alternative spatial priors, model additional cancer sites and to expand methods to countries with only local incidence data.

The methods described in this dissertation were developed to provide useful information for guiding programmatic actions, planning and resource allocation. In my next professional position I have been tasked with combining surveillance data, household surveys, and administrative data to aid in planning interventions for the polio eradication campaign. Surely, the specific methodological challenges in my future work will be different, but one thing is clear, I will continue to build on the work that is described in this dissertation to develop and validate methods to combine data from multiple sources in resource-limited settings.

BIBLIOGRAPHY

Alkema, L., J. R. New, et al. (2014). Global estimation of child mortality using a bayesian b-spline bias-reduction model. *The Annals of Applied Statistics 8*(4), 2122–2149.

Allison, P. (1984). *Event History Analysis: Regression for Longitudinal Event Data*. Number 46. Sage Publications Inc.

Anscombe, F. (1948). The transformation of Poisson, binomial and negative-binomial data. *Biometrika 35*, 246–254.

Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics – Theory and Methods 35*, 439–460.

Bell, R. and M. Cohen (2007). Comment on "Struggles with survey weighting and regression modeling". *Statistical Science 22*, 165–167.

Benkeser, D. (2016). Personal Communication.

Besag, J. and C. Kooperberg (1995). On conditional and intrinsic auto-regressions. *Biometrika 82*, 733–746.

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

Binder, D. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review 51*, 279–292.

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

Breidt, F. and J. Opsomer (2007). Comment on "Struggles with survey weighting and regression modeling". *Statistical Science 22*, 168–170.

Browne, W. and D. Draper (2006a). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis 1*, 473–514.

Browne, W. and D. Draper (2006b). A comparison of Bayesian and likelihood-based methods for fitting multilevel models (rejoinder). *Bayesian Analysis 1*, 547–550.

Browne, W. J. (2004). An illustration of the use of reparameterisation methods for improving mcmc efficiency in crossed random effect models. *Multilevel modelling newsletter 16*(1), 13–25.

Byass, P., A. Worku, A. Emmelin, and Y. Berhane (2007). Dss and dhs: longitudinal and cross-sectional viewpoints on child and adolescent mortality in ethiopia. *Population Health Metrics 5*(1), 12.

Carpenter, B., A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell (2015). Stan: a probabilistic programming language. *Journal of Statistical Software*.

CDC (2011, accessed 2016-022-10). *Behavioral Risk Factor Surveillance System (BRFSS) Fact Sheet: Raking*. `http://health.mo.gov/data/brfss/BRFSSweightingmethod.pdf`.

CDC (2016). What are the risk factors for breast cancer?

Chen, C., J. Wakefield, and T. Lumely (2014). The use of sampling weights in bayesian hierarchical models for small area estimation. *Spatial and spatio-temporal epidemiology 11*, 33–43.

Clark, S. J., J. Wakefield, T. McCormick, and R. Michelle (2012). Hyak mortality monitoring system innovative sampling and estimation methods: Proof of concept by simulation.

Technical Report 118, Center for Statistics and the Social Sciences (CSSS), University of Washington, https://www.csss.washington.edu/Papers/wp118.pdf.

Clayton, D. (1996). Generalized linear mixed models. In W. Gilks, S. Richardson, and D. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*, pp. 275–301. Chapman and Hall.

Congdon, P. and P. Lloyd (2010). Estimating small area diabetes prevalence in the US using the behavioral risk factor surveillance system. *Journal of Data Science 8*, 235–252.

Dean, C., M. Ugarte, and A. Militino (2001). Detecting interaction between random region and fixed age effects in disease mapping. *Biometrics 57*(1), 197–202.

Demographic and Health Surveys (1992). *Demographic Health Survey 1991/1992*. Bureau of Statistics Planning Commission.

Demographic and Health Surveys (1997). *Tanzania Demographic and Health Survey 1996*. Bureau of Statistics [Tanzania] and Macro International Inc.

Demographic and Health Surveys (2000). *Tanzania Demographic and Health Survey 1999*. Bureau of Statistics [Tanzania] and Macro International Inc.

Demographic and Health Surveys (2005). *Tanzania Demographic and Health Survey 2004-05*. National Bureau of Statistics (NBS) [Tanzania] and ORC Macro.

