Methods, Models, and Interpretations for Spatial-Temporal Public Health Applications

Aaron Osgood-Zimmerman

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

Doctor of Philosophy

University of Washington

2022

Reading Committee:

Jon Wakefield, Chair

Alex Luedtke

Bobby Reiner

Program Authorized to Offer Degree:

Statistics

University of Washington

Abstract

Methods, Models, and Interpretations for
Spatial-Temporal Public Health Applications

Aaron Osgood-Zimmerman

Chair of the Supervisory Committee:
Dr. Jon Wakefield
Biostatistics and Statistics

Improving the health of communities and individuals around the world is one of the great challenges of this densely connected global era which finds itself rife with disparity. In order to make the best use of our limited resources, spatially-resolved and time-specific estimates of health indicators are required to make well-informed decisions regarding resource allocation and policy implementation. The availability of the health data used to make these estimates is often too limited compared to the spatial and temporal heterogeneity of the population of interest for traditional methods to produce reliable estimates. To address these difficulties, different data types and sources are frequently harmonized to achieve reliable high-fidelity estimates and their associated uncertainty. This dissertation describes methods and models used to perform Bayesian spatial-temporal smoothing to leverage the complete set of sparse health data to make granular predictions across the space-time domain of interest. In particular, we provide a statistical introduction to Template Model Builder, a flexible inferential tool for mixed effects model estimation that proves to be well-suited to spatial-temporal applications, and use it to jointly estimate European breast cancer incidence and mortality with data from local cancer registries and national databases. We conclude with a discussion of the limitations of interpreting mixed effects model uncertainty intervals and propose a novel unbiased coverage probability estimator that can be used to aid in dissemination and interpretation of the results from these models.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

AD: Automatic (or algorithmic) Differentiation

AR1: Autoregressive process of order 1

BHM: Bayesian Hierarchical Model

BYM: Besag-York-Mollie

BYM2: A modern extension to the original BYM

CP: (Frequentist) Coverage Probability

EB: Empirical Bayes

GMRF: Gaussian Markov Random Field

GP: Gaussian Process

IARC: International Agency for Research on Cancer

ICAR: Intrinsic Conditional Autoregressive

IID: Independent and Identically Distributed

INLA: Integrated Nested Laplace Approximation

LA: Laplace Approximation

LGM: Latent Gaussian Model

MCMC: Markov Chain Monte Carlo

MMAP: Marginal Maximum a Posteriori

MEM: Mixed Effects Model

MI: Mortality-Incidence (ratio)

MLE: Maximum Likelihood Estimation

MSE: Mean-Square Error

SAE: Small Area Estimation

ACKNOWLEDGMENTS

First, I would like to thank my advisor, Jon Wakefield, for his guidance, insight, enthusiasm, and good humor across the years. In addition, I would like to thank him for reaching out to me at a time of great difficulty in my life during which I was unsure if I would continue or complete this dissertation and my PhD. It is no small stretch to say that this project would not be what it is without his help, and it is not at all impossible that it may never have been completed except for his kindness and generosity.

Next, I would like to thank my committee members. I would like to thank Mark Ellis for his engagement and interest despite the distance between parts of this work and his own area of expertise. I would like to thank Simon Hay for his practical insights and thoughts and for helping keep the theory and methods from getting in the way of the real-world applicability of these chapters. I would like to thank Jim Hughes for helping connect parts of my work with relevant ideas and concepts that I was unaware of and for gently pushing me to deepen my understanding even further. I would like to thank Alex Luedtke for his clarity and precision, and for his encouragement to improve my own. Lastly, I would like to thank Bobby Reiner for teaching me to interpret all aspects of my estimates, particularly through the use of well thought out and clear visualizations. I would also like to thank Simon and Bobby for our years of work together during which I learned an immense amount from them and during which I believe we completed some very good work.

From the Department of Statistics at the University of Washington, I would like to thank Peter Guttorp and June Morita for their support and mentorship across the years. I would like to thank Ellen Reynolds, Vickie Graybeal, and Mee Ling Hon for their kindness, understanding, and help navigating the Graduate School and Departmental requirements

and for keeping me from falling off the rails. I would also like to thank my fellow students from Statistics and, in particular, Jan Irvahn for his friendship, for his demonstration of a successful and healthy work/life balance, and for our many enjoyable rounds of disc golf.

I would like to thank my collaborators Laina Mercer at PATH and Jacques Ferlay, Marytn Plummer and Freddie Bray from the International Agency for Research on Cancer for sharing their data, their data, and their expertise on cancer and cancer surveillance data, and their deep interest in advancing global health research.

To the extended 5210 squad, thank you for always being there, for the many adventures, amazing conversations, relaxed afternoons and long evenings, and for making my time in Seattle one of the most loved times of my life. To my friends from IHME, thank you for your excellence on and dedication to our projects, for making a (sometimes stressful) workplace a place I enjoyed, and most importantly for the many amazing lunches and dinners we shared together. In particular, I would like to thank Roy Burstein for his friendship, for rolling up his sleeves and diving into countless problems with me, for showing me how to fearlessly code, for his generosity sharing his global health knowledge and most everything else, for classical ethical and far futuristic conversations together, and for sharing your space with me even though I cannot help but chew loudly.

For their love, encouragement, and for never once doubting that I would finish my doctorate, I would like to thank my sisters, Hannah, Samantha, and Eliana. I would like to thank Wayne for reminding me again and again that an only good thesis is a done thesis. During the final year of this effort I was extremely fortunate to meet a new family member: our son Asa. I would like to thank you for helping me find slowness and wonder in a busy year and for providing me the encouragement I needed to get each day's work done to ensure I could spend time with you. Day after day, this yielded real progress, and thinking about our future together helped me finish laying this foundation for our family's long-term trajectory.

I would like to thank my wife, Logan Osgood-Zimmerman, who has been exceedingly

patient with me. I could not wish for a more caring, supportive and encouraging friend and partner, and I am extremely grateful to have had you by my side to celebrate with me during my successes and to lift me up in the darker moments. Thank you for letting me finish at my own pace but for encouraging me to get it done, thank you for helping me to see how to be proud in myself, thank you for taking charge on a cross-country move (with an infant!) when I was still wrapping this work up, and thank you for concurrently cheering me through to the finish while I needed it the most.

There are many other friends who touched my life during this time. Thank you for the love and fun you shared with me.

DEDICATION

In memory of my father, George Richard Zimmerman.
Thank you for your love and thank you for your enthusiasm and excitement in my studies.
When I thought I was out of all three, you showed me how to find them again.

And, in memory of the best couch companion to ever grace this plane.
Cosmo, you spent more hours encouraging me
than anyone can imagine.

Chapter 1
INTRODUCTION

Statistics is the study and application of collecting, analyzing, and interpreting data. When applied without clear motivation and guidance, these actions are capable of producing noise or even misinformation. In this third decade of the third millennium of the common era, we have the pleasure of seeing more well thought out statistical analyses than ever before while simultaneously having the need to navigate through the highest density of poorly produced statistics humankind has ever output. By traversing this explosion of information, the global community has found great success in many endeavors, but there are many challenges that we continue to face. One of the challenges we are obligated to address, as set forth in Article 25 of the Universal Declaration of Human Rights, is the mandate to promote and protect the health of all people:

> *Everyone has the right to a standard of living adequate for the health and well being of [themself] and of [their] family, including food, clothing, housing and medical care.*

With nearly 8 billion people alive today, this is an immense but not insurmountable task. The requisite foundation of this effort and any attempt to effectively improve the health and wellbeing of a population is an understanding of the current state of health within the population. Better yet would be to have an understanding of how multiple important health indicators vary across the geography in which the population resides and how these indicators have changed across time. Ideally, every person could have their health assessed and a clear understanding of the population's health would follow. In practice, these measurements are expensive to undertake and the volume and accuracy of the available health data is rarely sufficient to produce detailed estimates at any particular point in space and time. Spatial-

temporal statistics, including the subfield of *geostatistics* for continuous indexed geographies and the subfield of *small area estimation* for discretely indexed geographies, provides a suite of methods and models whose aim is to extrapolate between observations indexed in space and or time to densely predict at unobserved locations. These methods allow, for example, estimation of population characteristics at granular resolution even if some or many of the areas of interest were insufficiently sampled to produce estimates using traditional statistical methods.

This dissertation follows in the footsteps of the many well-established spatial-temporal disease mapping methods which have been used to synthesize disparate data sources. Their aim, and ours, is to produce the best understanding, via estimates and their associated uncertainty, we have of the current state of health indicators and how they vary across populations, locations and times.

1.1 Methodological Contributions of Dissertation

There are two primary contributions presented in Chapter 3. The first is a detailed statistical description of a random effects model fitting tool named Template Model Builder (TMB). TMB was developed recently and in the last few years it has received significant interest and use, particularly in ecological fields in which its authors are active. While exceptionally flexible, powerful, and fast compared to many other software packages which implement random effects models, the authors, not being statisticians, have never described it in significant statistical detail and this has limited its uptake among statisticians. The second contribution is a series of large-scale simulation studies which assess TMB's viability as an inferential tool for spatial statistics models. In addition, the simulations make up the largest study of the approximate stochastic partial different equations (SPDE) approach to Gaussian Process inference. In total, the result of this chapter is a descriptive study of TMB demonstrating its utility and appropriateness in a wide-range of spatial modeling situations.

The primary methodological contribution of the work described in Chapter 4 is to develop and implement a modeling framework which can jointly model cancer incidence and

mortality using data from local and national sources and which is indexed across age, space, and time. The underlying model formulation is taken from a previously proposed framework (Mercer, 2016, Chapter 5) which described a joint model of breast cancer incidence, mortality conditional on incidence, and unconditional mortality. At the time the initial framework was developed, the authors were unable to satisfactorily fit their full model proposal using the inferential tools available to them. We extend the model formulation to allow for interactions between age, space, and time and successfully implement the complete model, which includes a nonlinear interaction between incidence and unconditional mortality, using Template Model Builder. The result is a descriptive and flexible model of cancer incidence and mortality which coherently shares information across age groups, countries, and years to produce dense estimates and reliable uncertainty intervals for all 18 ages, 18 years, and 39 countries, regardless of the age-country-time strata's available data. Incidence estimates are typically most useful for planning purposes but the incidence reporting lags behind available mortality data. We demonstrate the models ability to use recent mortality data to backcast years with no incidence data.

The primary contributions of Chapter 5 focus on improving the interpretability of mixed effects model uncertainty intervals and are methodological and pedagogical in nature. We propose a novel tool to aid in the interpretation of mixed effects model uncertainty interval including Bayesian credible intervals: estimating and disseminating the frequentist coverage probability of the uncertainty intervals. For the small area estimation Fay-Herriot model with Gaussian sampling distribution, we demonstrate that the naive estimator for the coverage probability of area-specific intervals is biased and develop an unbiased estimator for the coverage probability. We conclude with an applied example demonstrating the use of the estimator and show how it provides relevant information that is frequently overlooked. The result is a discussion on the need to improve general understanding and interpretation of many popular mixed effects model uncertainty intervals along with a proposal and demonstration on how we can use estimated coverage probability as a means to this end.

1.2 Organization of Dissertation

Chapter 2 provides a review of the foundational statistical concepts which are used throughout the remainder of this dissertation. This includes an overview of Bayesian methods, Gaussian Markov Random Fields, and Gaussian Processes. The heart of the dissertation follows. In Chapter 3 we describe the statistical methods that underly Template Model Builder and conduct extensive spatial simulation studies. The joint estimation approach for estimating breast cancer incidence and mortality in the European Union is detailed and applied in Chapter 4. Chapter 5 demonstrates the utility in including the frequentist coverage probability of uncertainty intervals alongside publication of the intervals and develops an unbiased estimator for the coverage probability of the group means in Gaussian sampling random effects models. Finally, in Chapter 6, we conclude with a discussion of the dissertation and our plans for the future.

Chapter 2

BACKGROUND

2.1 Bayesian Inference and Computational Approaches

Bayesian inference aims to study the parameters of interest, $\boldsymbol{\theta}$, after having observed data, $\boldsymbol{y}$. Inference proceeds via the posterior distribution of the parameters of interest, which is derived using Bayes' Rule applied to a prior probability distribution on the parameters and a likelihood function defined by the sampling model of the observed data:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})} \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{2.1}$$

where the prior $p(\boldsymbol{\theta})$ is a probability distribution that encodes the modeler's beliefs about the parameters before data is collected and the likelihood $p(\boldsymbol{y}|\boldsymbol{\theta})$ defines the joint probability of the data observations conditional on the parameters. Generally, it is not guaranteed that a closed form expression for the posterior distribution will be available. Much of the difficulty of performing Bayesian inference can be attributed to finding (computationally) efficient means to study the posterior despite its lack of an analytical form.

The simple model and posterior defined in (2.1) by the sampling layer and prior can be viewed as two layers of a Bayesian model, and we refer to such models as *Bayesian hierarchical models* (BHMs). These can be extended in many ways. In the context of time series modeling, Berliner (1996) introduced a general three-layer description of BHMs which

is widely applicable and which we find useful throughout this dissertation:

$$\text{Stage 1 - Data:} \quad [\text{data} \mid \text{process},\ \text{parameters}] \tag{2.2}$$

$$\text{Stage 2 - Process:} \quad [\text{process} \mid \text{parameters}] \tag{2.3}$$

$$\text{Stage 3 - Hyperpriors:} \quad [\text{parameters}]\,. \tag{2.4}$$

Together these three stages form a BHM, and the posterior distribution, our target of inference, can be found from them using Bayes' Rule:

$$\begin{aligned} \text{Posterior:} \quad & [\text{process},\ \text{parameters} \mid \text{data}] \\ &= \frac{[\text{data} \mid \text{process},\ \text{parameters}]\,[\text{process} \mid \text{parameters}]\,[\text{parameters}]}{[\text{data}]} \\ &\propto [\text{data} \mid \text{process},\ \text{parameters}]\,[\text{process} \mid \text{parameters}]\,[\text{parameters}]\,. \end{aligned} \tag{2.5}$$

One all purpose tool used to analyze the posterior distribution is Markov Chain Monte Carlo (MCMC) sampling algorithms (Hastings, 1970; Gelfand and Smith, 1990; Robert and Casella, 2011). These methods are typically used to calculate numerical approximations to multi-dimensional integrals. The study of MCMC for Bayesian inference is quite rich and beyond the scope of this dissertation. The primary reason we do not go into further detail on MCMC is that inference for spatial random effects models is notoriously difficult, due to the dimensionality of the parameter space and the dependencies in the likelihood×prior surface. Inference using MCMC can be difficult to implement and even with helpful techniques, such as block updating, MCMC may be quite slow to run and convergence of the sample chains can be difficult to diagnose (Knorr-Held and Rue, 2002; Filippone *et al.*, 2013; Margossian *et al.*, 2020). Modern sampling techniques, such as those implemented in `Stan` (Carpenter *et al.*, 2017), have improved the feasibility of using MCMC for discrete spatial models inference, but continuous spatial models continue to lack a scalable implementation. Chapter 3 details two different approximate techniques for Bayesian inference which, as we will see, perform

well for spatial and temporal models. Those two tools, Template Model Builder (TMB) and Integrated Nested Laplace Approximations (INLA) will be used throughout this dissertation, but we leave their main introduction to Chapter 3.

2.2 Gaussian Markov Random Fields

Gaussian Random Fields (GRFs) are simple constructs consisting of a finite-dimensional random vector which follows a multivariate Gaussian distribution. There are many ways to incorporate spatial smoothing structure into such models. One of the most popular methods, pioneered by Besag *et al.* (1991), is to assume a more restrictive form of GRF which satisfies additional *conditional independence* constraints. The conditional independence assumptions encode a type of memory-less property in the random vector such that the state of elements in the vector only depends on the state of its neighbors and is independent, conditional on the neighbors, from all non-neighbor elements. The addition of such a constraint to a GRF restricts it to the subclass of *Gaussian Markov Random Fields* (GMRFs).

Consider a random vector $\boldsymbol{x} = (x_1, x_2, x_3)^T$. Definitionally, x_1 and x_3 are conditionally independent given x_2 if, having already observed x_2 to take on a particular value, observing x_3 provides no new information about the distribution of x_1. That is, conditional independence implies that the density $p(x_1|x_2, x_3)$ simplifies to and is identical to $p(x_1|x_2)$ and we denote this property by $x_1 \perp\!\!\!\perp x_3|x_2$.

Now, let $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ have a normal distribution with mean $\boldsymbol{u}$ an $n \times 1$ vector and $n \times n$ covariance matrix $\boldsymbol{\Sigma}$ and precision matrix $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$ such that the density of $\boldsymbol{x}$ is:

$$\pi(\boldsymbol{x}) = (2\pi)^{-n/2}|\boldsymbol{Q}|^{1/2}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{u})^T\boldsymbol{Q}(\boldsymbol{x}-\boldsymbol{u})\right), \quad \boldsymbol{x} \in \mathbb{R}^n. \tag{2.6}$$

Consider an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with labeled vertices $\mathcal{V} = \{1, \ldots, n\}$ and edges $\mathcal{E}$ defined such that there is no edge between vertices i and j if and only if $x_i \perp\!\!\!\perp x_j|\boldsymbol{x}_{-ij}$ where $\boldsymbol{x}_{-ij}$ denotes all elements of $\boldsymbol{x}$ except the i^{th} and j^{th}. If such a graph $\mathcal{G}$ exists for $\boldsymbol{x}$, then we say that $\boldsymbol{x}$ is a GMRF with respect to $\mathcal{G}$.

The power and utility of the Markov property of a GMRF becomes apparent with the following theorem:

Theorem 1. *If* $\boldsymbol{x}$ *is normally distributed with mean* $\boldsymbol{u}$ *and precision matrix* $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$*, then for* $i \neq j$*,*

$$x_i \perp\!\!\!\perp x_j | \boldsymbol{x}_{-ij} \iff Q_{ij} = 0.$$

See, for example, (Rue and Held, 2005, Section 2.2) for a proof. In practice, this is extremely useful since the conditional independence of the GMRF vector is encoded in its precision matrix. For problems where n is large, but the number of neighbors for each element is relatively small, this formulation yields sparse precision matrices which can be used for efficient computation. Furthermore, the GMRF assumption makes calculations to find the conditional mean and conditional precision straightforward and computationally efficient:

$$\begin{aligned} \mathrm{E}[x_i | \boldsymbol{x}_{-i}] &= \mu_i - Q_{ii}^{-1} \sum_{j:j\sim i} Q_{ij}(x_j - \mu_j), \\ \mathrm{Prec}[x_i | \boldsymbol{x}_{-i}] &= Q_{ii}^{-1}, \end{aligned}$$

where $j \sim i$ denotes that vertex j is a neighbor of i and thus $\{i, j\} \in \mathcal{E}$.

2.2.1 Autoregressive process of order 1

As a first example, and one that we will use throughout the dissertation for discretely indexed one-dimensional indices (time and age), we consider an *autoregressive process of order 1* (AR1) with standard normal errors. The name comes from a formulation of the process where the element at the next index is a function only of the element at the current index:

$$x_t = \phi x_{t-1} + \epsilon_t, \quad \epsilon_t \overset{iid}{\sim} N(0, 1),$$

where $|\phi| < 1$ is the autocorrelation parameter and index t is a discrete index with $t = \{1, \ldots, n\}$. The graph for the AR1 GMRF may be found at the bottom of Figure B.3.1. To ensure that each element of $\boldsymbol{x}$ has the same marginal variance, the marginal distribution of the initial state is taken to be $x_1 \sim N(0, (1-\phi^2)^{-1})$. The joint distribution of this process can then be written as

$$\begin{aligned} \pi(\boldsymbol{x}) &= \pi(x_1)\pi(x_2|x_1)\ldots\pi(x_n|x_{n-1}) \\ &= (2\pi)^{-n/2}|\boldsymbol{Q}|^{1/2}\exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x}\right), \end{aligned} \tag{2.7}$$

with precision matrix

$$\boldsymbol{Q} = \begin{pmatrix} 1 & -\phi & & & & & \\ -\phi & 1+\phi^2 & -\phi & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & & -\phi & 1+\phi^2 & -\phi & \\ & & & & -\phi & 1 \end{pmatrix}.$$

The precision is tridiagonal and quite sparse, since all non-adjacent indices being conditionally independent given the indices between them. As we mentioned, the sparse nature of the precision matrices offers the opportunity for fast computation even though the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{Q}^{-1}$ is dense. An alternative but identical formulation using the full conditionals (as opposed to the directed conditionals shown in (2.7) explicitly shows the smoothing

structure that this AR1 GMRF provides within its conditional expectations:

$$x_t|\boldsymbol{x}_{-t} \sim \begin{cases} N(\phi x_{t+1}, 1), & \text{if } t = 1 \\ N\left(\frac{\phi}{1+\phi^2}(x_{t-1} + x_{t+1}),\ (1+\phi^2)^{-1}\right), & \text{if } 1 < t < n \\ N(\phi x_{n-1}, 1), & \text{if } t = n. \end{cases} \tag{2.8}$$

We see that the conditional expectation of t, in the case where it has two neighbors, is the scaled sum of its two neighbors. As ϕ approaches 1, the conditional expectation approaches the average of its neighbors and as it approaches zero we see the process devolves into iid noise. For the elements that only have one neighbor, we see that the conditional expectation smooths between a random effect of zero and its neighbor, effectively smoothing to the fixed effects mean of the model. The structured GMRFs (and Gaussian Processes) we use in this dissertation all exhibit some form of this "smoothing". Their precision encodes that random elements that are nearby (or that are neighbors) should be more alike and this carries through to the forms of the conditional expectation.

2.2.2 Besag-York-Mollié

As we move beyond one-dimensional indices, we very often find ourselves in situations where there is no natural ordering of the indices. Relevant examples include indexing countries within Europe or indexing latitude-longitude point-referenced observations within a continuous domain. In these situations, it is no longer particularly useful to factor the joint density of $\boldsymbol{x}$ in some order, as we saw in the one-dimensional setting in (2.7), and commonly the joint distribution is implicitly specified by the set of the n full conditional distributions. This approach was made popular by Besag *et al.* (1991) when they proposed the *intrinsic conditional autoregressive* (ICAR) smoothing model:

$$u_i|\boldsymbol{u}_{-i}, \tau_u \sim N\left(\frac{1}{n_i}\sum_{j:j\sim i} u_j,\ (\tau_u n_i)^{-1}\right),$$

where τ_u is the global precision of the field and n_i the number of neighbors of index i. We see that the conditional mean again implies smoothing across neighbors since it is the average value of element i's neighbors and the conditional variance of u_i is the global precision appropriately scaled by the number of neighbors, n_i, which contribute to the conditional mean.

The ICAR model, while quite useful and popular in small area estimation applications, has a number of drawbacks. First, the precision of the ICAR is rank-deficient so using the ICAR prior does not lead to a proper posterior density. Generally, we will use the term *intrinsic* to denote models with rank-deficient precision. Frequently, an extra independent and identically distributed (iid) Gaussian term, $v_i \sim N(0, \sigma_v^2 = \tau_v^{-1})$, is added to each index so that the field total for each index i is $b_i = u_i + v_i$. The iid term helps ensure that not all of the unstructured error in the observations is attributed to the ICAR process which generally helps avoid overfitting. With this addition, the total variance now consists of the sum of the ICAR and iid variances:

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

where $\boldsymbol{Q}^-$ denotes the generalized inverse of the precision. This formulation consisting of the sum of the ICAR and iid components is popularly referred to as the Besag-York-Mollié (BYM) model. While the iid term helps attenuate overfitting, it comes at a cost. Since the structured and unstructured components are summed and cannot be observed independently from one another, they are not generally not identifiable, and it can be difficult for inferential procedures to tease these variances apart. A second difficulty with this model is that, τ_v^{-1} and τ_u^{-1} represent different scales of variability with

- τ_v^{-1} interpreted as the *marginal* variance of the unstructured random effects, and
- τ_u^{-1} interpreted as controlling the variability of the structured components, u_i, *conditional* on the effects in its neighboring areas.

To address these and other issues, Riebler *et al.* (2016) proposed an updated version of the BYM model which they refer to as the BYM2. Others (Leroux *et al.*, 2000; Dean *et al.*, 2001) have attempted to address the identifiability issue by reparameterizing the BYM model to have one parameter that controls the total variance, and a second that controls the decomposition of the total variance into structured and unstructured components. Neither of these models solve the second issue of scaling.

The second issue with ICAR (and, more generally, all intrinsic GMRF) models involves difficulties in appropriately handling the scaling of the structured variance component. Scaling is crucial in the assignment of hyperpriors within and across models and without appropriate consideration for the scaling the hyperpriors may have unexpected influence. From Riebler *et al.* (2016):

> the Besag model is an intrinsic GMRF which penalizes local deviation from its null space, which is a constant level in the case of one connected component. The hyperprior will control this local deviation and, as such, influence the smoothness of the estimated spatial effects. If the estimated field is too smooth, the precision is large and potential spatial variation might be blurred. On the other hand, if the precision is too small the model might overfit due to large local variability.

In general, the marginal variances for regions, $\tau_b^{-1}[\boldsymbol{Q}_{ii}^{-1}]$ depend on the graph structure encoded in $\boldsymbol{Q}$. Riebler *et al.* (2016) demonstrate this by noting that the generalized variance of GMRF $\boldsymbol{u}$,

$$\begin{aligned}\sigma^2_{GV}(\boldsymbol{u}) &= \exp\left(\frac{1}{n_s}\sum_{i=1}^{n_s}\log\left(\frac{1}{\tau_b}[\boldsymbol{Q}^{-1}]_{ii}\right)\right)\\ &= \frac{1}{\tau_b}\exp\left(\frac{1}{n_s}\sum_{i=1}^{n_s}\log\left([\boldsymbol{Q}^{-1}]_{ii}\right)\right),\end{aligned} \tag{2.9}$$

defined to be the geometric mean of the marginal variances (though they could have used any other standard measure of central tendency) can be different for two different graph structures, even if the graphs have the same number of regions.

If a scaling of the ICAR component could be found such that $\text{Var}(u_i) \approx 1$ for all vertices

i, then the BYM model with structured and unstructured components could be reparameterized with single interpretable variance component. Generally, unscaled ICAR models do not exhibit this variance behavior and based on (2.9), Riebler *et al.* (2016) suggest multiplicatively scaling the ICAR precision matrix by the geometric mean of the unscaled variance components shown in (2.9).

Letting $\boldsymbol{Q}_\star$ represent precision of the BYM ICAR scaled by (2.9), the total (combined structured and unstructured components) BYM2 model can be written as

$$\boldsymbol{b} = \frac{1}{\sqrt{\tau_b}}\left(\sqrt{1-\varphi}\boldsymbol{v} + \sqrt{\varphi}\boldsymbol{u}^\star\right) \tag{2.10}$$

with covariance

$$\text{Var}\left(\boldsymbol{b}|\tau_b, \varphi\right) = \tau_{b^{-1}}\left((1-\varphi)\boldsymbol{I} + \varphi\boldsymbol{Q}_\star^{-1}\right), \tag{2.11}$$

where $\tau_b \geq 0$ is the global precision of the field, $\varphi \in [0,1]$ models how much of the total variance is spatially structured (instead of unstructured), $\boldsymbol{v} \sim N(0, \boldsymbol{I})$ is the unstructured iid random effect with variance 1, and $\boldsymbol{u}^\star$ is the ICAR component scaled such that $\text{Var}(u_i^\star) \approx 1$ for all i.

With the variance components appropriately scaled, it is possible to interpret and set meaningful priors on the combined variance parameter, τ_b, and the field decomposition parameter, φ, in the BYM2. We use the BYM2 model for discrete spatial modeling in Chapters 3 and 4.

2.3 Gaussian Processes

A *stochastic process* is often defined as a family of random variables defined on the same probability space. A *Gaussian Process* (GP) is a stochastic process which is continuously indexed (typically in space or in time) and for which every finite collection of those random variables has a multivariate Gaussian distribution with the density shown in (2.6). The full

distribution of the GP is the joint distribution of all the random variables and accordingly it is a distribution over continuous-domain functions.

Consider a GP $\{u(s), s \in \mathcal{S}\}$ over continuous index s in domain $\mathcal{S}$. For a spatial model, typically one assumes $\mathcal{S} \subset \mathbb{R}^2$, though other options exist such as considering a two-dimensional process on the surface of a three-dimensional sphere. Definitionally, for every finite set of indices $s_1, \ldots, s_n$ within the index set $\mathcal{S}$, $\boldsymbol{u}(\boldsymbol{s}) = (u(s_1), \ldots, u(s_n))^T$ must be a multivariate Gaussian random variable. We denote its mean by $\boldsymbol{\mu}(\boldsymbol{s}) = (\mu(s_1), \ldots, \mu(s_n))^T$ and its covariance with $\boldsymbol{\Sigma}(\boldsymbol{s})$. The covariance encodes the relationship between the process at different indices and it is used to define the smoothness inherent in the process. In spatial and temporal statistics, an assumption is often made that observations that are "closer" together should be more alike. In practice, this often means a covariance kernel, $C(u(s_i), u(s_j)) = \Sigma_{ij}$, is selected where the correlation between two indices decays as the distance between them increases. The covariance also encodes other features of the process including *stationarity* and *isotropy*. If the covariance kernel is unaffected by additive shifts, that is if $C(u(s_i), u(s_j)) = C(u(s_i + \delta), u(s_j + \delta))$, the process is said to be stationary. If the covariance kernel depends only on the distance between two indices, the process is said to be *isotropic*. That is, the process is isotropic if for the chosen distance function $d(\cdot)$, $C(u(s_i), u(s_j)) = C\left(d(s_i, s_j)\right)$. Covariance functions which are non-stationary and anisotropic are more complex to parameterize and requires larger volumes of data to estimate them well when compared to stationary isotropic models. The health applications which are the focus of this dissertation often have relatively sparse observations and we restrict ourselves to working with stationary isotropic covariances.

In particular, we use the Matérn covariance which was popularized by Stein (1999). This covariance is quite flexible and with a small number of parameters it allows the smoothness (in terms of mean-square differentiability), magnitude, and autocorrelation of the process to be specified. The Matérn is stationary and isotropic and defines the covariance between two

spatial locations at distance $||s_i - s_j||$ from one another to be:

$$C(u(s_i), u(s_j)) = \frac{\sigma_{\mathrm{m}}^2}{2^{\nu-1}\Gamma(\nu)} \left(\kappa||s_i - s_j||\right)^{\nu} K_{\nu} \left(\kappa||s_i - s_j||\right), \tag{2.12}$$

where σ_{m}^2 is the variance, $\kappa > 0$ is a scaling parameter related to the range, $r_m = \frac{\sqrt{8\nu}}{\kappa}$ which is defined to be the distance at which the spatial correlation drops to 0.1, $\nu > 0$ is related to the smoothness of the field, and K_ν is the modified Bessel function of the second kind.

2.4 Higher-Dimensional Models

So far we have considered spatial models and temporal models. These GPs and GMRFs that have been discussed could be summed to allow, for example, a model with a linear predictor which has spatial variation which is constant across time in addition to temporal variation which is constant across space. We could also desire a model that allows independent spatial processes at different time points, or one that allows independent temporal processes at different spatial locations, or one that allows the space and time components to interact such that the spatial process evolves slowly in time (or, similarly, that a temporal process smoothly varies across space). To build these sorts of dimensional interactions, we follow the suggestion of Clayton (1996) and build processes for the higher order interactions by taking the Kronecker product of covariances of the main effects processes to be the covariance of the interaction process. The Kronecker product is a generalization of an outer product and results in a block matrix. For matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, their Kronecker product, denoted with $\otimes$, is

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix}.$$

As an example, if $\boldsymbol{\Sigma}_s$ is the covariance of a spatial GMRF and $\boldsymbol{\Sigma}_t$ the covariance of a

temporal GMRF, then we can build a space-time GMRF by setting its covariance to be $\mathbf{\Sigma}_s \otimes \mathbf{\Sigma}_t$. By building interactions of unstructured and structured covariances, we can build any of the types of interactions previously mentioned. For example, an iid spatial covariance interacted with a structured temporal covariance will yield spatially independent temporal effects. The 4 different types of space-time interactions that can be created by taking the Kronecker product of structured and unstructured space and time effects are described by Knorr-Held (2000). In Chapter 4 we extend this idea to interactions across age, space, and time.

Chapter 3

A STATISTICAL REVIEW OF TEMPLATE MODEL BUILDER: A FLEXIBLE TOOL FOR SPATIAL MODELING

3.1 Introduction

The integrated nested Laplace approximation (INLA) is a well-known and popular technique for spatial modeling with a user-friendly interface in the `R-INLA` package. Unfortunately, only a certain class of latent Gaussian models are amenable to fitting with INLA. In this chapter we review Template Model Builder (`TMB`), an existing technique and software package which is well-suited to fitting complex spatio-temporal models. `TMB` is relatively unknown to the spatial statistics community, but it is a flexible random effects modeling tool which allows users to define customizable and complex mixed effects models through `C++` templates. After contrasting the methodology behind `TMB` with INLA, we provide a large-scale simulation study assessing and comparing `R-INLA` and `TMB` for continuous spatial models, fitted via the Stochastic Partial Differential Equations (SPDE) approximation. The results show that the predictive fields from both methods are comparable in most situations even though `TMB` estimates for fixed or random effects may have slightly larger bias than `R-INLA`. We also present a smaller discrete spatial simulation study, in which both approaches perform well. We conclude with an applied example of a joint analysis of breast cancer incidence and mortality data using a model which cannot be fit with INLA. The breast cancer example is greatly expanded in Chapter 4.

Random effects models, broadly speaking, follow an archetypical structure:

$$\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\phi}_1 \sim p_1(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\phi}_1) \tag{3.1}$$

$$\boldsymbol{b}|\boldsymbol{\phi}_2 \sim p_2(\boldsymbol{b}|\boldsymbol{\phi}_2) \tag{3.2}$$

where $\boldsymbol{y}$ represents n data observations, p_1 and p_2 are the likelihood and random effects distributions, respectively, $\boldsymbol{\beta}$ represent fixed effects, $\boldsymbol{b}$ random effects and $\boldsymbol{\phi} = [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2]$ variance-covariance parameters, with $\boldsymbol{\phi}_1$ appearing in the likelihood and $\boldsymbol{\phi}_2$ in the prior. The random effects, $\boldsymbol{b}$, may be split into a set of spatial random effects $\boldsymbol{u}$ and non-spatial random effects $\boldsymbol{v}$ so that $\boldsymbol{b} = [\boldsymbol{u}, \boldsymbol{v}]$. We consider Gaussian spatial random effects, commonly used across many applications, so that $\boldsymbol{b}|\boldsymbol{\phi}_2$ falls within the class of *latent Gaussian models* (LGMs). If a Bayesian approach to inference is desired, the model, comprised of (3.1) and (3.2), is completed by adding the hyperprior, $p_3(\boldsymbol{\beta}, \boldsymbol{\phi}) = p_3(\boldsymbol{\beta})p_3(\boldsymbol{\phi})$, and we assume the prior for $\boldsymbol{\beta}$ is Gaussian. A detailed summary of the dimensionality of these variables is provided in Table 3.1.

Spatial statistics concerns itself with the analysis of data which consists, at minimum, of geographical information associated with the measurement. The geographic label itself may be of intrinsic interest, or it may be used to map between a stochastic model for the phenomenon of interest and the data observations. In both cases, the spatial model component, specified by $\boldsymbol{u}|\boldsymbol{\phi}_2$, introduces spatial dependence between measurements which accounts for heterogeneity across the modeling region. Traditionally, a distinction is made between settings which use discrete (areal) spatial locations, for example data measured across counties, and processes which use continuous (point-referenced) spatial locations, like GPS coordinates (Gelfand *et al.*, 2010; Banerjee *et al.*, 2014). In both, a measure of distance between locations is required, and, typically, correlation between the spatial process at two locations is assumed to decrease as the distance between the locations increase.

Inference for spatial random effects models is notoriously difficult, due to the dimensionality of the parameter space and the dependencies in the likelihood$\times$prior surface. Inference using Markov chain Monte Carlo (MCMC) techniques can be difficult to implement, requires large amounts of memory (RAM), and even with block updating run times remain long and diagnosing convergence can be difficult (Knorr-Held and Rue, 2002; Filippone *et al.*, 2013; Margossian *et al.*, 2020). Modern sampling techniques, such as those implemented in `Stan` (Carpenter *et al.*, 2017), have improved the feasibility of using MCMC for discrete spatial

models inference, but continuous spatial models continue to lack a scalable implementation. In the absence of viable MCMC solutions, inference typically relies on analytic approximation methods, including variational methods (Ren *et al.*, 2011) and those based on the *Laplace approximation* (LA) popularized by Tierney and Kadane (1986). Spatial LGMs have received much attention due to the computational advantages they proffer and a suite of approximations specific to these models have been developed. Markov assumptions often make discrete models computationally tractable by inducing sparsity in the precision matrices. In the continuous spatial setting, approximate models tend to focus on increasing the sparsity of the covariance or precision matrices and approaches include covariance tapering, fixed rank kriging, lattice kriging, discrete process convolutions, predictive processes, and a *stochastic partial differential equations* (SPDE) approximation (see Heaton *et al.*, 2019 for a thorough overview and comparison). In practice, many of these model approximation techniques, such as lattice kriging (Nychka *et al.*, 2015) and the SPDE approach (Lindgren *et al.*, 2011) may be used in conjunction with approximate methods to perform efficient inference on the desired (approximate) model formulation.

Note the distinction between approximate models and approximate inferential methods. In this chapter, we focus on two approximate methods that use the LA to marginalize over the latent Gaussian variables. These methods perform inference on the desired (exact or approximate) model, and both are capable of leveraging parallelization across processors and sparse matrix routines if sparsity is present in the model. Laplace approximations have a long history in Bayesian computation (Tierney and Kadane, 1986) and were revitalized with the advent of the integrated nested Laplace approximation (INLA) method implemented in the `R-INLA R` package (Rue *et al.*, 2009; Martins *et al.*, 2013). Since its introduction, INLA has grown in popularity and is now the method of choice for many spatial and spatio-temporal analyses. The method is very well-documented (Blangiardo *et al.*, 2013; Lindgren and Rue, 2015; Rue *et al.*, 2017; Martino and Riebler, 2020) and its popularity is evidenced by a number of book-length treatments specifically for spatial data (Blangiardo and Cameletti, 2015; Krainski *et al.*, 2018; Moraga, 2019; Gomez-Rubio, 2020) along with numerous applications.

However, the R implementation of INLA does not offer complete flexibility with respect to the likelihood specification.

In this chapter we provide a detailed statistical overview of the lesser known LA random effects modeling R package, Template Model Builder (TMB), developed by Kristensen *et al.* (2016). TMB is exceptionally flexible, allowing the user to define custom models within a C++ template, is computationally efficient, and is applicable to a wide class of sampling models. Strongly inspired by AD Model Builder (ADMB) (Fournier *et al.*, 2012; Bolker *et al.*, 2013), base TMB, the focus of this chapter, provides a powerful platform to calculate deterministic approximations to marginal densities. Like ADMB, TMB also leverages automatic differentiation (AD) to improve the efficiency of its computations. TMB's general applicability has led to its use in other software development such as glmmTMB (Brooks *et al.*, 2017), which allows convenient fitting of generalized linear mixed models, tmbstan (Monnahan and Kristensen, 2018), which can stochastically sample from target distributions defined with TMB templates using no U-turn sampling (NUTS) MCMC and allows for fully Bayesian inference, and gllvm (Niku *et al.*, 2019), which allows for rapid fitting of generalized linear latent variable models to multivariate abundance data.

We focus on TMB since it is less well known to statisticians and the methods have received little attention, even though it is a popular tool in the ecological literature (Albertsen *et al.*, 2015; Thorson and Kristensen, 2016; Bolstad *et al.*, 2017; Thygesen *et al.*, 2017; Niku *et al.*, 2017; Free *et al.*, 2019). Our main contributions are a detailed review of the TMB methodology and the estimates it generates using statistical terms – such a complete description has been missing from the literature – and a detailed simulation study to vet the method. In addition, we provide source code to implement a variety of the most popular spatial models. The spatial simulation studies we use to assess TMB and R-INLA add to a growing body of literature showing the strengths and limitations of these methods (Taylor and Diggle, 2014; Ferkingstad *et al.*, 2015; Auger-Méthé *et al.*, 2017). They also extend the limited number of studies validating the stochastic partial differential equations (SPDE) approach to fitting continuous spatial models in non-Gaussian data settings (Teng *et al.*, 2017; Righetto *et al.*,

2020).

The structure of this chapter is as follows. In Section 3.2 we describe the models that we are considering, along with an overview of inferential approaches. Sections 3.3, 3.4 and 3.5 describe INLA and `TMB` and summarize their differences, respectively. Section 3.6 compares the two approaches, via extensive simulations, and Section 3.7 considers the modeling of European breast cancer data using an incidence/mortality model that can be fitted in `TMB`, but not in INLA. We conclude with a discussion in Section 3.8.

3.2 Inferential Overview

Table 3.1 provides an overview of the variables defined in (3.1) and (3.2). This includes the nomenclature used INLA an `TMB`, as well as the dimensions of variable vectors, which can greatly impact computation scaling.

			`TMB`	
			fixed effects $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\phi}_1, \boldsymbol{\phi}_2)$ $\dim(\boldsymbol{\theta}) = m + p + q$	random effects $\boldsymbol{b} = (\boldsymbol{u}, \boldsymbol{v})$ $\dim(\boldsymbol{b}) = l + k$
INLA	latent field	$\boldsymbol{B} = (\boldsymbol{\beta}, \boldsymbol{u}, \boldsymbol{v})$ $\dim(\boldsymbol{B}) = m + l + k$	$\boldsymbol{\beta}$ $\dim(\boldsymbol{\beta}) = m$	$\boldsymbol{u}$ $\quad$ $\boldsymbol{v}$ $\dim(\boldsymbol{u}) = l$ $\quad$ $\dim(\boldsymbol{v}) = k$
	hyper params	$\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2)$ $\dim(\boldsymbol{\phi}) = p + q$	$\boldsymbol{\phi}_1$ $\quad$ $\boldsymbol{\phi}_2$ $\dim(\boldsymbol{\phi}_1) = p$ $\quad$ $\dim(\boldsymbol{\phi}_2) = q$	

Table 3.1: Summary of the parameter combinations as used in `TMB` and `R-INLA` based on the two-stage model defined in (3.1) and (3.2). The bottom-right quadrant contains all the variables defined in these equations, the top-right quadrant defines how they are partitioned within `TMB`, and the bottom-left quadrant defines how they are partitioned within `R-INLA`. There are n data points, m regression coefficients, l spatial random effects, k non-spatial random effects, and $p + q$ hyperparameters.

The automatic differentiation (AD) at the heart of `TMB` allows various inferential approaches. Using the general random effects model formulated in (3.1) and (3.2), we define

$$f(\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\phi}) = \log p_1(y|\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\phi}_1) + \log p_2(\boldsymbol{b}|\boldsymbol{\phi}_2).$$

The simplest inferential approach for this model is maximum likelihood estimation (MLE) for the fixed effects and variance-covariance parameters (REML is also available for $\boldsymbol{\phi}$). The MLEs maximize the marginal likelihood,

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\phi}) = \int \exp\left[f(\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\phi})\right] \ \mathrm{d}\boldsymbol{b}, \tag{3.3}$$

which is often not available in a closed form. Inference may proceed via the asymptotic normal distribution of the MLE with uncertainty expressed via the observed information evaluated at the MLE, as in `TMB`'s methodological ancestor, ADMB (Fournier *et al.*, 2012). Within `TMB` one may also maximize the marginal posterior,

$$p(\boldsymbol{\beta}, \boldsymbol{\phi}|\boldsymbol{y}) \propto \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\phi}) \times p_3(\boldsymbol{\beta}) \times p_3(\boldsymbol{\phi}), \tag{3.4}$$

to produce marginal maximum a posteriori (MMAP) estimates for the fixed effects and variance parameters, and where inference will be based on the asymptotic distribution of the posterior. In either setting, inference for the random effects then occurs through empirical Bayes by maximization of $f(\widehat{\boldsymbol{\beta}}, \boldsymbol{b}, \widehat{\boldsymbol{\phi}})$ having conditioned on the parameter estimates, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\phi}}$.