Demographic and Health Surveys (2010). *Tanzania Demographic and Health Survey 2010*. National Bureau of Statistics (NBS) [Tanzania] and ICF Macro.

Duane, S., A. D. Kennedy, B. J. Pendleton, and D. Roweth (1987). Hybrid monte carlo. *Physics letters B 195*(2), 216–222.

Dwyer-Lindgren, L., F. Kakungu, P. Hangoma, M. Ng, H. Wang, A. D. Flaxman, F. Masiye, and E. Gakidou (2014). Estimation of district-level under-5 mortality in zambia using birth history data, 1980–2010. *Spatial and spatio-temporal epidemiology 11*, 89–107.

Dyba, T. and T. Hakulinen (2000). Comparison of different approaches to incidence prediction based on simple interpolation techniques. *Statistics in medicine 19*(13), 1741–1752.

Ferlay, J., D. Parkin, and E. Steliarova-Foucher (2010). Estimates of cancer incidence and mortality in Europe in 2008. *European Journal of Cancer 46*, 765–781.

Ferlay, J., H.-R. Shin, F. Bray, D. Forman, C. Mathers, and D. M. Parkin (2010). Estimates of worldwide burden of cancer in 2008: Globocan 2008. *International journal of cancer 127*(12), 2893–2917.

Ferlay, J., E. Steliarova-Foucher, J. Lortet-Tieulent, S. Rosso, J. Coebergh, H. Comber, D. Forman, and F. Bray (2013). Cancer incidence and mortality patterns in europe: estimates for 40 countries in 2012. *European Journal of Cancer 49*, 1374–1403.

Fitzmaurice, C., D. Dicker, A. Pain, H. Hamavid, M. Moradi-Lakeh, M. F. MacIntyre, C. Allen, G. Hansen, R. Woodbrook, C. Wolfe, et al. (2015). The global burden of cancer 2013. *JAMA oncology 1*(4), 505–527.

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

Foreman, K. J., R. Lozano, A. D. Lopez, C. Murray, et al. (2012). Modeling causes of death: an integrated approach using codem. *Population Health Metrics 10*(1).

Forouzanfar, M., K. Foreman, A. Delossantos, R. Lozano, A. Lopez, C. Murray, and M. Naghavi (2011). Breast and cervical cancer in 187 countries between 1980 and 2010: a systematic analysis. *The Lancet 378*, 1461–1484.

Fottrell, E., F. Enquselassie, and P. Byass (2009). The distribution and effects of child mortality risk factors in ethiopia: a comparison of estimates from dss and dhs. *Ethiopian Journal of Health Development 23*(2).

Gakidou, E., K. Cowling, R. Lozano, and C. J. Murray (2010). Increased educational attainment and its effect on child mortality in 175 countries between 1970 and 2009: a systematic analysis. *The Lancet 376*(9745), 959–974.

Gelfand, A. E., S. K. Sahu, and B. P. Carlin (1995). Efficient parametrisations for normal linear mixed models. *Biometrika 82*(3), 479–488.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis 1*, 515–534.

Gelman, A. (2007a). Rejoinder to "Struggles with survey weighting and regression modeling". *Statistical Science 22*, 184–188.

Gelman, A. (2007b). Struggles with survey weighting and regression modeling. *Statistical Science 22*, 153–164.

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

Hammer, G. P., B. Kouyaté, H. Ramroth, and H. Becher (2006). Risk factors for childhood mortality in sub-saharan africa: a comparison of data from a demographic and health survey and from a demographic surveillance system. *Acta Tropica 98*(3), 212–218.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika 57*(1), 97–109.

Held, L., B. Schrödle, and H. Rue (2010). Posterior and cross-validatory predictive checks: A comparison of MCMC and INLA. In T. Kneib and G. Tutz (Eds.), *Statistical Modeling and Regression Structures – Festschrift in Honour of Ludwig Fahrmeir*, pp. 91–110. Physica-Verlag.

Hogan, M., K. Foreman, M. Naghavi, S. Ahn, M. Wang, S. Makela, A. Lopez, R. Lozano,

and C. Murray (2010). Maternal mortality for 181 countries, 1980–2008: a systematic analysis of progress towards millennium development goal 5. *The Lancet 375*, 1609–1623.

Homan, M. D. and A. Gelman (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *The Journal of Machine Learning Research 15*(1), 1593–1623.

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

HUD (2012, accessed 2015-11-05). *U.S. Department of Housing and Urban Development*. `http://www.huduser.org/portal/datasets/usps_crosswalk.html`.

INDEPTH Network (2014, accessed 2014-10-20). *Health and Demographic Surveillance Systems*. `http://www.indepth-network.org/index.php?option=com_content&task=view&id=1798&Itemid=501`.

Jenkins, S. P. (1995). Easy estimation methods for discrete-time duration models. *Oxford Bulletin of Economics and Statistics 57*(1), 129–136.

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

Knorr Held, L. and H. Rue (2002). On block updating in markov random field models for disease mapping. *Scandinavian Journal of Statistics 29*(4), 597–614.

Knutson, K., W. Zhang, and F. Tabnak (2008). Applying the small-area estimation method to estimate a population eligible for breast cancer detection services. *Preventing chronic disease 5*(1).

Lambert, P. (2006). Comment on article by Browne and Draper. *Bayesian Analysis 1*, 543–546.

Leroux, B. G., X. Lei, and N. Breslow (2000). Estimation of disease rates in small areas: A new mixed model for spatial dependence. In *Statistical models in epidemiology, the environment, and clinical trials*, pp. 179–191. Springer.

Lindgren, F. and H. Rue (2015). Bayesian spatial modelling with r-inla. *Journal of Statistical Software 63*(19).

Little, R. (2007). Comment on "Struggles with survey weighting and regression modeling". *Statistical Science 22*, 171–174.

Little, R. J. and D. B. Rubin (2002). *Statistical Analysis with Missing Data*. Wiley.

Lohr, S. (2007). Comment on "Struggles with survey weighting and regression modeling". *Statistical Science 22*, 175–178.

Lohr, S. (2009). *Sampling: Design and Analysis*. Cengage Learning.

Longford, N. (1996). Model-based variance estimation in surveys with stratified clustered design. *Australian Journal of Statistics 38*, 333–352.

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

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis using R*. Hoboken, Jersey: John Wiley and Sons.

Lunn, D. J., A. Thomas, N. Best, and D. Spiegelhalter (2000). Winbugs-a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing 10*(4), 325–337.

Mathers, C., T. Boerma, and D. Ma Fat (2009). Global and regional causes of death. *British Medical Bulletin 92*, 7–32.

Mercer, L., J. Wakefield, C. Chen, and T. Lumley (2014). A comparison of spatial smoothing methods for small area estimation with sampling weights. *Spatial Statistics 8*, 69–85.

Mercer, L. D., J. Wakefield, A. Pantazis, A. M. Lutambi, H. Masanja, and S. Clark (2015). Space–time smoothing of complex survey data: Small area estimation for child mortality. *The Annals of Applied Statistics 9*(4), 1889–1905.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics 21*(6), 1087–1092.

Møller, B., H. Fekjær, T. Hakulinen, H. Sigvaldason, H. H. Storm, M. Talbäck, and T. Haldorsen (2003). Prediction of cancer incidence in the nordic countries: empirical comparison of different approaches. *Statistics in medicine 22*(17), 2751–2766.

Paris21 (2014, accessed 2014-10-20). *Paris21: Partnership for Statistics in Development in the 21st Century*. `http://www.paris21.org`.

Parkin, D., F. Bray, J. Ferlay, and P. Pisani (2001). Estimating the world cancer burden: Globocan 2000. *International Journal of Cancer 94*(2), 153–156.

Patil, A., D. Huard, and C. J. Fonnesbeck (2010). Pymc: Bayesian stochastic modelling in python. *Journal of statistical software 35*(4), 1.

Pedersen, J. and J. Liu (2012, August). Child mortality estimation: Appropriate time periods for child mortality estimates from full birth histories. *PLoS Medicine 9*(8).

Pfefferman, D. (2007). Comment on "Struggles with survey weighting and regression modeling". *Statistical Science 22*, 179–183.

Pfeffermann, D., C. Skinner, D. Holmes, H. Goldstein, and J. Rasbash (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, Series B 60*, 23–40.

Plummer, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics 9*, 523–539.

Potthoff, R., M. Woodbury, and K. Manton (1992). "Equivalent sample size" and "equivalent degrees of freedom" refinements for inference using survey weights under superpopulation models. *Journal of the American Statistical Association 87*, 383–396.