`TMB`, initially developed for frequentist inference, differentiates between random and non-random effects, $\boldsymbol{b}$ and $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\phi})$. `R-INLA`, developed for Bayesian inference, differentiates layers of the hierarchical model. Table 3.1 illustrates the different ways these methods partition the parameters. All the parameters in the linear predictor (the right-hand side of the link function in (3.5)) are designated as latent field parameters, $\boldsymbol{B} = (\boldsymbol{\beta}, \boldsymbol{b})$, and all parameters defined in subsequent layers or stages, such as (3.6), are designated as hyperparameters, $\boldsymbol{\phi}$.

3.3 Integrated Nested Laplace Approximation

INLA has been well described in a number of statistical venues, including recently in Martino and Riebler (2020) and Rue *et al.* (2017). We provide a brief overview of its details here for comparison against the details of `TMB` presented in Section 3.4 and to aid in comparing the INLA and `TMB` simulation results presented in Section 3.6.

We first note that INLA does not use the "standard" Laplace approximation (which is used in `TMB`, as described in Section 3.4), but rather implements the Laplace ratio approximation (LRA) described in Tierney and Kadane (1986, Section 4.1) which benefits from some cancellation of approximation error.

INLA was introduced by Rue *et al.* (2009) to provide a quick option for Bayesian computation for the class of additive LGMs (ALGMs) and is ideally suited to handle problems where the number of hyperparameters, $\dim(\boldsymbol{\phi}) = p + q$, is kept small. In an ALGM, the general hierarchical model formulation described in (3.1) and (3.2) is restricted to models where the conditional expectation of the observations can be related to a linear combination of the fixed and random effects via a known link function $g()$:

$$\mathrm{E}[y_i|\boldsymbol{\beta}, \boldsymbol{b}] = g(\eta_i) = g\left(\beta_0 + \sum_{j=1}^{J} \beta_j z_{ij} + \sum_{k=1}^{K} b_i^{(k)}\right), \tag{3.5}$$

for observed covariates, z_{ij}, associated with the fixed effects, $\beta_j, j = 1, \ldots, J$, random effects, $\{b_i^{(k)},\ k = 1, \ldots, K\}$, and with η_i the linear predictor for each observation i. All the parameters of the linear predictor are assumed to be Gaussian, completing the LGM definition.

We introduce a slight change of model formulation, to accommodate the consolidation of like terms. Specifically, we collect together the Gaussian fixed and random effects, from (3.2), to write

$$p_2(\boldsymbol{\beta}, \boldsymbol{b}|\boldsymbol{\phi}_2) = p_2(\boldsymbol{b}|\boldsymbol{\phi}_2) \times p_3(\boldsymbol{\beta}).$$

Defining these terms to be $\boldsymbol{B} = \left[\boldsymbol{\beta}, \boldsymbol{b}^{(1)}, \ldots, \boldsymbol{b}^{(K)}\right]$, in the LGM setting we then have

$$\boldsymbol{B}|\boldsymbol{\phi}_2 \sim \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{Q}^{-1}(\boldsymbol{\phi}_2)\right) \tag{3.6}$$

where $\boldsymbol{Q}^{-1}(\boldsymbol{\phi}_2)$ is the precision matrix for the Gaussian field. This ensures that the linear predictor $\boldsymbol{\eta}$ is Gaussian as well. To complete the Bayesian model specification, the hyperprior, $p_3(\boldsymbol{\phi})$ is specified.

The primary targets of inference for the INLA algorithm are univariate posterior densities for the latent field parameters, $p(B_i|\boldsymbol{y})$, and the joint posterior of the hyperparameters, $p(\boldsymbol{\phi}|\boldsymbol{y})$. INLA approximates these in three steps:

1. Explore and discretize $\boldsymbol{\phi}$-space via an approximation,

$$\widetilde{p}(\boldsymbol{\phi}|\boldsymbol{y}) \propto \left.\frac{p_1(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\phi}) p_2(\boldsymbol{\beta}, \boldsymbol{b}|\boldsymbol{\phi}) p_3(\boldsymbol{\phi})}{\widetilde{p}_G(\boldsymbol{\beta}, \boldsymbol{b}|\boldsymbol{\phi}, \boldsymbol{y})}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*(\boldsymbol{\phi}), \boldsymbol{b}=\boldsymbol{b}^*(\boldsymbol{\phi})}, \tag{3.7}$$

   where $\widetilde{p}_G(\boldsymbol{\beta}, \boldsymbol{b}|\boldsymbol{\phi}, \boldsymbol{y})$ is the Gaussian approximation to the conditional distribution obtained by numerically finding and matching the mode, $\{\boldsymbol{\beta}^*(\boldsymbol{\phi}), \boldsymbol{b}^*(\boldsymbol{\phi})\}$, and curvature at the mode, for given $\boldsymbol{\phi}$.

The approximation in (3.7) is equivalent to the Laplace Ratio Approximation (LRA) for the posterior marginal proposed by Tierney and Kadane (1986). As the hyperparameter space is explored, (3.7) is evaluated at L high-density points to generate an approximate discretization: $\widetilde{p}(\boldsymbol{\phi}^{(l)}|\boldsymbol{y})$ at points $\{\boldsymbol{\phi}^{(1)}, \ldots, \boldsymbol{\phi}^{(L)}\}$. `R-INLA` has three available approaches for selecting $\boldsymbol{\phi}^{(l)}$: the empirical Bayes (EB) option uses only the modal value as a single integration point, the grid method develops a regular grid on the primary orthogonal axis of the hyperparameter space, and the central composite design (CCD) approach which efficiently selects the modal value and a group of 'star points' surrounding the center. The default within `R-INLA` selects the grid option for small numbers of hyperparameters and otherwise selects the CCD method

2. Approximate $p(B_i|\boldsymbol{\phi}^{(l)}, \boldsymbol{y})$ for $l = 1, \ldots, L$ using one of three approximations: Gaussian, Laplace, or Simplified Laplace (SL).

In the Gaussian approximation, $\widetilde{p}_G(B_i|\boldsymbol{\phi}^{(l)}, \boldsymbol{y})$ is calculated directly as the marginal of $\widetilde{p}_G(\boldsymbol{\beta}, \boldsymbol{b}|\boldsymbol{\phi}, \boldsymbol{y})$, from (3.7). While very fast, this approximation is often not particularly good (Blangiardo and Cameletti, 2015, Section 4.7.2). In the Laplace approximation, a computationally optimized version of the LRA of Tierney and Kadane (1986) is used,

$$\widetilde{p}_L(B_i|\boldsymbol{\phi}^{(l)}, \boldsymbol{y}) \propto \frac{p_1(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\phi}^{(l)}) p_2(\boldsymbol{\beta}, \boldsymbol{b}|\boldsymbol{\phi}^{(l)}) p_3(\boldsymbol{\phi})}{\widetilde{p}_G(\boldsymbol{B}_{-i}|B_i, \boldsymbol{\phi}^{(l)}, \boldsymbol{y})}\Bigg|_{\boldsymbol{B}_{-i}=\boldsymbol{B}^*_{-i}(B_i, \boldsymbol{\phi}^{(l)})}, \tag{3.8}$$

where $\widetilde{p}_G(\boldsymbol{B}_{-i}|B_i, \boldsymbol{\phi}, \boldsymbol{y})$ is the Gaussian Laplace approximation to $p(\boldsymbol{B}_{-i}|B_i, \boldsymbol{\phi}, \boldsymbol{y})$ with mode $\boldsymbol{B}^*_{-i}(B_i, \boldsymbol{\phi})$. This approximation often works very well since the conditional distribution of the latent field parameters are generally close to Gaussian, but it is computationally expensive since it must be recomputed for all desired combinations of $\boldsymbol{B}$ and $\boldsymbol{\phi}$. The SL approximation, $\widetilde{p}_{SL}(B_i|\boldsymbol{\phi}^{(l)}, \boldsymbol{y})$, uses a Taylor-series approximation of $\widetilde{p}_L(B_i|\boldsymbol{\phi}^{(l)}, \boldsymbol{y})$. This approximation is quick and accurate for many applications and is the default option within `R-INLA`. More details on this approximation can be found in Rue *et al.*, 2009, Section 3.2.

3. Approximate the marginal using numerical integration,

$$\widetilde{p}(B_i|\boldsymbol{y}) = \sum_{l=1}^{L} \widetilde{p}(B_i|\boldsymbol{\phi}^{(l)}, \boldsymbol{y}) \times \widetilde{p}(\boldsymbol{\phi}^{(l)}|\boldsymbol{y}) \times \Delta_l, \tag{3.9}$$

over the integration points, $\boldsymbol{\phi}^{(l)}$, appropriately scaled by their associated weights, Δ_l.

Although INLA returns the univariate marginals, in general we may be interested in functions of the parameters and `R-INLA` provides a method to sample from the approximate joint posterior, implemented in their sampling function, `inla.posterior.sample()`, using the mixture:

$$\widetilde{p}(\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\phi}|\boldsymbol{y}) \approx \sum_{l=1}^{L} \widetilde{p}_G(\boldsymbol{\beta}, \boldsymbol{b}, |\boldsymbol{y}, \boldsymbol{\phi}^{(l)}) \times \widetilde{p}(\boldsymbol{\phi}^{(l)}|\boldsymbol{y}).$$

For each draw, d, first a sample, $\boldsymbol{\phi}^{(d)}$ is drawn from the discretized hyperparameter posterior, $\widetilde{p}(\boldsymbol{\phi}^{(l)}|\boldsymbol{y})$, and then a sample, $\{\boldsymbol{\beta}^{(d)}, \boldsymbol{b}^{(d)}\}$, is drawn from a Gaussian approximation to the joint conditional latent distribution, $\widetilde{p}_G(\boldsymbol{\beta}, \boldsymbol{b}, |\boldsymbol{y}, \boldsymbol{\phi}^{(d)})$ which is found to match the mode and curvature at the mode conditional on the specific hyperparameter draw. While there is no guarantee that this joint approximation will lead to the same approximate univariate marginals from the full INLA algorithm, by default `R-INLA` corrects both the mean and the skew of the Gaussian marginals sampled from the joint posterior my mapping to a SkewNormal distribution using the better-approximated marginal posteriors (Wakefield *et al.*, 2016). We note that even when the EB method for the hyperparameter integration is chosen in INLA and there is only one 'integration point' for $\boldsymbol{\phi}$, this approximate joint distribution can still account for skew.

`R-INLA` only allows for a limited set of built-in likelihoods, and while INLA theoretically can be used on any ALGM, it does have some pragmatic suggestions to ensure reasonable computing times. In particular, a reasonably small dimension of $\boldsymbol{\phi}$ is required (Rue *et al.* (2017) suggests 2-5 and not more than 20) to minimize the computational burden of the numerical integration in (3.9). Crucially, sparsity in the precision of the latent field parameters can be leveraged at numerous points in the algorithm by `R-INLA` to greatly reduce the computational cost and speed up the approximation. The class of ALGMs is the target of the INLA method because they often permit sparsity and because the posteriors can often be well approximated with the LRA.

Spatial statistics applications commonly use Gaussian Markov Random Field (GMRF) model specifications which maintain high levels of sparsity in the precision. A multivariate Gaussian random variable is a GMRF on an undirected graph $G = \{V, E\}$ with vertices, V, and edges E if non-zeros in the precision matrix correspond to the edges of G. GMRF models are used in the following simulations and in the cancer application.

3.4 Template Model Builder

`TMB`, rooted in frequentist inference, requires the user to differentiate between random and

non-random effects, $\boldsymbol{b}$ and $\boldsymbol{\theta}$. We focus our description of `TMB` in the context of a frequentist treatment while noting the differences for Bayesian inference. In contrast to INLA, `TMB` uses a single LA to integrate out the random effects from the full joint distribution to obtain an approximation to the marginal likelihood in (3.3). Defining the conditional mode,

$$\widehat{\boldsymbol{b}}(\boldsymbol{\beta}, \boldsymbol{\phi}) := \underset{\boldsymbol{b}}{\operatorname{argmin}} \; -f(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{b}), \tag{3.10}$$

`TMB` approximates (3.3) using the LA to marginalize over the random effects:

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\phi}) \approx \tilde{\mathcal{L}}(\boldsymbol{\beta}, \boldsymbol{\phi}) = (2\pi)^{n/2} |\boldsymbol{\mathcal{H}}(\boldsymbol{\beta}, \boldsymbol{\phi})|^{1/2} \exp[-f(\boldsymbol{\beta}, \widehat{\boldsymbol{b}}(\boldsymbol{\beta}, \boldsymbol{\phi}), \boldsymbol{\phi})], \tag{3.11}$$

where $\boldsymbol{\mathcal{H}}(\boldsymbol{\beta}, \boldsymbol{\phi})$ is the Hessian of $f(\boldsymbol{\beta}, \widehat{\boldsymbol{b}}(\boldsymbol{\beta}, \boldsymbol{\phi}), \boldsymbol{\phi})$, with (j, k)-th element

$$\frac{\partial^2}{\partial b_j \partial b_k} f(\boldsymbol{\beta}, \widehat{\boldsymbol{b}}(\boldsymbol{\beta}, \boldsymbol{\phi}), \boldsymbol{\phi}). \tag{3.12}$$

Estimation in `TMB` is performed through a two-stage nested optimization procedure which searches for the vector $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}})$ maximizing $\tilde{\mathcal{L}}(\boldsymbol{\beta}, \boldsymbol{\phi})$ as defined in (3.11). To evaluate $\tilde{\mathcal{L}}(\boldsymbol{\beta}, \boldsymbol{\phi})$, we need to evaluate both $\widehat{\boldsymbol{b}}$ in (3.10) and $\boldsymbol{\mathcal{H}}(\boldsymbol{\beta}, \boldsymbol{\phi})$ in (3.12), but neither are usually available in closed-form. While $\widehat{\boldsymbol{b}}(\boldsymbol{\beta}, \boldsymbol{\phi})$ may be evaluated through nonlinear optimization of f, the crux of the work is to evaluate $\boldsymbol{\mathcal{H}}(\boldsymbol{\beta}, \boldsymbol{\phi})$, and `TMB` is capable of operating on very high-dimensional problems as long as the Hessian is a sparse matrix. `TMB` computes gradients and the Hessian using automatic differentiation (Appendix A.1 presents an overview) performed via `CppAD` (Bell, 2007) and at the same time can auto-detect sparsity in the model (Kristensen *et al.*, 2016, Section 4.2) in order to use the `C++ Eigen` library (Guennebaud *et al.*, 2010) to leverage sparse matrix routines. `TMB` reuses the LA and the automatically generated Hessian to produce estimates of the joint covariance between the fixed effects estimates, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\phi}}$, and the random effects predictors, $\widehat{\boldsymbol{b}}$, as described in Section 3.4.3. Inference then proceeds assuming asymptotic normality for the joint distribution of the estimated fixed effects and predicted random effects.

3.4.1 The TMB Estimation Algorithm

`TMB` implements a two-step nested optimization routine to iteratively search for the fixed effects estimates while also generating the random effects predictions. Given initial starting values, the routine performs the following steps at each evaluation in the search:

1. Given current values of the fixed effects parameters, $(\boldsymbol{\beta}^\star, \boldsymbol{\phi}^\star)$, perform nonlinear optimization to find updated modal values of the random effects, $\widehat{\boldsymbol{b}}(\boldsymbol{\beta}^\star, \boldsymbol{\phi}^\star)$, as in (3.10), and set this to be the current value of the random effects, $\boldsymbol{b}^\star$.

2. Given current values of the random effects parameters, $\boldsymbol{b}^\star$, find the modal values of the Laplace approximation to the marginal likelihood, $\tilde{\mathcal{L}}(\boldsymbol{\beta}, \boldsymbol{\phi})$, shown in (3.11), and set them to the current value of the fixed effects, $(\boldsymbol{\beta}^\star, \boldsymbol{\phi}^\star)$. In addition, evaluate the gradient of the marginal likelihood to assess stopping conditions.

3. If the maximum gradient component (MGC) of the marginal likelihood is below a stopping threshold the routine stops, otherwise go to step 1.

If a stopping criteria was reached, then the final values of the fixed effect and random effects are returned as the fixed effects estimates, $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}) = (\boldsymbol{\beta}^\star, \boldsymbol{\phi}^\star)$, and the random effects predictions are updated one final time to yield $\widehat{\boldsymbol{b}}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}})$.

3.4.2 The TMB Estimators

In frequentist settings, `TMB` produces the marginal MLEs of the fixed effects, $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$, from the approximate marginal likelihood defined in (3.11). In this scenario, the random effects predictions are empirical Bayes predictors, taken to be the mode of their conditional distribution, $p(\boldsymbol{b}|\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}_1, \boldsymbol{y})$, using 'plug-in' estimates of the fixed effects MLEs (Carlin and Louis, 2000; Cressie *et al.*, 2009).

In Bayesian settings (fixed effects models with priors or random effects models with hyperpriors), `TMB` produces MMAP estimates of the fixed effects, again using the Laplace

approximation from (3.11) applied to the marginal posterior shown in (3.4). In this scenario, `TMB` provides parameteric empirical Bayes (PEB) predictions (Carlin and Louis, 2000, Chapter 3.3) for the random effects, using the estimated posterior, $p(\boldsymbol{b}|\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}, \boldsymbol{y})$, where the hyperparameters have been estimated using the data. In essence, this replaces the posterior,

$$p(\boldsymbol{b}|\boldsymbol{y}) = \int p(\boldsymbol{b}|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\phi}) p(\boldsymbol{\beta}, \boldsymbol{\phi}|\boldsymbol{y}) \ \mathrm{d}\boldsymbol{\beta} \, \mathrm{d}\boldsymbol{\phi},$$

with $p(\boldsymbol{b}|\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}, \boldsymbol{y})$ where $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\phi}}$ are MMAP estimates. The EB option for the numerical integration in (3.9) in the INLA algorithm makes the same tradeoff: skipping the computational complexity of the integration by not accounting for the uncertainty in the (hyper)parameters.

While the random effects predictors, conditional on the fixed effects estimates, are fast to compute via optimization, these EB predictions of random effects generally have no guarantees of consistency (Thorson and Kristensen, 2016). In order to improve the predictions for the random effects and differentiable functions of the mixed effects, $\boldsymbol{d}(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{b})$, `TMB` implements a generic bias-correction adjustment termed the "epsilon" method (Thorson and Kristensen, 2016). The correction is applicable to quantities which depend on random effects, $\boldsymbol{d}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}, \boldsymbol{b})$, and has no effect on the fixed effects estimates. This correction is based an approximation to moment-generating functions, using $M(\boldsymbol{\epsilon}) = E[\exp\{\boldsymbol{\epsilon} \cdot \boldsymbol{d}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}, \boldsymbol{b})\}]$, with $\cdot$ denoting the dot product, proposed by Tierney *et al.*, 1989, Section 3.1. Introducing an auxiliary parameter vector, $\boldsymbol{\epsilon}$, of dimension equal to the dimension of $\boldsymbol{d}()$, the correction is motivated by considering a new function with the auxiliary parameters:

$$\boldsymbol{e}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}, \boldsymbol{b}, \boldsymbol{\epsilon}|\boldsymbol{y}) = \log\Big(\int \exp\big(f(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}, \boldsymbol{b}) + \boldsymbol{\epsilon} \cdot \mathrm{d}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}, \boldsymbol{b}) \big) \ \mathrm{d}\boldsymbol{b} \Big). \tag{3.13}$$

Employing the namesake method of moment-generating functions, the form of the bias-

corrected estimator is found by differentiating with respect to $\boldsymbol{\epsilon}$ before evaluation at $\boldsymbol{\epsilon} = \mathbf{0}$:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\epsilon}} \left(\boldsymbol{e}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}, \boldsymbol{b}, \boldsymbol{\epsilon} | \boldsymbol{y}) \right) \Big|_{\boldsymbol{\epsilon}=\mathbf{0}} &= \frac{\int \exp\left(f(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}, \boldsymbol{b}) + \boldsymbol{\epsilon} \cdot \boldsymbol{d}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}, \boldsymbol{b}) \right) \boldsymbol{d}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}, \boldsymbol{b}) \ \mathrm{d}\boldsymbol{b}}{\int \exp\left(f(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}, \boldsymbol{b}) + \boldsymbol{\epsilon} \cdot \boldsymbol{d}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}, \boldsymbol{b}) \right) \ \mathrm{d}\boldsymbol{b}} \Big|_{\boldsymbol{\epsilon}=\mathbf{0}} \\
&= \frac{\int \exp\left(f(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}, \boldsymbol{b}) \right) \boldsymbol{d}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}, \boldsymbol{b}) \ \mathrm{d}\boldsymbol{b}}{\int \exp\left(f(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}, \boldsymbol{b}) \right) \ \mathrm{d}\boldsymbol{b}} \\
&= \mathrm{E}[\boldsymbol{d}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}, \boldsymbol{b}) | \boldsymbol{y}].
\end{aligned} \tag{3.14}$$

This identity allows `TMB` to improve the predictions for the random effects (if $\boldsymbol{d}(\cdot)$ is taken to be the identity function) and nonlinear functions of the mixed effects using the gradient of the approximate marginal likelihood with respect to the auxiliary parameters, $\boldsymbol{\epsilon}$. `TMB` can efficiently calculate this quantity using `CppAD` to automatically evaluate derivatives and the LA to approximate the integrals. Thorson and Kristensen (2016) note that this is only a bias correction algorithm because the LA used to evaluate both numerator and denominator in (3.14) will be inexact unless the likelihood, conditional on the fixed effects, is multivariate Gaussian.

3.4.3 Variance of TMB Estimators

`TMB` approximates the covariance of the fixed effects using the inverse of the observed Hessian of the log-likelihood:

$$\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}} = \mathrm{Cov}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}) = \left(-\nabla^2 \log \tilde{\mathcal{L}}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\phi}}) \right)^{-1}. \tag{3.15}$$

In models that include random effects, the joint covariance of the fixed and random effects

is approximated using an application of the law of total variance and a linearization:

$$\begin{aligned}\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\phi}},\widehat{\boldsymbol{b}}} = \operatorname{Cov}\begin{pmatrix}\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\phi}}\\ \hat{\boldsymbol{b}}\end{pmatrix} &= \operatorname{E}\left[\operatorname{Cov}\left(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\phi}},\hat{\boldsymbol{b}}\middle|\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\phi}}\right)\right] + \operatorname{Cov}\left[\operatorname{E}\left(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\phi}},\hat{\boldsymbol{b}}\middle|\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\phi}}\right)\right]\\ &\approx \begin{pmatrix}0 & 0\\ 0 & \boldsymbol{\mathcal{H}}_{bb}^{-1}(\boldsymbol{\beta},\boldsymbol{b},\boldsymbol{\phi})\end{pmatrix} + \boldsymbol{J}\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\phi}}}\boldsymbol{J}^{\mathrm{T}}\end{aligned} \tag{3.16}$$

where $\boldsymbol{\mathcal{H}}_{bb}(\boldsymbol{\beta},\boldsymbol{b},\boldsymbol{\phi})$ is the random effects sub-matrix of the full joint Hessian of $f(\boldsymbol{\beta},\boldsymbol{b},\boldsymbol{\phi})$, and $\boldsymbol{J}$ is the Jacobian of the vector $\left(\boldsymbol{b},\boldsymbol{\phi},\hat{\boldsymbol{b}}(\boldsymbol{b},\boldsymbol{\phi})\right)^{\mathrm{T}}$ with respect to $(\boldsymbol{b},\boldsymbol{\phi})$. The δ-method is used to find the joint covariance of differentiable functions of the mixed effects:

$$\operatorname{Cov}\left(\boldsymbol{d}(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\phi}},\hat{\boldsymbol{b}})\right) = \nabla\boldsymbol{d}\,\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\phi}},\hat{\boldsymbol{b}}}\nabla\boldsymbol{d}^{\mathrm{T}}, \tag{3.17}$$

where $\nabla\boldsymbol{d}$ is the Jacobian of $\boldsymbol{d}$. For models with only fixed effects, this simplifies to

$$\operatorname{Cov}\left(\boldsymbol{d}(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\phi}})\right) = \nabla\boldsymbol{d}\;\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\phi}}}\nabla\boldsymbol{d}^{\mathrm{T}}. \tag{3.18}$$

Much like the correction for the random effects predictions, `TMB` can improve the covariance estimator of the random effects prediction and functions of the mixed effects shown in (3.17) using the second derivative of $\boldsymbol{e}(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\phi}},\boldsymbol{u},\boldsymbol{\epsilon}|\boldsymbol{y})$ from (3.13). The form of the improved variance estimator again uses the law of total variance:

$$\begin{aligned}\operatorname{Cov}\left(\boldsymbol{d}(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\phi}},\widehat{\boldsymbol{b}})\right) = &\left[\frac{\partial^2}{\partial^2\boldsymbol{\epsilon}}\left(\boldsymbol{e}(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\phi}},\boldsymbol{b},\boldsymbol{\epsilon}|\boldsymbol{y})\right) + \right.\\ &\left.\frac{\partial}{\partial\boldsymbol{\theta}}\frac{\partial}{\partial\boldsymbol{\epsilon}}\left(\boldsymbol{e}(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\phi}},\boldsymbol{b},\boldsymbol{\epsilon}|\boldsymbol{y})\right)^{\mathrm{T}}\;\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\phi}}}\;\frac{\partial}{\partial\boldsymbol{\theta}}\frac{\partial}{\partial\boldsymbol{\epsilon}}\left(\boldsymbol{e}(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\phi}},\boldsymbol{b},\boldsymbol{\epsilon}|\boldsymbol{y})\right)\right]_{\boldsymbol{\epsilon}=\boldsymbol{0}}\end{aligned} \tag{3.19}$$

where the first term on the right-hand side works out to be the standard variance estimator for random effects conditional on the fixed effects (demonstrated analogously to the derivation in (3.14)), and the second term accounts for having conditioned on the fixed effects estimators. This is an improvement over the naive EB variance estimators which ignore the conditioning

on the fixed effects (Carlin and Louis, 2000, Chapter 3.5). The entire vector of fixed effects has been denoted by $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\phi})$, as in Table 3.1.

3.5 Contrasting TMB and INLA

For clarity, Table 3.2 provides a summary of the primary differences - and similarities - between `TMB` and INLA that were discussed in detail in the previous two sections.

`TMB` and INLA will yield the most similar results under the following conditions:

- `TMB` is coded to optimize the marginal posterior using the same priors and with the same internal parameter representations used by `R-INLA`,
- `TMB` uses both bias and variance corrections to functions of REs,
- `R-INLA` uses the Gaussian approximations and the empirical Bayes 'integration' strategy,
- `R-INLA` does not use the bias and skew corrections when drawing posterior samples.

In addition, in cases where the posterior is expected to be close to Gaussian, TMB should perform as well as INLA.

Despite how well TMB performs in the following simulation studies, we do note that the INLA algorithm, with its best suite of approximations, should be capable of better approximations than TMB. So, if given the choice to use either method, it seems reasonable to choose INLA. That said, there are a number of situations where TMB will be preferred: primarily when the model is not permitted within `R-INLA` or the computational efficiency of TMB is required.

3.6 Spatial Simulation Study

We performed two simulation studies on popular continuous and discrete spatial smoothing models to assess the ability of `TMB` and `R-INLA` to estimate the total spatial field effects, the

Table 3.2: Summarization of primary differences between `TMB` and INLA. Entries in (parentheses) indicate outcomes from `TMB` for models that include priors and indicate outcomes from INLA under eb 'numerical integration' over the hyperparameters. The table is split into sections corresponding to methods, approximations, post-model sampling, and computation. FE=fixed effects, RE=random effects, MMAP=marginal maximum a posterior estimates, PEB = parametric Empirical Bayes, LRA = Laplace ratio approximation.

	`TMB`	`R-INLA`
Inferential Method: FE	Frequentist: MLEs from marginal likelihood (or MMAP from marginal posterior)	Full Bayesian (or PEB)
Inferential Method: RE	Frequentist: EB (or PEB)	Full Bayesian (or PEB)
Laplace approximation	Standard Gaussian approximation	LRA from Tierney and Kadane (1986)
Bias Corrections	Yes, to functions of REs	Applied to align joint posterior samples with marginal distributions
Var. Corrections	Yes, to functions of REs	No
Skew Corrections	No	Applied to align joint posterior samples with marginal distributions
Hyperpar. Integration	No	Yes (no)
Inferential sampling	Gaussian centered at point estimates and covariance from observed information and linearization	Gaussian approximation to joint posterior with mean and skew corrections applied to marginals
Hessian Evaluation	Automatic Differentiation	Finite Differences
Sparse Matrices	Yes	Yes
Parallelization	Yes	Yes

parameters and hyperparameters, and how their computational performance scales as data volume and random effects dimension increase.

The second purpose of the continuous simulation study was to perform a thorough assessment of the popular stochastic partial differential equations (SPDE) representation for fitting GPs which uses a finite element method over a triangulation to solve a particular SPDE whose solution is known to have Matérn covariance. An overview of the SPDE ap-

proach may be found in Appendix A.2.1 and details on the SPDE approximation can be found in Lindgren *et al.* (2011) and Miller *et al.* (2020). The discrete simulation study was included in part to assess and verify inference in `TMB` using models that require hard constraints on the random effects parameters.

For each study, a grid of experiments, shown in Tables 3.3 and 3.4, was defined. Each level of each experiment was replicated 25 times to obtain Monte Carlo errors on the validation metrics. For each replicate within each experiment, completely new spatial fields, sampling locations, and observations were generated. Once data was generated, `TMB` and `R-INLA` algorithms were run, and inference and validation was performed using 500 joint samples drawn from each model. In the continuous simulations, joint parameters samples were projected to 5×5 km^2 raster grids, and in both simulations posterior draws of the parameters and spatial fields were compared against against the truth. For both models, the internal representation of parameters in `TMB` were coded to align with those used by `R-INLA`. `TMB` was always run using its bias correction method and the improved variance estimates, and `R-INLA` joint estimates were generated using its available mean bias correction. All of the simulation analyses use the empirical distribution taken across the 25 replicates of each experimental level.

3.6.1 Continuous Spatial Simulation

This simulation was designed with respect to three governing motivations that dictated the choice of models, covariates, and true parameter ranges: (1) vetting `TMB` and INLA inference for spatial GPs while (2) assessing the SPDE approximation in a variety of settings (3) using simulated risk fields and data akin to those commonly see in public health settings. Although motivated by public health applications, the broad range the parameters, such as the maximum number of simulated data locations, greatly extend this study's applicability beyond any one applied domain.

We consider Gaussian process (GP) models (see to Section 2.3) with mean $\mu(s)$ at location s, and the Matérn covariance function of (2.12). That is, for any finite collection

of locations within the domain, $\{s_1, \ldots, s_n\} \in \mathcal{S}$, the random vector $\boldsymbol{u} = [u(s_1), \ldots, u(s_n)]$ has multivariate Gaussian distribution with covariance matrix $\boldsymbol{\Sigma}$, with $\Sigma_{i,j} = C(u(s_i), u(s_j))$ from (2.12), and precision matrix $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$.

The simulated data, observed at locations $s_i,\ i = 1, \ldots, n_s$, taken to be within the domain defined by Nigeria's border, and selected using a stratified spatial sampling design, arise from the following hierarchical model:

$$\begin{aligned}
\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\phi}_1 &\sim p_1(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\phi}_1) \\
\mathrm{E}[y_i|\boldsymbol{\beta}, u_i, v_i] &= g^{-1}\left(\alpha + \boldsymbol{z}_i^T\boldsymbol{\beta} + u_i + v_i\right) \\
\boldsymbol{u} &\sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{Q}(r_{\mathrm{m}}, \sigma^2_{\mathrm{m}})) \\
\boldsymbol{v} &\sim \mathrm{N}(\boldsymbol{0}, \mathrm{I}_{n_s}\sigma^2_{\mathrm{clust}}).
\end{aligned}$$

where α is the intercept and is fixed to -1 across all simulations, I_{n_s} is the $n_s \times n_s$ identity matrix and the last two lines correspond to $p_2(\boldsymbol{b}|\boldsymbol{\phi}_2)$ with $\boldsymbol{b} = [\boldsymbol{u}, \boldsymbol{v}]$ and $\boldsymbol{\phi}_2 = [\sigma^2_{\mathrm{m}}, r_{\mathrm{m}}, \sigma^2_{\mathrm{clust}}]$. The precision for the spatial GP, $\boldsymbol{Q}$, is Matérn with range r_{m} and standard deviation σ_{m}. In some models, two spatially varying covariates were included: access time to heath care and malaria incidence. While the SPDE representation is used to fit the spatial fields, the true fields are simulated directly on a higher resolution regular grid using the `RandomFields R` package (Schlather *et al.*, 2015).

The spatial region, covariates and stratified sampling scheme were chosen to represent public health datasets, such as the Demographic and Health Surveys (DHS) and UNICEF Multiple Indicator Cluster Surveys, which are increasingly being used to predict continuous spatial(-temporal) maps of health outcomes. The covariates, access (travel time) to health care and malaria incidence, are both reasonable choices to be correlated with heath risks, and the magnitude of the GP and total field were selected to yield moderately rare outcomes. Plots of the covariates are shown in Figure 3.1. Our stratified cluster sampling design mimics the one used by DHS which stratifies by regions (administrative level 1 units) and urban/rural status and usually collects observations at 250-750 clusters, with typically 25-35 households

sampled within each cluster. Across our simulations, the expected sample size per cluster was set to 35. An example of the stratified cluster locations is shown in Figure 3.2a. An example of a simulated risk field, on the linear predictor scale, consisting of the linear combination of the access covariate, malaria incidence covariate, and simulated Matérn GP is shown in Figure 3.2b.

Both Binomial and Gaussian data are commonly collected in a variety of applications (Poisson likelihoods are examined in the discrete simulation), and the binary data provide an extra challenge for both `TMB` and INLA. The Matérn range was selected to represent medium-small and medium-large spatial ranges over the approximate 10×10 (degrees latitude-longitude) area of Nigeria. The Matérn variance takes two values representing small- and large-scale spatial effects relative to the covariate effects and the small, medium, and large values of the iid cluster variance and Gaussian observation variance. The number of vertices in the SPDE triangulation were selected to represent medium, fine, and very fine meshes, relative to the scale of Nigeria and resolution of the 5×5km raster representation, and to assess the computational scaling of `TMB` and `R-INLA`. The three different resolution triangulation meshes are shown in Figure 3.3. While the INLA method with empirical Bayes integration and Gaussian approximations are closest to the methods in `TMB`, we also evaluated some of the more accurate options in order to asses the default (and better) approximations available in `R-INLA`.

To complete the model specification, the following priors were used:

$$\begin{aligned} \boldsymbol{\beta} &\sim \mathrm{N}(0, 5^2) \\ r_{\mathrm{m}}, \sigma_{\mathrm{m}} &\sim \mathrm{PCspde}(u_r = 10, \alpha_r = .95, u_\sigma = 1, \alpha_\sigma = .05) \\ \tau_{\mathrm{clust}} = \sigma_{\mathrm{clust}}^{-2} &\sim \mathrm{PCprec}(u = .5, \alpha = .05) \end{aligned}$$

The penalized complexity (PC) priors (Simpson *et al.*, 2017; Fuglstad *et al.*, 2019) shrink towards a base model and are set such that $\mathrm{Prob}(r_{\mathrm{m}} < 10^\circ) = .95$, $\mathrm{Prob}(\sigma_{\mathrm{m}} > 1) = .05$ and $\mathrm{Prob}(\sigma_{\mathrm{clust}} > .5) = .05$. Internally, `TMB` was coded to use the same internal representations

of the Matérn parameters used in `R-INLA`: $\log(\kappa)$, and $\log(\tau_*) = \log(\sigma_m^{-2})$ from (2.12).

Simulations were performed with $n_i \sim$ Poisson(35) observations taken at each spatial location $s_i,\ i = 1, \ldots, n_s$. Experiments were run on Gaussian data with $\text{Var}(y_i) = \sigma^2_{\text{obs}}/n_i$, using an identity link function, and with the same PCprec prior on the observation variance, $\sigma^2_{\text{obs}} \in \boldsymbol{\phi}_1$, as is used for σ^2_{clust}. Experiments run on binomial data have no σ^2_{obs} and use a logit link function.

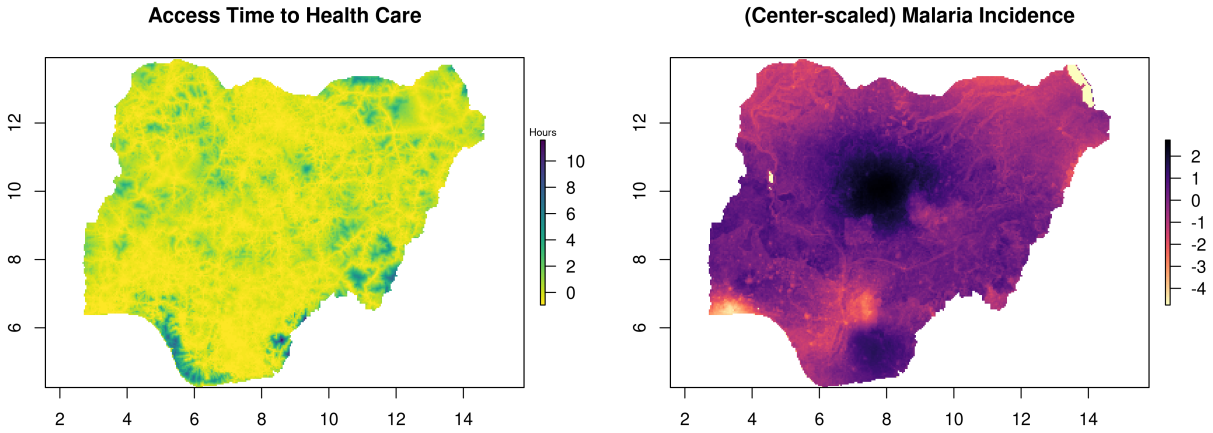
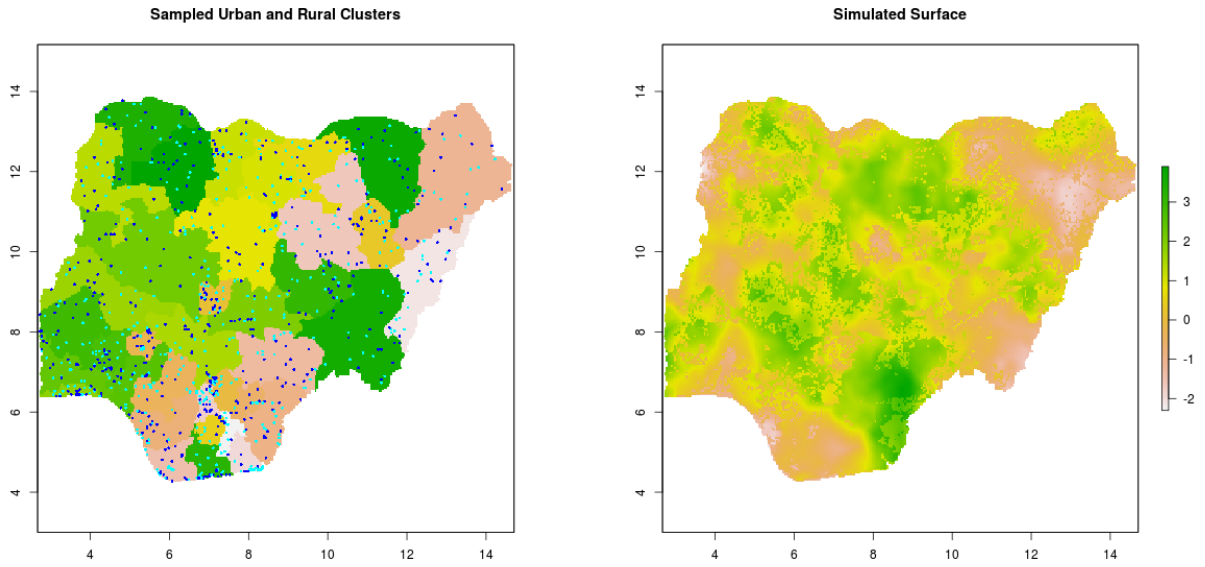


Figure 3.1: Covariates used in the continuous simulation studies: access (time in hours) to healthcare (Weiss *et al.*, 2018) and malaria incidence (Weiss *et al.*, 2019) in Nigeria.

The (nearly) full combinatorial grid of simulation parameters shown in Table 3.3 comprised the set of 16128 experiments (`R-INLA` with 8000 clusters and the full Laplace approximation was excluded due to computational time constraints and the Gaussian variances were not varied for Binomial observations).

Selected Continuous Simulation Results

The simulation results presented in this section compare results from `TMB` against those



(a) Stratified spatial observations

(b) 'True' latent surface

Figure 3.2: Examples of (a) simulated stratified random cluster locations where the locations have been stratified by states within Nigeria and by urban (dark blue) and rural (light blue) demarcations within states, and (b) a simulated latent surface comprised of a linear combination of covariates and GP.

from `R-INLA` using the default options: the simplified Laplace approximation and the CCD numerical integration scheme with mean and skew corrections to the marginals of the joint posterior sampling distribution (see Section 3.3). While using the grid numerical integration scheme and the full Laplace approximations could offer an overall better approximation, and the empirical Bayes integration with Gaussian approximations would be closest to the approximations in `TMB`, the default `R-INLA` options provide a nice balance between computation and accuracy and are what many users explicitly or implicitly choose to use. For these reasons we felt this setting to be a useful and fair benchmark for which to contrast `TMB`. The selection of extended results which include comparisons with other combinations of INLA approximations and results from the different SPDE mesh triangulations are included in Appendix A.2.

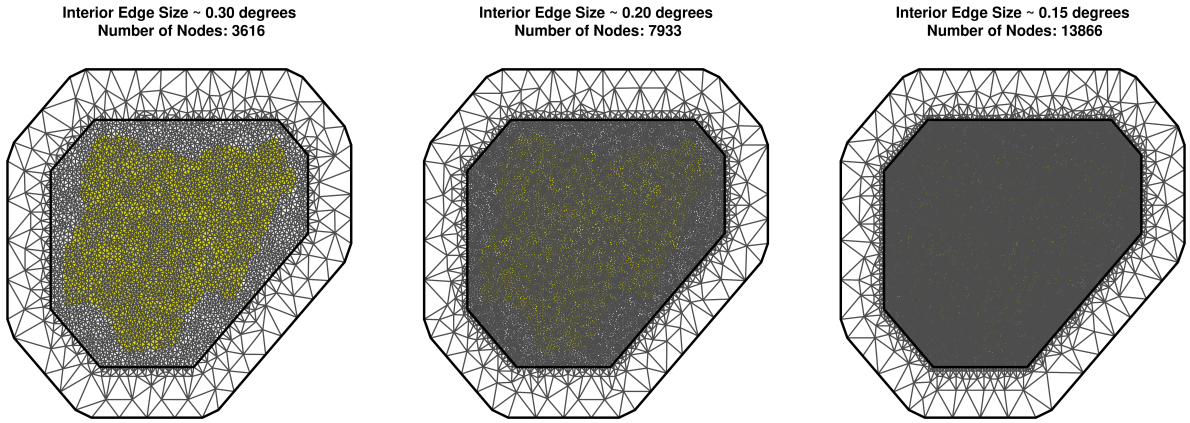


Figure 3.3: Coarse, medium, and fine resolution Delauney triangulations used in the SPDE approximation to the GP. The outline of the spatial domain, Nigeria, is shown beneath the mesh for reference.