Rabe-Hesketh, S. and A. Skrondal (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society, Series A 169*, 805–827.

Raftery, A. E., D. Madigan, and J. A. Hoeting (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association 92*(437), 179–191.

Raghunathan, T., D. Xie, N. Schenker, V. Parsons, W. Davis, K. Dood, and E. Feuer (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk factos and screening. *Journal of the American Statistical Association 102*, 474–486.

Rajaratnam, J., J. Marcus, A. Flaxman, H. Wang, A. Levin-Rector, L. Dwyer, M. Costa, A. Lopez, and C. Murray (2010). Neonatal, postneonatal, childhood, and under-5 mortality for 187 countries, 1970–2010: a systematic analysis of progress towards millennium development goal 4. *The Lancet 375*, 1988–2008.

Rajaratnam, J., J. Marcus, A. Levin-Rector, A. Chalupka, H. Wang, L. Dwyer, M. Costa, A. Lopez, and C. Murray (2010). Worldwide mortality in men and women aged 15–59 years from 1970 to 2010: a systematic analysis. *The Lancet 375*, 1704–1720.

Rao, J. (2003). *Small Area Estimation*. New York: John Wiley.

Roberts, G., N. Rao, and S. Kumar (1987). Logistic regression analysis of sample survey data. *Biometrika 74*(1), 1–12.

Roberts, G. O. and J. S. Rosenthal (2009). Examples of adaptive mcmc. *Journal of Computational and Graphical Statistics 18*(2), 349–367.

Rue, H. (2001). Fast sampling of gaussian markov random fields. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63*(2), 325–338.

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

Rue, H., S. Martino, and N. Chopin (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. O. and G. Rojas (2006). Tanzania demographic and health survey 1996. *Calverton, Maryland: ORC Macro*.

Schrödle, B. and L. Held (2011). Spatio-temporal disease mapping using inla. *Environmetrics 22*(6), 725–734.

Skinner, C. (1989). Domain means, regression and multivariate analysis. In C. Skinner, D. Holt, and T. Smith (Eds.), *Analysis of Complex Surveys*, pp. 59–87. Chichester: Wiley.

Song, L., L. D. Mercer, J. Wakefield, A. Laurent, and D. Solet (2016). 'using small area estimation to estimate the prevalence of smoking by sub-county geographies in king county, washington. *Preventing Chronic Disease 13*.

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

Spiegelhalter, D., N. Best, B. Carlin, and A. V. D. Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B 64*, 583–639.

Spiegelhalter, D., N. Best, B. Carlin, and A. V. D. Linde (2014). The deviance information criterion: 12 years on (with discussion). *Journal of the Royal Statistical Society: Series B 64*, 485–493.

Stan Development Team (2015a). Stan: A c++ library for probability and sampling, version 2.8.0.

Stan Development Team (2015b). *Stan Modeling Language User's Guide and Reference Manual, Version 2.8.0*.

Uhry, Z., A. Belot, M. Colonna, N. Bossard, A. Rogel, J. Iwaz, N. Mitton, P. Grosclaude, and L. Remontet (2013). National cancer incidence is estimated using the incidence/mortality ratio in countries with local incidence data: Is this estimation correct? *Cancer epidemiology 37*(3), 270–277.

UN (2000, accessed 2014-10-20). *Millennium Development Goals*. `http://www.un.org/millenniumgoals/`.

UN (2014a, accessed 2014-10-20). *Civil Registration and Vital Statistics Coverage*. `http://unstats.un.org/unsd/demographic/CRVS/CR_coverage.htm`.

UN (2014b, accessed 2014-10-20). *Data Revolution for Sustainable Development*. `http://www.un.org/apps/news/story.asp?NewsID=48594#.VEVQpoctuvJ`.

UN (2014c, accessed 2014-10-20). *Millennium Development Goal number 4: Reduce by two thirds, between 1990 and 2015, the under-five mortality rate*. `http://www.un.org/millenniumgoals/childhealth.shtml`.

UN (2014d, accessed 2014-10-20). *The Post-2015 Development Agenda*. `http://www.post2015hlp.org/the-report/`.

UN (2014e, accessed 2014-10-20). *Sustainable Development Goals*. `http://sustainabledevelopment.un.org/owg.html`.

United Nations, D. o. E. and P. D. Social Affairs (2015). World population prospects: The 2015 revision, key findings and advance tables. *Working Paper No. ESA/P/WP 241*.

USAID (2014, accessed 2014-10-20). *Demographic and Health Surveys*. `http://www.dhsprogram.com`: United States Agency for International Development.

Wakefield, J. (2009). Multi-level modelling, the ecologic fallacy, and hybrid study designs. *International Journal of Epidemiology 38*, 330–336.

Wakefield, J. C., N. G. Best, and L. A. Waller (2000). Bayesian approaches to disease mapping. In P. Elliott, J. C. Wakefield, N. G. Best, and D. Briggs (Eds.), *Spatial Epidemiology: Methods and Applications*, pp. 104–27. Oxford: Oxford University Press.

Wang, H., C. A. Liddell, M. M. Coates, M. D. Mooney, C. E. Levitz, A. E. Schumacher, H. Apfel, M. Lannarone, B. Phillips, K. T. Lofgren, et al. (2014). Global, regional, and national levels of neonatal, infant, and under-5 mortality during 1990–2013: a systematic analysis for the Global Burden of Disease study 2013. *The Lancet 384*(9947), 957–979.

World Bank and World Health Organization (2014, accessed 2014-10-20). *Global Civil Registration and Vital Statistics Scaling Up Investment Plan 2015-2024*. `http://www.worldbank.org/en/topic/health/publication/global-civil-registration-vital-statistics-scaling-up-investment`.

Ye, Y., M. Wamukoya, A. Ezeh, J. B. Emina, and O. Sankoh (2012). Health and demographic surveillance systems: a step towards full civil registration and vital statistics system in sub-sahara africa? *BMC Public Health 12*(1), 741.

Appendix A

APPENDIX FOR CHAPTER 2

A.1 Block Updating Details

$$
\begin{aligned}
q(\mathbf{u}|y) &\propto \exp\left\{-\frac{\tau}{2}\left(\mathbf{u}^T\mathbf{R}\mathbf{u}\right)\right\} \times \prod_{i=1}^{n} \text{Pois}\left(N_i p_i\right) \\
&\propto \exp\left\{-\frac{\tau}{2}\left(\mathbf{u}^T\mathbf{R}\mathbf{u}\right)\right\} \times \prod_{i=1}^{n} (N_i p_i)^{y_i} \exp(-N_i p_i) \\
&\propto \exp\left\{-\frac{\tau}{2}\left(\mathbf{u}^T\mathbf{R}\mathbf{u}\right)\right\} \times \prod_{i=1}^{n} \exp(\alpha + u_i)^{y_i} \exp(-N_i \exp(\alpha + u_i)) \\
&\propto \exp\left\{-\frac{\tau}{2}\left(\mathbf{u}^T\mathbf{R}\mathbf{u}\right) + \sum_{i=1}^{n} [y_i u_i - N_i \exp(\alpha + u_i)]\right\} \\
&\approx \exp\left\{-\frac{\tau}{2}\left(\mathbf{u}^T\mathbf{R}\mathbf{u}\right) + \sum_{i=1}^{n} \left[y_i u_i - N_i \exp(\alpha) \underbrace{\left(B(u_{i,0})u_i + \frac{1}{2}C(u_{i,0})u_i^2\right)}_{\text{Taylor expansion of } \exp(u_i) \text{ around } u_{i,0}}\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\tau\mathbf{u}^T\mathbf{R}\mathbf{u} + \exp(\alpha)\mathbf{u}^T\text{diag}\left(\mathbf{N}C(\mathbf{u}_0)\right)\mathbf{u}_0\right) + \left(\mathbf{y} - \exp(\alpha)\mathbf{N}B(\mathbf{u}_0)\right)^T\mathbf{u}\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\mathbf{u}^T\left[\tau\mathbf{R} + \exp(\alpha)\text{diag}(\mathbf{N}C(\mathbf{u}_0))\right]\mathbf{u}\right) + \left(\mathbf{y} - \mathbf{N}B(\mathbf{u}_0)\right)^T\mathbf{u}\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\mathbf{u}^T\mathbf{Q}(\mathbf{u}_0)\mathbf{u}\right) + \mathbf{b}^T(\mathbf{u}_0)\mathbf{u}\right\} \\
&\propto N_C\left(\mathbf{b}(\mathbf{u}_0), \mathbf{Q}(\mathbf{u}_0)\right)
\end{aligned}
$$