Table 3.3: Parameters varied across the continuous simulation experiments. The total number of experiments was 16128 (Gaussian variance was not varied for Binomial experiments), each replicated 25 times.

Parameter	Simulation Values
Data Observations	Binomial, Gaussian
Gaussian Observation Variance, σ^2_{obs}	$0.1^2, 0.2^2, 0.4^2$
Covariates	None, $(-.25 \times \text{access} + .25 \times \text{Malaria Incid.})$
Number of Clusters, $n_{\boldsymbol{s}}$	250, 500, 750, 1000, 2000, 4000, 8000
Expected Samples per Cluster, E$[n_i]$	35
Spatial Range (lat-lon degrees)	1, $\sqrt{8}$
Spatial Variance	$0.25^2, 0.5^2$
Cluster Variance, σ^2_{clust}	0, $0.1^2, 0.2^2, 0.4^2$
Num. Nodes in SPDE Mesh	3631, 7922, 13869 (low, medium, high resolution)
`R-INLA` Integration Strategy	Empirical Bayes (`EB`), Central Composite Design (`CCD`)
`R-INLA` Approximation Strategy	`Gaussian`, `Simplified Laplace`, `Laplace`

The following three figures display:

- Figure 3.4: Binomial scenarios' parameter bias for the intercept, access and malarian incidence coefficients, cluster variance, and the Matérn standard deviation and range.

- Figure 3.5: Binomial, medium-resolution mesh, scenarios' mean pixel coverage, strati-

fied by the value of the true GP, and faceted by cluster variance and number of spatial observations.

- Figure 3.6: Normal $\sigma^2_{obs} = 0.04$, medium-resolution mesh, scenarios' mean pixel coverage, stratified by the value of the true GP, and faceted by cluster variance and number of spatial observations.

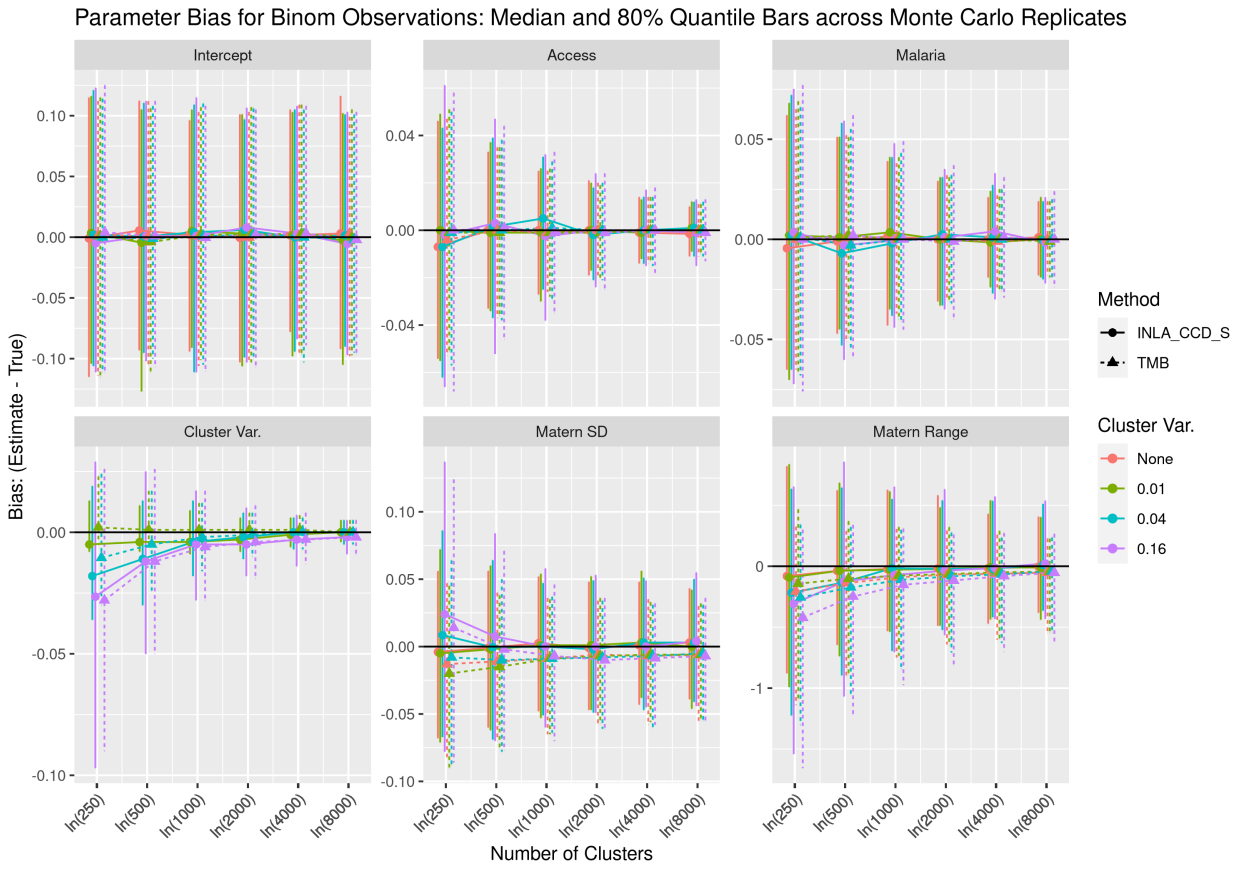


Figure 3.4: Comparison of the estimated parameter bias from `TMB` (dashed lines) and `R-INLA` using CCD hyperparameter integration and simplified Laplace approximations (solid lines) plotted against the number of cluster observations for Binomial observation experiments. Colors represent different cluster (iid nugget) variances used in an experiment. Each point is the median bias of 3 experiments (coarse, medium, and fine SPDE triangulation), calculated across 75 replicates, and the bars represent the middle 80% quantile range of the bias across replicates.

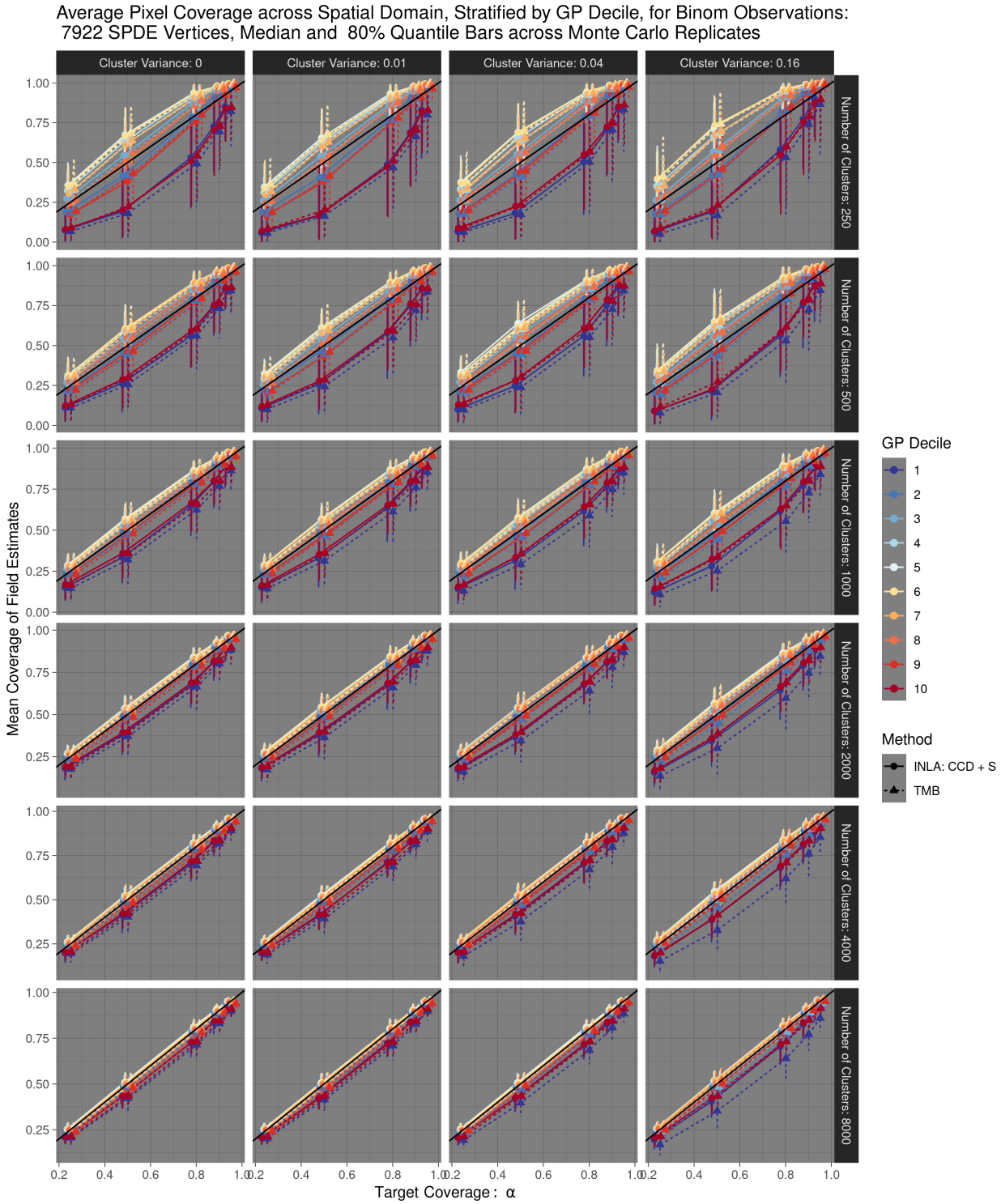


Figure 3.5: Comparison of the average estimated field coverage of the simulated truth, faceted by cluster (iid nugget) variance and the number of clusters, from `TMB` (dashed lines) and `R-INLA` using CCD hyperparameter integration and simplified Laplace approximations (solid lines) plotted against the target nominal coverage, α, for Binomial observation experiments with the medium resolution SPDE triangulation. Colors stratify pixels included in the average coverage calculation by the decile of the true GP for the experiment replicate. Each point is the median average coverage of an experiment, calculated across 25 replicates, and the bars represent the middle 80% quantile range of the average coverage across replicates.

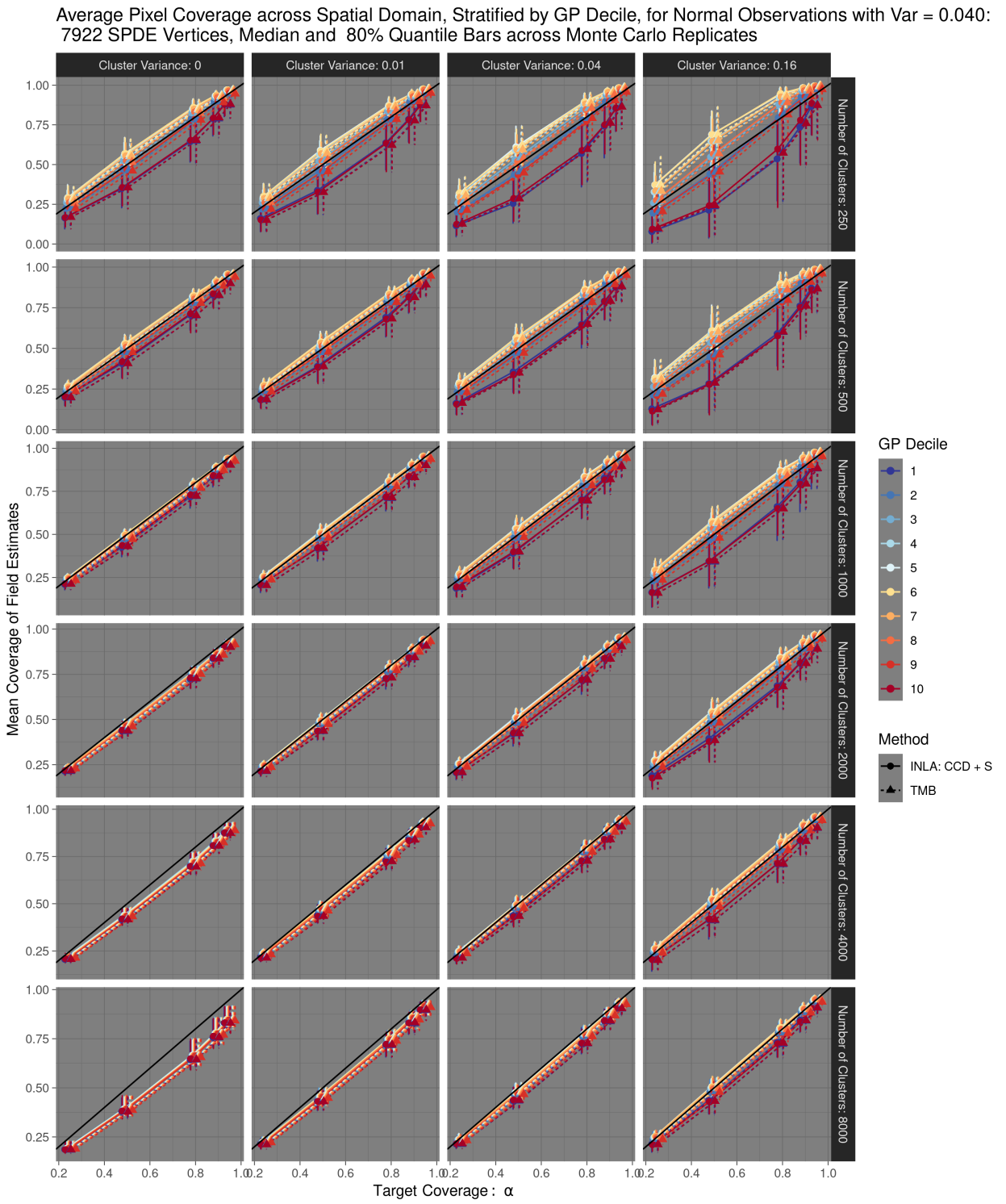


Figure 3.6: Comparison of the average estimated field coverage of the simulated truth, faceted by cluster (iid nugget) variance and the number of clusters, from `TMB` (dashed lines) and `R-INLA` using CCD hyperparameter integration and simplified Laplace approximations (solid lines) plotted against the target nominal coverage, α, for Gaussian observation experiments with $\sigma^2 = 0.04$ and the medium resolution SPDE triangulation. Colors stratify pixels included in the average coverage calculation by the decile of the true GP for the experiment replicate. Each point is the median average coverage of an experiment, calculated across 25 replicates, and the bars represent the middle 80% quantile range of the average coverage across replicates.

While the overall results are quite similar, `TMB` generally has larger bias in the fixed effects estimators, particularly hyperparameters which may deviate further from Gaussianity, as shown in the continuous Binomial experiments, Figures 3.4, and the continuous Gaussian experiments, Figure A.2.1. This trend appears consistently across a variety of `R-INLA` options, Figures A.2.4-A.2.9.

In spatial statistics settings, the hyperparameters (and sometimes all parameters) may not be of inferential interest. Figures 3.5 and 3.6 demonstrate that `TMB` consistently yields results very similar to those from `R-INLA` at the spatial field level, and that the results are similar across all ranges of the spatial effect. Figures A.2.2 and A.2.3 show these results collapsed across the GP magnitude. In contrast to `R-INLA`, `TMB` seems to consistently have slightly lower coverage which could be attributed to the lack of integration over the hyperparameters even though the covariance estimator in (3.19) attempts to account for this.

In the Binomial experiment, we saw no notable differences in the spatial field coverage across the different resolutions of the SDDE triangulation suggesting that the approximation was appropriately resolved. In the Gaussian data setting, we observed that the coarser meshes undercovered the field estimates in experiments with small σ^2_{clust} and large sample sizes - but this was mostly remedied at the finer triangulation resolution. Interestingly, we saw more severe field undercoverage for larger numbers of spatial observations. These patterns were observed in results from both `TMB` and `R-INLA`. See the lower left plots of Figures A.2.10-A.2.12.

Timing comparisons

The main set of continuous simulation experiments, shown in Table 3.3, were run restricting both `TMB` and `R-INLA` to use a single CPU thread. This was done in order to better leverage the particulars of the computing cluster which was used. The median and 80% percent timing quantiles, taken across the 25 replicates of each experimental level, are shown

in Figure 3.7. The plot breaks down the timing by the task, the algorithmic method, the number of cluster observations, whether the observed data were Binomial or Gaussian, and the dimension of the spatial random effects. While it would be unusual for people under normal circumstances to restrict either `TMB` or `R-INLA` to use a single core, this plot nonetheless offers some indication of the total computational burden of each of the methods.

In addition, a small experiment was designed and implemented to give a sense of the timing under more typical, parallelized, computation. The design of this experiment was very similar to those of the Binomial data continuous GP experiments but only the number of clusters (250, 2500, 10000), the number of spatial random effects (low, medium, and high resolution SPDE triangulations with 3616, 7933, and 13866 vertices, respectively), and the number of CPU threads (1, 2, 4) were varied. This created 27 experimental levels, each of which was replicated 5 times. The mean of the fit, prediction, and total times are summarized in Figure 3.8. `R-INLA` was run using the PARDISO library (Alappat *et al.*, 2020; Bollhöfer *et al.*, 2020; Bollhöfer *et al.*, 2019), and it was forced to use the requested number of threads (it has built-in logic that, by default, is capable of using fewer than the max threads if it believes that will be more efficient). `TMB` was run using METIS to create fill-reducing orderings of sparse matrices and parallelization was enabled by setting the available `OpenMP` threads.

In both single-thread and multi-threading scenarios we see that `TMB` scales very well compared to `R-INLA` which is already known to be very computationally efficient.

Details on the software versions and hardware used in this study can be found in Supplement A.5.

3.6.2 Discrete Spatial Simulation

For the discrete simulation study, we considered the BYM2 model (a modern formulation of the classic Besag-York-Mollie model developed by Riebler *et al.*, 2016). This model requires a sum-to-zero hard constraint which was implemented in `TMB` using appropriate conditional densities (Gelfand *et al.*, 2010, Section 12.1.7.4) to match the linear constraint used in the

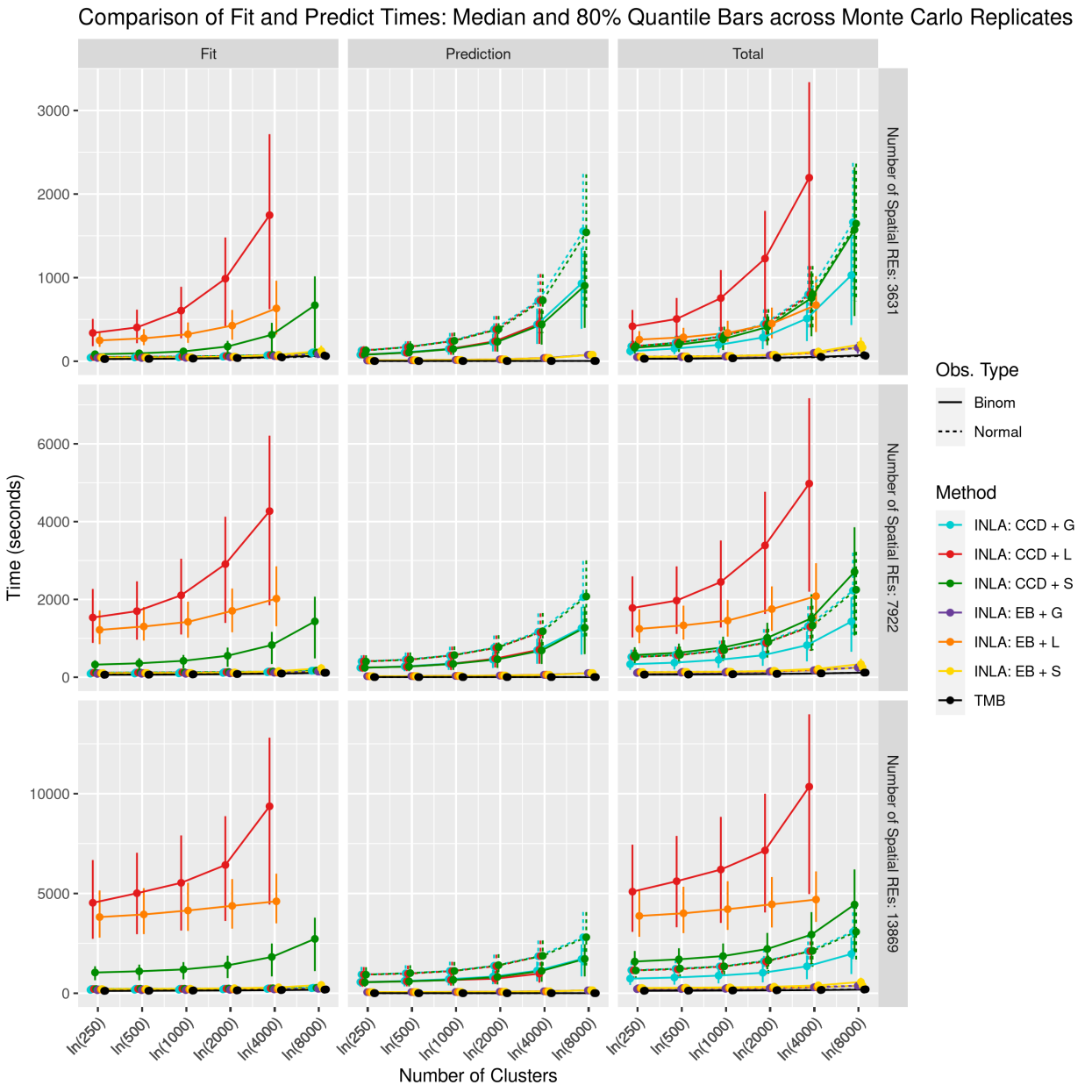


Figure 3.7: Comparison of the average fit, predict, and total times from `TMB` and `R-INLA`, faceted by the number of spatial random effects, plotted aginst the number of clusters. Each point is the median time of an experiment, calculated across 25 replicates, and the bars represent the middle 80% quantile range of the bias across replicates.

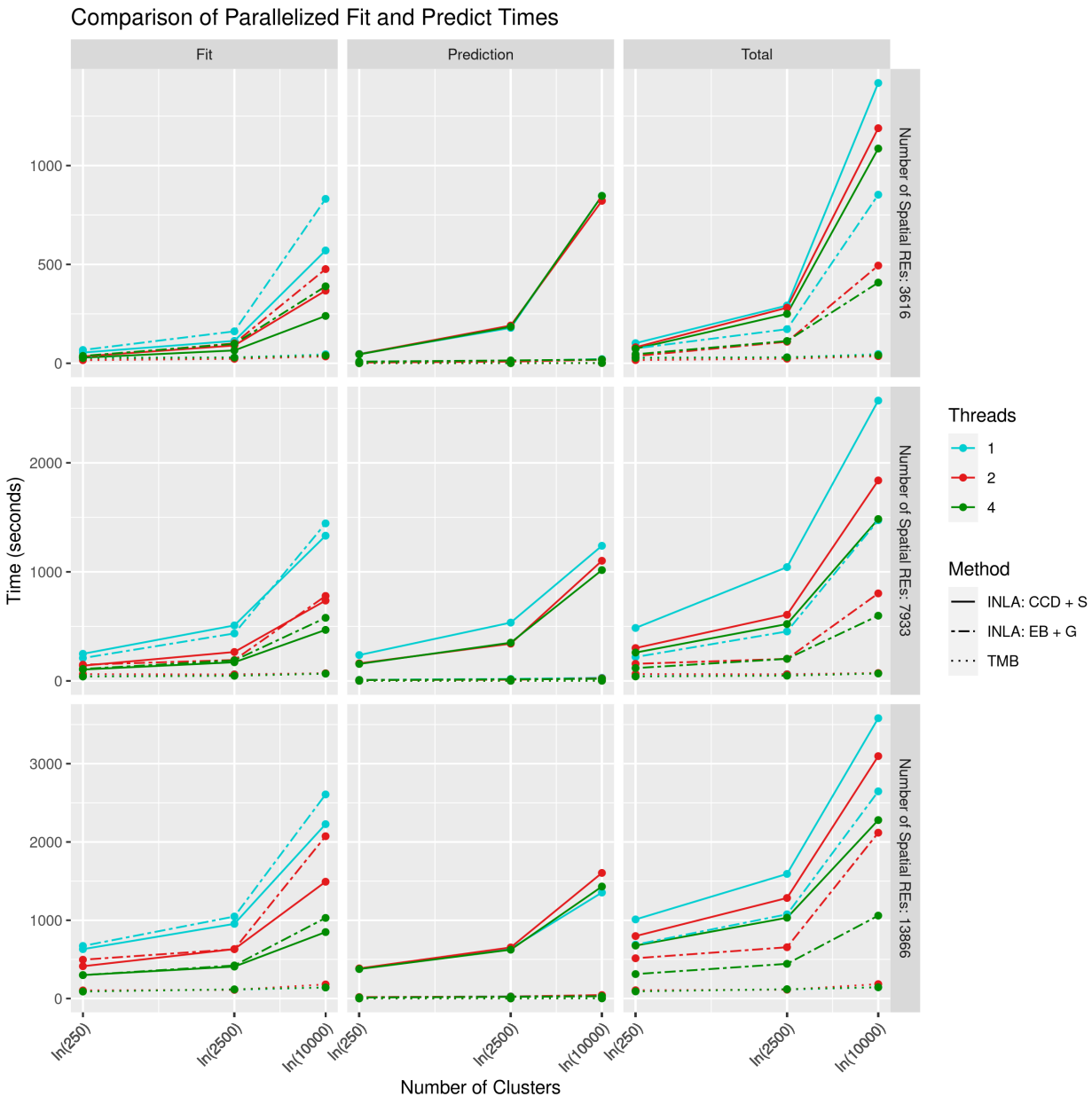


Figure 3.8: Comparison of the average fit, predict, and total times from `TMB` and `R-INLA`, faceted by the number of spatial random effects, plotted against the number of clusters. Each point is the mean time across 5 replicates of a Binomial data experiment. The colors represent the number of threads available for parallelization within each method.

`R-INLA` BYM2 model formulation.

The discrete model was implemented over the 37 regions (first-level administrative units) of Nigeria, and neighbors were defined by to be regions with shared boundaries. Within each region, the population, n_s, for that area was first sampled from iid Poisson distributions. Conditional on the population, data was simulated from another Poisson distribution under the hierarchical model:

$$\begin{aligned} y_i|n_s, \eta_i &\sim \text{Poisson}(n_s \times \eta_i) \\ \eta_i &= \exp\left(\alpha + \boldsymbol{b}_i\right) \\ \boldsymbol{b} &= \frac{1}{\sqrt{\tau}}\left(\sqrt{1-\varphi}\boldsymbol{v} + \sqrt{\varphi}\boldsymbol{u}_\star\right) \\ \boldsymbol{v} &\sim N\left(\boldsymbol{0}, \boldsymbol{I}\right) \\ \boldsymbol{u}_\star &\sim N\left(\boldsymbol{0}, \boldsymbol{Q}_\star^{-1}\right), \text{s.}t. \sum_{i=1}^{37} u_{\star_i} = \boldsymbol{0}, \end{aligned} \tag{3.20}$$

with α the GMRF intercept, fixed at -3 across simulations, BYM2 field $\boldsymbol{b}$ with total variance τ^{-1}, mixing parameter φ controlling the contribution of $\boldsymbol{v}$, the unstructured iid portion of the BYM2 field, and $\boldsymbol{u}_\star$, the scaled spatially structured component of the BYM2. The structured portion of the BYM2 is specified with precision $\boldsymbol{Q}_\star$, a scaled version of the ICAR precision from the classic BYM model, and is constrained to sum to zero.

The effect of the constraint was correctly included in `TMB` by conditioning on the constraint:

$$p(\boldsymbol{u}_\star|\boldsymbol{A}\boldsymbol{u}_\star) = \frac{p(\boldsymbol{A}\boldsymbol{u}_\star|\boldsymbol{u}_\star)p(\boldsymbol{u}_\star)}{p(\boldsymbol{A}\boldsymbol{u}_\star)} \tag{3.21}$$

where, generally, $\boldsymbol{A}\boldsymbol{u}_\star = \boldsymbol{e}$ encodes the linear constraints and specific to this example, $\boldsymbol{A}\boldsymbol{u}_\star = \sum \boldsymbol{u}_\star = 0$.

To complete the model specification, the following priors are included:

$$\begin{aligned} \alpha &\sim \text{N}(0, 5^2) \\ \varphi &\sim \text{Beta}(.5, .5) \\ 1/\sqrt{\tau} = \sigma &\sim \text{N}(0, 5^2)\mathbb{1}_{\sigma>0}. \end{aligned}$$

Internally, `TMB` was coded to use the same internal representations of parameters for the optimization step: $\text{logit}(\varphi)$ and $\log(\tau)$.

The full combinatorial grid of simulation parameters shown in Table 3.4 comprised the set of 20 discrete simulation experiments, with each experiment replicated 25 times to obtain Monte Carlo errors on the validation metrics.

Table 3.4: Parameters varied across the discrete simulation experiments. The total number of experiments was 20, and each was replicated 25 times.

Parameter	Simulation Values
Data Observations	Poisson
Mean Observations per Region, $\text{E}[n_{\boldsymbol{s}}]$	16, 36, 49, 100, 400
BYM2 φ	$0.25, 0.5, 0.75, 0.9$
BYM2 Variance, (τ^{-1})	0.5
GMRF Intercept	-3
`R-INLA` Integration Strategy	Central Composite Design (`CCD`)
`R-INLA` Approximation Strategy	Simplified Laplace

Discrete Simulation Results

The results from the discrete simulations, including the summaries parameter bias in Figure 3.9 and the summaries of spatial field coverage in Figure 3.10, share many similarities to those observed in the continuous simulations of Section 3.6.1. Both `TMB` and `R-INLA`

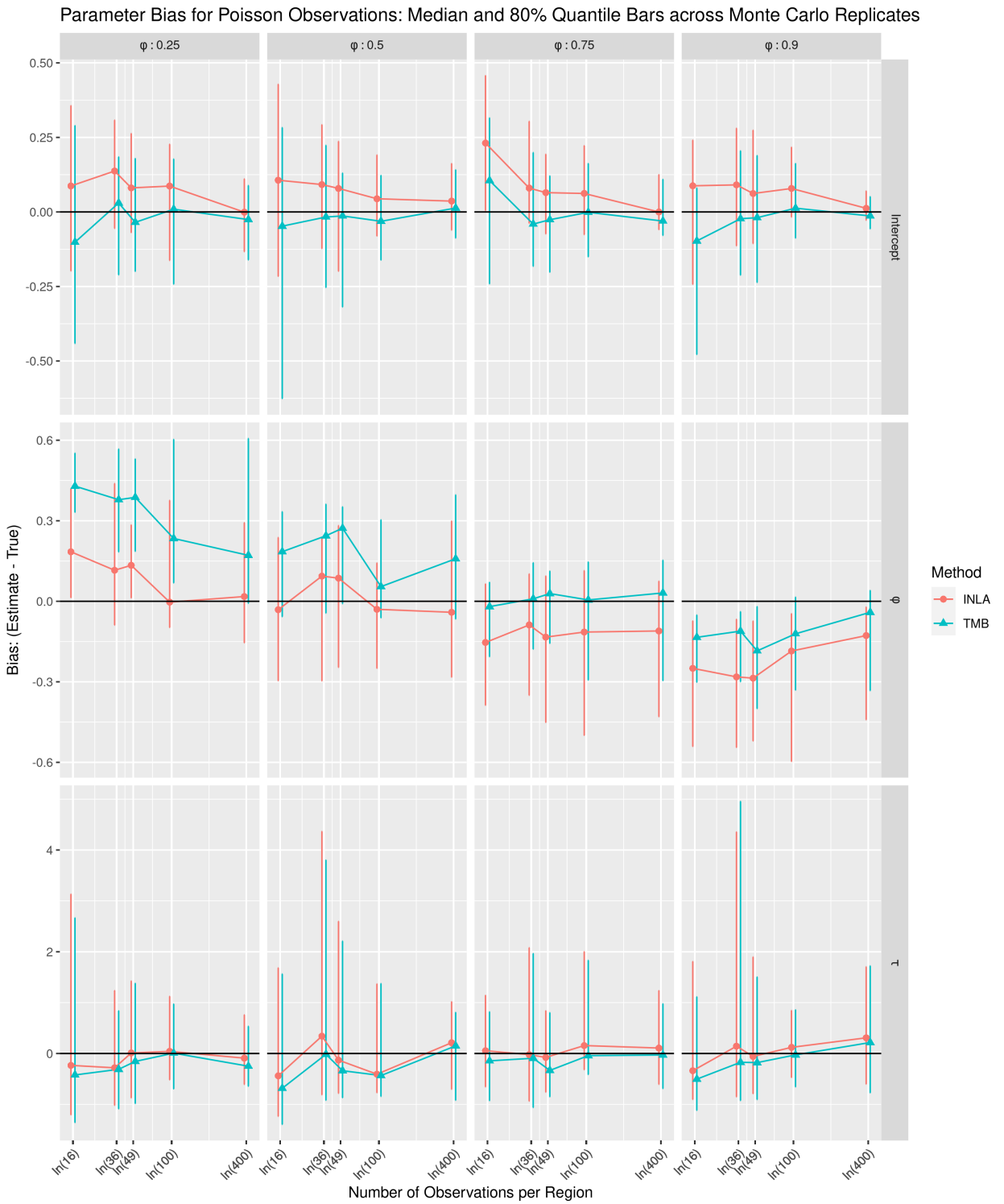


Figure 3.9: Comparison of the estimated parameter bias from `TMB` (red) and `R-INLA` (blue) plotted against the number of observations per region for Poisson data experiments with varying values of the true BYM2 φ. Each point is the median bias of 1 experiments, calculated across 25 replicates, and the bars represent the middle 80% quantile range of the bias across replicates.

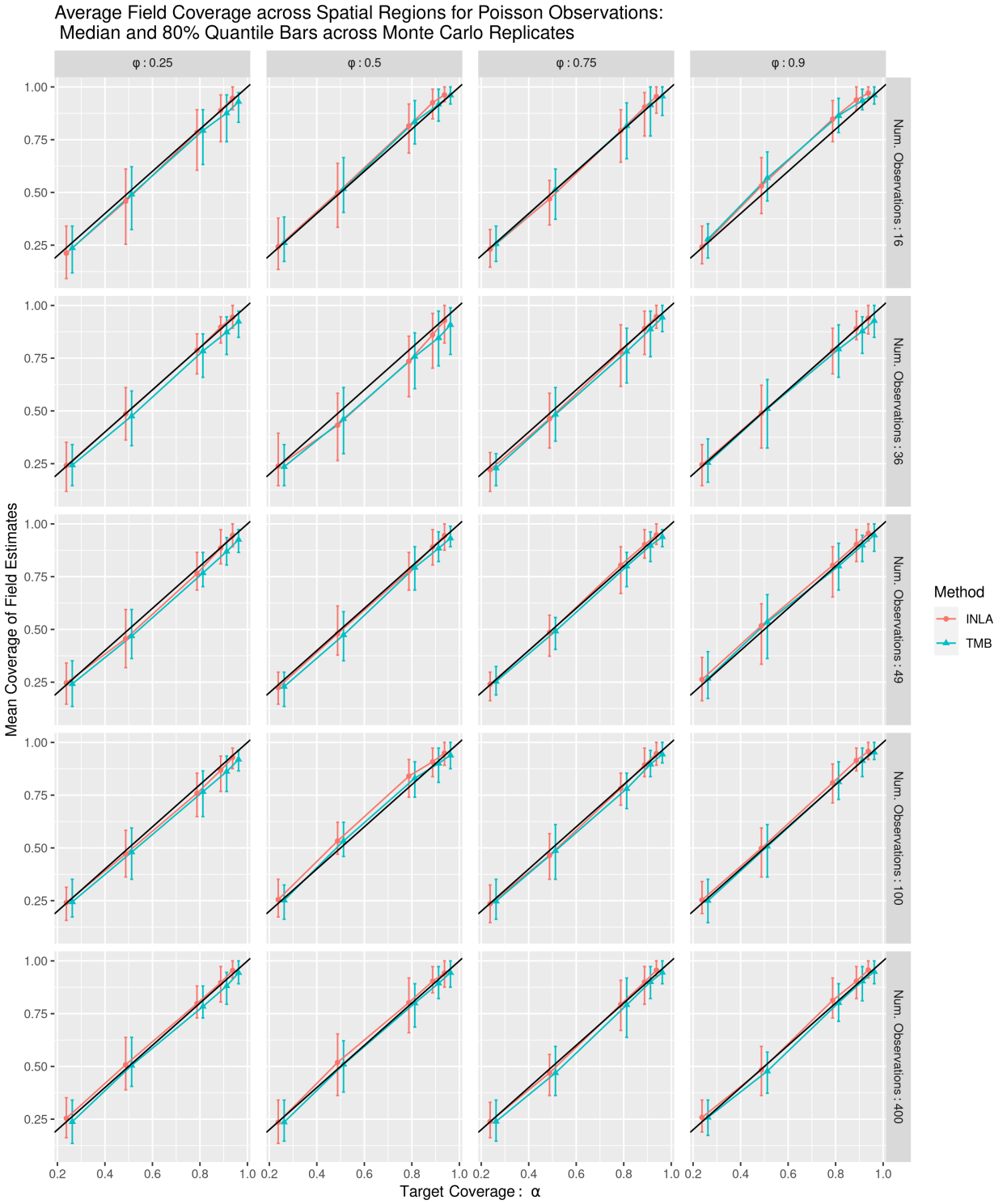


Figure 3.10: Comparison of the average estimated region coverage of the simulated truth, faceted by values of the true BYM2 φ and number of observations per region, from `TMB` (red) and `R-INLA` (blue), plotted against the target nominal coverage, α, for Poisson observation experiments. Each point is the median average coverage of an experiment, calculated across 25 replicates, and the bars represent the middle 80% quantile range of the average coverage across replicates.

produce very similar spatial field estimates though `TMB` appears more likely to have larger bias for some parameters.

3.7 European Breast Cancer Application Example

In this section, we provide a real spatial application which uses TMB. This section also serves as a gentle introduction to the full cancer application contained in Chapter 4. The base model used in this application was originally proposed by Mercer (2016) and requires a multiplicative interaction between a Poisson rate and a Binomial probability, each modeled with canonical links and their own linear predictors. They tried to implement this model at scale using Hamiltonion Monte Carlo methods implemented via `Stan` (Carpenter *et al.*, 2017) and found the problem intractable due to the induced correlation between the two spatial random fields. We use the same base model here, and later greatly extend it in Chapter 4, to demonstrate one of the clear advantages `TMB` has over `R-INLA`: `TMB's` flexible user-defined template allows for a larger class of models including these sorts of nonlinear interactions. Both the Poisson and the Binomial parameters are modeled with discrete spatial random effects over countries in and around Europe.

Data quality for cancer monitoring vary significantly across the EU and range from complete registry coverage and high quality mortality estimates from vital registration to countries with no data. We use `TMB` to fit a two-level nonlinear model which first assumes a Poisson model for cancer incidence and then, conditional on cancer counts, models deaths as a binomial outcome. This model appropriately handles the variety of cancer data present in the EU in part because the conditional two-level Poison-Binomial form induces a Poisson model for unconditional mortality to account for countries without incidence data from registries. Using data provided by the International Agency for Research on Cancer (IARC), we implement a Bayesian spatial smoothing model to borrow strength between countries to provide estimates of national incidence and mortality, along with measures of uncertainty.

The approach directly models mortality (which is more universally available) and the mortality-incidence (MI) ratio, to estimate both incidence and mortality for all countries. In

this example, we use a subset of the complete model implemented in Chapter 4. The model synthesizes four types of country data: type I countries with national incidence and mortality data, type II countries with sub-national incidence and mortality data (from registries) in addition to national mortality, type III countries with only national mortality, and type IV countries with no available data. While there is reason to think cancer incidence may be spatially correlated across countries, for example due to environmental and lifestyle risks, different preventions and screening strategies may result in large variability between nearby countries. For this reason, the BYM2 model presented in Section 3.6.2 was used to model a combination of spatial and iid country effects. We would also expect some smoothness in mortality over space due to similarities in GDP, and therefore healthcare, in close by countries.

For countries, c, that have both national mortality and incidence data (type I) we assume a Poisson process with rate p_c for cancer incidence, where Y_c is the number of reported individuals with breast cancer from a population of N_c. Conditional on having cancer, we model total mortality, Z_c, as a binomial outcome with probability of death r_c for each of the Y_c individuals with cancer. This induces a Poisson process for mortality when incidence is unobserved with rate $p_c \times r_c$. For illustration, we work with only with data from 2008 in women aged 50-54. Our base model for type I countries is:

$$Y_c|N_c, p_c \sim \text{Poisson}(N_c \times p_c), \quad p_c = \exp(\alpha_c^I) \tag{3.22}$$

$$Z_c|Y_c, r_c \sim \text{Binomial}(Y_c, r_c), \quad r_c = \frac{\exp(\alpha_c^{MI})}{1 + \exp(\alpha_c^{MI})} \tag{3.23}$$

which implies the unconditional mortality model:

$$Z_c|N_c, p_c \sim \text{Poisson}(N_c q_c), \quad q_c = p_c \times r_c. \tag{3.24}$$

We assume log- and logit-linear models for incidence and conditional mortality and we assign

the following forms:

$$\alpha_c^I = \alpha^I + b_c^I \tag{3.25}$$

$$\alpha_c^{MI} = \alpha^{MI} + b_c^{MI} \tag{3.26}$$

where p_c is the reported incidence, r_c is the reported mortality, α^I and α^{MI} are global intercepts, and b_c^I and $b_c^{MI}$are country random effects that are assumed to have BYM-2 structure comprising of a spatially correlated term as well as an unstructured (iid) country specific term. Additional details on the data and the model may be found in Appendix A.3. Maps of $\hat{b}_c^I$, $\hat{b}_c^{MI}$, $\widehat{p}_c$, and $\widehat{r}_c$, with measures of uncertainty, are presented in Figure 3.11.

The nonlinear unconditional mortality rate, $q_c = p_c \times r_c$, is necessary to include countries with incomplete or missing mortality data and prohibits this model from being fit within INLA. A similar model that used country-level fixed effects without the spatial random effect would be possible in INLA (Meehan *et al.*, 2020) but without complete data in each country this is not feasible.

3.8 Discussion

We were pleasantly surprised to find near concurrence in the distribution of spatial field estimates from `TMB` and `R-INLA` – in both continuous and linearly constrained discrete model settings – across a wide range of simulation parameters. We suspect the generally smaller parameter bias of the `R-INLA` results is due to the integration over the fixed effects, similar to restricted maximum likelihood (REML) inference which is known to reduce bias, in contrast to the ML inference performed in `TMB`. One possible remedy for this could be to effectively enable REML in `TMB` by adding all linear fixed effects parameters to the list of 'random' parameters defined by the modeler.