Where

$$\begin{aligned}
\mathbf{b}(\mathbf{u}_0) &= \mathbf{y} - \exp(\alpha)\mathbf{N}B(\mathbf{u}_0) \\
\mathbf{Q}(\mathbf{u}_0) &= \tau\mathbf{R} + \exp(\alpha)\mathrm{diag}(\mathbf{N}C(\mathbf{u}_0)) \\
B(a) &= \exp(a)(1-a) \\
C(a) &= \exp(a)
\end{aligned}$$

and $N_C(\mathbf{b}, \mathbf{Q})$ represents the Canonical parameterization of a GMRF (Definition 2.2, Rue & Held). For a GMRF $\mathbf{x}$ with respect to some graph G and canonical parameters $\mathbf{b}$ and precision matrix $\mathbf{Q} > 0$ the density of x is

$$\pi(\mathbf{x}) \propto \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{b}^T\mathbf{x}\right)$$

written as $\mathbf{x} \sim N_C(\mathbf{b}, \mathbf{Q})$ which is equivalent to $N(\mu = \mathbf{Q}^{-1}\mathbf{b}, \Sigma = \mathbf{Q}^{-1})$.

Appendix B

Table B.1: Data availability, source, and quality by country in Europe.

		Registries		Nat. Inc.	Nat. Mort.	Category	
European Region	Country	Number	Years	Years	Years	Score	Type
Southern	Albania			NA	95-04	G3	III
	Bosnia-Herzegovina	1	2009	NA	2011	D5	II
	Croatia			88-07	90-09	A2	I
	Cyprus			98-07	04-09	A3	I
	Greece			NA	90-09	G3	III
	Italy	17	03-07	NA	00-03,06-09	B2	II
	Malta			00-09	90-09	A1	I
	Montenegro			NA	NA	G6	IV
	Portugal	2	03-07	NA	00-03,07-09	C3	II
	Serbia	1	03-07	NA	01-09	B2	II
	Slovenia			89-08	90-09	A1	I
	Spain	12	03-07	NA	90-09	B2	II
Western	Austria			90-09	90-09	A2	I
	Belgium			08-09	05-10	A2	I
	France	11	03-07	NA	00-09	B2	II
	Germany	9	00-08	NA	90-09	B2	II
	Luxembourg			NA	90-09	D2	III
	Netherlands			89-08	90-09	A2	I
	Switzerland	7	03-07	NA	00-10	B2	II
Northern	Denmark			90-09	90-09	A2	I
	Estonia			88-07	90-09	A1	I
	Finland			90-09	90-09	A1	I
	Iceland			90-09	90-09	A1	I
	Ireland			94-07	90-09	A1	I
	Latvia			88-07	01-10	A1	I
	Lithuania			88-07	01-10	A1	I
	Norway			90-09	90-09	A2	I
	Sweden			90-09	90-09	A2	I
	United Kingdom			93-07	90-09	A1	I
Eastern & Central	Belarus			88-07	07-09	A2	I
	Bulgaria			94-08	90-09	A2	I
	Czech Republic			89-08	90-09	A2	I
	Hungary			NA	90-09	G1	III
	Moldova			NA	01-10	G1	III
	Poland	5	03-07	NA	01-10	C3	II
	Romania	1	06-08	NA	01-10	E1	II
	Russia			94-08	90-10	D2	I
	Slovakia			88-07	90-09	A1	I
	Ukraine			03-08	08-09	A2	I

VITA

Laina Mercer was born to Cheryl Mercer and Brad Hendren in Anchorage, Alaska. She earned an A.S. in Geology from Skagit Valley College in 2004. She worked as a Teaching Assistant at Sea Mar Visions in Bellingham, Washington between 2004 and 2006. In 2008 she earned her B.S. in Mathematics with a minor in Chemistry from Western Washington University. In 2011 she earned her M.S. in Biostatistics from the University of Washington under the supervision of Drs. Robyn McClelland and Ken Rice. She worked as a Statistical Research Associate at Seattle Children's Research Institute between 2011 and 2013. In 2016, she earned her Ph.D. in Statistics from the University of Washington under the supervision of Dr. Jon Wakefield.