The field coverage results shown in Figures 3.5 and 3.6 (and additional figures shown in Appendix 3.6.1), clearly demonstrate that random effects coverage is a function of the magnitude of the effect (Yu and Hoff, 2018) and should serve as a warning for those trying

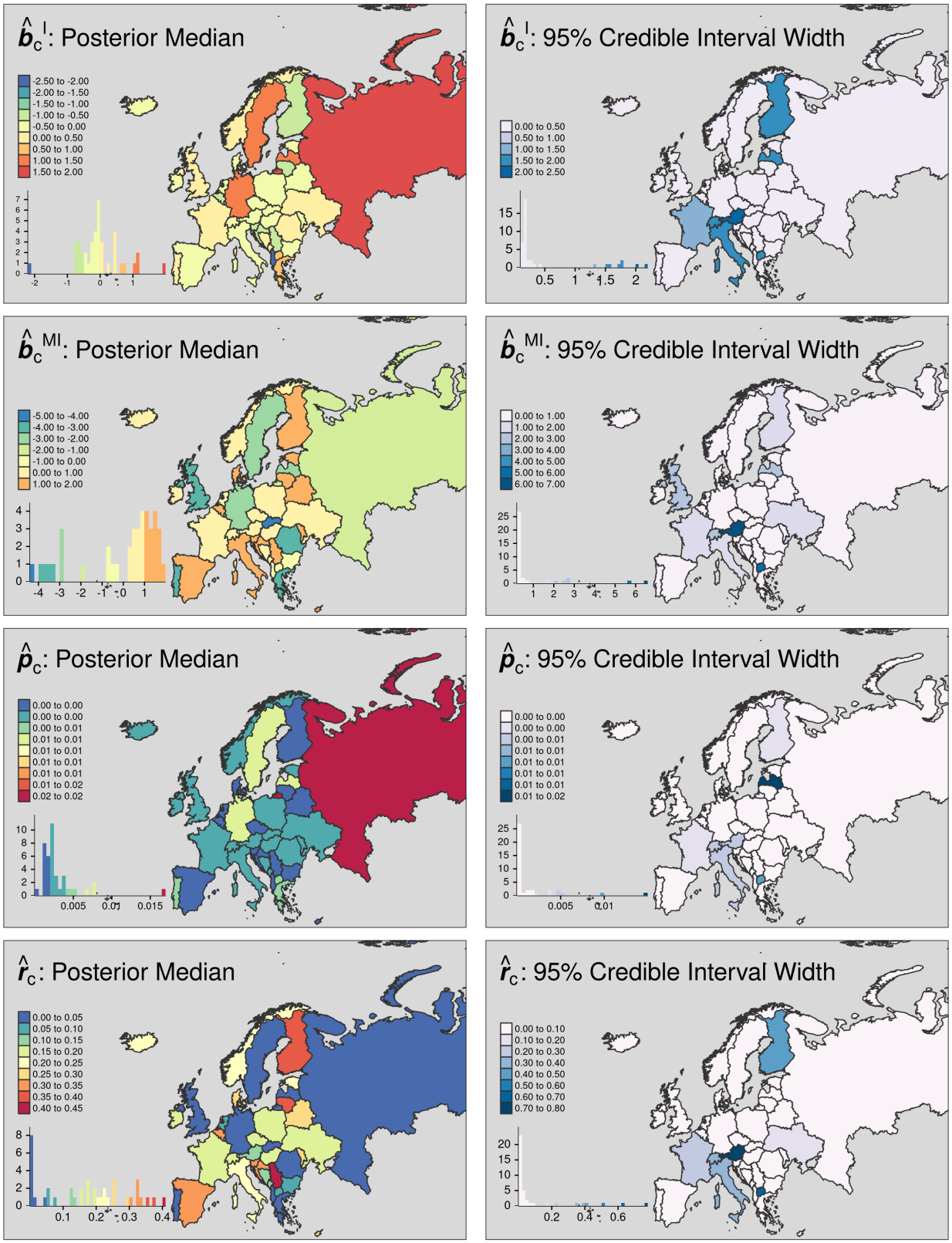


Figure 3.11: Each row consists of a pair of posterior median and 95% credible interval widths for the estimated quantity. Row 1: the BYM2 country random effects for incidence, row 2: the BYM2 country random effects for MI ratio, row 3: estimates country incidence rates, and row 4: estimated country mortality—incidence probabilities.

to interpret coverage of specific portions of estimated spatial fields. We discuss this feature in far greater detail in Chapter 5. The discrete Poisson experiment showed excellent recovery of the BYM2 field and the continuous models, while the SPDE approximation to the GP performed nearly as well in both the Gaussian and binomial data simulations. In fact, we see that the SPDE approximation in the binomial scenarios appears more robust than in the Gaussian and the undercoverage observed under extremely large data volumes warrants further investigation. Even so, the SPDE approximation performs very well across a wide range of experiments, and the user should be able to confidently determine when the approximation is resolved by increasing the density of the mesh until the results no longer change. We believe these to be the largest-scale simulation studies on the SPDE approach to fitting GPs in `TMB` or in `R-INLA`, particularly for non-Gaussian data, and we hope that the demonstrated success encourages continued use of this convenient approach.

The simulations also provided an opportunity to assess the relative computational burdens of `TMB` and `R-INLA` and Section 3.6.1 detailed both serial (one CPU thread) and parallelized timing experiments. Figures 3.7 and 3.8 demonstrate that `TMB` scales extremely well, even in comparison to `R-INLA` which was designed, in part, to be a computationally efficient and quick alternative to MCMC sampling. We note that timing tests attempting to replicate 'real world' experience are quite difficult and these timing results only provide some crude measures of computational cost. Nonetheless, `TMB` appears to perform quite favorably and, at minimum, it could be used to quickly test and iterate models before final inference is performed using the modeler's method of choice.

One of the main limitations of this study is the lack of tuning of any particular simulation and inference. Due to the sheer number of experiments, a set of default model parameters (for example, the starting values) were used across simulations. The simulation code was built to catch and relaunch convergence issues in both `TMB` and `R-INLA` that may, with human interaction, have been remedied. While sensible tuning decisions were attempted, this also implies a lack of sensitivity analysis on the prior selection for any one simulated dataset or experiment. Furthermore, these simulation studies use correctly specified models and do not

attempt to study either tool under model misspecification.

The novel incidence-mortality model from Section 3.7 demonstrates the utility of the flexibility TMB provides. Simulation results for this model shown in Figure A.3.3 indicate that TMB is capable of recovering parameter estimates from nonlinear mixed effects models that cannot be fit within R-INLA or easily with MCMC. One of the main tradeoffs in return for this flexibility is relative difficulty of constructing the C++ templates. There has already been efforts, such as the glmmTMB R-package (Brooks *et al.*, 2017) which allows users to run TMB from within R using standard R notation for mixed effects models, to streamline the use of TMB, but none will likely be able to offer the freedom of directly coding within the C++ templates. We note that there are numerous tutorial documents and examples available on the TMB github page and the TMB authors have worked to make it easier for R coders to use. Example code for fitting spatial GPs via the SPDE approach is available in Appendix A.4 to provide a sense of modeling in TMB and R-INLA, and the full code used in this study is available online to serve as starting points for those interested in spatial modeling in TMB. Furthermore, others have previously used TMB in spatial settings, such as classic spatial models (for example, Dwyer-Lindgren *et al.*, 2016), and more recent work to account for missing spatial information (Wilson and Wakefield, 2020; Wilson and Wakefield, 2021; Marquez and Wakefield, 2021), and their code is available too.

The primary restriction of TMB's applicability lies in the assumptions underlying the LA. Although many models may be fit within the TMB framework, it is clear that the LA will not perform equally well in all cases - though an intelligently chosen reparameterization, like those used for the hyperparameters in both simulations, can help significantly. While the quality of the LA may be difficult to assess, the recent availability of the tmbstan package permits MCMC sampling from TMB models - with or without the integration of the random effects performed by LA. By comparing results run with TMB and the LA against those run with MCMC without the LA, it is possible to assess the quality of the LA in particular applications. The speed of TMB makes it a useful option for iterating through models during exploration phases of research and it may provide opportunities to fit models that could not

otherwise be fit in reasonable amounts of time or at all.

After conducting these extensive simulations, we have found `TMB` to perform more than adequately in comparison to existing well-trusted tools. We hope that this unified and detailed explanation of `TMB` and the additional new model demonstrations will improve understanding, instill confidence, and generate further interest and use in a compelling and generally unknown statistical computational tool.

Chapter 4

JOINT MODELING OF CANCER INCIDENCE AND MORTALITY: ESTIMATING RATES OF BREAST CANCER IN EUROPE

4.1 Introduction

Cancer incidence reporting lags behind the availability of mortality data but is arguably more important for planning. In this chapter we jointly model breast cancer incidence and mortality in women across 39 European countries, 18 age groups, and 18 years of data. The core idea is to use a trio of dependent likelihoods for incidence, mortality conditional on incidence and unconditional mortality in order to leverage estimated mortality:incidence ratios and mortality to backcast recent missing years of incidence data. Structured random effects across age, location, and time are used within a Bayesian hierarchical model to jointly estimate incidence, the MI ratio, and mortality via a non-linear relationship between the three quantities. Inference is performed using Template Model Builder. A final model is selected from a pool of candidates and we demonstrate its appropriateness. We conclude with a discussion comparing our estimates against those from the International Agency for Research on Cancer and estimates from the Institute of Health Metrics and Evaluation's 2019 Global Burden of Disease study, which is non-reproducible and employs ad-hoc methods with unknown statistical properties.

Estimates from GLOBOCAN 2020 (Sung *et al.*, 2021) and the Global Burden of Disease 2019 update (Kocarnik *et al.*, 2021) show that cancer is one of the leading causes of worldwide deaths, that cancer incidence and mortality have been growing during the last decade, and that the global burden of cancer is expected to continue increasing during the next decades (Bray *et al.*, 2012; Foreman *et al.*, 2018). The increase in incidence and mortality attributed

to cancer is likely to be most noticeable in places with aging populations or societies where mortality rates due to communicable diseases are decreasing and life expectancy is improving. As a consequence, accurate estimation of cancer incidence and mortality is of particular interest in these places and it is important for research, planning, and evaluation of cancer control programs the world over.

Cause-specific mortality estimation is often troubled by incomplete and missing data. One advantage to studying cancer incidence rates is that cancer registries are relatively common and were designed, in part, to address this issue. More than half of all countries globally only have local incidence data available to pair with mortality data which ranges from complete vital registration to none. When local registry and/or mortality data is incomplete, the observed relationship between incidence and mortality can be used to estimate the missing counts. The mortality-incidence (MI) ratio is often used to infer the relationship between incidence and mortality (Forouzanfar *et al.*, 2011; Uhry *et al.*, 2013), while other methods use mixtures of survival analyses, MI ratios, and incidence-mortality (IM) ratios (Ferlay *et al.*, 2010b, 2013, 2018; Sung *et al.*, 2021).

The International Agency for Research on Cancer (IARC), the specialized cancer agency of the WHO, provides worldwide estimates and predictions of cancer incidence and mortality for the major types of cancer through the GLOBOCAN project (Parkin *et al.*, 2001; Ferlay *et al.*, 2010b, 2013, 2018; Sung *et al.*, 2021) and has provided statistics on cancer incidence and mortality for the countries in Europe since 1980. Available cancer data quality and sources vary widely between countries and even within countries across time, and these estimates have been used globally by public health organizations. Data availability ranges from complete registry coverage (for example, Scandinavia), providing high-quality subnational data resolution, to incomplete or nonexistent national mortality data in countries without vital registration. The current IARC modeling procedure informally borrows information from neighboring countries when data quality is low. In order to handle the various data qualities and availabilities in different countries, the IARC approach utilizes a number of different models. Prior to 2018, their methods did not provide any measures of uncertainty,

but their recent publications now include such measures (Ferlay *et al.*, 2019).

In the last decade the Institute for Health Metrics and Evaluation (IHME) has done extensive work to compile large amounts of health data and complete descriptive analyses on many public health and demographic indicators. Their efforts are wide-ranging including estimates of child mortality (Rajaratnam *et al.*, 2010a; Wang *et al.*, 2014, 2017; Burstein *et al.*, 2019), maternal mortality (Hogan *et al.*, 2010; Kassebaum *et al.*, 2016), all cause mortality by gender (Rajaratnam *et al.*, 2010b; Roth *et al.*, 2018), worldwide education levels (Gakidou *et al.*, 2010), breast and cervical cancer (Forouzanfar *et al.*, 2011), the global burden of all cancers (Fitzmaurice *et al.*, 2015, 2017), and many more. Across the IHME analyses they process and pool together a large quantity of data, often from diverse sources, to generate the desired country-specific annual rates. The IHME methods provide estimates with uncertainty intervals, but their methods used to generate the intervals are complex and contain ad hoc steps which ensure statistical properties, such as frequentist coverage, of their uncertainty intervals are unknown. Their process partitions a total global envelope for all-cause mortality into mortality by cause. This ensures consistency in global mortality counts, but it induces correlation across all different causes of mortality and incidence which means their work is typically not reproducible. Further, due to the volume of estimates they regularly update, they often use the same standard method for a variety of endpoints, which one would expect to be suboptimal.

We use data provided by IARC to generate estimates of reported country- and age-specific rates of cancer cases and deaths in women for 39 countries in Europe from 2000-2017. In 2008, the most data abundant year, 37 of the 39 countries have some mortality data, and 28 of them have national mortality data. In 2008, only 20 countries have national incidence data, but 30 countries have at least some incidence and mortality data. Twenty countries have both national incidence and mortality, and ten have both types of data available from local registries. Greece and Ukraine have no data whatsoever. A summary of the quantity of available data by country is available in Table B.1.1.

IARC has standardized their estimation approach by creating an alphanumeric scoring

system to classify the availability and quality of incidence and mortality data in each country. These scores are available for all countries included in the GLOBOCAN database. The methods described in this chapter collapse these scores into six data availability categories determined by both the availability and the geographic resolution (national and/or subnational) of incidence and mortality data sources.

Cancer is a complex collection of distinct diseases, and incidence for each type of cancer varies across space, time and different populations, such as those with different age structures. In this study, we focus on estimating incidence of and mortality due to breast cancers (malignant tumors which develop in cells of the breast). Although we expect differences across space and time to generally be small, we avoid assuming constant rates to allow for known differences in cultural practices and environmental conditions. In addition, we know that the probability of incident cases appearing in surveillance can vary greatly by country and across time within countries, and, general improvements in health systems have led to higher reported rates of incidence and lower reported mortality rates. Common underlying risk factors that lead to cancer suggest we would expect similar incidence rates across space which led us to hypothesize the need for spatial smoothing terms within our model. Temporal changes in cultural and environmental risk factors are likely to be small, or at least slowly varying, similarly suggesting the inclusion of temporal smoothing. Finally, we expect rates in adjacent age groups to be relatively similar, warranting the inclusion of smoothing terms over age groups.

Our aim is to use Bayesian space-time-age smoothing models to borrow strength between observations that are recorded for close by spatial locations, years, and age groups in order to make predictions for all country-time-age combinations, including those with no data observations. By modeling mortality, incidence, and the mortality-incidence relationship in places where both mortality and incidence data are available, we can leverage the MI ratio to predict missing mortality or incidence when only one data type is available. Since mortality data is often more quickly available than incidence sources, this is particularly useful to aid in backcasting recent missing years of incidence, which is important for near-term planning.

In addition to borrowing strength across similar locations, ages, times, and using the MI ratio to impute missing cases or deaths, this approach also provides measures of uncertainty for the mortality and incidence estimates across the complete space-time-age cube.

Our overall approach to modeling incidence and mortality is to directly model mortality (which is more universally available at the national level in Europe) 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 many only have subnational incidence data available from local registries. For countries with subnational 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 subnational registries, is inferred based on the difference between the national and subnational mortality counts. 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. The national and subnational registry data provided by IARC is described in Section 4.2. The joint model for mortality rates and the MI ratio is developed in Section 4.3.1, the joint likelihood across all data types is defined in Section 4.3.2, and a spatial simulation study is detailed in Section 4.3.3. Section 4.3.4 presents a suite of ten models with various age, space and time interactions to be considered in our full analysis of European breast cancer data, which is conducted in Section 4.4 and discussed in Section 4.5.

4.2 European Breast Cancer Data

IARC assigns incidence data quality scores based on the availability and coverage of registries within the country. The quality scores range between A (high quality national or regional data with greater than 50% coverage) to G (no data). Similarly, mortality data scores are also assigned based on completeness and quality, and they range from 1 (high quality complete vital registration) to 6 (no data). Complete definitions for the IARC quality scores are shown in Appendix B.1.2 and Appendix Tables B.1.2 and B.1.3.

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 six types of data; (I) national incidence and mortality, (II) subnational incidence and mortality and national mortality, (III) only national mortality, (IV) only national incidence, (V) only subnational data, and (VI) no available data. The relationship between the IARC scores and our categories are shown in Table B.1.4. The available data type for each country and year is shown in the top of Figure 4.1. Complete details about the number of registries, years of incidence data, years of mortality data, and data types for the 39 European countries are provided in Appendix B.1 .

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 article, each different subnational registry constitutes a data observation that enters the model. 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.

4.3 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 national or local incidence data by providing an explicit link between mortality and incidence.

4.3.1 Joint incidence, MI ratio, and mortality model

We begin by establishing notation with a indexing 5-year age-groups, c indexing country, t indexing time in years, L indicating local (registry) data from the L^{th} subnational registry in country c at time t, and R denoting the remainder data (not covered by local registries). The number of subnational registries varies by countries and times but we suppress that notation and allow $L \in \{1, \ldots, n_{ct}^L\}$ as appropriate for any country-time index ct.

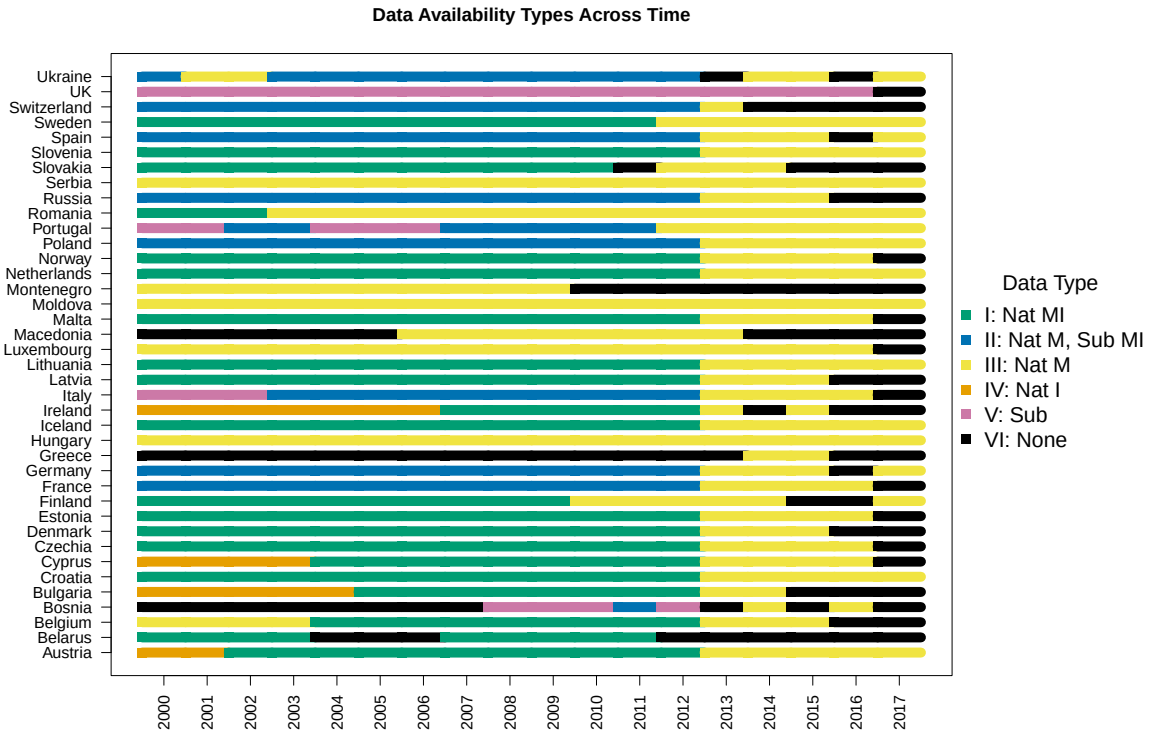


Figure 4.1: Data type by country from 2000–2017.

For age group a, country c, and time t we define:

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

- Y_{act}^{L} = reported cases (incidence) in time t, age group a, and local area L in country c (cases reported at registry L)

- Z_{act}^{L} = reported deaths (mortality) in time t, age group a, and local area L in country c (deaths reported at registry L),

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

- Y^R_{act} = reported cases (incidence) remainder in time t, age group a, and country c among the remainder population (cases not covered by registries),

- Z^R_{act} = reported deaths (mortality) remainder in time t, age group a, and country c among the remainder population (deaths not covered by registries),

- $Y_{act} = \sum_L Y^L_{act} + Y^R_{act}$ = all reported cases in time t, age group a and country c,

- $Z_{act} = \sum_L Z^L_{act} + Z^R_{act}$ = all reported deaths in time t, age group a and country c,

- $N_{act} = \sum_L N^L_{act} + N^R_{act}$ = total population in time t, age group a and country c,

- $p_{act} = P(\text{incident case}|a, c, t)$ = reported incidence risk in time t, age group a and country c,

- $q_{act} = P(\text{ death }|a, c, t)$ = reported unconditional mortality risk in time t, age group a and country c, and

- $r_{act} = \Pr(\text{death }|\text{incident case}, a, c, t)$ = the MI ratio in time t, age group a and country c.

We first describe the overarching model that we assume underlies all incidence, mortality conditional on incidence, and mortality count data across all observed data sources. Putting aside, for the moment, differences between national and local registry data, our approach assumes the following two conceptual models for incidence and mortality given incidence, respectively:

$$Y_{act}|N_{act}, p_{act} \sim \text{Poisson}(N_{act} \times p_{act}) \tag{4.1}$$

$$Z_{act}|Y_{act}, r_{act} \sim \text{Binomial}(Y_{act},\ r_{act}). \tag{4.2}$$

The log linear model in (4.1) acknowledges the rare disease assumption. This pair of models imply the (unconditional) mortality model:

$$Z_{act}|N_{act}, q_{act} \sim \text{Poisson}(N_{act} \times q_{act}), \qquad q_{act} = p_{act} \times r_{act}. \tag{4.3}$$

Due to the constrained relationship between the incidence rate, p, the mortality conditional on incidence probability, r, which is the MI ratio, and the unconditional mortality rate q, we only model two of there parameters directly and infer the third. Since one of the primary objectives of this chapter is to use mortality counts to fill in missing incidence observations, we choose to model the MI ratio and the mortality risk. Using canonical link functions, we assume the following forms for the linear predictor:

$$\log(q_{act}) = \eta^M_{act} = \alpha^M + (\beta^M_a a + b^M_a) + b^M_c + (\beta^M_t t + b^M_t) + \dots \tag{4.4}$$

$$\text{logit}(r_{act}) = \eta^{MI}_{act} = \alpha^{MI} + (\beta^{MI}_a a + b^{MI}_a) + b^{MI}_c + (\beta^{MI}_t t + b^{MI}_t) + \dots, \tag{4.5}$$

which gives the incidence rate as:

$$p_{act} = \frac{q_{act}}{r_{act}} = \frac{\exp(\eta^M_{act})(1 + \exp(\eta^{MI}_{act}))}{\exp(\eta^{MI}_{act})}. \tag{4.6}$$

The ellipses are included to denote that additional higher-order random effect interactions are included in some of the hypothesized model formulations. These are described in detail in Section 4.3.4. In practice, for observations from a specific age-country-time strata, the women dying from breast cancer are not, in general, women who were diagnosed with breast cancer in the same year. There is a lead time from diagnosis to death that may span years. Our model relies on the simplifying assumption that the incidence will not change dramatically from one year to the next.

The linear predictors for mortality, η^M, and the MI ratio, η^{MI} are comprised of global intercepts, α, log-linear terms for age and time, β_a and β_t, and structured random effects, $b^{\bullet}_{\bullet}$, with BYM2 priors developed by Riebler *et al.* (2016) and described in Appendix B.3.3, across

5-year age-groups, countries, and years. Different combinations of higher-order random effects interactions were assessed to select a final model from a suite of candidate models, as described in Section 4.3.4. This model formulation and parameterization is applicable to all country-time-ages. The appropriate likelihood for each of the five different observed data types (I-V) can be constructed from components of the underlying model described in (4.1)–(4.6).

4.3.2 Likelihoods for the five data types

Next, as an illustrative example, we describe how the model from the previous section can be used to construct the data likelihood for type II data observations (subnational incidence and mortality and national mortality). This provides an example of the three components necessary to construct the likelihood for observations belonging to all data types.

For country-times with local incidence and mortality data from registries and national mortality data, our goal is to use the observed local MI ratio in conjunction with the national mortality to estimate the total national incidence. Our approach relies on jointly modeling the local incidence and mortality in addition to the national mortality counts that are not contained in the registries. Our MI ratio model assumes that the underlying local registry and national rates are identical for each age-country-time strata.

For the local data, data from each registry, L, is included as a data observation,

$$Y^L_{act}|N^L_{act}, p^L_{act} \sim \text{Poisson}(N^L_{act} \times p_{act}), \qquad p_{act} = \frac{q_{act}}{r_{act}} \tag{4.7}$$

$$Z^L_{act}|Y^L_{act}, r^L_{act} \sim \text{Binomial}(Y^L_{act},\ r_{act}), \qquad r_{act} = \frac{\exp(\eta^{MI}_{act})}{1+\exp(\eta^{MI}_{act})}, \tag{4.8}$$

as is the unconditional mortality remainder,

$$Z^R_{act}|N^R_{act}, q_{act} \sim \text{Poisson}(N^R_{act} \times q_{act}), \quad q_{act} = \exp(\eta^M_{act}). \tag{4.9}$$

The linear predictor parameters are assumed to be of the form (4.4) and (4.5). We see

that the available incidence observations, (4.7), available mortality data when conditioned on incidence, (4.8), and available mortality data without incidence (4.9), take the corresponding general forms put forth in model described by (4.1)–(4.3).

Our model formulation models the observed incidence, using the quotient of two parameters, q_{act} and r_{act}. Both parameters are influenced by all the data: the mortality parameter, q_{act} enters directly into the likelihood for the national mortality data but it also features in the likelihood for the local incidence data. Likewise, the MI parameter is directly influenced by the local incidence and mortality data but the joint formulation, made explicit by the nonlinear relationship used to define the incidence parameter, p_{act}, means the national mortality remainder also contributes to the MI parameters estimates. By combining the modeled MI ratio, estimated from the local registries, with the remainder mortality, the model allows inference on the unobserved national incidence rate. Depending on the registry coverage for a country-time, as $\sum_L N^L/N$ approaches one, the parameters will be more heavily influenced by data from the registries, and when $\sum_L N^L/N$ is small the parameters will be more heavily influenced by the national mortality data. Let ω_2 index country-times with subnational incidence and mortality data from registries and national mortality data, and then define:

- $\boldsymbol{N}^{(2)L} = \{N^L_{act} : \{c,t\} \in \omega_2\}$ local population served by registry L in age a in type II country-times,
- $\boldsymbol{N}^{(2)R} = \{N^R_{act} : \{c,t\} \in \omega_2\}$ national remainder population not covered by any local registry in age a in type II country-times,
- $\boldsymbol{y}^{(2)L} = \{y^L_{act} : \{c,t\} \in \omega_2\}$ local incidence data from registry L in age a in type II country-times,
- $\boldsymbol{z}^{(2)L} = \{z^L_{act} : \{c,t\} \in \omega_2\}$ local mortality data from registry L in age a in type II country-times,

- $\boldsymbol{z}^{(2)R} = \{z^R_{act} : \{c,t\} \in \omega_2\}$ national mortality remainder data in age a in type II country-times,
- $\boldsymbol{b}^{(2)M} = \{b^M_\bullet : \{c,t\} \in \omega_2\}$ age-country-time-specific random effects for mortality in type II countries, and
- $\boldsymbol{b}^{(2)MI} = \{b^{MI}_\bullet : \{c,t\} \in \omega_2\}$ age-country-time-specific random effects for MI in type II countries.

The likelihood of the data from type II countries is

$$p(\boldsymbol{y}^{(2)L}, \boldsymbol{z}^{(2)L}, \boldsymbol{z}^{(2)R} | \boldsymbol{N}^{(2)L}, \boldsymbol{N}^{(2)R}, \alpha^I, \beta^M_a, \beta^M_t, \boldsymbol{b}^{(2)I}, \alpha^{MI}, \beta^{MI}_a, \beta^{MI}_t, \boldsymbol{b}^{(2)MI}) =$$

$$\underbrace{p(\boldsymbol{y}^{(2)L} | \boldsymbol{N}^{(2)L}, \alpha^I, \beta^M_a, \beta^M_t, \boldsymbol{b}^{(2)M}, \alpha^{MI}, \beta^{MI}_a, \beta^{MI}_t, \boldsymbol{b}^{(2)MI})}_{\prod_l \prod_{\{c,t\}\in\omega_2} \prod_a \text{Poisson}(N^{(2)L}_{act} \times p_{act})} \times \tag{4.10}$$

$$\underbrace{p(\boldsymbol{z}^{(2)L} | \boldsymbol{y}^{(2)L}, \alpha^{MI}, \beta^{MI}_a, \beta^{MI}_t, \boldsymbol{b}^{(2)MI})}_{\prod_l \prod_{\{c,t\}\in\omega_2} \prod_a \text{Binomial}(Y^L_{act},\ r_{act})} \times \tag{4.11}$$

$$\underbrace{p(\boldsymbol{z}^{(2)R} | \boldsymbol{N}^{(2)R}, \alpha^I, \beta^M_a, \beta^M_t, \boldsymbol{b}^{(2)M})}_{\prod_{\{c,t\}\in\omega_2} \prod_a \text{Poisson}(N^{(2)R}_{act} \times q_{act})}, \tag{4.12}$$

and in the event that the registry L only reports incidence data, that Binomial likelihood term for conditional mortality is left out.

The individual likelihood densities that comprise (4.10)–(4.12) can be used to construct the likelihoods for data from all data types. For example, the likelihood for data from type I country-times, those with national incidence and mortality data and no unconditional mortality counts, will take the from of the product of (4.10) and (4.12), where as the likelihood for data from type III country-times, with only national mortality, will take the form of (4.12). The likelihoods for each of the six data types, along with pertinent parameters, corresponding definitions, and all other model details, including prior and hyperprior specification, are described in Appendix B.3.

4.3.3 Spatial Simulation

This section details a brief simulation study designed to assess the model's ability to estimate the true incidence and mortality rates for a space-only version of the mortality, incidence, and MI ratio model described in Section 4.3.

Generation of Spatially Correlated Data

Five hundred synthetic data sets were simulated to be similar to the observed data for women age 50–54 in Europe for a single year, 2008. Population sizes (N_c and N_c^L) were taken to be the populations in from the real data. Populations ranged from 8,562 in Iceland to 5,299,659 in Russia. The observed data type (I-VI) for each country in 2008 was recreated in the synthetic dataset, as was the number of local registries and their respective populations. True parameters for each synthetic dataset were selected to generate mortality rates (q_c) and MI ratios (r_c), and thus also incidence rates ($p_c = q_c/r_c$), with distributions similar to the observed rates for this age and year. Specifically, mortality rates were generated using:

$$p_c = \exp\left(\alpha^M + \frac{1}{\sqrt{\tau^M}}\left(\sqrt{1-\phi_c}v_c^M + \sqrt{\phi_c}\tilde{u}_c^M\right)\right)$$

where $\alpha^M = -7.5,\ \tau^M = 0.2, \phi^M = 0.8$, and $\boldsymbol{b}^M = \frac{1}{\sqrt{\tau^{MI}}}\left(\sqrt{1-\phi_c}v_c^{MI} + \sqrt{\phi_c}\tilde{u}_c^{MI}\right)$ is defined to be the BYM2 GMRF defined in Appendix B.3.3. Similarly, the MI ratios were generated using:

$$r_c = \text{expit}\left(\alpha^{MI} + \frac{1}{\sqrt{\tau^{MI}}}\left(\sqrt{1-\phi_c}v_c^{MI} + \sqrt{\phi_c}\tilde{u}_c^{MI}\right)\right)$$

where $\alpha^{MI} = -1.25, \tau^{MI} = 0.2, \phi^{MI} = 0.8$, and $\boldsymbol{b}^{MI} = \frac{1}{\sqrt{\tau^{MI}}}\left(\sqrt{1-\phi_c}v_c^{MI} + \sqrt{\phi_c}\tilde{u}_c^{MI}\right)$ is defined as independent BYM2. There were 20 type I countries, 9 type II, 7 type III, 0 type IV, 2 type V, and 1 type VI. Registry coverage (N_c^L/N_c) in type II countries varied between 1.8%–63%. Conditional on the synthetic parameters, national cases were generated from Poisson($N_c^R \times p_c$) and registry cases from a Poisson($N_c^L \times p_c$). For observations with

observed incidence, deaths were generated from Binomial(Y_c^R, r_c) and Binomial(Y_c^L, r_c) distributions, otherwise they were generated from Poisson($N_c^R \times q_c$) or Poisson($Y_c^L \times q_c$) distributions. An example of the simulated fields used to generate one synthetic dataset is illustrated in Figure 4.2.

Coverage of Credible Intervals

A space-only model, simplified from the age-space-time methods described in Section 4.3, was implemented on the simulated data using a Bayesian approach with computation via Template Model Builder (TMB) (Kristensen *et al.*, 2016). Figure 4.3 plots a summary of the posterior distributions for incidence rates, mortality rates, and the MI ratio from a single simulated data set. Table 4.1 summarizes the mean bias, absolute error, and 3 different target posterior coverage probabilities averaged across the 1000 simulated datasets. Both show strong concordance between the estimates and the simulated observations, suggesting that the our nonlinear model formulation between mortality, the MI ratio, and incidence as implemented via TMB can adequately recover the observed rates. Across the 1000 simulations, the true incidence was contained within 94.5% of the nominal 95% credible intervals, the true mortality in 94.4%, and the true MI Ratio in 94.3%. As expected, the true incidence and MI ratio is harder to recover in type III countries, which only have national mortality data, and this is generally reflected in the width of their credible intervals. We also see expected results for type II countries, those with local incidence and mortality data and national mortality, with tight mortality intervals and MI ratio and incidence intervals which depend on the local registry coverage, N_c^L/N_c.

4.3.4 Modeling Higher-Order Interactions

The primary dimensions across which we aim to estimate national incidence are age, location, and time. After the joint model for mortality and the MI ratio was formulated and vetted via simulation, the primary remaining modeling decisions relate to the specification of the linear predictor across these three dimensions. As previously discussed, we want to avoid constant

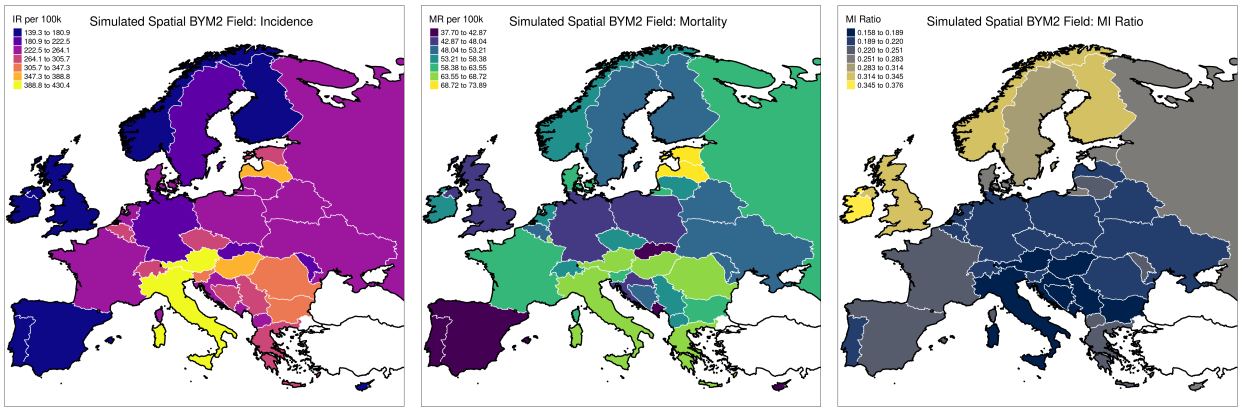


Figure 4.2: One realization of the simulated BYM2 fields for mortality and the MI ratio, along with the incidence field that they imply.

Measure	Obs. per 100k	Bias per 100k	Abs. Err. per 100k	$PC^{.8}$	$PC^{.9}$	$PC^{.95}$
Mortality	56.453	-0.163	3.686	0.795	0.894	0.944
MI Ratio	0.224	-0.001	0.015	0.792	0.893	0.943
Incidence	258.406	-0.749	15.671	0.798	0.895	0.945

Table 4.1: Summary metrics averaged across 1000 synthetic datasets. For mortality and incidence, the median simulated observation, bias and absolute error are shown per 100k population. Posterior coverage (PC), averaged across the 39 locations and 1000 synthetic datasets is shown for nominal 80%, 90%, and 95% probabilities.

rates across any of these dimensions. Exploratory work using independent generalized linear models (GLM) on mortality and the MI ratio suggested that linear terms in age and time did not suitably capture the variation in these dimensions, which led to the inclusion of independent structured random effects priors across the three dimensions and the base model proposed in (4.4) and (4.5). Fitting the base model in TMB and plotting the residuals across age, space, and time clearly showed structure correlation remaining across interactions of these dimensions, with more complex structures observed for the mortality rates. Residual plots from the base model and the final selected model are plotted in Figures B.5.4–B.5.7. These initial attempts helped shape a suite of ten candidate models from which we perform

Figure 4.3: Posterior median and 95% credible intervals plotted against the true simulated incidence (left) and mortality (center) and MI ratio (right) for a single realization of synthetic data.

model assessment and select a final model.

In the base model, which is nested in all other models, structured random effects across five-year age bands, denoted by b_a^M, b_a^{MI}, structured random effects across years, denoted by b_t^M, b_t^{MI}, and structured spatial random effects across countries, denoted b_c^M, b_c^{MI} are all given independent BYM2 priors defined in Appendix B.3.3. Higher-order structured random effect interactions, denoted by $\delta_{i,j}^M$, and $\delta_{i,j}^{MI}$ for $i, j \in \{a, c, t\}$, are given mean zero GMRF priors with covariance defined to be the Kronecker product between BYM2 covariances for the relevant pair of dimensions. For example, $\text{Cov}(\delta_{ac}^M) = \text{Cov}(\delta_{a\in ac}^M) \otimes \text{Cov}(\delta_{c\in ac}^M)$, denotes the country-age structured random effect interaction with covariance equal to the product between the covariance of a BYM2 prior over age, $\delta_{a\in ac}^M$, and the covariance of a BYM2 prior over location, $\delta_{c\in ac}^M$. Each of the component BYM2s used to generate the interactions, $\delta_{k\in ij}^\bullet$ for $k \in \{i, j\} \in \{a, c, t\}$, are parameterized independently from all other BYM2s across age, space, or time. In order to ensure identifibiliby, only a single scaling parameter is used in the covariance for each higher-order interaction. These interaction terms permit deviations from the main random effects across age, space and time. Including δ_{ac}, for example, allows for country-specific age effects where nearby countries are more likely to have similar age-

structures (or, equivalently age-specific spatial effects where similar age-groups are more likely to have similar spatial fields).

The linear predictors for all ten models are shown in Table 4.2. Models II-V add to the base model different interactions to the linear predictor for the MI ratio, often including the age-country interaction which the preliminary GLM explorations indicated was likely the most important interaction for both mortality and the MI ratio. An earlier model comparison we performed selected all two-way interactions for the mortality linear predictor leading to Models VI-X which repeat the MI ratio linear predictors from models I-V, but include all two-way interactions for mortality.

Table 4.2: Candidate specifications of the linear predictors for $\log(q_{act})$ and $\text{logit}(r_{act})$.

Model	Linear Predictor for $\log(q_{act})$	Linear Predictor for $\text{logit}(r_{act})$
I (base)	$\alpha^I + (\beta_a^I a + b_a^I) + b_c^I + (\beta_t^I t + b_t^I)$	$\alpha^{MI} + (\beta_a^{MI} a + b_a^{MI}) + b_c^{MI} + (\beta_t^{MI} t + b_t^{MI})$
II	base	base $+ \delta_{ac}^{MI}$
III	base	base $+ \delta_{ac}^{MI} + \delta_{ct}^{MI}$
IV	base	base $+ \delta_{ac}^{I} + \delta_{at}^{MI}$
V	base	base $+ \delta_{ac}^{I} + \delta_{at}^{MI} + \delta_{ct}^{MI}$
VI	base $+ \delta_{ac}^{M} + \delta_{at}^{M} + \delta_{ct}^{M}$	base
VII	base $+ \delta_{ac}^{M} + \delta_{at}^{M} + \delta_{ct}^{M}$	base $+ \delta_{ac}^{MI}$
VIII	base $+ \delta_{ac}^{M} + \delta_{at}^{M} + \delta_{ct}^{M}$	base $+ \delta_{ac}^{MI} + \delta_{ct}^{MI}$
IX	base $+ \delta_{ac}^{M} + \delta_{at}^{M} + \delta_{ct}^{M}$	base $+ \delta_{ac}^{MI} + \delta_{at}^{MI}$
X	base $+ \delta_{ac}^{M} + \delta_{at}^{M} + \delta_{ct}^{M}$	base $+ \delta_{ac}^{MI} + \delta_{at}^{MI} + \delta_{ct}^{MI}$

4.4 Application to European Breast Cancer

Using TMB, the ten models described in Table 4.2 were fit to breast cancer incidence and mortality data from the 39 European countries pictured in Figure 4.1, from 2000–2017, with eighteen age groups, (0-4, 5-9,..., 80-84, 85+). Posterior samples were drawn using a multivariate Gaussian approximation to the posterior as is standard when using TMB, and then were appropriately corrected to satisfy the sum to zero constraints. For models I–X,

1000 draws from the posterior distribution were taken. Each model required approximately 1.5–3 hours of computation time to fit and predict estimates using an Intel I7-8550U 1.8GHz CPU with 8 threads and 16GB of RAM.

In addition to the age-specific rates, we also calculated the European age-standardized rate (ASR) for both incidence and mortality, using a standard population structure to collapse our estimates across age in order to make rates from different populations more comparable. The ASR summary is further described in Eq. (B.4.1) and is used by IARC in their European-specific publications (Ferlay *et al.*, 2013). The remainder of this section details the model selection and assessment procedure before presenting summaries of the posterior distribution, estimated breast cancer incidence and mortality rates, and a brief comparison with comparable results published by IARC and GBD.

Model evaluation and assessment was performed using two likelihood metrics, the deviance information criteria (DIC) and the log pseudo-marginal likelihood (LPML), as well as metrics comparing the fitted model to the observed data such as the bias and root mean square error (RMSE) of the posterior median and the posterior predictive coverage (PPC) for both the observed incidence and mortality ASRs. The bias, RMSE, and coverage metrics were calculated across space-time, space, time, and collapsed to a single comprehensive summary metric. Additional details on these calculations is provided in Appendix B.3.6.

Figure 4.4 displays the DIC, effective number of parameters, and the LPML across the ten models. Table 4.3 contains the bias, standard deviation, RMSE, and $\text{PPC}^{.9}$ of the ASRs for incidence and mortality. Based on the model selection criteria, Model IX was selected. It demonstrated DIC, LPML, and posterior predictive coverage comparable those of model X, the most complicated model and the one which technically performs best in these metrics. Generally, Model IX performed well across all metrics, exhibiting the lowest mortality ASR bias and RMSE. While model IX did not perform best in any of the incidence metrics, it has the second best and reasonable $\text{PPC}^{.9}$, indicating that it adequately describes the variability in observed incidence ASR. Other models that performed better in incidence ASR bias, standard deviation, and RMSE have more severe PPC undercoverage. As the final deciding

factor, we note that the effective number of parameters for model IX is approximately 140 less than that of model X, and that model IX has a notably smaller bias, standard deviation and RMSE for the incidence ASR as compared to model X. The rest of this study focuses on the fit of model IX to the European breast cancer data from 2000–2017.

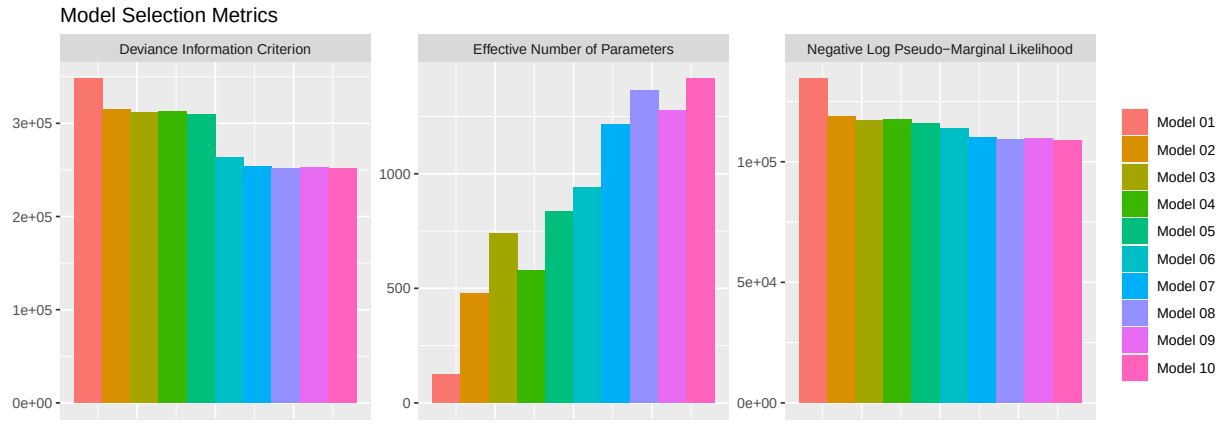


Figure 4.4: Comparison of model selection metrics, DIC, effective number of parameters, and LPML, across the ten models described in Table 4.2.

4.4.1 Validation of Projected European Breast Cancer

We performed an out-of-sample validation exercise to verify that the selected model from our candidate pool was capable of estimating missing incidence data. This check ensures that our model is not only best among the candidate models but that it is capable of backcasting recent missing incidence data. This validation was performed by fitting model IX to the European breast cancer data after withholding all incidence data from 2012 and beyond, the last year with substantial incidence data. The coverage of observed ASRs in 2012 by the 95% posterior predictive intervals was used to assess the backcasting capability of the model. There are a total of 34 countries with national mortality data in 2012, 16 countries with national incidence, 8 with subnational incidence, and three countries with no data.

Table 4.3: DIC and LPML from the posterior, bias, standard deviation and RMSE relative to the posterior ASR medians, and PPC of ASR for models described in Table 4.2.

Model	Incidence				Mortality			
	Bias	Std. Dev.	RMSE	$PPC^{.9}$	Bias	Std. Dev.	RMSE	$PPC^{.9}$
I	34.734	**1.749**	34.764	0.531	1.534	0.054	1.535	0.668
II	37.846	4.887	38.088	0.678	1.509	0.053	1.51	0.671
III	38.638	5.721	38.96	0.705	1.501	0.054	1.502	0.67
IV	38.576	5.392	38.848	0.685	1.508	0.054	1.509	0.669
V	39.801	6.463	40.197	0.723	1.508	**0.052**	1.509	0.671
VI	**34.217**	1.873	**34.253**	0.745	1.203	0.069	1.204	0.826
VII	36.065	3.416	36.185	0.772	1.164	0.07	1.165	0.872
VIII	37.112	4.177	37.293	0.779	1.197	0.065	1.198	0.888
IX	37.158	3.712	37.299	0.779	**1.157**	0.067	**1.158**	0.877
X	38.45	4.645	38.658	**0.786**	1.176	0.065	1.178	**0.894**

Eighty-five percent of incidence ASRs (22/26) and 94% (34/36) of observed mortality ASRs were contained in the 95% posterior predictive intervals.

4.4.2 Summarizing the Posterior Distribution

Figure 4.6 shows the mean estimated age effects for the incidence, mortality, and the MI ratio. Corresponding mean observations for incidence and mortality are shown in blue and illustrate strong concordance between our estimates and the observed data. In the top left, the average estimated incidence rate per 100k population, derived from the mortality and MI ratio models, is shown for each age group. The estimates, having averaged across all 39 locations and 18 years of estimates are plotted in red with the corresponding 95% posterior credible intervals on the incidence rate scale. Incidence rates below the age of 25 are extremely low, and climb steadily until around age 60 at which point they flatten out. In the top right, we show a similar plot for the average estimated mortality rate per 100k population for each age group, again having averaged over the estimates from all modeled locations and years. Mortality rates stay low for longer, starting to notably rise around age

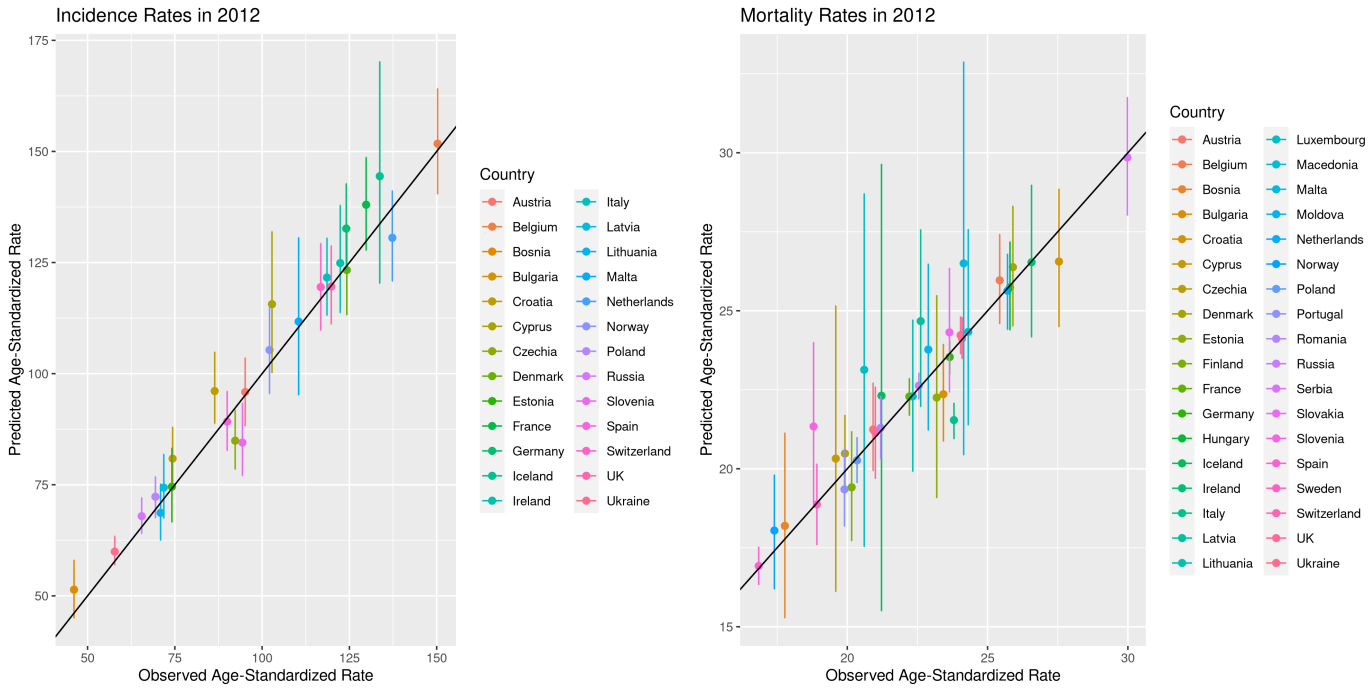


Figure 4.5: Posterior predicted estimates and 95% credible intervals the incidence and mortality ASRs in 2012 from model IX fit to incidence data from 2000-2011 and mortality data from 2000-2017, compared against observed values. All data in the incidence plot was withheld from the model.

30, before steadily increasing. The mean estimated MI ratio is seen to gradually increase across ages, with larger increases at older ages. We note that the credible intervals for mortality, and to a lesser extent the MI ratio, appear quite tight in these plots. This is due to the large population totals behind the mortality observations and the relative consistency in age-specific observed unconditional mortality rates across space and time. Despite this, the posterior predictive credible intervals for our mortality estimates appear to be well calibrated as shown in Table 4.3 and as shown in the country-specific age-time plots of Figure 4.11.

Figure 4.7 shows the average trends across time. We see the mean estimated incidence rate increasing, while the MI ratio drops fast enough to drive unconditional mortality rates down. The 95% posterior credible intervals in Figure 4.7 are much wider than those of

Figure 4.6, indicating greater variability across our age-country estimates compared to the country-time variability. This corroborates our initial exploratory findings which indicated that the age-country interactions were most important for capturing observed variability.

Figure 4.8 plots the average incidence, mortality, and MI ratio across age groups, stratified by year. Over time, we see that the increase in incidence appears across most ages, but is predominantly driven by increased incidence reporting in females over the age 60. The MI ratio drops most rapidly among the middle age groups, with the smallest gains in the youngest and oldest age groups. In combination, these have slightly driven down the unconditional mortality rate, most notably among the middle-ages of 40-65.

Figure 4.9 maps the posterior medians of the structured (top) and total (bottom) spatial random effects for mortality, $\boldsymbol{b}^M$ (left), and the MI ratio, $\boldsymbol{b}^{MI}$ (right). The structured effects, shown in the linear predictor scale, show that, all else being equal, the smooth spatial field indicates higher unconditional mortality rates in Western and Southern Europe and higher MI ratios in Eastern Europe and in the UK and Ireland. The total spatial effects, the weighted sum of the scaled structured and unstructured spatial components, show similar patterns for the mortality rate and MI ratio, though inclusion of the unstructured component allows larger deviations from neighboring countries, as particularly demonstrated by the total unconditional mortality rate in Ukraine.

4.4.3 Estimates and Projections of European Breast Cancer

In Figure 4.7 we show the mean incidence ASR across all countries from 2013 to 2017, backcasting the recent years with missing incidence data. We see the width of the credible intervals for the incidence ASRs and MI ratio widen after 2012, as we would expect at the end of the observed time-series as the model transitions from estimation to imputation.

Figure 4.10 maps the ASRs for incidence, mortality and the MI ratio in 2000, averaged across all years, and in 2017. The largest differences from Figure 4.9 are due to variations between country-specific populations and the standard population used to calculate the ASRs. In both plots, we see higher MI ratios in Eastern Europe. In contrast, Figure 4.10

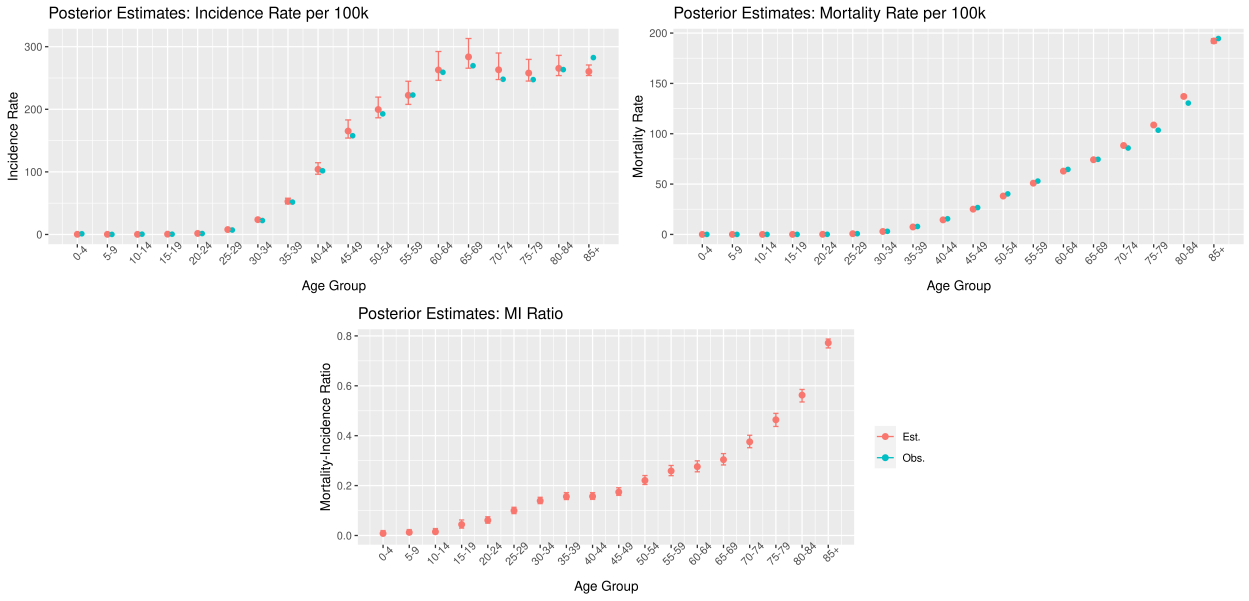


Figure 4.6: Estimated age-specific incidence rates per 100k, mortality rates per 100k, MI ratio. Estimates are calculated by averaging across all country and time predictions and plotted along with 95% credible intervals and the corresponding mean data observations for incidence and mortality.

shows higher incidence rates in Western Europe. The large differences in Eastern and Western breast cancer incidence and MI probabilities 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 incidence and mortality rates of breast cancer.

Between 2000 and 2009, approximately the midpoint of the time series, and again between 2009 and 2017 we see decreases in the MI ratio in all countries. Likewise, we estimate every country to have increased incidence ASRS between 2000 and 2017, though Denmark was estimated to have a decrease between 2009 and 2017. Unconditional mortality rates have a more varied narrative. Most countries are seen to improve, though the improvements are not as uniformly distributed throughout the continent, and some countries with increasing incidence rates and slowly decreasing MI ratios appear to have increasing unconditional

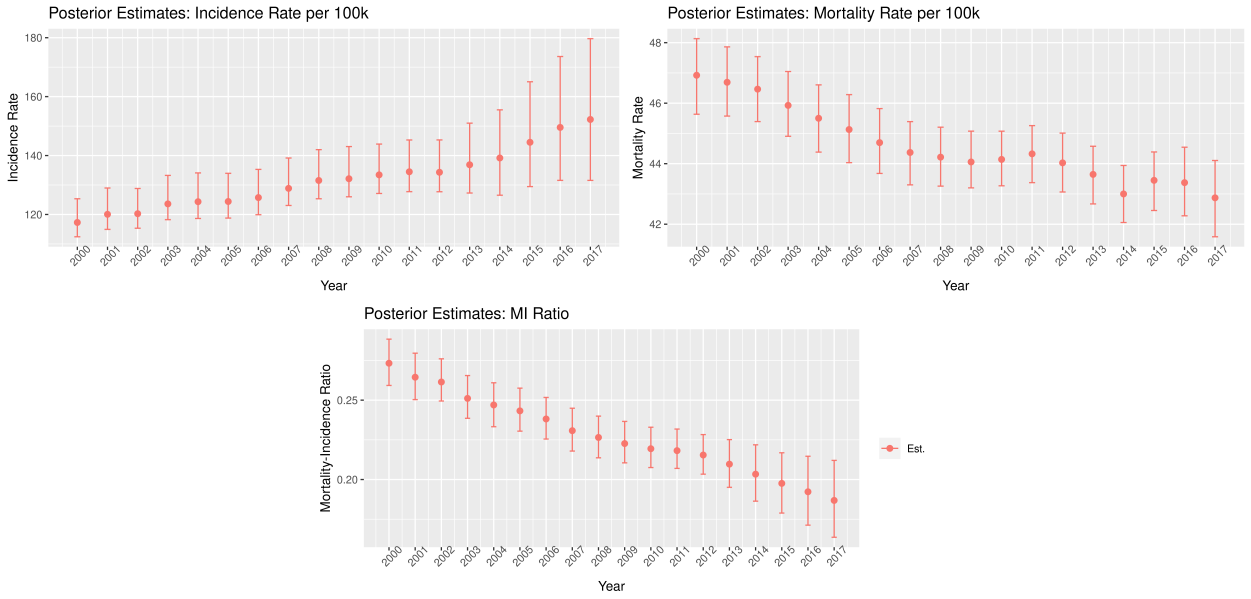


Figure 4.7: Estimated incidence ASRs, mortality ASRs, and MI ratio over time. Estimates are calculated by averaging across all countries and plotted along with 95% credible intervals.

mortality rates. Between 2000 and 2009, Denmark, Moldova, Romania, Russia, Serbia, Slovakia are estimated to have increased mortality ASRs, though by 2017 Denmark, Russia and Slovakia have recovered with their rates have dropping below their 2000 levels. During the second half of the time series, Moldova's mortality rates drop but still end above 2000 levels, whereas Romania and Serbia mortality ASRs continue to rise higher again. In addition, Bosnia, Bulgaria, Latvia, and Poland are estimated to have increases between 2009 and 2017.

Figure 4.11 displays the fitted age-specific incidence and mortality rates per 100k for the Netherlands, Germany, Hungary, and Ireland, starting from the top-left and moving clockwise. They were selected, respectively to serve as examples of a type I, type II, type III, type IV country, though Ireland is only type IV for the first half of the time-series. The posterior median and 95% posterior predictive credible intervals are shown in addition to the mean national data observation for each country-age-time. We see, as confirmed in

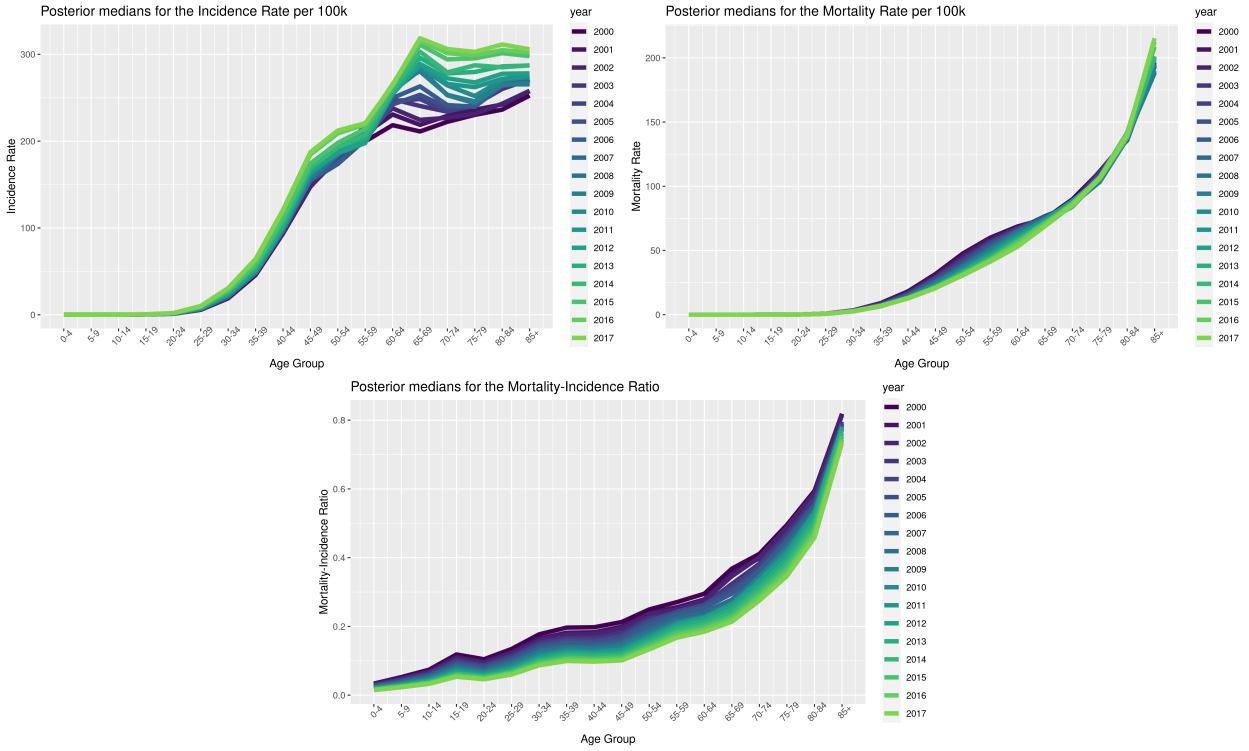


Figure 4.8: Estimated incidence ASRs, mortality ASRs, and MI ratio across age groups and time. Estimates are calculated by averaging across.

Table 4.3, that the posterior predictive credible intervals appropriately capture the variation in the incidence and mortality observations and that they respect the available information: widening when incidence or mortality are unavailable or when the time series is noisier.

Appendix B.4 details IARC and GBD methods and contrasts them against the methods used in this project. Figures B.5.1–B.5.3 plot our 2017 estimates against IARC 2018 estimates and 2017 estimates from the GBD 2019 study. We use the same data used by IARC and we see very similar incidence estimates even though the methods are quite different. GBD uses very different methods and data and we see similar mortality estimates. The GBD incidence estimates are consistently lower than our estimates and their uncertainty intervals are unexpectedly narrow compared to the mortality estimates from which they are

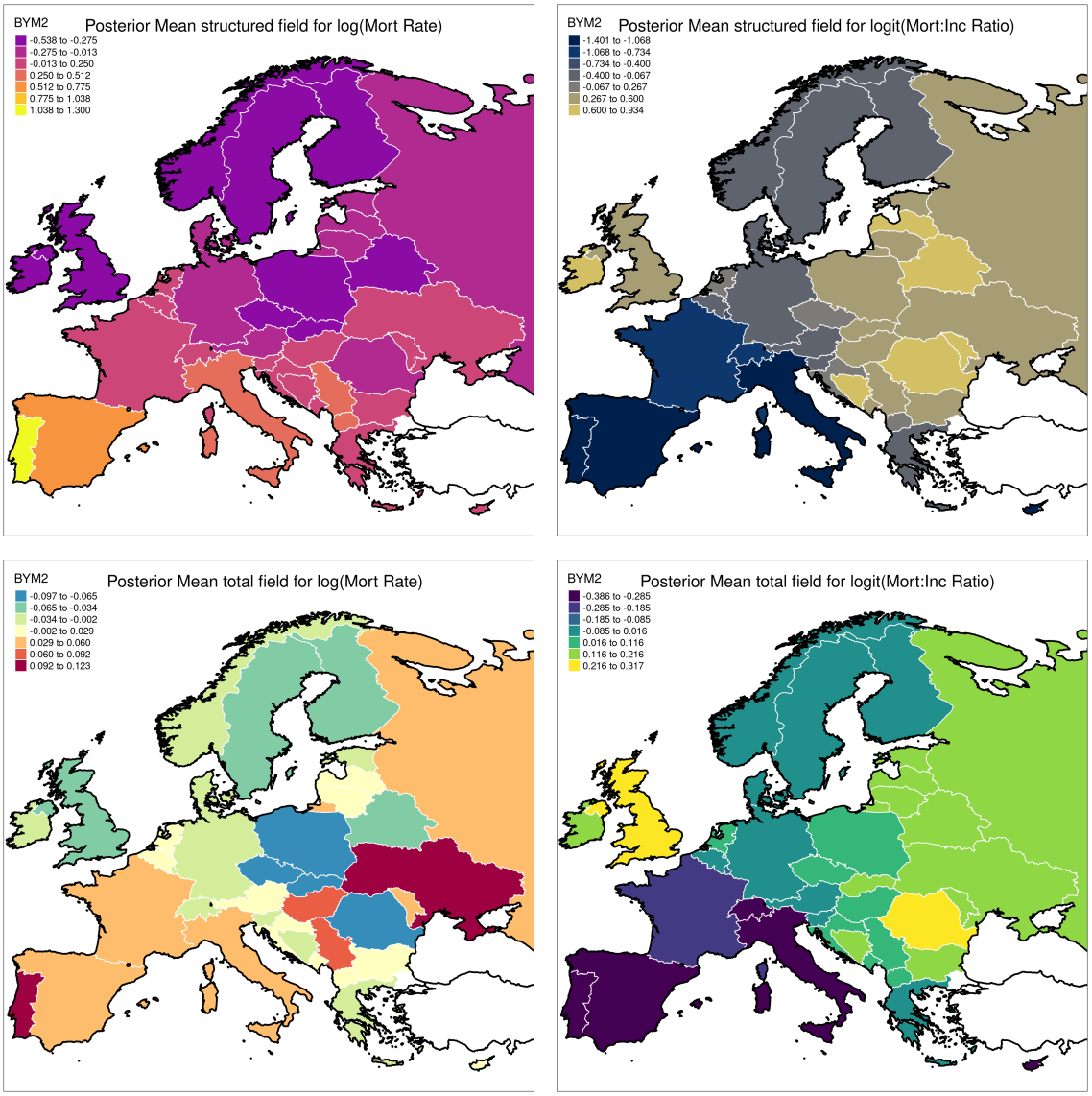


Figure 4.9: Posterior medians of the scaled structured component of the main spatial random effects, $\tilde{U}_c^{\bullet}$ (top), and the main total spatial random effects, $\frac{1}{\sqrt{\tau^{\bullet}}}\left(\sqrt{1-\phi_c}V_c^{\bullet}+\sqrt{\phi_c}\tilde{U}_c^{\bullet}\right)$ (bottom), for the Mortality model (left) and MI ratio model (right).

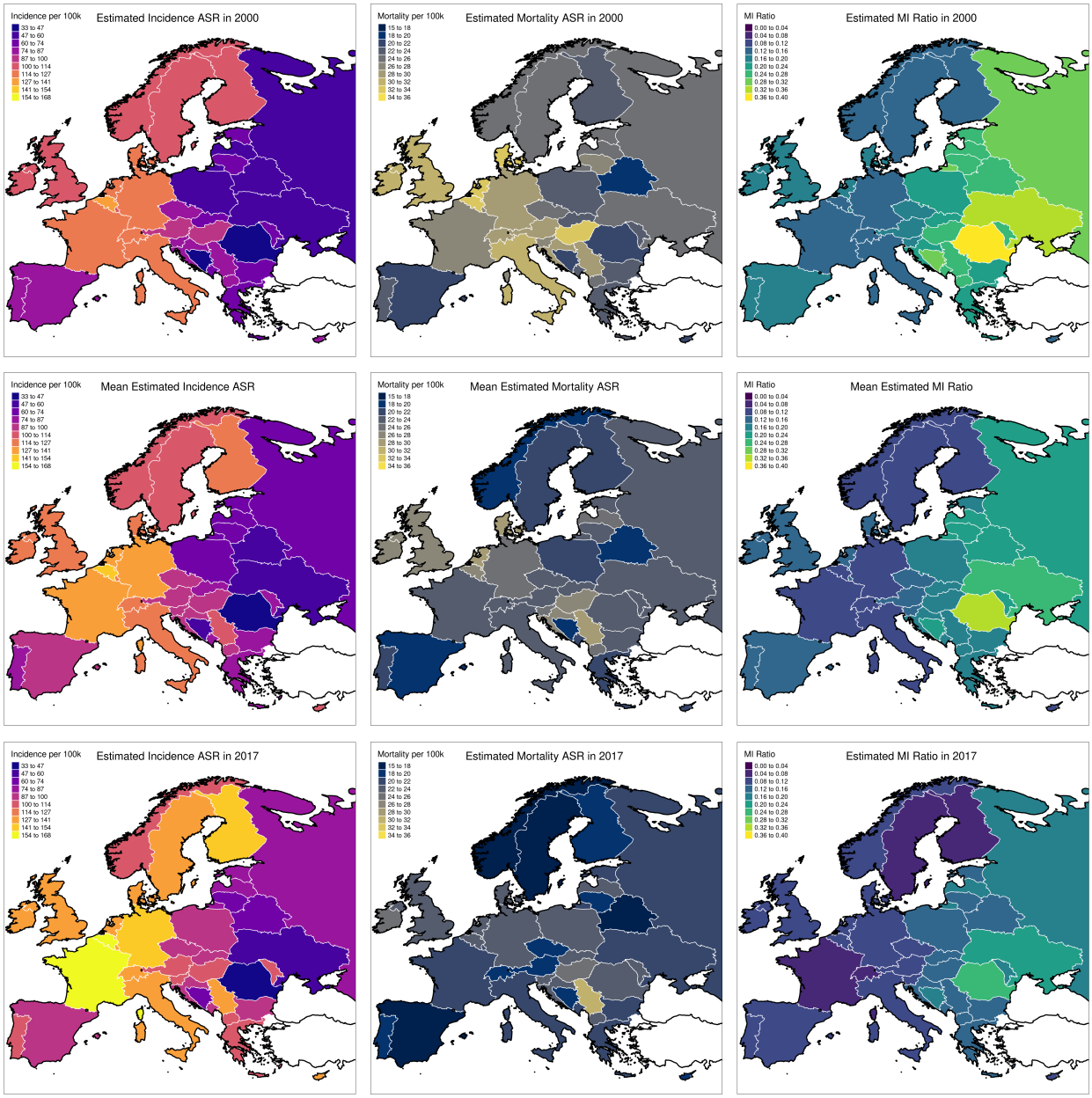


Figure 4.10: Posterior medians of each countries estimated incidence ASRs (left), MI Ratio (center), and mortality ASRs (right), in 2000 (top), averaged across all time (middle), and in 2017 (bottom).

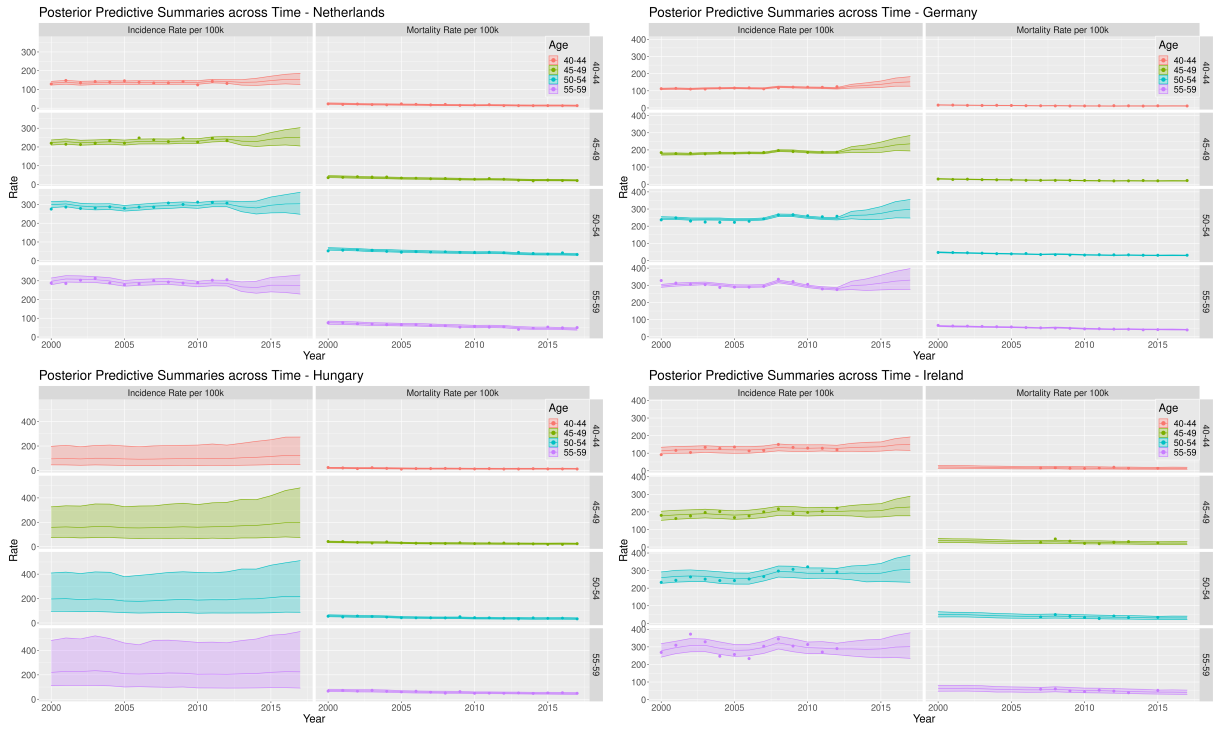


Figure 4.11: Posterior predictive medians and 95% credible intervals for age-specific breast cancer rates in the Netherlands, Germany, Hungary, and Ireland, moving clockwise from the top-left. Average national observations are shown as points.

derived.

4.5 Discussion

We proposed, implemented, and vetted a joint model for breast cancer incidence and mortality in 39 European countries which allows strength to be borrowed across ages, times, locations, and between mortality and incidence. Comparison of our point estimates for breast cancer rates align closely with those previously published IARC, who used the same data but different methods. Recent methods improvements at IARC have resulted in the ability to produce uncertainty intervals for their estimate, but they aim to use country-specific models and lose predictive power by disregarding correlation structures in the data. The current

IARC approach requires cancer, gender, and country specific models. We have proposed a unified model that can be fit jointly across all locations, decreasing the modeling burden by a factor of 40, and our Bayesian joint model automatically generates uncertainty intervals around our estimates. Our methodology is straightforward, reproducible, and demonstrates adequate uncertainty interval properties - none of which can be said for the GBD estimates.

The work presented in this chapter has a number of limitations which would benefit from additional investigation. First, we acknowledge that the present model is only capable of predicting incidence levels as would be reported from cancer registries and cancer surveillance within each country. It is not currently capable of accounting for constant or varying levels of underreporting of incidence or mortality. On the other hand, it produces estimates of incidence rates that are akin to those that planners working with registry data frequently use. To help account for this and to improve predictions in data sparse countries and times, a natural extension to this work would incorporate covariates. Second, we would like to extend the model to account for subnational heterogeneity of rates within countries. There are a number of ways that this could be done, but, as with modeling the national rates, the primary difficulty relates to finding a parsimonious model formulation that appropriately models this variation across age, location, and time. Lastly, we would like to apply and validate our methods to generate estimates for other cancers in Europe, and extend this work to be used in the lower and middle income countries where incidence data is often more available than cause-specific mortality data (Ferlay *et al.*, 2010b; Forouzanfar *et al.*, 2011).

In summary we developed a unified and principled approach for generating national estimates, including backcasting missing years of incidence, and uncertainty intervals for breast cancer incidence and mortality in Europe. Our approach is based on jointly modeling incidence, mortality conditional on incidence (through the MI ratio), and mortality. This approach is generalizable, able to simultaneously handle countries with national incidence data, registry incidence data, or no incidence data. The flexible random effects specification across age, country, is able to account for varying trends across different populations while simultaneously pooling information to impute where data is missing.

Chapter 5

FREQUENTIST COVERAGE OF BAYESIAN INTERVALS

5.1 Introduction

One of the uses of mixed effects models (MEMs) is to allow for pooling of data from heterogeneous groups to borrow information from across the groups to better estimate each of them. Performing inference using these methods provides a number of advantages over using direct group-specific estimators which only use data from one group at a time. Notably, by pooling information from across the entire dataset, intervals from MEMs are often (much) narrower than intervals constructed for the direct estimates. A set of $1-\alpha$ confidence intervals from a MEM will, averaging across the groups (marginally), achieve the nominal confidence level, but the coverage rate for any particular group (conditionally) depends on the magnitude of the group random effect and it may be far from the nominal $1-\alpha$ level (Yu and Hoff, 2018). Frequentist coverage probabilities are ubiquitous and conditional mixed effects intervals and credible intervals with unknown coverage probabilities can be misleading and difficult to interpret by those expecting or assuming that the credible intervals will have uniform frequentist properties. There is a growing body of work that provides alternative inferential approaches to construct intervals with the target coverage in some MEM settings. None of these methods offer universal fixes. As an alternative, we propose that the frequentist coverage probability of conditional MEM confidence intervals and Bayesian credible intervals be estimated to allow additional interpretation of the modeled uncertainty. We develop an unbiased estimator for the coverage probability of MEM intervals in Gaussian likelihood settings to allow valid frequentist interpretation of intervals for conditional estimators and suggest that the estimated coverage probability be reported alongside of conditional interval results known to have non-uniform coverage probabilities. The suggested procedure is

examined through simulations and on a real-world application.

Mixed effects models (MEMs), including many Bayesian Hierarchical Models (BHMs), provide a model-based approach to pooling information across heterogeneous groups, locations, times, or strata of observations. By allowing information to be shared across these groups, the uncertainty of estimators for each group can be smaller than direct group estimates which are estimated independently, ignoring the other observations. One of the costs associated with these methods and the increased precision of the estimators is that they introduce bias to the estimators which pull or shrink them towards the mean of all the observations. Despite the introduction of this bias, the precision is often so improved as to make these "shrinkage estimators" more appealing (in terms of mean-square error) than using independent unbiased estimators. In the simplest iid MEMs, the amount of shrinkage, and thus the amount of introduced bias, is a function of the distance between the mean of the observation of the mean across all observations. As a result and as we will see, the coverage properties of shrinkage estimator intervals depend on this distance as well. This chapter aims to improve our understanding of the coverage properties of these sorts of estimators and intervals.

As in Casella and Hwang (2012), we define a confidence procedure $C(\theta, y)$ to be a set in the product space of $\Theta \times \mathcal{Y}$ and define the y-section and θ-section, respectively, to be:

- $C(y) \subseteq \Theta$, the confidence set for the true underlying parameter θ^0 given observed data y, which is defined as follows:

$$C(y) = \{\theta : y \in C(\theta)\}, \text{ and}$$

- $C(\theta) \subseteq \mathcal{Y}$, the acceptance region of the test $H_0 : \theta^0 = \theta$ against the complementary alternative, which is defined as follows:

$$C(\theta) = \{y : \theta \in C(y)\},$$

where the superscript in θ^0 denotes the true value of that parameter as opposed to possible values it might take, θ. These two sections are intricately related with $\theta^0 \in C(y)$ if and only if $y \in C(\theta^0)$. This allows us to simplify calculations such as $P_{Y|\theta^0}(\theta^0 \in C(y))$ with the often simpler calculation of $P_{Y|\theta^0}(y \in C(\theta^0))$. A familiar confidence set example involves the problem of estimating the mean of a multivariate normal with known variance σ^2. In this scenario the p-dimensional $1-\alpha$ confidence set is usually taken to be:

$$C(\theta, y) = \{\theta : |\theta - y| \leq c\sigma\}, \tag{5.1}$$

given observation $Y = y$ from a p-dimensional normal distribution $Y \sim N(\theta, \sigma^2 I)$ with σ known and I the $p \times p$ identity matrix.

Many classic confidence procedures and corresponding confidence sets are designed to ensure that θ^0 will fall in $C(y)$ with uniform probability, with respect to Y, for all possible θ^0 values. Using the example in (5.1), if c is selected such that c^2 is the $1-\alpha$ quantile of a chi-squared distribution with p degrees of freedom, then the probability that sets generated in this fashion will contain the true mean vector will be $1-\alpha$, regardless of the value of the mean vector. More generally, for a confidence set $C(y)$, the *coverage probability* (CP) is calculated as:

$$CP = \mathrm{P}_{Y|\theta^0}\left(\theta^0 \in C(y)\right). \tag{5.2}$$

Note that the coverage probability definition conditions on the unknown parameter of interest, $\boldsymbol{\theta}^0$. In other applications where confidence sets are not guaranteed to have uniform coverage probability, such as confidence intervals from MEMs or credible intervals from Bayesian inferences, the coverage probability is typically a function of the unknown parameter. As such, the coverage probability will necessarily need to be estimated. To the best of our knowledge, estimators for coverage probabilities in MEMs have not been studied before.

As an intuitive example, consider a simple model with an independent and identically

distributed random effect for the group means:

$$Y_j|\theta_j, \sigma_j \sim N(\theta_j, \sigma_j^2), \qquad j \in \{1, \ldots, m\} \tag{5.3}$$

$$\theta_j|\boldsymbol{\beta}, \sigma_\theta \sim N(\mu_{\theta_j} = \boldsymbol{x}_j^T\boldsymbol{\beta}, \sigma_\theta^2), \tag{5.4}$$

where (5.3) describes the sampling model of the observations and (5.4) defines the latent model for the unobservable parameters which define across-group heterogeneity. In this sampling model there are conditionally independent observations Y_j, from m groups, with group means θ_j, and within-group variances σ_j^2. The common one-way experimental design follows this model if we view each Y_j observation as the mean of n_j $Y_{1,j}, \ldots, Y_{n_j,j}$ iid Gaussian observations, with $\sigma_j^2 = \sigma^2/n_j$.

The latent process layer allows model-based estimators to borrow information from other groups and in this formulation allows it to take advantage of available group-level covariates. More generally, the mean of the latent process layer need not take the form of a linear model. The latent process layer of (5.4) is sometimes referred to as a "linking model" in the small area estimation literature. The particular formulation of a sampling layer combined with a linking model shorn in (5.3) and (5.4) is an example of the Fay-Herriot model which assumes that the group (or area) means are distributed independently but with common parameters $\boldsymbol{\beta}$ and σ_θ (Fay and Herriot, 1979). A Bayesian modeler would complete the Bayesian hierarchical model (BHM) specification of this model by assigning priors to the remaining parameters: $\boldsymbol{\beta}, \sigma_j$, and σ_θ. To simplify computation, a conjugate family of distributions, such as inverse-gamma distributions for the variance parameters, could be selected.

Burris and Hoff (2020) describe three inferential techniques that produce three different intervals that could be constructed for θ_j^0. First, relying only on the sampling layer (5.3) and ignoring information from other groups, the *direct method* would use the unbiased observation (or sample mean) y_j as an estimator for θ_j. Confidence interval construction would proceed by inversion of the uniformly most powerful unbiased test to generate either the standard z-interval (if σ_j is known) or t-interval (if σ_j is unknown). That is, assuming σ_j is known,

one could construct the direct $1-\alpha$ confidence interval for θ_j^0:

$$C_D^j(\boldsymbol{y}) = \{\theta_j : y_j + \sigma_j z_{\alpha/2} < \theta_j < y_j + \sigma_j z_{1-\alpha/2}\}, \tag{5.5}$$

where z_p is the p^{th} quantile of the standard normal distribution. The direct method leads to an unbiased estimator for θ_j^0 and by construction it has uniform confidence level and coverage probability, regardless of the values of θ_j^0 and σ_j, across all groups:

$$1-\alpha = P(\theta_j \in C_D^j(\boldsymbol{y})|\boldsymbol{\theta}). \tag{5.6}$$

Burris and Hoff (2020) call this property *area-specific coverage*. While area-specific coverage is an intuitively desirable property, there are many real-world situations, for example, when working with survey designs which have small sample sizes, when the direct confidence intervals may be too wide to be practically useful. To increase the precision of the estimator and decrease the width of the interval, the second layer of the model is introduced to allow information to be shared across groups.

The MEM inferential approach leverages data from all the groups by adding the latent process layer of (5.4) to make inferences about each θ_j^0. If $\boldsymbol{\beta}$ and σ_θ are known, the latent layer (5.4) reduces to a conjugate $N(\boldsymbol{x}_j^T\boldsymbol{\beta}, \sigma_\theta^2)$ prior for θ_j. This prior, combined with the sampling model of (5.3) and assuming known variance, σ_j, yields the following posterior distribution for θ_j:

$$\theta_j|y_j \sim N\left(\frac{\sigma_\theta^2 y_j + \sigma_j^2 \boldsymbol{x}_j^T\boldsymbol{\beta}}{\sigma_j^2+\sigma_\theta^2}, \frac{\sigma_j^2\sigma_\theta^2}{\sigma_j^2+\sigma_\theta^2}\right). \tag{5.7}$$

The mean of this posterior is the classic shrinkage estimator for θ_j^0, which can be seen as a weighted combination of the observation y_j and the mean of the random effect, $\boldsymbol{x}_j^T\boldsymbol{\beta}$. This form of the shrinkage estimator highlights the bias that the MEM introduces, but it also shows why it is so often used. Letting $\tilde{\theta}_j$ denote the mean of (5.7), we see that the mean-

square error (MSE) of this estimator is $\mathrm{E}[(\tilde{\theta}_j - \theta_j)^2] = \frac{\sigma_j^2 \sigma_\theta^2}{\sigma_j^2 + \sigma_\theta^2}$, whereas the MSE of the direct estimator is $\mathrm{E}[(y_j - \theta_j)^2] = \sigma_j^2$. The shrinkage estimator is thus guaranteed to have smaller MSE (unless $\sigma_j = 0$), by a factor of $\frac{\sigma_\theta^2}{\sigma_j^2 + \sigma_\theta^2} < 1$. Recalling that MSE is the sum of bias and variance of an estimator, we see that the increase in precision gained from using information across all the groups is worthwhile, under MSE loss, compared to the magnitude of the bias that this procedure introduces. The other, less obvious, cost associated with using the MEM shrinkage framework is that, for a nominal confidence level $1 - \alpha$, the coverage rate is not constant across all m groups.

Following the same logic that led to the direct confidence interval, the posterior distribution of (5.7) allows for convenient calculation of the equal-tail Bayesian credible interval:

$$C_B^j(\boldsymbol{y}) = \{\theta : \tilde{\mu}_j + \tilde{\sigma}_\theta z_{\alpha/2} < \theta < \tilde{\mu}_j + \tilde{\sigma}_\theta z_{1-\alpha/2}\}, \tag{5.8}$$

where $\tilde{\mu}_j$ and $\tilde{\sigma}_\theta^2$ denote the conditional mean and variance of $\theta_j | y_j$ shown in (5.7). Unfortunately, it is unusual for $\boldsymbol{\beta}$ and σ_θ to be known a priori. In practice, priors will be placed on all parameters that appear in the model and full Bayesian inference can be performed, typically requiring integration over the posterior distribution of these parameters. If this is too computationally expensive or the modeler is philosophically opposed to including priors, an empirical Bayes (EB) approach (Morris, 1983) may be used where estimates for the parameters, $\hat{\boldsymbol{\beta}}$, and $\hat{\sigma}_\theta$, are calculated and substituted into the direct interval of (5.5) to construct an empirical Bayes confidence interval:

$$C_{EB}^j(\boldsymbol{y}) = \{\theta : \hat{\mu}_j + \hat{\sigma}_\theta z_{\alpha/2} < \theta < \hat{\mu}_j + \hat{\sigma}_\theta z_{1-\alpha/2}\}, \tag{5.9}$$

where $\hat{\mu}_j$ represents the conditional mean of θ_j given y_j using the "plug-in" estimates $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_\theta$, and $\hat{\sigma}_j$ estimated using the entire set of observations. In contrast to the full Bayes construction, the EB method ignores the uncertainty of the estimated plug-in quantities.

Unlike the area-specific coverage property of the direct confidence intervals, Burris and

Hoff (2020) note that the Bayesian credible intervals only have *population-level coverage*:

$$1-\alpha = \int \mathrm{P}(\theta_j \in C_B^j(\boldsymbol{y})|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\beta},\sigma_\theta)\,\mathrm{d}\theta_j, \tag{5.10}$$

where π is the prior for $\boldsymbol{\theta}$. They also note that the EB intervals exhibit population-level coverage asymptotically in the number of groups when using consistent plug-in estimators. That is, the standard intervals constructed when using these methods only achieve nominal coverage marginally, integrating over the prior for θ_j. For any specific group j, both C_{EB}^j and C_B^j lack the desired $1-\alpha$ coverage since the intervals are centered around biased shrinkage estimators of θ_j^0. To demonstrate this, again consider the Fay-Herriot example with fixed $\boldsymbol{\beta}$, σ_θ, and σ_j. Since the conjugate nature of this model formulation has a simple posterior, the coverage probability can be found in a fashion very similar to the classic two-step construction of confidence interval: first writing down the probability that θ_j^0 falls between two quantiles, and then inverting the inequality to yield a probability statement about values of y_j falling between some upper and lower bound.

Given the posterior distribution in (5.7) , the standard symmetric $1-\alpha$ credible interval can be constructed as in (5.8):

$$C_B^j(\boldsymbol{y}) = \left\{\theta : \frac{\sigma_\theta^2 y_j + \sigma_j^2 \boldsymbol{x}_j^T\boldsymbol{\beta}}{\sigma_j^2+\sigma_\theta^2} + \frac{\sigma_j\sigma_\theta}{\sqrt{\sigma_j^2+\sigma_\theta^2}} z_{\alpha/2} < \theta < \frac{\sigma_\theta^2 y_j + \sigma_j^2 \boldsymbol{x}_j^T\boldsymbol{\beta}}{\sigma_j^2+\sigma_\theta^2} + \frac{\sigma_j\sigma_\theta}{\sqrt{\sigma_j^2+\sigma_\theta^2}} z_{1-\alpha/2}\right\}. \tag{5.11}$$

From this, the coverage probability, that is the probability that after observing a particular y_j an interval constructed as defined by (5.11) contains the true value of θ_j, may be calculated:

$$\mathrm{P}_Y(\theta_j^0 \in C_B^j(\boldsymbol{y})|\theta_j^0) = \mathrm{P}\left(\frac{\sigma_\theta^2 y_j + \sigma_j^2 \boldsymbol{x}_j^T\boldsymbol{\beta}}{\sigma_j^2+\sigma_\theta^2} + \frac{\sigma_j\sigma_\theta}{\sqrt{\sigma_j^2+\sigma_\theta^2}} z_{\alpha/2} < \theta_j^0 < \frac{\sigma_\theta^2 y_j + \sigma_j^2 \boldsymbol{x}_j^T\boldsymbol{\beta}}{\sigma_j^2+\sigma_\theta^2} + \frac{\sigma_j\sigma_\theta}{\sqrt{\sigma_j^2+\sigma_\theta^2}} z_{1-\alpha/2}\right). \tag{5.12}$$

Denoting the prior mean $\mathrm{E}[\theta_j^0] = \boldsymbol{x}_j^T\boldsymbol{\beta}$ more generally by μ_{θ_j} and solving the inequality for the pivotal quantity $\frac{y_j-\theta_j^0}{\sigma_j} \sim N(0,1)$ yields the area-specific coverage probability:

$$
\begin{aligned}
&CP(\theta_j^0, \mu_{\theta_j}, \sigma_j, \sigma_\theta, \alpha) \\
&= \Phi\left(\frac{\theta_j^0 - (\mu_{\theta_j} - z_{1-\alpha/2}\sqrt{1+\sigma_j^2/\sigma_\theta^2}\frac{\sigma_\theta^2}{\sigma_j})}{\sqrt{\frac{\sigma_\theta^4}{\sigma_j^2}}}\right) - \Phi\left(\frac{\theta_j^0 - (\mu_{\theta_j} - z_{\alpha/2}\sqrt{1+\sigma_j^2/\sigma_\theta^2}\frac{\sigma_\theta^2}{\sigma_j})}{\sqrt{\frac{\sigma_\theta^4}{\sigma_j^2}}}\right) \qquad (5.13) \\
&= \Phi\left(\frac{\sigma_j}{\sigma_\theta^2}(\theta_j^0 - \mu_{\theta_j}) + z_{1-\alpha/2}\sqrt{1+\frac{\sigma_j^2}{\sigma_\theta^2}}\right) - \Phi\left(\frac{\sigma_j}{\sigma_\theta^2}(\theta_j^0 - \mu_{\theta_j}) + z_{\alpha/2}\sqrt{1+\frac{\sigma_j^2}{\sigma_\theta^2}}\right), \qquad (5.14)
\end{aligned}
$$

with Φ denoting the standard normal CDF. We note that this coverage probability is a function of the magnitude of the group random effect, $\theta_j^0 - \boldsymbol{x}_j^T\boldsymbol{\beta}$. Shrinkage estimators tend to shrink larger magnitude effects further, biasing those estimates more, and causing more severe undercoverage. As it turn out, groups with random effects close to zero display the opposite effect, exhibiting overcoverage. The theoretical curve drawn in Figure 5.1 plots the coverage probability shown in (5.14) for select values of the variance parameters. As the figure demonstrates, the over- or under-coverage of a nominal $1-\alpha$ credible interval may be quite severe. From this example it is clear that any decision process based on mixed effects intervals or credible intervals which assumes the intervals have a nominal frequentist coverage probability of $1-\alpha$ will, with probability 1, be misinterpreting the uncertainty intervals. The misinterpretation could easily lead to incorrect decisions compared to those made with full understanding of the frequentist properties of the intervals.

This example highlights a few salient features of the standard method to calculate coverage probabilities:

1. Coverage probabilities assess if a random interval, as function of data observations $\boldsymbol{y}$ and conditional on the true parameter θ_j^0, covers θ_j^0.

2. The construction of a confidence interval or calculation of a coverage probability typi-

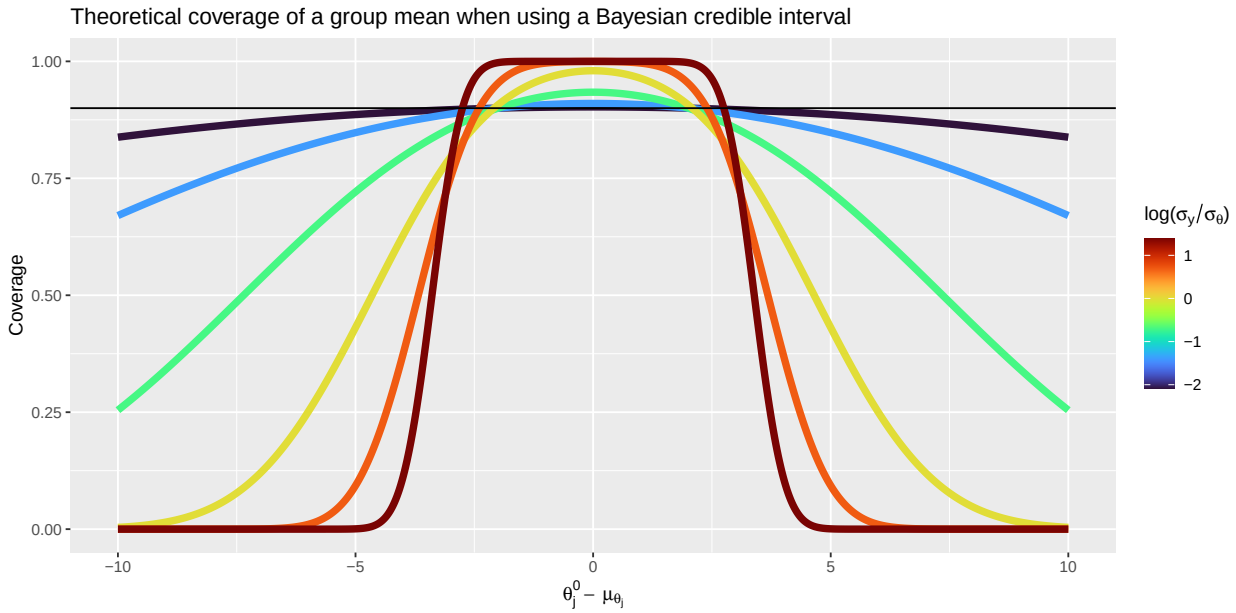


Figure 5.1: Theoretical coverage probability of θ_j^0 as shown in (5.14) plotted as a function of the distance between the mean of the sampling distribution and the mean of the latent process distribution which is equal to the magnitude of the random effect. The horizontal line at 0.90 denotes the nominal coverage probability and the various curves are shown for different values of the ratio σ_j/σ_θ given $\sigma_\theta = 2$.

cally relies on the functional form of the density of $\boldsymbol{y}$.

3. The coverage probability may be a function of the unknown parameter.

Acknowledging the first two observations and in spite of the third, for the Fay-Herriot setting we develop an unbiased estimator for the area-specific coverage and provide the variance of the proposed estimator. If needed, for example to construct appropriate confidence intervals for the coverage estimator that lie between 0 and 1, it is simple to sample from the distribution of the estimator.

The remainder of the chapter is organized as follows. In Section 5.2 we provide additional background information on alternate approaches to constructing frequentist intervals in a variety of settings and including a discussion on why a Bayesian should be interested in the

frequentist properties of their credible intervals. Section 5.3 develops the analytical form of unbiased estimator of coverage probabilities for conditional group intervals in the Fay-Herriot setting. Sections 5.4 and 5.5 provide simulation results and a real world example, and we conclude with a discussion in Section 5.6.

5.2 Background

The fact that shrinkage model intervals have non-uniform coverage probabilities is well-known (see, for example, Snijders and Bosker, 2011 and Yu and Hoff, 2018). A variety of EB posterior intervals calculations have been proposed which yield narrower intervals than the direct z- or t-interval, but still do not demonstrate area-specific coverage (Laird and Louis, 1987; Morris, 1983; He, 1992; Hwang *et al.*, 2009).

More recently, Yu and Hoff (2018), extended by Hoff and Yu (2019) and Burris and Hoff (2020), have built upon work by Pratt (1963) to develop Bayes-optimal confidence interval procedures. Intervals from these procedures, which they call *Frequentist assisted by Bayes* or *Frequentist and Bayes* (FAB) intervals, minimize the expected width of the intervals with respect to a selected prior among all confidence interval procedures which have $1 - \alpha$ freqentist coverage. In the multigroup setting, they select a prior based on data from all groups other than group j for which inference is being performed. Parameters for the model can still be estimated using data from all groups and their procedure produces uniform coverage confidence intervals for the group means which are typically narrower than the direct intervals (Yu and Hoff, 2018). These methods yield intervals with the correct coverage as long as the sampling distribution is correctly specified – even if the prior is misspecified. While we find this work exciting and inspirational, so far the results are limited to settings where analytic posteriors can be evaluated.

Conformal prediction is another class of techniques which have become increasingly popular in the machine learning and statistics community over the last two decades. Made popular by in the definitive book of Vovk *et al.* (2005), conformal prediction consists of a set of distribution-free algorithms which are capable of producing $1 - \alpha$ confidence intervals

around point predictions from a wide class of models. Conformal methods typically rely on the data being exchangeable, which is violated for many mixed effects models. There has been recent significant interest in extending conformal techniques beyond exchangeability. These extensions typically focus on specific model formulations and rely on assumptions related to local or approximate exchangeability (Tibshirani *et al.*, 2019, Xu and Xie, 2021, Dunn *et al.*, 2022). Ongoing work by Barber *et al.* (2022) aims to relax these assumptions even further, but the undercoverage gap (the distance between nominal coverage and actual coverage) drops as the data becomes less exchangeable. In our experience, for spatial models with even modest correlation, this forces the modeler to make a choice between very wide intervals or severe undercoverage. The very wide intervals are minimally informative, and intervals with non-negligible undercoverage leave us in the same predicament: having created intervals with unknown coverage properties.

Unlike the large variety of frequentist interval techniques, the choice of Bayesian interval procedure is very often taken to be the central $100 \times (1 - \alpha)$ credible interval. This can be computed as the $\alpha/2$ and $1 - \alpha/2$ quantiles from the posterior, and, in practice, these are frequently approximated using draws from the posterior generated with Monte Carlo Markov Chain sampling or samples from an approximate posterior. The posterior distributions condition on the observations and encode the subjective beliefs of the modeler coherently with information from observations. The credible interval, then, is a probability statement that encodes the modelers belief, augmented by information present in the observed data, on where the parameter of interest may lie. While the pure Bayesian will believe the probability statement of the interval and be content to interpret it, there are many reasons that a pragmatic Bayesian may also wish to study the frequentist properties of their credible interval. As Rubin (1984) notes, "frequency calculations can be Bayesianly justifiable if conceptualized properly. The only requirement is that we condition on observed values and calculate the distribution of unobserved quantities." More specifically, he writes that:

> [F]requency calculations are useful for making Bayesian statements scientific, scientific in the sense of capable of being shown wrong by empirical test; here the technique is

> calibration of Bayesian probabilities to the frequencies of actual events. ... Consider the following scheme. Given observed data X_{obs}, what would we expect to see in hypothetical replications of the study that generated X_{obs}? Intuitively, if the model specifications are appropriate, we would expect to see something similar to what we saw this time, at least similar in "relevant ways". This statement, which is essentially a fundamental premise of frequency inference, seems to me so basic that it needs no defense.

Repeatability, which is inherently an exercise in frequentist-style repetition, is of primary importance to the scientific method, and as such will be of interest to many researchers using Bayesian models and methods.

In summary, the existing methods for MEM confidence intervals typically produce models without area-specific coverage. Existing methods that guarantee area-specific coverage, including FAB intervals and conformal prediction algorithms, can only be applied in specific settings and may come at the cost of increased interval sizes when compared to credible interval procedures. In addition, a Bayesian who believes in the probability statements of their credible intervals may not be interested in using a different inferential procedure to produce area-specific intervals. With all this in mind, we propose that a useful and generally applicable alternative (or addition) to these specific methods is to provide the frequentist coverage probability of intervals alongside the uncertainty intervals. As we will see, there are situations where theoretical coverage probability may be analytically derived but finding these curves does not necessarily translate into obtaining good estimates of the coverage probability in data applications. To the best of our knowledge, developing unbiased estimators for coverage probability intervals has not been proposed before, nor have the estimators been developed. In the following sections we demonstrate this process for the Fay-Herriot model.

5.3 Methods

Our goal is to derive an unbiased estimator for coverage in settings where the coverage probability of a confidence-type procedure is unknown. Examples of such settings include determining the coverage probability of confidence intervals for conditional linear predictor

shrinkage estimators (that is, group-level linear predictors as opposed to the marginal overage probability averaging across all groups) and the coverage probability of Bayesian confidence intervals. In both cases, unlike traditional coverage probability calculations, we can no longer analytically work out pivotal quantities for the coverage calculation which means that the coverage probability calculations condition on the point distribution of the truth. The true parameter for which we are constructing an interval is unknown in real data applications, necessitating the development of a coverage probability estimator if we wish to understand the frequentist properties of the (conditional) random effect intervals. As a first foray into these calculations, we develop an unbiased estimator of the area-specific coverage probability shown in (5.14) for the Fay-Herriot setting.

5.3.1 Fay-Herriot Coverage with unknown truths

Consider the Fay-Herriot model defined by (5.3) and (5.4). Given data observations, $\boldsymbol{Y}$, one might consider replacing the unknown value of θ_j^0 in the coverage probability calculation of (5.14) with a consistent and unbiased estimator of the group mean, $Y_j = \frac{1}{n_j}\sum_i^{n_j} Y_{ij} \sim N(\mu_{\theta_j}, \sigma_j^2)$:

$$\widehat{CP}_N\left(Y_j, \mu_{\theta_j}, \sigma_j, \sigma_\theta, \alpha\right) =$$
$$\Phi\left(\frac{\sigma_j}{\sigma_\theta^2}(Y_j - \mu_{\theta_j}) - z_{1-\alpha/2}\sqrt{1+\sigma_j^2/\sigma_\theta^2}\right) - \Phi\left(\frac{\sigma_j}{\sigma_\theta^2}(Y_j - \mu_{\theta_j}) - z_{\alpha/2}\sqrt{1+\sigma_j^2/\sigma_\theta^2}\right). \quad (5.15)$$

Unfortunately, as we will prove, the naive estimator is biased. The following lemma simplifies the calculation of the expectation of the naive estimator.

Lemma 1. *Let $X \sim N(\mu, \sigma)$ with pdf $f_X(x)$, and let $\Phi(\cdot)$ represent the CDF of a standard normal. Then*

$$\int \Phi\left(\frac{x+a}{b}\right) f_X(x)\,\mathrm{d}x = \Phi\left(\frac{\mu+a}{\sqrt{b^2+\sigma^2}}\right). \quad (5.16)$$

Proof. Let Y be a standard normal random variable, independent of X, with $P(Y \leq y) =$

$\Phi(y)$. Then,

$$\begin{aligned}\int \Phi\left(\frac{x+a}{b}\right) f_X(x)\,\mathrm{d}x &= \int P\left(Y \leq \frac{x+a}{b}\right) f_X(x)\,\mathrm{d}x \\ &= \int P\left(Y \leq \frac{X+a}{b} | X = x\right) f_X(x)\,\mathrm{d}x \\ &= P\left(Y \leq \frac{X+a}{b}\right) = P\left(bY - X \leq a\right),\end{aligned}$$

where the last line follows from the law of total probability. Since Y and X are both Gaussian, their linear combination is also Gaussian and $W = bY - X \sim N(-\mu, b^2+\sigma^2)$. Standardizing W yields

$$P(W \leq a) = P\left(\frac{W+\mu}{\sqrt{b^2+\sigma^2}} \leq \frac{a+\mu}{\sqrt{b^2+\sigma^2}}\right) = \Phi\left(\frac{a+\mu}{\sqrt{b^2+\sigma^2}}\right).$$

□

Corollary 1.1. *The naive coverage probability estimator proposed in (5.15) is biased and*

$$\begin{aligned}&E\left[\widehat{CP}_N(Y_j, \mu_{\theta_j}, \sigma_j, \sigma_\theta, \alpha)\right] = \\ &\Phi\left(\frac{\theta_j^0 - (\mu_{\theta_j} - z_{1-\alpha/2}\sqrt{1+\sigma_j^2/\sigma_\theta^2}\frac{\sigma_\theta^2}{\sigma_j})}{\sqrt{\frac{\sigma_\theta^4}{\sigma_j^2}+\sigma_j^2}}\right) - \Phi\left(\frac{\theta_j^0 - (\mu_{\theta_j} - z_{\alpha/2}\sqrt{1+\sigma_j^2/\sigma_\theta^2}\frac{\sigma_\theta^2}{\sigma_j})}{\sqrt{\frac{\sigma_\theta^4}{\sigma_j^2}+\sigma_j^2}}\right). \quad (5.17)\end{aligned}$$

Proof. The expectation of the naive coverage probability estimator is

$$\begin{aligned}\mathrm{E}\left[\widehat{CP}_N(Y_j, \mu_{\theta_j}, \sigma_j, \sigma_\theta, \alpha)\right] = &\int \Phi\left(\frac{y - (\mu_{\theta_j} - z_{1-\alpha/2}\sqrt{1+\sigma_j^2/\sigma_\theta^2}\frac{\sigma_\theta^2}{\sigma_j})}{\sqrt{\frac{\sigma_\theta^4}{\sigma_j^2}}}\right) f_{Y_j}(y)\,\mathrm{d}y - \\ &\int \Phi\left(\frac{y - (\mu_{\theta_j} - z_{\alpha/2}\sqrt{1+\sigma_j^2/\sigma_\theta^2}\frac{\sigma_\theta^2}{\sigma_j})}{\sqrt{\frac{\sigma_\theta^4}{\sigma_j^2}}}\right) f_{Y_j}(y)\,\mathrm{d}y.\end{aligned}$$

For $Y_j \sim N(\theta_j^0, \sigma_j^2)$, applying Lemma 1 to each integral yields (5.17). By comparing (5.17) to the theoretical coverage shown in (5.13) we see that the expected value of the naive estimator differs from the true coverage only by the addition of an extra σ_j^2 in the radical of the denominator in both CDF arguments. As a result, the naive estimator will almost surely be biased unless no shrinkage occurs because either $\sigma_j^2 = 0$ or $\sigma_\theta^2 \to \infty$. □

Figure 5.2 plots the bias for select variance parameters. While the bias can be significant, the similarity of the form between (5.17) and (5.13) and the result from Lemma 1 suggest the possibility to bias-correct the naive estimator. The form of the bias also provides insight into the cause of the bias. Relative to the theoretical coverage probability calculation which can be seen as an expectation with respect to the point distribution of the truth, the expected value of the naive estimator spreads out the theoretical coverage curve by integrating with respect to the sampling distribution of Y_j. This fattens and lifts the tails of the coverage probability curve and flattens and squashes the center of the curve.

Fay-Herriot unbiased coverage estimator

For $Y_j \sim N(\theta_j^0, \sigma_j^2)$, we know from Lemma 1 that

$$\mathrm{E}\left[\Phi\left(\frac{\tilde{a}_j + Y_j}{\tilde{b}_j}\right)\right] = \Phi\left(\frac{\tilde{a}_j + \theta_j^0}{\sqrt{\tilde{b}_j^2 + \sigma_j^2}}\right).$$

Identifying $\tilde{a}_j$ and $\tilde{b}_j$ such that $\mathrm{E}\left[\Phi\left(\frac{\tilde{a}_j+Y_j}{\tilde{b}_j}\right)\right]$ equals either of the terms involved in the theoretical area-specific coverage shown in (5.14) will permit the construction of an unbiased estimator. That is, we need to determine $\tilde{a_j}$ and $\tilde{b_j}$ such that:

$$\mathrm{E}\left[\Phi\left(\frac{\tilde{a}_j + Y_j}{\tilde{b}_j}\right)\right] = \Phi\left(\frac{\tilde{a}_j + \theta_j^0}{\sqrt{\tilde{b}_j^2 + \sigma_j^2}}\right) = \Phi\left(\frac{\theta_j^0 - (\mu_{\theta_j} - z_p\sqrt{1 + \sigma_j^2/\sigma_\theta^2}\frac{\sigma_\theta^2}{\sigma_j})}{\frac{\sigma_\theta^2}{\sigma_j}}\right).$$

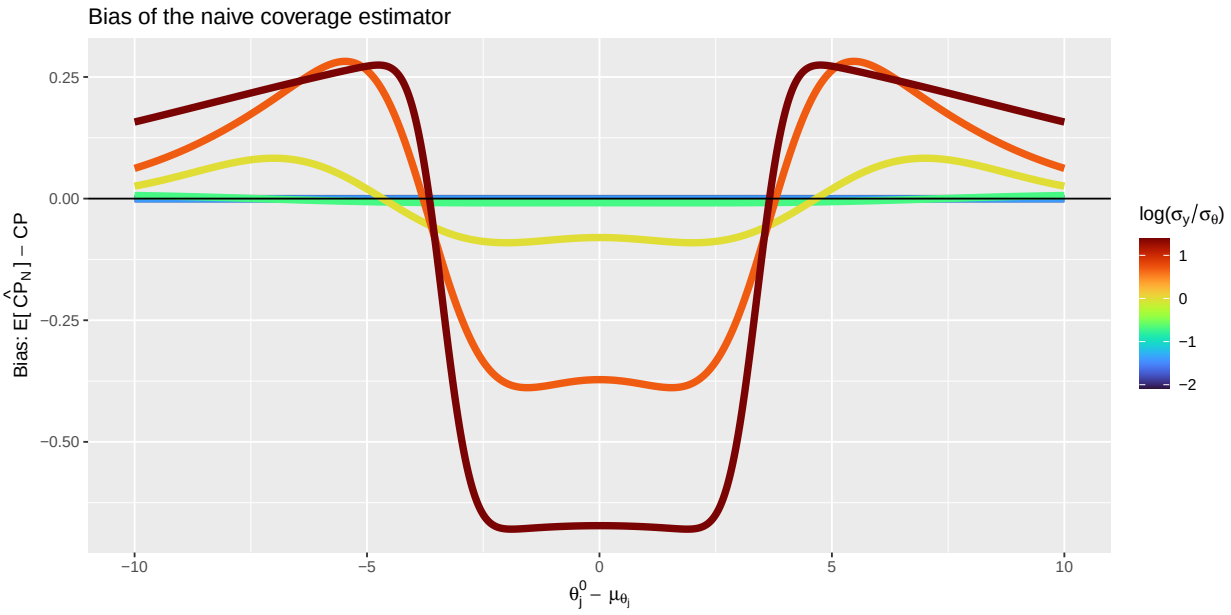


Figure 5.2: The bias of the naive coverage estimator of (5.15) as a function of the distance between the sampling mean θ_j^0 and the latent process mean μ_{θ_j}. The various bias curves are shown for different values of the ratio σ_j/σ_θ given $\sigma_\theta = 2$ and align with those shown in Figure 5.1.

Equating the arguments inside these two standard normal CDFs and solving for $\tilde{a}_j$ yields:

$$\tilde{a}_j = \theta_j^0 \times \left(\frac{\sigma_j \sqrt{\tilde{b}_j^2 + \sigma_{y_j}^2}}{\sigma_\theta^2} - 1 \right) + \left(z_p \sqrt{1 + \frac{\sigma_j^2}{\sigma_\theta^2}} - \frac{\sigma_j}{\sigma_\theta^2} \mu_{\theta_j} \right) \sqrt{\tilde{b}_j^2 + \sigma_j^2},$$

which, unfortunately, depends on θ_j^0 – and we require that our estimator and thus $\tilde{a}_j$ and $\tilde{b}_j$ do not rely on the unknown parameter. Fortunately, we are presented with one option for $\tilde{b}_j$ which leads to functions for both $\tilde{a}_j$ and $\tilde{b}_j$ which are θ_j^0-free. To do this, we set

$$\frac{\sigma_j \sqrt{\tilde{b}_j^2 + \sigma_{y_j}^2}}{\sigma_\theta^2} = 1,$$

which yields

$$\tilde{b}_j = \sqrt{\frac{\sigma_\theta^4 - \sigma_j^4}{\sigma_j^2}}, \qquad \text{and} \tag{5.18}$$

$$\tilde{a}_j = -\frac{\sigma_j}{\sigma_\theta^2}\mu_{\theta_j} + z_p\sqrt{\frac{\sigma_\theta^4}{\sigma_j^2} + \sigma_\theta^2}. \tag{5.19}$$

This representation of $\tilde{b}_j$ in (5.18) shows the main limitation of this estimator: it is only available if $\sigma_\theta^2 \geq \sigma_j^2$. That is, the within-group variance divided by the sample size of the group must be less than the across-group variance. Given at least modest sample sizes for the groups, this does appear to be an overly restrictive requirement. We also note that this derivation is easily extended to the non-iid prior setting since it only requires a conjugate Gaussian-Gaussian model. In Section 5.5 we demonstrate this by using a spatially correlated covariance for discrete areal units.

Theorem 2. *If* $Y_j \sim N(\theta_j^0, \sigma_j^2)$ *and* $\theta_j^0 \sim N(\mu_{\theta_j}, \sigma_\theta^2)$*, then*

$$\widehat{CP}_U(Y_j, \mu_{\theta_j}, \sigma_j, \sigma_\theta, \alpha) =$$

$$\Phi\left[\frac{Y_j - \frac{\sigma_j}{\sigma_\theta^2}\mu_{\theta_j} + z_{1-\alpha/2}\sqrt{\frac{\sigma_\theta^4}{\sigma_j^2} + \sigma_\theta^2}}{\sqrt{\frac{\sigma_\theta^4 - \sigma_j^4}{\sigma_j^4}}}\right] - \Phi\left[\frac{Y_j - \frac{\sigma_j}{\sigma_\theta^2}\mu_{\theta_j} + z_{\alpha/2}\sqrt{\frac{\sigma_\theta^4}{\sigma_j^2} + \sigma_\theta^2}}{\sqrt{\frac{\sigma_\theta^4 - \sigma_j^4}{\sigma_j^4}}}\right] \tag{5.20}$$

is an unbiased estimator for

$$CP(\theta_j^0, \mu_{\theta_j}, \sigma_j, \sigma_\theta, \alpha) =$$

$$\Phi\left((\theta_j^0 - \mu_{\theta_j})\frac{\sigma_j}{\sigma_\theta^2} + z_{1-\alpha/2}\sqrt{1 + \sigma_j^2/\sigma_\theta^2}\right) - \Phi\left((\theta_j^0 - \mu_{\theta_j})\frac{\sigma_j}{\sigma_\theta^2} + z_{\alpha/2}\sqrt{1 + \sigma_j^2/\sigma_\theta^2}\right).$$

Proof. Taking the expected value of (5.20) and applying Lemma 1 yields $CP(\theta_j^0)$ as shown in (5.13). □

We also develop the variance of this estimator. The following lemma simplifies the calculation.

Lemma 2. *Let $Y \sim N(\mu, \sigma^2)$ with pdf $f(y)$, and let $\Phi(\cdot)$ and $\phi(\cdot)$ represent the CDF and PDF of a standard normal. Then*

$$\int \Phi\left(\frac{y+a}{b}\right) \Phi\left(\frac{y+c}{d}\right) f(y)\,\mathrm{d}y = F_{MVN}\left(x = \begin{bmatrix} a/b \\ c/d \end{bmatrix}; \mu = \begin{bmatrix} -\mu/b \\ -\mu/d \end{bmatrix}, \Sigma = \begin{bmatrix} 1+b^{-2}\sigma^2 & (bd)^{-1}\sigma^2 \\ (bd)^{-1}\sigma^2 & 1+d^{-2}\sigma^2 \end{bmatrix}\right), \tag{5.21}$$

where F_{MVN} is the CDF of a multivariate normal distribution.

Proof. Let Y and Z be independent standard normal random variables, both independent of X, with $P(Y \leq y) = \Phi(y)$ and $P(Z \leq z) = \Phi(z)$. Then,

$$\begin{aligned}
\int \Phi\left(\frac{x+a}{b}\right) \Phi\left(\frac{x+c}{d}\right) f(x)\,\mathrm{d}x &= \int P\left(Y \leq \frac{x+a}{b}\right) P\left(Z \leq \frac{x+c}{d}\right) f(x)\,\mathrm{d}x \\
&= \int P\left(Y \leq \frac{X+a}{b},\ Z \leq \frac{X+c}{d} | X = x\right) f(x)\,\mathrm{d}x \\
&= P\left(Y \leq \frac{X+a}{b},\ Z \leq \frac{X+c}{d}\right) \\
&= P\left(Y - X/b \leq a/b, Z - X/d \leq c/d\right),
\end{aligned}$$

where the last line proceeds from the previous line by the law of total probability. Since Y, Z, and X are all Gaussian, the pair of linear combinations is multivariate Gaussian. Straightforward calculations of the mean, variance, and covariance between $W = bY - X$ and $U = dZ - X$ yields their joint distribution to have the mean and covariance matrix shown in (5.21). □

Using this Lemma 2, we can now proceed to derive the variance of the unbiased estimator of (5.20).

Theorem 3. *If* $Y_j \sim N(\theta_j^0, \sigma_j^2)$ *and* $\theta_j^0 \sim N(\mu_{\theta_j}, \sigma_\theta^2)$*, then*

$$
\begin{aligned}
&Var\left(\Phi\left[\frac{Y_j - \frac{\sigma_j}{\sigma_\theta^2}\theta_j^0 + z_{1-\alpha/2}\sqrt{\frac{\sigma_\theta^4}{\sigma_j^2}+\sigma_\theta^2}}{\sqrt{\frac{\sigma_\theta^4-\sigma_j^4}{\sigma_j^4}}}\right] - \Phi\left[\frac{Y_j - \frac{\sigma_j}{\sigma_\theta^2}\theta_j^0 + z_{\alpha/2}\sqrt{\frac{\sigma_\theta^4}{\sigma_j^2}+\sigma_\theta^2}}{\sqrt{\frac{\sigma_\theta^4-\sigma_j^4}{\sigma_j^4}}}\right]\right) = \\
&F_{MVN}\left(x = \begin{bmatrix} a/b \\ a/b \end{bmatrix}; \mu = \begin{bmatrix} -\theta_j^0/b \\ -\theta_j^0/b \end{bmatrix}, \Sigma = \begin{bmatrix} 1+b^{-2}\sigma_j^2 & b^{-2}\sigma_j^2 \\ b^{-2}\sigma_j^2 & 1+b^{-2}\sigma_j^2 \end{bmatrix}\right) + \\
&F_{MVN}\left(x = \begin{bmatrix} c/b \\ c/b \end{bmatrix}; \mu = \begin{bmatrix} -\theta_j^0/b \\ -\theta_j^0/d \end{bmatrix}, \Sigma = \begin{bmatrix} 1+b^{-2}\sigma_j^2 & b^{-2}\sigma_j^2 \\ b^{-2}\sigma_j^2 & 1+b^{-2}\sigma_j^2 \end{bmatrix}\right) - \\
&2 \times F_{MVN}\left(x = \begin{bmatrix} a/b \\ c/b \end{bmatrix}; \mu = \begin{bmatrix} -\theta_j^0/b \\ -\theta_j^0/b \end{bmatrix}, \Sigma = \begin{bmatrix} 1+b^{-2}\sigma_j^2 & b^{-2}\sigma_j^2 \\ b^{-2}\sigma_j^2 & 1+b^{-2}\sigma_j^2 \end{bmatrix}\right) - \\
&CP(\theta_j^0, \mu_{\theta_j}, \sigma_j, \sigma_\theta, \alpha)^2,
\end{aligned}
\tag{5.22}
$$

where

$$
\begin{aligned}
a &= \frac{\sigma_j}{\sigma_\theta^2}\theta_j^0 + z_{1-\alpha/2}\sqrt{\frac{\sigma_\theta^4}{\sigma_j^2}+\sigma_\theta^2} \\
b &= \sqrt{\frac{\sigma_\theta^4-\sigma_j^4}{\sigma_j^4}} \\
c &= \frac{\sigma_j}{\sigma_\theta^2}\theta_j^0 + z_{\alpha/2}\sqrt{\frac{\sigma_\theta^4}{\sigma_j^2}+\sigma_\theta^2}
\end{aligned}
$$

Proof. Expanding the variance operator simplifies the calculation:

$$
\begin{aligned}
Var(A-B) =& E\left[(A-B)^2\right] - E^2\left[A-B\right] \\
=& E\left[A^2\right] + E\left[B^2\right] - 2E\left[AB\right] - E^2\left[A-B\right].
\end{aligned}
$$

Letting A and B equal the left and right terms inside left-hand side of variance operator in (5.21), and applying Lemma 2 yields the first three terms on the right-hand side of (5.22). By Theorem 2, $A-B$ is an unbiased estimator for the coverage probability, which provides

the final term. □

To demonstrate the variability of the estimator, in Figure 5.3 we plot the unbiased coverage estimator along with the 5^{th} and 95^{th} quantiles and ribbons representing the middle 90% of the distribution of the estimator. The quantiles were found by generating 10,000 samples from Y_j, applying the unbiased coverage estimator of (5.20) to each draw, and then finding the quantiles of the transformed draws. Unlike directly plotting the estimator plus and minus the (scalar multiplied) standard deviation, these quantiles will lie between zero and one. Only curves where the ratio $\sigma_j/\sigma_\theta < 1$ can be shown due to the limitation of the unbiased estimator.

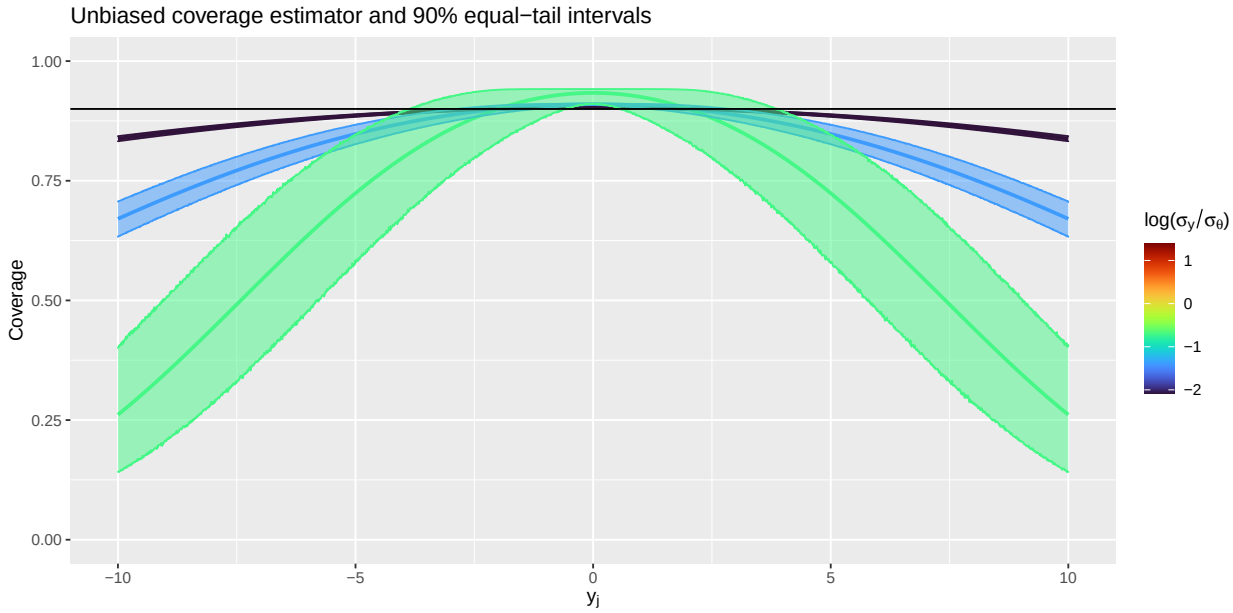


Figure 5.3: Curves of the unbiased estimator as a function of the (mean of the) observation for group j, y_j. The middle 90% of the distribution of the estimator is also shown. The various curves are shown for different values of the ratio σ_j/σ_θ given $\sigma_\theta = 2$ and align with those shown in Figures 5.1 and 5.2.

5.3.2 Proposed Use

We suggest that in the case of non-uniform coverage probability intervals, an estimate of the coverage probability be included along with any uncertainty interval. In particular, the coverage we are suggesting is the coverage of mean of group j, conditional on all other observed data. We suggest this because the coverage of group j will depend on both the specified model and all other observations, $\boldsymbol{y}_{-j}$. The proposed estimated coverage for group j can then be interpreted as the long-term success rate at which the $1-\alpha$ credible interval for $\hat{\theta}_j^0$ captures θ_j^0 given the specific orientation of the other $\boldsymbol{y}_{-j}$ observations. To estimate the coverage of the mean of group j, we suggest the following steps:

1. Fit the full model using all data except that from group j and find the posterior for group j: $\pi(\theta_j^0|\boldsymbol{y}_{-j})$. For large computationally expensive models, this step may be performed using leave-one-out (LOO) importance sampling (Vehtari *et al.*, 2017).

2. If the posterior $\pi(\theta_j|\boldsymbol{y}_{-j})$ is Gaussian, proceed to the next step. Otherwise, approximate the posterior with a Gaussian by, for example, matching the mode and variance.

3. Treat the (approximate) Gaussian posterior from the previous step as the "prior" for θ_j so that $\mu_{\theta_j} = \mathrm{E}[\theta_j|\boldsymbol{y}_{-j}]$ and $\sigma_\theta^2 = \mathrm{Var}[\theta_j|\boldsymbol{y}_{-j}]$.

4. If σ_j^2 is not known, estimate σ_j^2 from the observed data from group j

5. Apply the unbiased coverage estimator (5.20).

In this fashion we can obtain the estimated frequentist coverage probability of group j conditional on the model choice and the observed data in the other groups. We note that the coverage estimator for each group j uses the complete dataset when making the estimate by using information from the data from all other groups in the "prior" in conjunction with the mean from group j, fulfilling Rubin's only requirement for Bayesianly justifiable frequency calculations.

5.4 Fay-Herriot Simulation Studies

We perform a simulation study to verify the unbiased coverage probability estimator. In each simulation, we simulate data from the following Fay-Herriot model:

$$
\begin{aligned}
Y_{ij} &\sim N(\theta_j^0, 1) \text{ for } i = 1 \dots n_j, \text{ and } j = 1, \dots, 32 \\
n_j &\sim Pois(8) \\
\theta_j^0 &\sim N(1 + X_j^1 + X_j^2, 1)
\end{aligned}
$$

where n_j is restricted to be at least 2 and the covariates the covariate X_j^1 and X_j^2 are simulated from a $Unif(0, 1)$ distribution. The posterior distribution is approximated using the R-INLA R package (Rue *et al.*, 2009; `www.rinla.org`). Using the posterior, 90% equal-tail credible intervals are found for each the 32 groups and binary coverage of the simulated θ_j^0's are checked. This process was repeated 256 times and the 256×32 simulated group means were split into 100 equal-count bins to average their binary coverage. In each iteration of the simulation, the unbiased coverage estimator of (5.20) was also calculated three different ways: assuming the variance parameters were known, using the posterior means of the variance parameters as plug-in estimates, and using the posterior means of the variance parameters as plug-in estimates and replacing the standard normal CDF functions of the unbiased estimator with the CDF of t-distributions with $n_j - 1$ degrees of freedom to somewhat account for the plug-in estimators variance. The results from these three different calculations of the unbiased estimator are shown in Figures 5.4 - 5.6 and the bias between them and the empirical binned coverage is shown in Figure 5.7. We see that the unbiased estimator using known variance parameters aligns quite well with the empirical binned coverage, as expected. Dropping in the plug-in estimates for the variances flattens out the estimated coverage compared to the truth, but using the t-distribution CDF corrects for a significant amount of the added bias. Due to the complex nature of the two CDF arguments of the unbiased estimator, we know that the t-distribution is not the exact distribution of

these terms when the variances are estimated, but the wider tails do appear to be more appropriate than those of the standard normal.

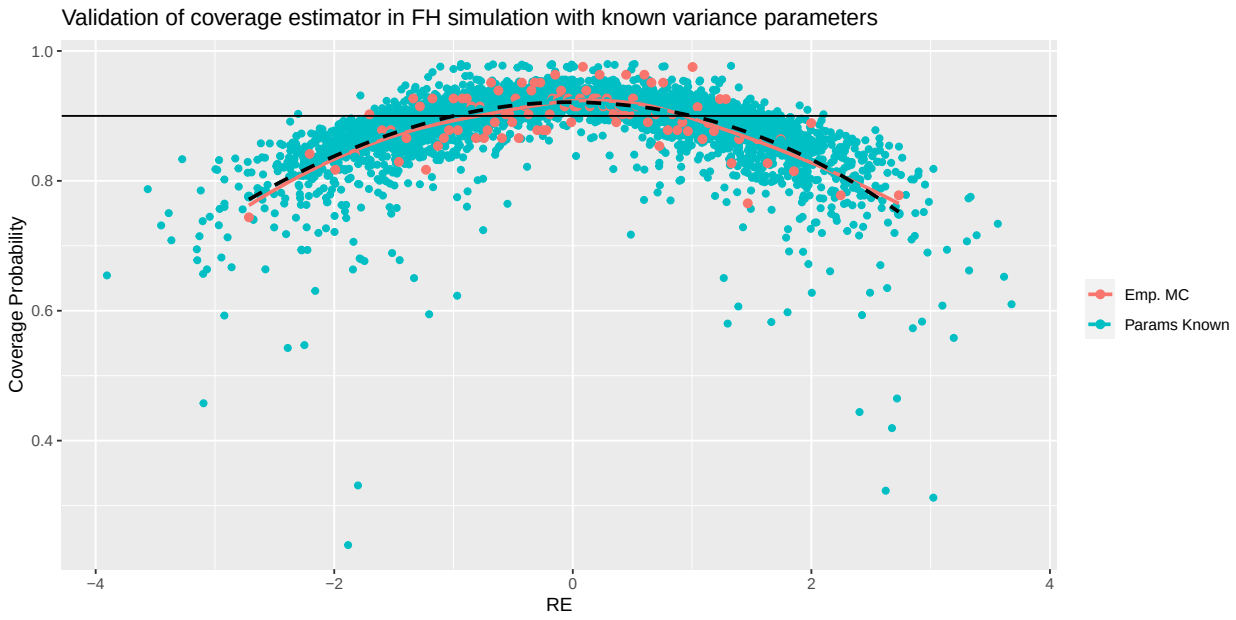


Figure 5.4: The red observations are the (binned) average of the binary indicators assessing if the true (simulated) group means were contained within the posterior credible intervals. In blue we have the (unbinned) unbiased coverage estimator for the group means calculated using known variances. For clarity, the smooth line through the estimated coverage cloud is shown in dashed black and the nominal coverage is drawn with a horizontal black line.

5.5 Empirical Example: Radon Levels

As an applied example, we consider data from the State Residential Radon Survey (SRRS) collected by the Environmental Protection Agency (EPA) during 1987-1988. Radon is a radioactive gas that has been observed in homes throughout the United States and elsewhere. It forms naturally in soil and rocks as one of the radioactive metals uranium, thorium, or radium break down within the material. It is the second leading cause of lung cancer and the United States EPA and Sugeon General's office estimate that radon is responsible for more

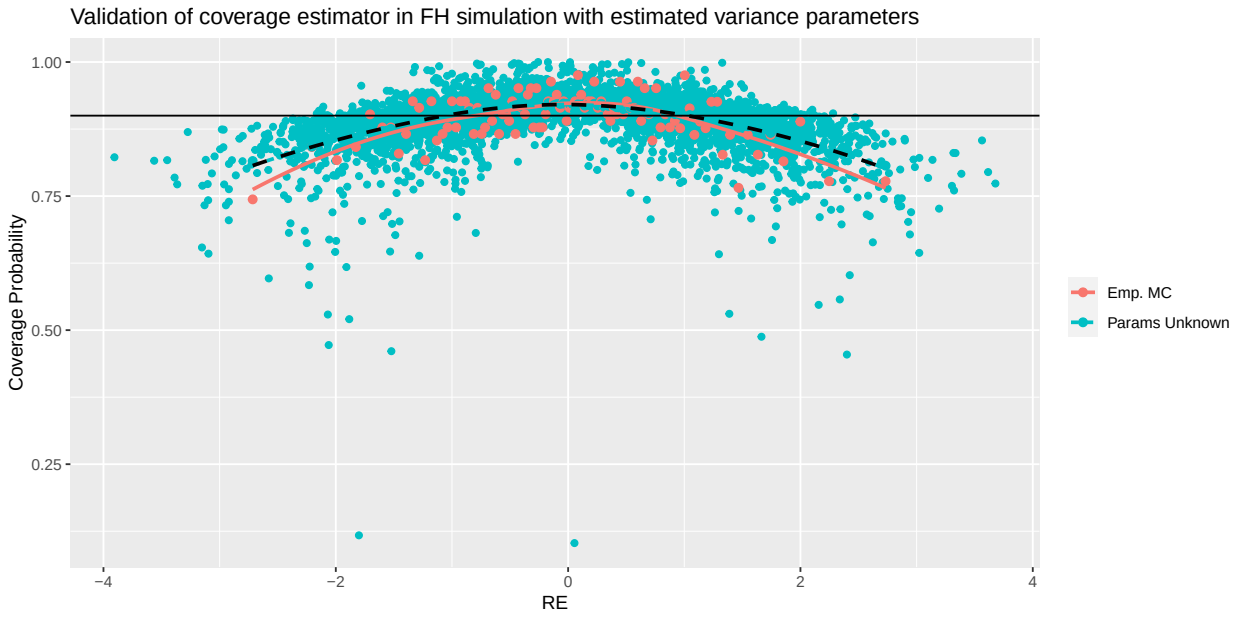


Figure 5.5: The points and lines have the same interpretation as in Figure 5.4, except all variance parameters used in the unbiased coverage estimation are estimated quantities.

than 20,000 lung cancer deaths each year in the US (CDC, 2022a). The SRRS collected a stratified random sample of radon concentrations present in 12,777 homes from across 472 counties and 9 states. This dataset was previously analyzed by Price *et al.* (1996) who developed a linear MEM to construct estimates and Bayesian credible intervals for county-specific geometric mean radon levels. The dataset was also used by Burris and Hoff (2020) to demonstrate their Fay-Herriot FAB intervals and we closely follow their analysis and apply our coverage probability estimator to the EB intervals that they also produced.

Following their lead, we only use a subset of the SRRS data, selecting observations from 3,767 households within 209 counties in four nearby US states: Indiana, Michigan, Minnesota, and Wisconsin. As in Price *et al.* (1996) and Burris and Hoff (2020), we use processed data that has been adjusted to minimize the oversized impact caused by very low concentration measurements that can be caused by measurement error. Once adjusted, the authors proceeded to model the household-level observations, $r_{ij} = \exp(y_{ij})$ within each

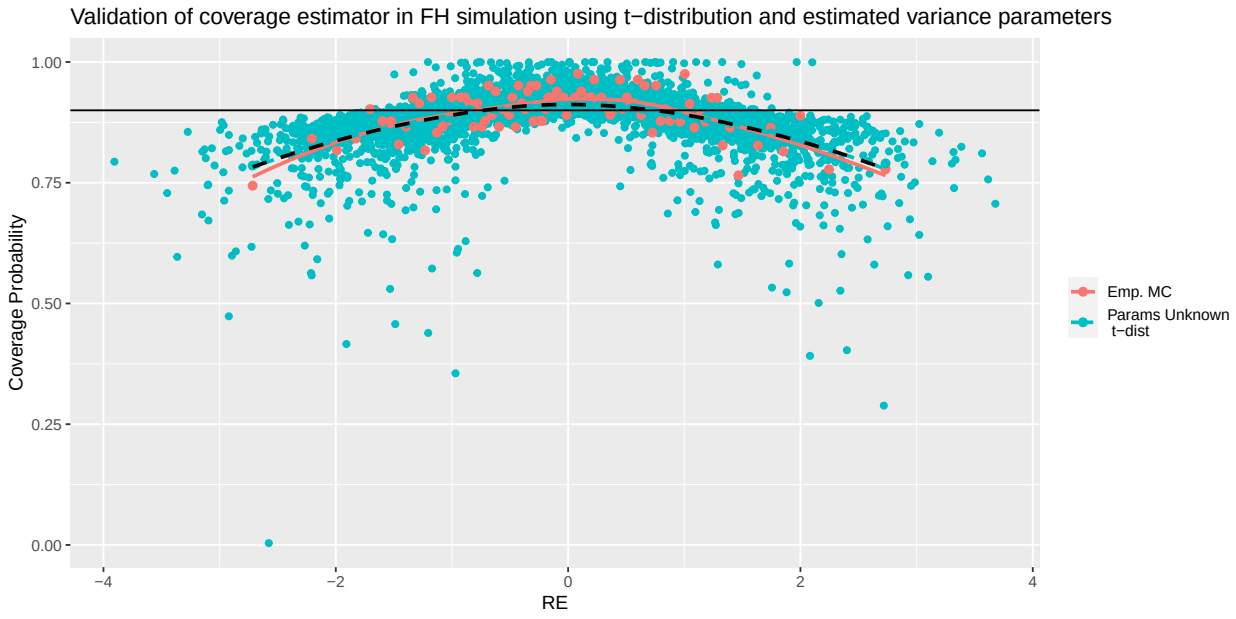


Figure 5.6: The points and lines have the same interpretation as in Figures 5.4 and 5.5, except all parameters used in the unbiased coverage estimation are estimated quantities and the standard normal CDF from the unbiased estimator in (5.20) were replaced with the CDF of a t-distribution with degrees of freedom equal to $n_j - 1$.

county with a log-normal sampling distribution. That is, for each county j, n_j normal observations $y_{1j}, \ldots, y_{n_j} \overset{iid}{\sim} N(\theta_j^0, \tau_j^2)$ are assumed to have been sampled where θ_j^0 is the true geometric mean radon concentration in county j (true in the sense that it would be the observed geometric mean if every household was sampled) and τ_j^2 is the unknown variance of the log radon observations within county j. The sample mean of the log-transformed measurements within a county is then assumed to be normal with $y_j \sim N(\theta_j^0, \sigma_j^2)$ where $\sigma_j^2 = \tau_j^2/n_j$.

Burris and Hoff (2020) fit a spatial Fay-Herriot model using a simultaneous autoregressive (SAR) model to account for the spatial relationship of the random effects (Singh *et al.*, 2005). In the SAR model, a $m \times m$ proximity matrix $\boldsymbol{W}$ and spatial correlation parameter ρ is used

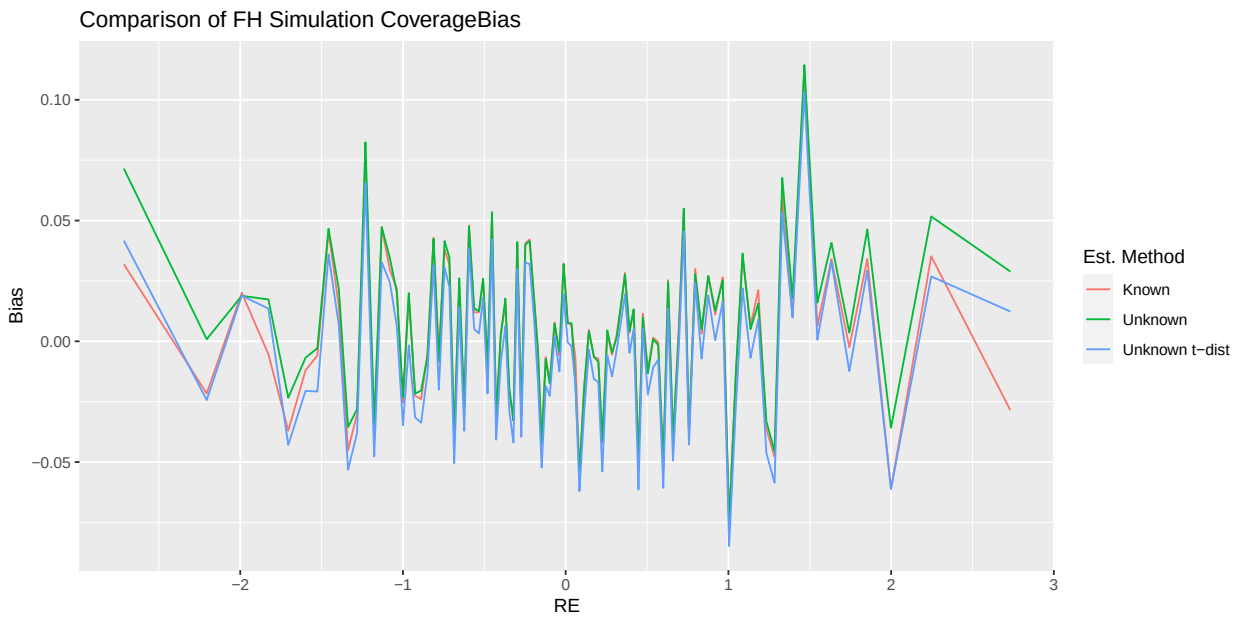


Figure 5.7: Comparison of the estimated coverage bias between the coverage probability estimates and the binned binary empirical coverage from the FH simulations shown in Figures 5.4, 5.5 and 5.6.

to smooth the spatial random effect $\boldsymbol{u}$ and relate it to itself:

$$\boldsymbol{u} = \rho \boldsymbol{W} \boldsymbol{u} + \boldsymbol{v},$$

which yields $\boldsymbol{u} = (\boldsymbol{I} - \rho \boldsymbol{W})^{-1} \boldsymbol{v}$ where $\boldsymbol{v}$ is a vector of iid Gaussian random variables with variance ω^2. Given this model for the structured spatial random effects yields the following linking model for $\boldsymbol{\theta}$:

$$\boldsymbol{\theta} \sim N\left(\boldsymbol{X}\boldsymbol{\beta},\ \omega^2 \left((\boldsymbol{I} - \rho \boldsymbol{W})(\boldsymbol{I} - \rho \boldsymbol{W}^T)\right)^{-1}\right).$$

A single county-level surface radium content measurement produced by the National Uranium Resource Evaluation was used as a covariate, and with an intercept these two terms make up the linear predictor: $\boldsymbol{x}_j^T \boldsymbol{\beta} = \beta_0 + radium_j \beta_1$. For $\boldsymbol{W}$ they use a row-standardized

squared exponential distance between the centroids of the counties: $W_{ij} = \frac{\exp(-d_{ij}^2)}{\sum_{j \neq i} \exp(-d_{ij}^2)}$, where d_{ij} is the distance between the centroids of country i and county j and $W_{ii} = 0$ for all i.

With the sampling and linking model defined, we now describe a real world application. For each county j, maximum likelihood estimation was performed on the Fay-Herriot model defined by the Gaussian county sampling model and GMRF SAR linking model with an intercept and radium covariate using all data except that from county j. Details of the MLE can be found in (Burris and Hoff, 2020, Appendix A.3). Using the LOO fit, both 90% FAB intervals and (nominal) 90% EB intervals were constructed. Using the LOO estimated parameters, the coverage probability of each EB interval was also calculated using the coverage probability estimator (5.20). The results are shown in Figure 5.8. Typically the EB intervals are narrower than the FAB intervals and we see the effect of the shrinkage in the intervals: due to local smoothing via the SAR process, the EB intervals for low county means tend to be higher than the corresponding FAB intervals and the EB intervals for high county means tend to be lower. The EB intervals are typically narrower (61% of the time in this example) than the corresponding FAB intervals, but as we clearly see, this comes at the cost of non-uniform frequentist coverage. Of particular policy import, those counties with the highest (and lowest) mean radon concentrations have the most severe undercoverge. If, for example, financial resources were to be distributed to counties proportionately to the magnitude of radon estimated from the Fay-Herriot formulation, then counties with the worst concentrations would likely not receive support commensurate to their needs.

5.6 Discussion

In this chapter we develop an unbiased estimator for the coverage probability of group means for the Fay-Herriot model when the within-group and across-group variances are known. While seemingly specific, this class of models encompasses any Gaussian-Gaussian conjugate model formulation with known variance components which could include popular spatial modeling methods such as small-area autoregressive models like the Besag-York-Mollie2

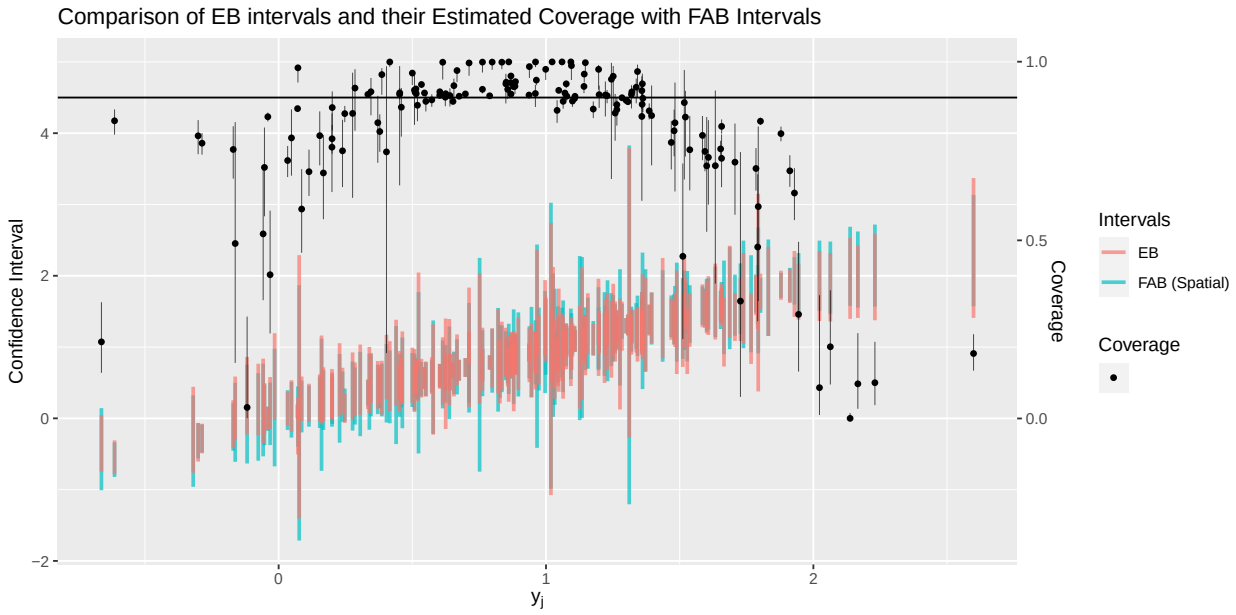


Figure 5.8: Comparison of the (nominal) 90% EB uncertainty intervals (red) and the 90% FAB intervals (blue) are shown using the left vertical axis plotted against the mean of the log adjusted household radon concentrations in each county. The estimated coverage probability of the EB intervals and the middle 90% of their sampling distribution is shown (black) plotted on the right vertical axis. The nominal coverage of the EB intervals is shown as a horizontal line on the right vertical axis.

(Riebler *et al.*, 2016) and geostatistical Gaussian process models (Gelfand and Schliep, 2016). While the known variance restriction is limiting, with sufficient data volumes and models where the within-group variances may be estimated using pooled data, the unbiased estimator will be approximately correct. The simulation results shown in Figure 5.6 also suggest that replacing the standard normal CDF with an appropriate t-distribution may attenuate bias caused from using unknown variances. The other primary limitation to the coverage probability estimates is the restriction that the unbiased estimator may only be applied in scenarios when the within-group variance is smaller than the across-group variance. This should not prevent application to problems with at least modest sample sizes.

As demonstrated, the area-specific frequentist coverage probability of standard intervals

for MEMs will not obtain nominal coverage levels even though they do exhibit population-level coverage. While there are alternative inferential techniques that can be used to produce uncertainty intervals with uniform area-specific coverage, these methods, so far, are restricted to particular applications. The coverage estimator we propose is also limited to a specific model family, but there is much opportunity to develop coverage estimators for other situations and as a future extension to this project we have plans to develop a non-analytical general-use coverage estimator for a wider class of MEMs.

To motivate the need and practicality of the coverage probabilities, we show via theoretical curves, simulation results, and an empricial study how significantly the coverage probability of group mean intervals may vary in MEMs. The pragmatic Bayesian may be interested in using the coverage estimator to help judge the validity of their methods across replication studies, and any modeler using shrinkage models may be interested in openly communicating how their uncertainty intervals may be interpreted to those more familiar with uniform coverage frequentist confidence intervals. As such, we hope that this chapter, in part, serves as a reminder and warning to those who use these methods and disseminate their results to the general reader. We believe that the inclusion of the coverage estimates alongside the uncertainty intervals will help clarify, or at minimum draw attention to, the non-uniform nature of the intervals from MEMs to help improve interpretation and decision making.

It is worth noting that sharing uncertainty intervals with others, including less statistically versed collaborators, that have a credible probability and a different frequentist confidence level could lead to confusion. At minimum, it may lead to questions about how someone should act on the apparent duality of the uncertainty interval. This is a deep question that relies on the statistical understanding and inclination of all parties involved. Our personal suggestion is that despite the unwieldiness of the frequentist confidence definition, it may be more interpretable than a subjective probability statement if it the frequentist confidence is that about as the long-term success rate of the interval to capture the truth given many replications of the experiment from start (data collection) to finish (uncertainty interval

calculation). That said, there may be situations, for example if a strongly motivated prior based on previous research can be used, where it is advantageous to lean more heavily on the credible probability interpretation.

Finally, I note that there is opportunity to extend these methods beyond the Gaussian likelihood and prior conjugate model that I explored. First, since this is primarily directed towards estimating the confidence of group means, non-Gaussian observations whose means are close to Gaussian fit nicely within the current paradigm. With this in mind, it may be possible to derive asymptotic results for the unbiased estimator of (5.20). Secondly, although I have not yet attempted it, I think it is likely that similar results are possible for other conjugate models such as beta-binomial priors and likelihoods. Computational methods should allow for the theoretical coverage function, like that in (5.14), to be calculated for many Bayesian models. Investigating whether or not a general technique to derive unbiased estimators from these non-analytical functions is possible is an open-ended question.

Chapter 6
DISCUSSION AND FUTURE WORK

Improving the health and well-being of a population requires an adequate understanding of the population's health indicators, their historical patterns and trends, and how they vary across geographies. More often than not, the available health data is limited and specialized tools and methods are required in order to make the best use of it. The available data often comes from disparate sources and may involve different types of measurements. The methods must be capable of leveraging all the available information to produce reliable granular estimates and their associated uncertainty. These quantities are necessary to evaluate the present state of facets of a population's health, in light of its historical progress, and to aid in future resource allocation. In this dissertation we addressed three aspects of the difficulties related to spatial public health estimation: the inferential tools, development of methods to combine different data types and sources for improved estimation, and we proposed a novel approach to allow improved interpretation of the uncertainty of the estimates.

In Chapter 3 we provided a detailed statistical review of Template Model Builder, contrasting it against the popular approximate Bayesian inferential technique, INLA, and demonstrating via extensive simulations its wide-ranging suitability for spatial effects modeling. Serving a secondary purpose, the simulations also studied the computationally efficient SPDE approach to GP modeling, and provided the most detailed simulation assessment of that approximation in non-Gaussian sampling scenarios. Both of these tools, TMB and the SPDE approach, have great potential to extend the possible set of models used in spatial public health and we hope that they allow researchers to implement the models their problems require instead of being limited to using the models their existing tools allow.

With a new inferential approach in hand, in Chapter 4 we overcame long-standing com-

putational challenges required to fit a joint model for breast cancer incidence and mortality which hinged upon a non-linear interaction between the two risks. We extended the model to allow it to flexibly capture nuanced differences in both incidence and mortality risk across 18 age groups in 39 European countries and over 18 recent years. We demonstrated that this approach is able to appropriately capture the variation present in the data and that its suited to use recently available mortality data to backcast incidence which lags behind in reporting.

In Chapter 5 we turned our attention to the interpretation of the uncertainty intervals of the spatial mixed effects models that we used in Chapters 3 and 4. Due to the sparsity of available health data, mixed effects models (MEMs) are commonly used to provide dense estimates of health indicators across space and time. Traditionally, MEMs are used for interpreting marginal effects, but in public health there is a real need to use the conditional, space and time (and age) specific, estimates. The shrinkage estimation inherent to these methods result in uncertainty intervals which do not have the frequentist coverage probability that many would expect them to have. In this chapter we argue the need to consider the under- or overcoverage of the strata-specific intervals and we demonstrate how the coverage probability can be estimated in Gaussian sampling models. We conclude with an example which demonstrates how neglecting the non-uniform coverage probabilities of uncertainty intervals from MEMs can lead to adverse decisions.

The topics covered in this dissertation were developed to better aid researchers as they tackle spatial-temporal public health modeling questions and work to disseminate the results. In addition, we hope that the by aiding the modelers in these ways, they will be able to provide useful information to policy makers as they evaluate the current health landscape and work towards improving it via resource planning and allocation. Of particular next interest is to extend the use cases for the coverage probability estimator in order to improve the interpretation of the uncertainty intervals from the mixed effects models which have become so ubiquitous in public health applications. In my next professional position I will be primarily tasked with providing an excellent statistics education to undergraduate students

and I look forward to continuing to improve statistical communication and literacy in that role. Concurrently, I plan to continue pushing forward advancements in spatial health modeling and I plan to find (and welcome) compelling applications and collaborations in that endeavor.

BIBLIOGRAPHY

Alappat, C., Basermann, A., Bishop, A. R., Fehske, H., Hager, G., Schenk, O., Thies, J., and Wellein, G. (2020). A Recursive Algebraic Coloring Technique for Hardware-Efficient Symmetric Sparse Matrix-Vector Multiplication. *ACM Transactions on Parallel Computing*, **7**(3).

Albertsen, C. M., Whoriskey, K., Yurkowski, D., Nielsen, A., and Flemming, J. M. (2015). Fast fitting of non-Gaussian state-space models to animal movement data via Template Model Builder. *Ecology*, **96**(10), 2598–2604.

Auger-Méthé, M., Albertsen, C. M., Jonsen, I. D., Derocher, A. E., Lidgard, D. C., Studholme, K. R., Bowen, W. D., Crossin, G. T., and Flemming, J. M. (2017). Spatiotemporal modelling of marine movement data using Template Model Builder (TMB). *Marine Ecology Progress Series*, **565**, 237–249.

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

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2022). Conformal prediction beyond exchangeability. *arXiv preprint arXiv:2202.13415*.

Bell, B. (2007). CppAD: a package for C++ algorithmic differentiation. http://www.coin-or.org/CppAD.

Berliner, L. M. (1996). Hierarchical Bayesian time series models. In *Maximum entropy and Bayesian methods*, pages 15–22. Springer.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, **43**(1), 1–20.

Blangiardo, M. and Cameletti, M. (2015). *Spatial and Spatio-Temporal Bayesian Models with R-INLA*. John Wiley & Sons.

Blangiardo, M., Cameletti, M., Baio, G., and Rue, H. (2013). Spatial and spatio-temporal models with R-INLA. *Spatial and Spatio-temporal Epidemiology*, **4**, 39–55.

Bolker, B. M., Gardner, B., Maunder, M., Berg, C. W., Brooks, M., Comita, L., Crone, E., Cubaynes, S., Davies, T., de Valpine, P., *et al.* (2013). Strategies for fitting nonlinear ecological models in R, AD Model Builder, and BUGS. *Methods in Ecology and Evolution*, **4**(6), 501–512.

Bollhöfer, M., Eftekhari, A., Scheidegger, S., and Schenk, O. (2019). Large-scale Sparse Inverse Covariance Matrix Estimation. *SIAM Journal on Scientific Computing*, **41**(1), A380–A401.

Bollhöfer, M., Schenk, O., Janalik, R., Hamm, S., and Gullapalli, K. (2020). State-of-the-art sparse direct solvers. In A. Grama and A. H. Sameh, editors, *Parallel Algorithms in Computational Science and Engineering*, pages 3–33. Springer.

Bolstad, G. H., Hindar, K., Robertsen, G., Jonsson, B., Sægrov, H., Diserud, O. H., Fiske, P., Jensen, A. J., Urdal, K., Næsje, T. F., *et al.* (2017). Gene flow from domesticated escapes alters the life history of wild Atlantic salmon. *Nature Ecology & Evolution*, **1**(5), 1–5.

Bray, F., Jemal, A., Grey, N., Ferlay, J., and Forman, D. (2012). Global cancer transitions according to the Human Development Index (2008–2030): a population-based study. *The Lancet Oncology*, **13**(8), 790–801.

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., and Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, **9**(2), 378–400.

Burris, K. C. and Hoff, P. D. (2020). Exact adaptive confidence intervals for small areas. *Journal of Survey Statistics and Methodology*, **8**(2), 206–230.

Burstein, R., Henry, N. J., Collison, M. L., Marczak, L. B., Sligar, A., Watson, S., Marquez, N., Abbasalizad-Farhangi, M., Abbasi, M., Abd-Allah, F., *et al.* (2019). Mapping 123 million neonatal, infant and child deaths between 2000 and 2017. *Nature*, **574**(7778), 353–358.

Carlin, B. P. and Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall/CRC,.

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

Casella, G. and Hwang, J. G. (2012). Shrinkage confidence procedures. *Statistical Science*, **27**(a), 51–60.

CDC (2022a). Radon and your health.

CDC (2022b). *What Are the Risk Factors for Breast Cancer?* https://www.cdc.gov/cancer/breast/basic_info/risk_factors.htm.

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

Cressie, N., Calder, C. A., Clark, J. S., Hoef, J. M. V., and Wikle, C. K. (2009). Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecological Applications*, **19**(3), 553–570.

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

Dunn, R., Wasserman, L., and Ramdas, A. (2022). Distribution-free prediction sets for two-layer hierarchical models. *Journal of the American Statistical Association*, (just-accepted), 1–29.

Dwyer-Lindgren, L., Bertozzi-Villa, A., Stubbs, R. W., Morozoff, C., Kutz, M. J., Huynh, C., Barber, R. M., Shackelford, K. A., Mackenbach, J. P., van Lenthe, F. J., *et al.* (2016). US county-level trends in mortality rates for major causes of death, 1980-2014. *The Journal of the American Medical Association*, **316**(22), 2385–2401.

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

Fay, R. E. and Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*, **74**(366a), 269–277.

Ferkingstad, E., Rue, H., *et al.* (2015). Improving the INLA approach for approximate Bayesian inference for latent Gaussian models. *Electronic Journal of Statistics*, **9**(2), 2706–2731.

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

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

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

Ferlay, J., Colombet, M., Soerjomataram, I., Dyba, T., Randi, G., Bettio, M., Gavin, A., Visser, O., and Bray, F. (2018). Cancer incidence and mortality patterns in Europe:

Estimates for 40 countries and 25 major cancers in 2018. *European Journal of Cancer*, **103**, 356–387.

Ferlay, J., Colombet, M., Soerjomataram, I., Mathers, C., Parkin, D. M., Piñeros, M., Znaor, A., and Bray, F. (2019). Estimating the global cancer incidence and mortality in 2018: Globocan sources and methods. *International journal of cancer*, **144**(8), 1941–1953.

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

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

Fitzmaurice, C., Allen, C., Barber, R. M., Barregard, L., Bhutta, Z. A., Brenner, H., Dicker, D. J., Chimed-Orchir, O., Dandona, R., Dandona, L., *et al.* (2017). Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 32 cancer groups, 1990 to 2015: a systematic analysis for the global burden of disease study. *JAMA Oncology*, **3**(4), 524–548.

Foreman, K. J., Marquez, N., Dolgert, A., Fukutaki, K., Fullman, N., McGaughey, M., Pletcher, M. A., Smith, A. E., Tang, K., Yuan, C.-W., *et al.* (2018). Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: reference and alternative scenarios for 2016–40 for 195 countries and territories. *The Lancet*, **392**(10159), 2052–2090.

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

Fournier, D. A., Skaug, H. J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M. N., Nielsen, A., and Sibert, J. (2012). AD Model Builder: using automatic differentiation

for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software*, **27**(2), 233–249.

Free, C. M., Thorson, J. T., Pinsky, M. L., Oken, K. L., Wiedenmann, J., and Jensen, O. P. (2019). Impacts of historical warming on marine fisheries production. *Science*, **363**(6430), 979–983.

Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2019). Constructing Priors that Penalize the Complexity of Gaussian Random Fields. *Journal of the American Statistical Association*, **114**(525), 445–452.

Gakidou, E., Cowling, K., Lozano, R., and Murray, C. J. (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. and Schliep, E. M. (2016). Spatial statistics and Gaussian processes: A beautiful marriage. *Spatial Statistics*, **18**, 86–104.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**(410), 398–409.

Gelfand, A. E., Diggle, P., Fuentes, M., and Guttorp, P. (2010). *Handbook of Spatial Statistics*. CRC press.

Gomez-Rubio, V. (2020). *Bayesian Inference with INLA*. CRC Press.

Griewank, A. and Walther, A. (2008). *Evaluating derivatives: principles and techniques of algorithmic differentiation*. Society for Industrial and Applied Mathematics, 2 edition.

Guennebaud, G., Jacob, B., *et al.* (2010). Eigen v3. http://eigen.tuxfamily.org.

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

He, K. (1992). Parametric empirical Bayes confidence intervals based on James-Stein estimator. *Statistics & Risk Modeling*, **10**(1-2), 121–132.

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M., *et al.* (2019). A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, **24**(3), 398–425.

Hoff, P. and Yu, C. (2019). Exact adaptive confidence intervals for linear regression coefficients. *Electronic Journal of Statistics*, **13**(1), 94–119.

Hogan, M., Foreman, K., Naghavi, M., Ahn, S., Wang, M., Makela, S., Lopez, A., Lozano, R., and Murray, C. (2010). Maternal mortality for 181 countries, 1980–2008: a systematic analysis of progress towards millennium development goal 5. *The Lancet*, **375**, 1609–1623.

Hwang, J. G., Qiu, J., and Zhao, Z. (2009). Empirical Bayes confidence intervals shrinking both means and variances. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**(1), 265–285.

Kassebaum, N. J., Barber, R. M., Bhutta, Z. A., Dandona, L., Gething, P. W., Hay, S. I., Kinfu, Y., Larson, H. J., Liang, X., Lim, S. S., *et al.* (2016). Global, regional, and national levels of maternal mortality, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet*, **388**(10053), 1775–1812.

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

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

Kocarnik, J. M., Compton, K., Dean, F. E., Fu, W., Gaw, B. L., and Global Burden of Disease 2019 Cancer Collaboration (2021). Global Burden of Disease 2019 Cancer Collaboration. Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and

Disability-Adjusted Life Years for 29 Cancer Groups From 2010 to 2019: A Systematic Analysis for the Global Burden of Disease Study 2019. *JAMA Oncology*.

Krainski, E. T. E. T., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F., and Rue, H. (2018). *Advanced Spatial Modeling with Stochastic Partial Differential Equations using R and INLA*. Chapman & Hall/CRC.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., and Bell, B. M. (2016). TMB : Automatic Differentiation and Laplace Approximation. *Journal of Statistical Software*, **70**(5), 1–21.

Laird, N. M. and Louis, T. A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *Journal of the American Statistical Association*, **82**(399), 739–750.

Leroux, B. G., Lei, X., and Breslow, N. (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*, pages 179–191. Springer.

Lindgren, F. and Rue, H. (2015). Bayesian Spatial Modelling with R-INLA. *Journal of Statistical Software*, **63**(19), 1–25.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B*, **73**(4), 423–498.

Margossian, C., Vehtari, A., Simpson, D., and Agrawal, R. (2020). Hamiltonian Monte Carlo using an adjoint-differentiated Laplace approximation: Bayesian inference for latent Gaussian models and beyond. *Advances in Neural Information Processing Systems*, **33**, 1–12.

Marquez, N. and Wakefield, J. (2021). Harmonizing child mortality data at disparate geographic levels. *Statistical Methods in Medical Research*, **30**(5), 1187–1210.

Martino, S. and Riebler, A. (2020). Integrated Nested Laplace Approximations (INLA). In *Wiley StatsRef: Statistics Reference Online*, pages 1–19. American Cancer Society.

Martins, T. G., Simpson, D., Lindgren, F., and Rue, H. (2013). Bayesian computing with INLA: New features. *Computational Statistics & Data Analysis*, **67**, 68–83.

Meehan, T. D., Michel, N. L., and Rue, H. (2020). Estimating Animal Abundance with N-Mixture Models Using the R-INLA Package for R. *Journal of Statistical Software, Articles*, **95**(2), 1–26.

Mercer, L. (2016). *Space-Time Smoothing Models for Surveillance and Complex Survey Data*. Ph.D. thesis, University of Washington.

Miller, D. L., Glennie, R., and Seaton, A. E. (2020). Understanding the Stochastic Partial Differential Equation Approach to Smoothing. *Journal of Agricultural, Biological and Environmental Statistics*, **25**(1), 1–16.

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

Monnahan, C. C. and Kristensen, K. (2018). No-U-turn sampling for fast Bayesian inference in ADMB and TMB: Introducing the adnuts and tmbstan R packages. *PLoS ONE*, **13**(5), e0197954.

Moraga, P. (2019). *Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny*. CRC Press.

Morris, C. N. (1983). Parametric Empirical Bayes Inference: Theory and Applications. *Journal of the American Statistical Association*, **78**(381), 47–55.

Niku, J., Warton, D. I., Hui, F. K., and Taskinen, S. (2017). Generalized linear latent

variable models for multivariate count and biomass data in ecology. *Journal of Agricultural, Biological and Environmental Statistics*, **22**(4), 498–522.

Niku, J., Hui, F. K., Taskinen, S., and Warton, D. I. (2019). gllvm: Fast analysis of multivariate abundance data with generalized linear latent variable models in R. *Methods in Ecology and Evolution*, **10**(12), 2173–2182.

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., and Sain, S. (2015). A multiresolution Gaussian process model for the analysis of large spatial datasets. *Journal of Computational and Graphical Statistics*, **24**(2), 579–599.

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

Pratt, J. W. (1963). Shorter confidence intervals for the mean of a normal distribution with known variance. *The Annals of Mathematical Statistics*, **34**(2), 574–586.

Price, P. N., Nero, A. V., and Gelman, A. (1996). Bayesian prediction of mean indoor radon concentrations for Minnesota counties. *Health Physics*, **71**, 922–936.

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

Ren, Q., Banerjee, S., Finley, A. O., and Hodges, J. S. (2011). Variational Bayesian methods for spatial data analysis. *Computational Statistics & Data Analysis*, **55**(12), 3197–3217.

Riebler, A., Sørbye, S. H., Simpson, D., and Rue, H. (2016). An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Statistical Methods in Medical Research*, **25**(4), 1145–1165.

Righetto, A. J., Faes, C., Vandendijck, Y., and Ribeiro Jr, P. J. (2020). On the choice of the mesh for the analysis of geostatistical data using R-INLA. *Communications in Statistics-Theory and Methods*, **49**(1), 203–220.

Robert, C. and Casella, G. (2011). A short history of Markov chain Monte Carlo: Subjective recollections from incomplete data. *Statistical Science*, **26**(1), 102–115.

Roth, G. A., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., Abbastabar, H., Abd-Allah, F., Abdela, J., Abdelalim, A., *et al.* (2018). Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*, **392**(10159), 1736–1788.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, **12**, 1151–1172.

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*, volume 104. Chapman & Hall/CRC, Boca Raton.

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

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., and Lindgren, F. K. (2017). Bayesian computing with INLA: A review. *Annual Review of Statistics and Its Application*, **4**(1), 395–421.

Schlather, M., Malinowski, A., Menck, P., Oesting, M., and Strokorb, K. (2015). Analysis, simulation and prediction of multivariate random fields with package RandomFields. *Journal of Statistical Software*, **63**(8), 1–25.

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

Singh, B. B., Shukla, G. K., and Kundu, D. (2005). Spatio-temporal models in small area estimation. *Survey Methodology*, **31**(2), 183.

Snijders, T. A. and Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage Publications, Los Angeles, CA, second ed. edition.

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

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

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, **71**(3), 209–249.

Taylor, B. M. and Diggle, P. J. (2014). INLA or MCMC? A tutorial and comparative evaluation for spatial prediction in log-Gaussian Cox processes. *Journal of Statistical Computation and Simulation*, **84**(10), 2266–2284.

Teng, M., Nathoo, F., and Johnson, T. D. (2017). Bayesian computation for log-Gaussian Cox processes: A comparative analysis of methods. *Journal of Statistical Computation and Simulation*, **87**(11), 2227–2252.

Thorson, J. T. and Kristensen, K. (2016). Implementing a generic method for bias correction in statistical models using random effects, with spatial and population dynamics examples. *Fisheries Research*, **175**, 66–74.

Thygesen, U. H., Albertsen, C. M., Berg, C. W., Kristensen, K., and Nielsen, A. (2017). Validation of ecological state space models using the Laplace approximation. *Environmental and Ecological Statistics*, **24**(2), 317–339.

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. (2019). Conformal prediction under covariate shift. *Advances in neural information processing systems*, **32**.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**(393), 82–86.

Tierney, L., Kass, R. E., and Kadane, J. B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association*, **84**(407), 710–716.

Uhry, Z., Belot, A., Colonna, M., Bossard, N., Rogel, A., Iwaz, J., Mitton, N., Grosclaude, P., and Remontet, L. (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.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, **27**(5), 1413–1432.

Vos, T., Lim, S. S., Abbafati, C., Abbas, K. M., Abbasi, M., Abbasifard, M., Abbasi-Kangevari, M., Abbastabar, H., Abd-Allah, F., Abdelalim, A., *et al.* (2020). Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*, **396**(10258), 1204–1222.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*. Springer Science & Business Media.

Wakefield, J., Simpson, D., and Godwin, J. (2016). Comment: Getting into space with a weight problem. *Journal of the American Statistical Association*, **111**(515), 1111–1118.

Wang, H., Liddell, C. A., Coates, M. M., Mooney, M. D., Levitz, C. E., Schumacher, A. E., Apfel, H., Lannarone, M., Phillips, B., Lofgren, K. T., *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.

Wang, H., Abajobir, A. A., Abate, K. H., Abbafati, C., Abbas, K. M., Abd-Allah, F., Abera, S. F., Abraha, H. N., Abu-Raddad, L. J., Abu-Rmeileh, N. M., *et al.* (2017). Global, regional, and national under-5 mortality, adult mortality, age-specific mortality, and life expectancy, 1970–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet*, **390**(10100), 1084–1150.

Weiss, D. J., Nelson, A., Gibson, H., Temperley, W., Peedell, S., Lieber, A., Hancher, M., Poyart, E., Belchior, S., Fullman, N., *et al.* (2018). A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature*, **553**(7688), 333–336.

Weiss, D. J., Lucas, T. C., Nguyen, M., Nandi, A. K., Bisanzio, D., Battle, K. E., Cameron, E., Twohig, K. A., Pfeffer, D. A., Rozier, J. A., *et al.* (2019). Mapping the global prevalence, incidence, and mortality of Plasmodium falciparum, 2000–17: a spatial and temporal modelling study. *The Lancet*, **394**(10195), 322–331.

Whittle, P. (1954). On Stationary Processes in the Plane. *Biometrika*, **41**(3/4), 434–449.

Wilson, K. and Wakefield, J. (2020). Pointless spatial modeling. *Biostatistics*, **21**, e17–e32.

Wilson, K. and Wakefield, J. (2021). Estimation of health and demographic indicators with incomplete geographic information. *Spatial and Spatio-temporal Epidemiology*, **37**, 100421.

Xu, C. and Xie, Y. (2021). Conformal prediction interval for dynamic time-series. In *International Conference on Machine Learning*, pages 11559–11569. PMLR.

Yu, C. and Hoff, P. D. (2018). Adaptive multigroup confidence intervals with constant coverage. *Biometrika*, **105**(2), 319–335.

Appendix A

APPENDIX FOR CHAPTER 3

A.1 Automatic Differentiation

This section is provided as a brief introduction and overview for those unfamiliar with automatic differentiation (AD), also known as algorithmic differentiation. AD comprises a set of computational techniques used to evaluate the derivate of functions within a computer framework. The methods generally work by noting that computers must break down even the most complicated functions into elementary or unary arithmetic operations in order to evaluate them, and that by (repeatedly) applying the chain rule to these operations, derivatives of different orders may also be numerically evaluated. Furthermore, it is well-established that the computational cost to evaluate the derivatives will be no more than a small multiplicative factor above the cost to evaluate the original function. To help provide some intuition with AD, we will use simple linear regression to provide an example of the fundamental AD process.

For $i \in \{1, \ldots, n\}$, assume a simple linear regression model:

$$y_i = a + b \times x_i + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2)$$

where the common least squares estimators for a and b are determined by minimizing the objective function:

$$RSS = \sum_{i=1}^{n} \Big(y_i - (a + b \times x_i) \Big)^2. \tag{A.1.1}$$

Unary Evaluation	Unary Operation	Symbolic Value	Partials
1	$u_1 = b \times x_i$	$b \times x_i$	$\frac{\partial u_1}{\partial b} = x_i$ $\frac{\partial u_1}{\partial x_i} = b$
2	$u_2 = a + u_1$	$a + b \times x_i$	$\frac{\partial u_2}{\partial a} = 1$ $\frac{\partial u_2}{\partial u_1} = 1$
3	$u_3 = y_i - u_2$	$y_i - (a + b \times x_i)$	$\frac{\partial u_3}{\partial y_i} = 1$ $\frac{\partial u_3}{\partial u_2} = -1$
4	$u_4 = u_3^2$	$\Big(y_i - (a + b \times x_i)\Big)^2$	$\frac{\partial u_4}{\partial u_3} = 2u_3$
5	$RSS_i = u_4$	$\Big(y_i - (a + b \times x_i)\Big)^2$	$\frac{\partial RSS_i}{\partial u_4} = 1$

Table A.1.1: Unary operations taken to evaluate the objective function defined in (A.1.1) as well as the numeric evaluation of the first partial derivatives shown in (A.1.2).

As a minimization problem, the solution may be determined by symbolically differentiating the function in (A.1.1) with respect to both a and b, setting both equations to zero and simultaneously solving for the two unknowns:

$$\begin{aligned}\frac{\partial RSS}{\partial a} &= \sum_{i=1}^{n} -2(y_i - (a + b \times x_i)) = 0 \\ \frac{\partial RSS}{\partial b} &= \sum_{i=1}^{n} -2x_i \times (y_i - (a + b \times x_i)) = 0. \end{aligned} \tag{A.1.2}$$

For a computer to compute these derivatives through AD, it would generate a list of all the unary operations needed to evaluate (A.1.1), as well as the derivatives for each unary operation, which can then be combined using the chain rule to arrive at the forms of the partial derivatives shown in (A.1.2). The elementary steps taken to do this are shown in Table A.1.1 and (A.1.3).

Once the function has been broken into its elementary operations and the partials have been derived for each of the basic elementary operations, the partials of the complete function

can be quickly evaluated through the chain rule:

$$\begin{aligned}
\frac{\partial RSS}{\partial a} &= \sum_{i=1}^{n} \frac{\partial RSS_i}{\partial a} = \sum_{i=1}^{n} \frac{\partial RSS_i}{\partial u_4}\frac{\partial u_4}{\partial u_3}\frac{\partial u_3}{\partial u_2}\frac{\partial u_2}{\partial a} \\
&= \sum_{i=1}^{n} 1 \times 2u_3 \times -1 \times 1 = \sum_{i=1}^{n} -2\Big(y_i - (a + b \times x_i)\Big) \\
\frac{\partial RSS}{\partial b} &= \sum_{i=1}^{n} \frac{\partial RSS_i}{\partial b} = \sum_{i=1}^{n} \frac{\partial RSS_i}{\partial u_4}\frac{\partial u_4}{\partial u_3}\frac{\partial u_3}{\partial u_2}\frac{\partial u_2}{\partial u_1}\frac{\partial u_1}{\partial b} \\
&= \sum_{i=1}^{n} 1 \times 2u_3 \times -1 \times 1 \times x_i = \sum_{i=1}^{n} -2x_i\Big(y_i - (a + b \times x_i)\Big).
\end{aligned} \tag{A.1.3}$$

After this formulation of the derivative has been built, it can be quickly evaluated for different values of the parameters, a, and b, conditional on the data, $\boldsymbol{y}$ and $\boldsymbol{x}$. Of course, higher order derivatives may then be calculated by iteratively applying AD.

When `TMB` is invoked to compile an objective function, often a likelihood or posterior distribution, defined in a `TMB C++` template, it also automatically returns executable functions to evaluate the gradient and the Hessian, both generated through AD. The gradient can be used in an optimization routine to more efficiently find the minimum of the objective function, or it could be used to sample directly from the posterior in, e.g., a Hamiltonion MCMC. The Hessian, as in (3.12), is needed in `TMB`'s implementation of the Laplace approximation shown in (3.11). Using AD, both the gradient and the Hessian can be quickly calculated and evaluated without the need for human-derived symbolic differentiation and explicit coding - which can be tedious and error prone.

See Fournier *et al.* (2012) and Kristensen *et al.* (2016) for further details about the AD methods used in `TMB` and Griewank and Walther (2008) for more general AD theory.

A.2 Continuous Spatial Simulation Study

This section provides additional information relevant to the continuous simulation discussed in Section 3.6.1, as well as an additional selection of results.

A.2.1 SPDE Details

Spatial models can be notoriously difficult to fit at scale and in this study we use the SPDE finite element method representation to approximate the GPs. Lindgren *et al.* (2011) prove that specific discretely indexed GMRF models defined on triangulations can approximate continuous spatial GPs with the Matérn covariance shown in (2.12). This relationship relies on the fact that solutions to a specific SPDE class have the Matérn covariance and that correctly chosen GMRF models are approximate solutions to the SPDE. Working with the GMRF approximations allows for fast and efficient sparse matrix operations permitted on GMRFs to applied to GP models. The specific SPDE of interest takes the form:

$$(\kappa^2 - \Delta)^{\alpha/2}(\tau x(\boldsymbol{s})) = \mathcal{W}(\boldsymbol{s}), \tag{A.2.1}$$

where Δ is the Laplacian, α controls the smoothness, $\kappa > 0$ the scale, τ the variance, and $\mathcal{W}(\boldsymbol{s})$ is a Gaussian spatial white noise process. Whittle (1954) showed that the exact stationary solution to this SPDE is a stationary Gaussian field with Matérn covariance, $x(\boldsymbol{s})$. There is a well-defined relationship between the parameters in (A.2.1) and (2.12) (Blangiardo and Cameletti, 2015, Section 6.5). The GMRF approximation uses a finite element method basis function representation defined on a triangulation over the domain:

$$x(\boldsymbol{s}) \approx \sum_{i=1}^{V} \psi_i(\boldsymbol{s}) w_i$$

with compact deterministic basis functions $\psi_i(\boldsymbol{s})$ and associated weights w_i summed over the V vertices in the triangulation. For specific basis functions, the vector of weights, $\boldsymbol{w}$,

are Gaussian with mean zero and sparse precision matrix that depends on the parameters in (A.2.1). See Lindgren *et al.* (2011) for details.

One of the key aspects of this approximation, as demonstrated by Righetto *et al.* (2020), is the choice of triangulation. Unlike their work, we have chosen to generate the triangulations to be constant density over the domain without using any of the data locations. The reason for this choice was two-fold. First, it allowed us to fix the mesh across the experiments and replications, removing one source of known variability. This allows us to more readily study the effect of the mesh density which has been shown to be a driver of oversmoothing (Teng *et al.*, 2017). Secondly, the discretization error varies by size of the triangles, and triangulations with pronounced variability in resolution may lead to different effects of the spatial field parameters across the domain.

In the continuous spatial simulation we used three different mesh resolutions characterized by the largest allowed triangle edge length: 0.15, 0.2, and 0.3 degrees which corresponded to 3616, 7933, and 13866 vertices, respectively. The triangulations were generated using the `inla.mesh.2d()` function from the `R-INLA` package. The centroid of every 5x5km pixel within Nigeria, the modeling domain, were used supplied `loc.domain` argument to define the extent, and the maximum allowed edge length outside the extent was set to 5 degrees. All other arguments were left as the defaults. Plots of the three triangulation meshes used in the continuous simulations are shown in Figure 3.3.

A.2.2 Additional Continuous Spatial Simulation Results

We provide a few additional plots, extending those shown in Section 3.6.1, contrasting `TMB` against the default `R-INLA` option results. Figure A.2.1 shows the parameter bias from these experiments analogous to Figure 3.4 in the main text. Figures A.2.2 and A.2.3 show collapsed versions of Figures 3.5 and 3.6 where averaging has been performed over the entire spatial field (as opposed to stratifying by the magnitude of the true GP).

Appendix A.2.2 includes parameter bias figures from the Binomial data setting for all combinations of `R-INLA` numerical integration and marginal approximations implemented in this study.

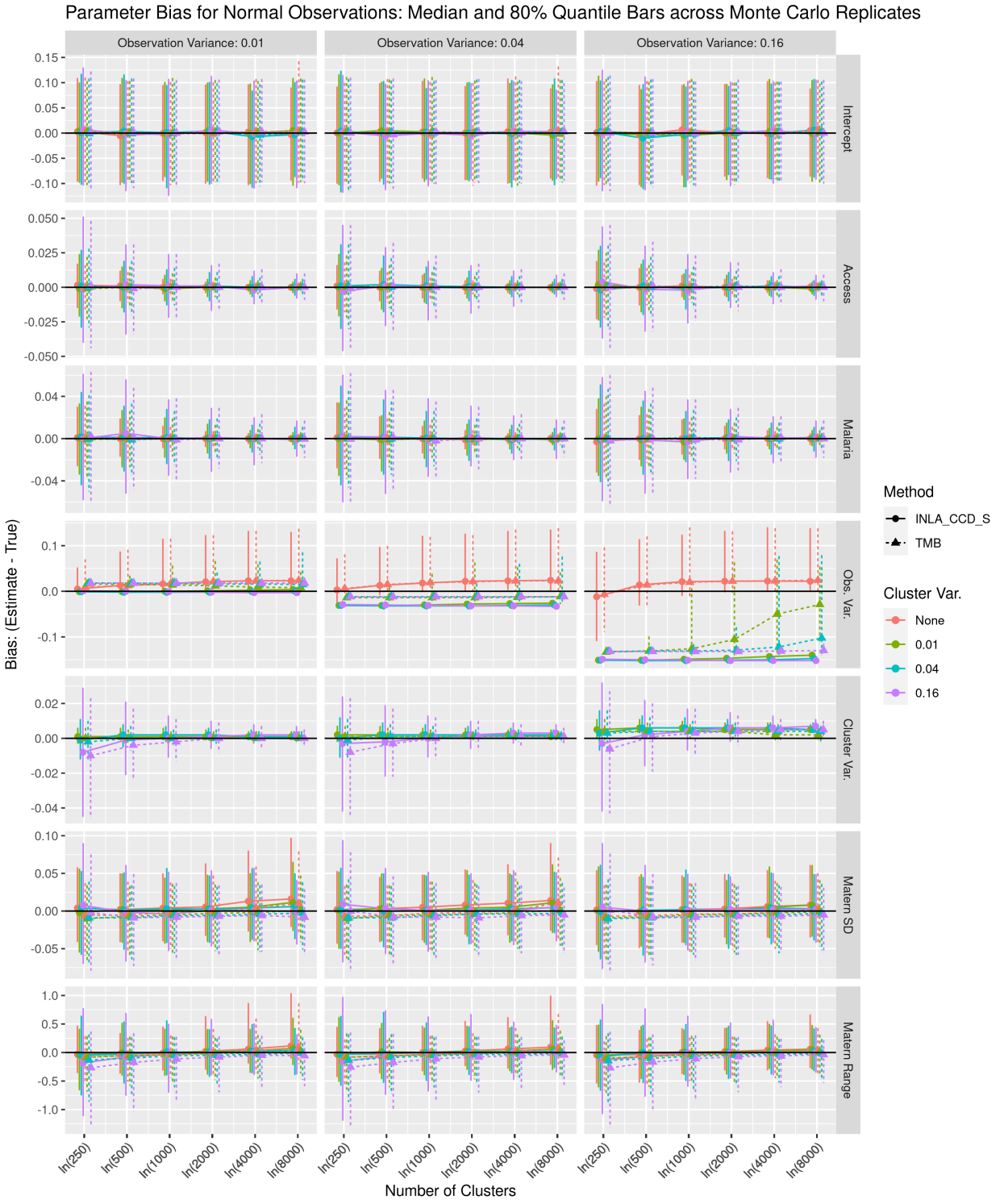


Figure A.2.1: Comparison of the estimated parameter bias from `TMB` (dashed lines) and `R-INLA` using CCD hyperparameter integration and simplified Laplace approximations (solid lines) plotted against the number of cluster observations for the Gaussian data experiments with varying observation variances. Colors represent different cluster (i.i.d nugget) variances used in an experiment. Each point is the median bias of 3 experiments (coarse, medium, and fine SPDE triangulation), calculated across 75 replicates, and the bars represent the middle 80% quantile range of the bias across replicates.

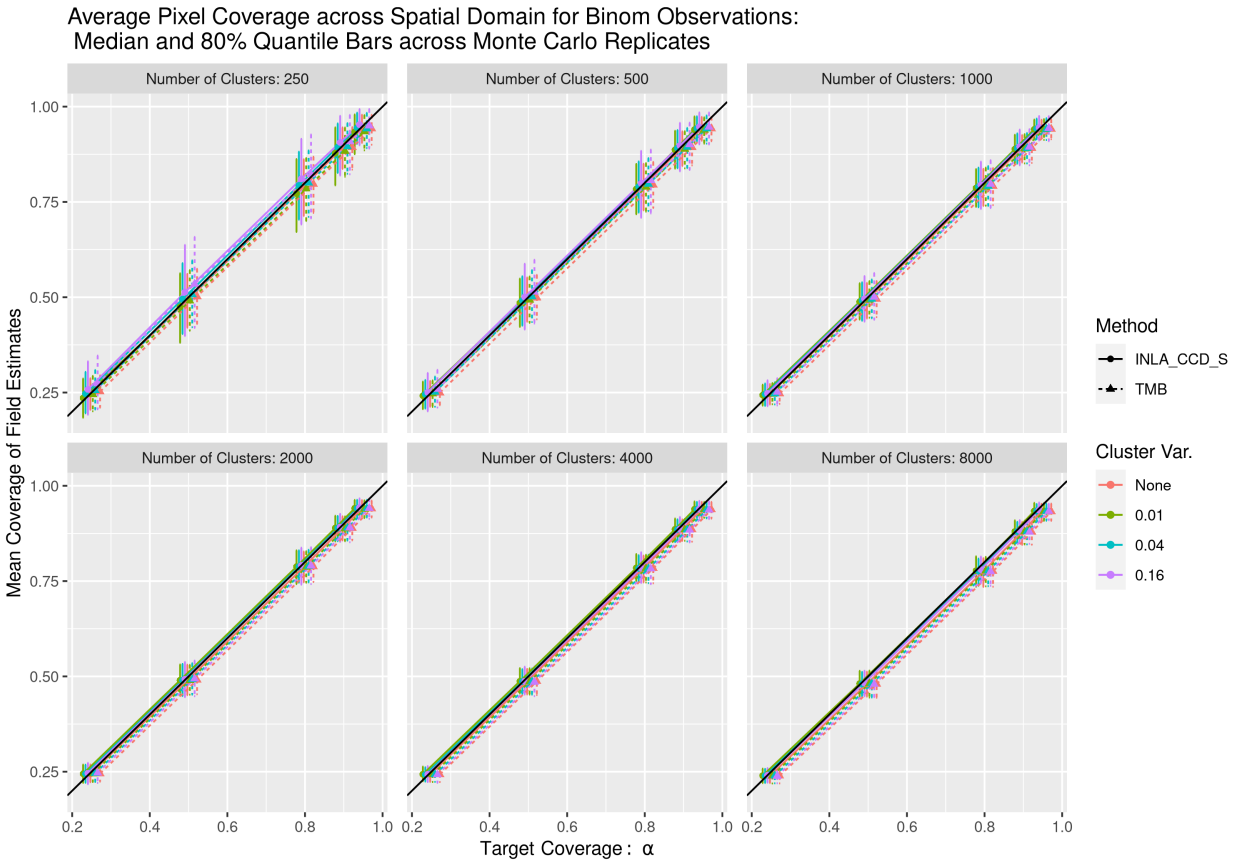


Figure A.2.2: Comparison of the average estimated field coverage of the simulated truth from `TMB` (dashed lines) and `R-INLA` using CCD hyperparameter integration and simplified Laplace approximations (solid lines) plotted against the target nominal coverage, α, for Binomial observation experiments. Colors represent different cluster (i.i.d nugget) variances used in an experiment. Each point is the median average coverage of 3 experiments (coarse, medium, and fine SPDE triangulation), calculated across 75 replicates, and the bars represent the middle 80% quantile range of the average coverage across replicates.

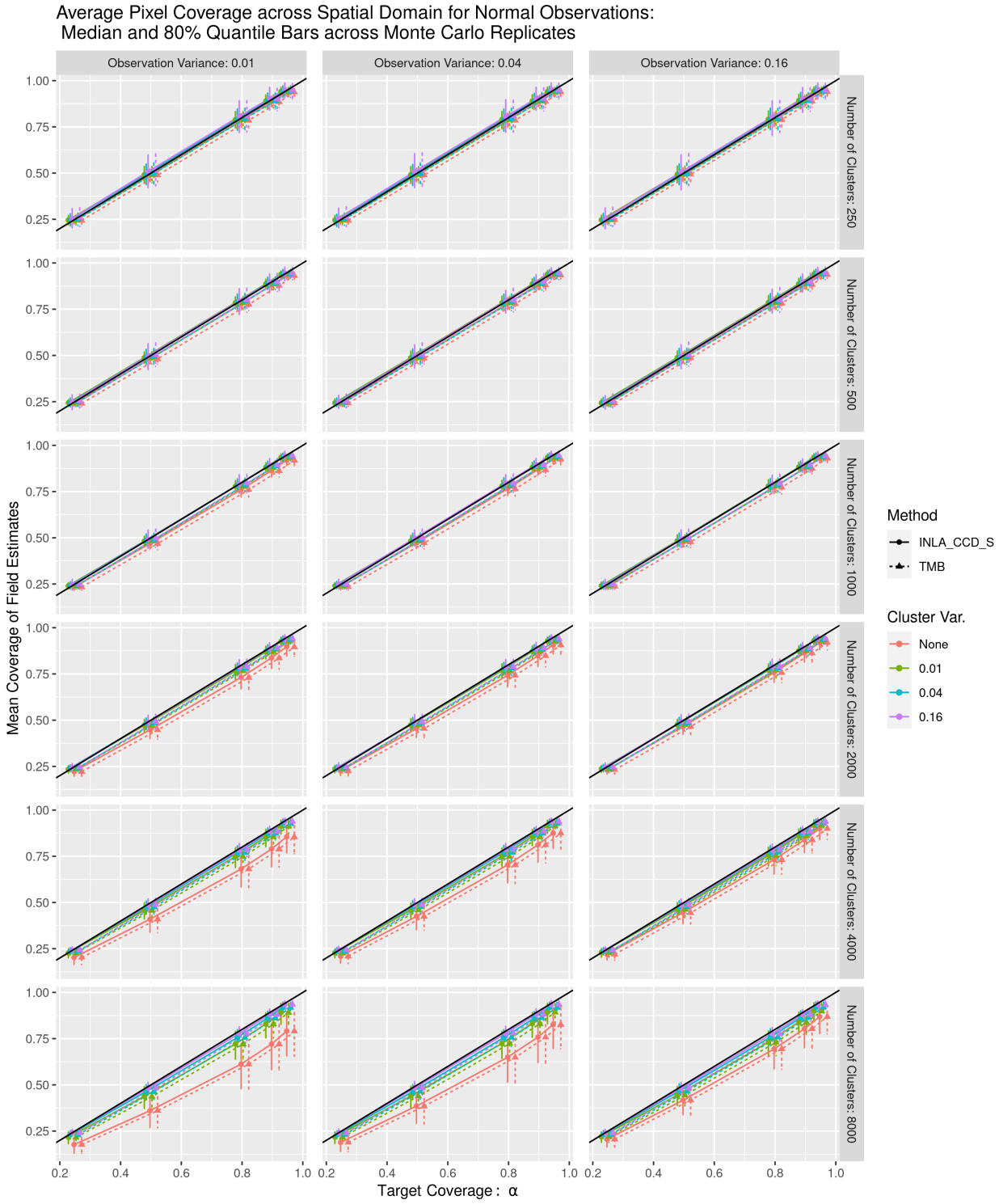


Figure A.2.3: Comparison of the average estimated field coverage of the simulated truth from `TMB` (dashed lines) and `R-INLA` using CCD hyperparameter integration and simplified Laplace approximations (solid lines) plotted against the target nominal coverage, α, for Gaussian observation experiments with varying observation variances. Colors represent different cluster (i.i.d nugget) variances used in an experiment. Each point is the median average coverage of 3 experiments (coarse, medium, and fine SPDE triangulation), calculated across 75 replicates, and the bars represent the middle 80% quantile range of the average coverage across replicates.

Alternate R-INLA approximation options

All figures shown in this section compare bias results from Binomial data experiments. Bias from `TMB` (dashed lines) and `R-INLA` results (solid lines), under a variety of approximation options, are plotted against the number of cluster observations for Binomial observation experiments. Colors represent different cluster (i.i.d nugget) variances used in an experiment. Each point is the median bias of 3 experiments (coarse, medium, and fine SPDE triangulation), calculated across 75 replicates, and the bars represent the middle 80% quantile range of the bias across replicates.

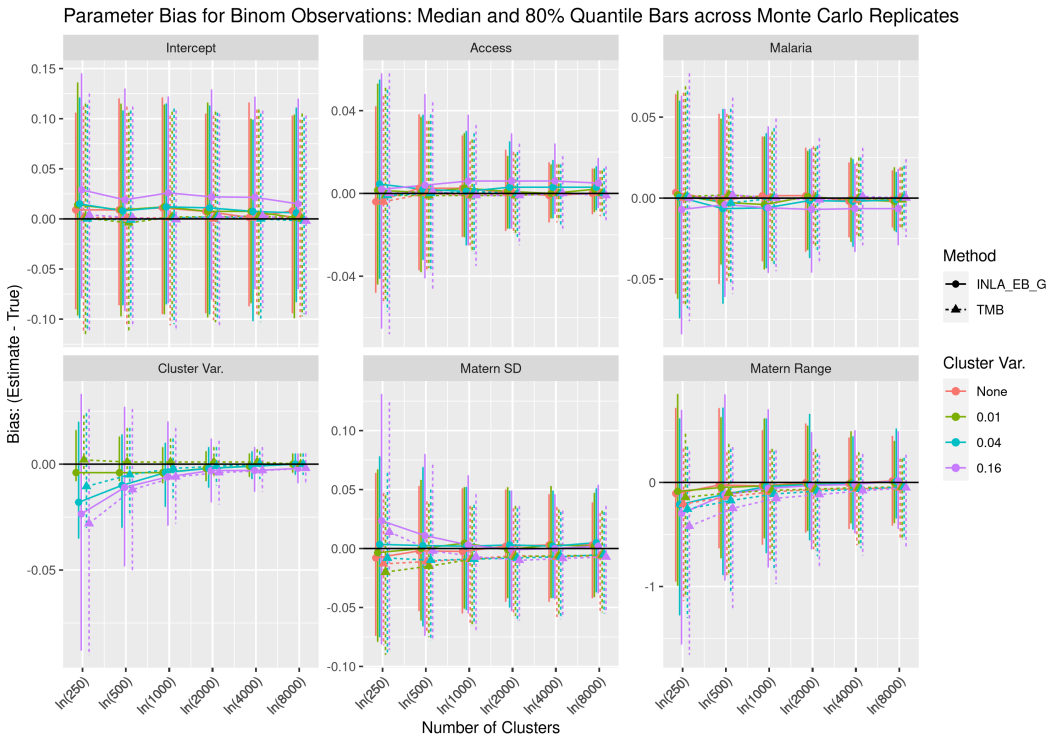


Figure A.2.4: Comparison of the estimated parameter bias from `TMB` (dashed) and `R-INLA` using EB 'integration' and Gaussian approximations (solid lines).

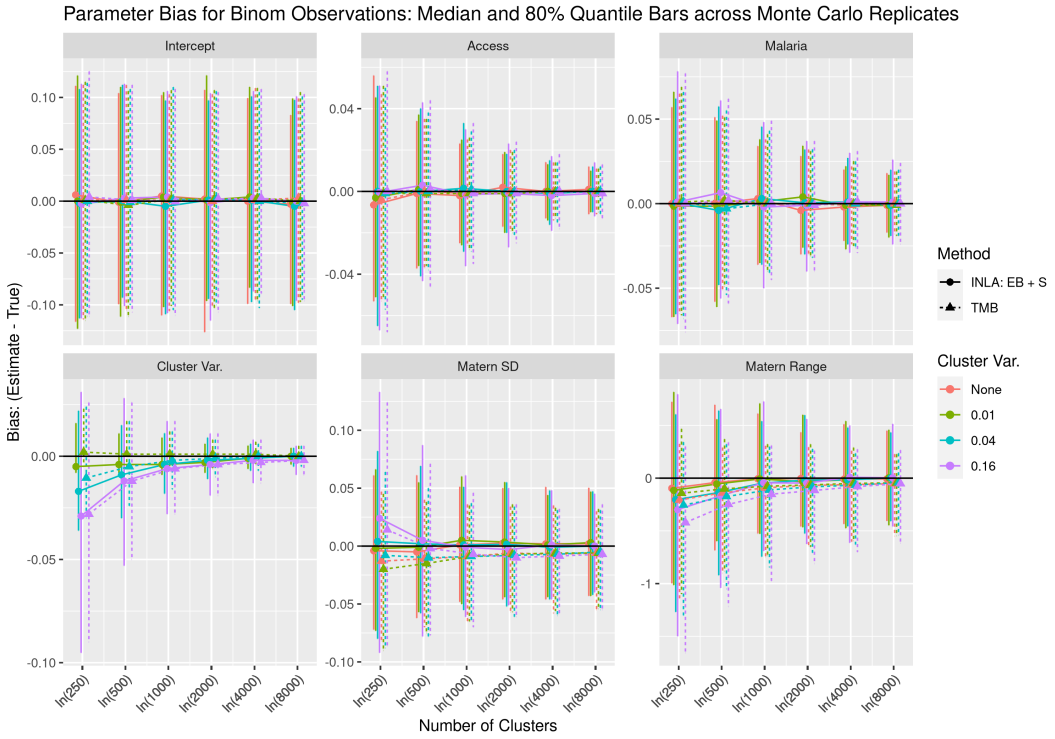


Figure A.2.5: Comparison of the estimated parameter bias from `TMB` (dashed lines) and `R-INLA` using EB 'integration' and simplified Laplace approximations (solid).

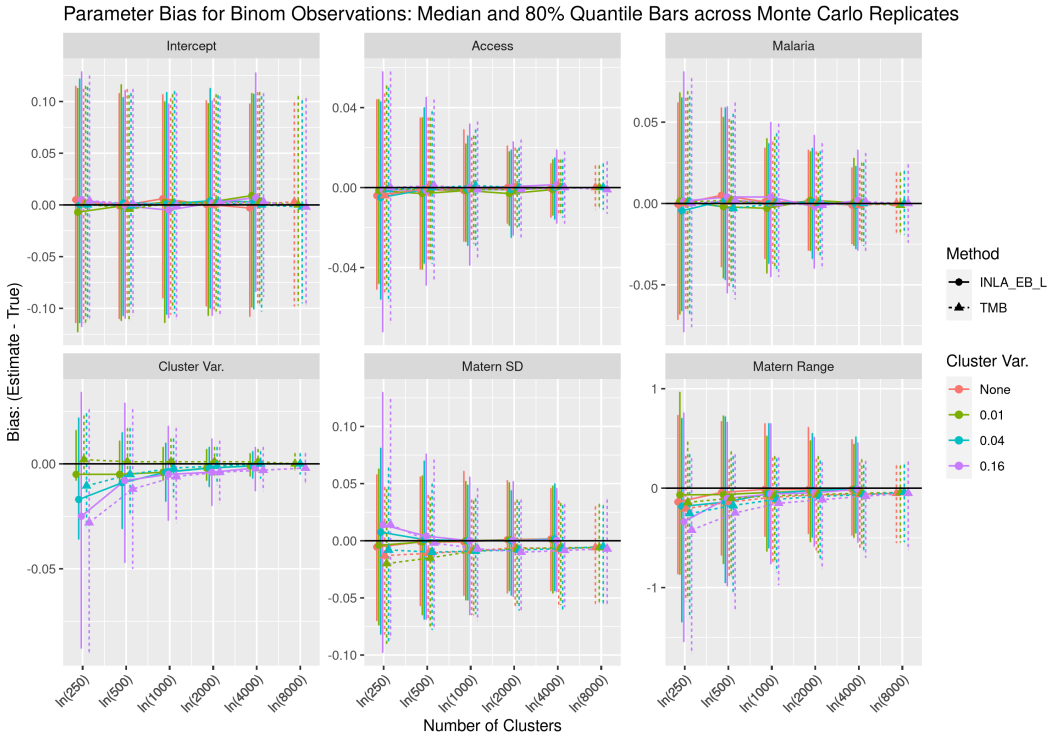


Figure A.2.6: Comparison of the estimated parameter bias from `TMB` (dashed lines) and `R-INLA` using EB 'integration' and full Laplace approximations (solid).

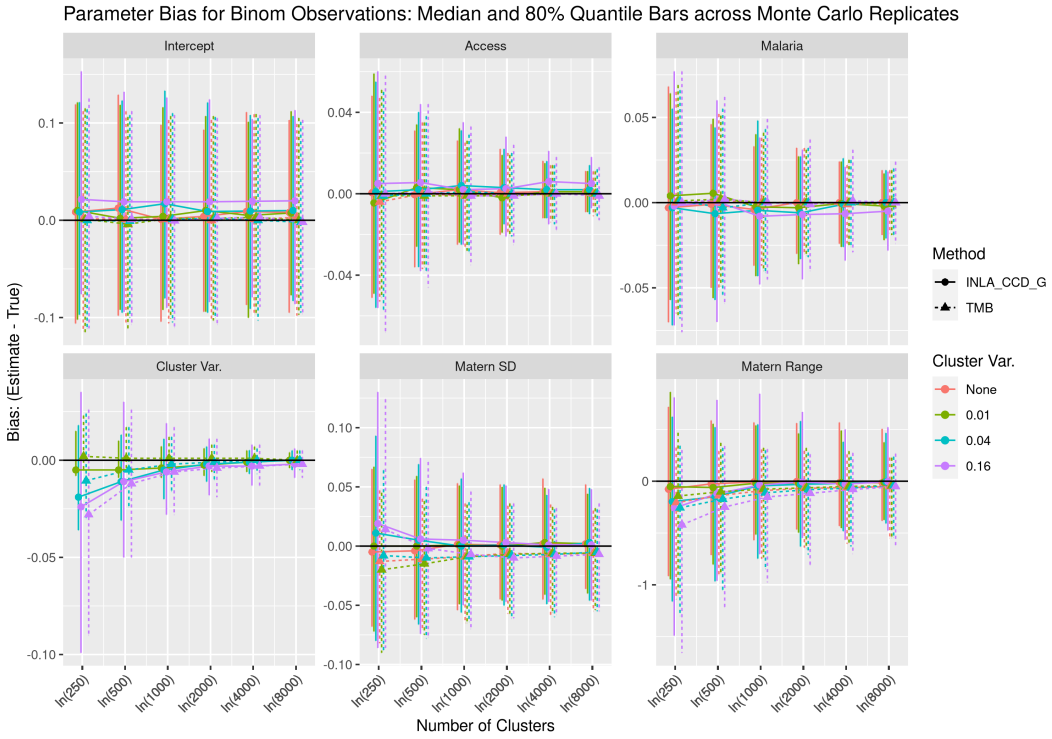


Figure A.2.7: Comparison of the estimated parameter bias from `TMB` (dashed lines) and `R-INLA` using CCD integration and Gaussian approximations (solid).

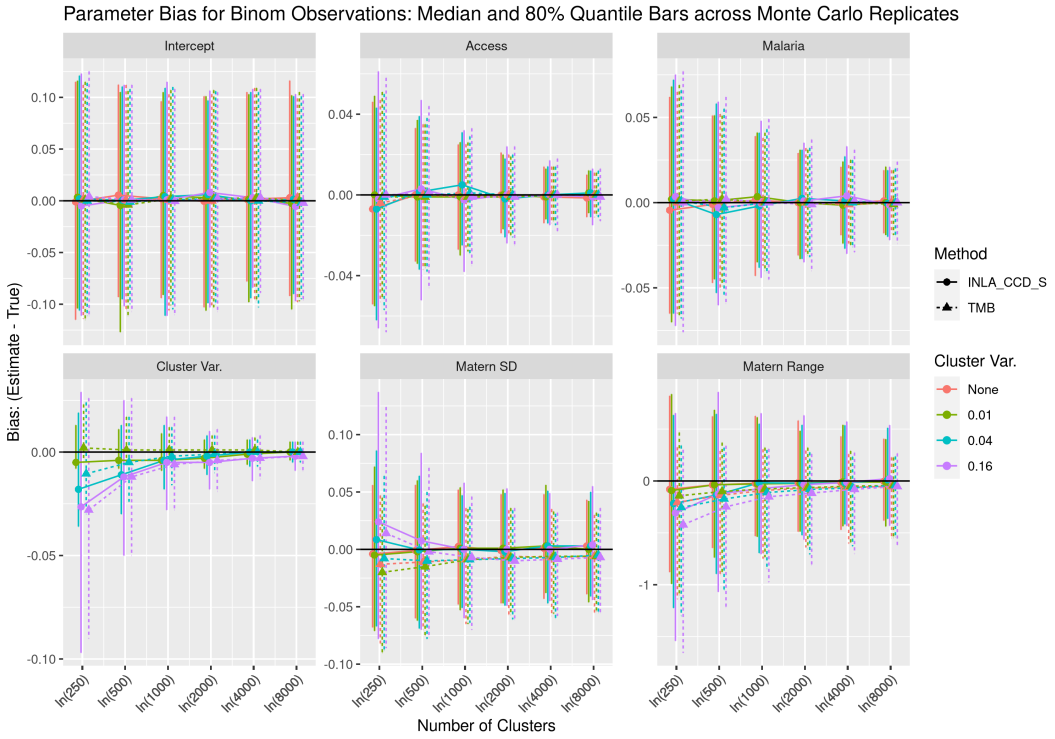


Figure A.2.8: Comparison of the estimated parameter bias from `TMB` (dashed lines) and `R-INLA` using CCD integration and simplified Laplace approx. (solid). Same as Figure 3.4.

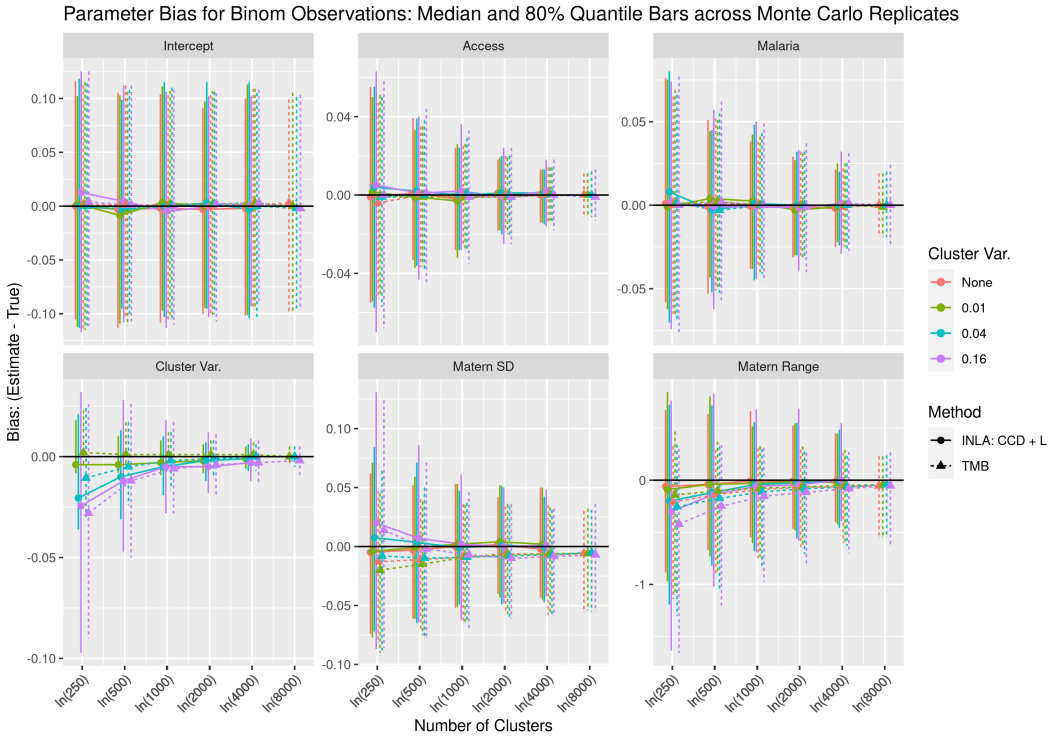


Figure A.2.9: Comparison of the estimated parameter bias from `TMB` (dashed lines) and `R-INLA` using CCD integration and full Laplace approximations (solid).

Differences due to SPDE triangulation resolution

Plots in this section show decreasing undercoverage of the spatial field of the true spatial field by the estimated field as the resolution of the triangulation mesh size increases (as the mesh becomes more dense with more vertices). The figures contrast results from `TMB` against those from `R-INLA` using the CCD integration and full Laplace approximations as the mesh density increases. These are the results from the best `R-INLA` approximations we evaluated, even though this pattern persists across all other `R-INLA` options tested, to demonstrate that this pattern appears to be a function of the SPDE approximation and not the `R-INLA` options or `TMB` algorithm.

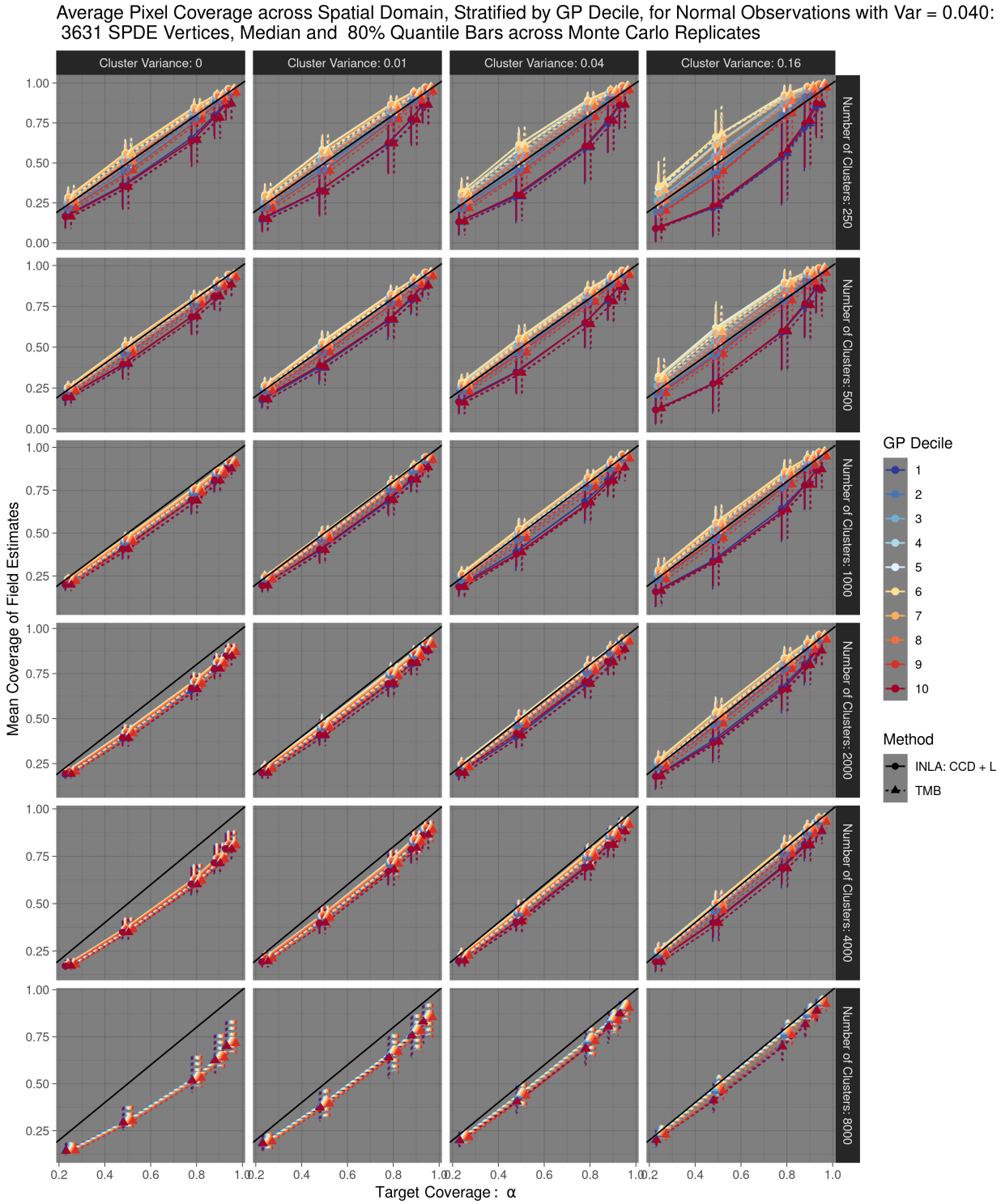


Figure A.2.10: Comparison of the average estimated field coverage of the simulated truth, faceted by cluster (i.i.d. nugget) variance and the number of clusters, from `TMB` (dashed lines) and `R-INLA` using CCD hyperparameter integration and full Laplace approximations (solid lines) plotted against the target nominal coverage, α, for Gaussian observation experiments with $\sigma^2 = 0.04$ and the coarse resolution SPDE triangulation. Colors stratify pixels included in the average coverage calculation by the decile of the true GP for the experiment replicate. Each point is the median average coverage of an experiment, calculated across 25 replicates, and the bars represent the middle 80% quantile range of the average coverage across replicates.

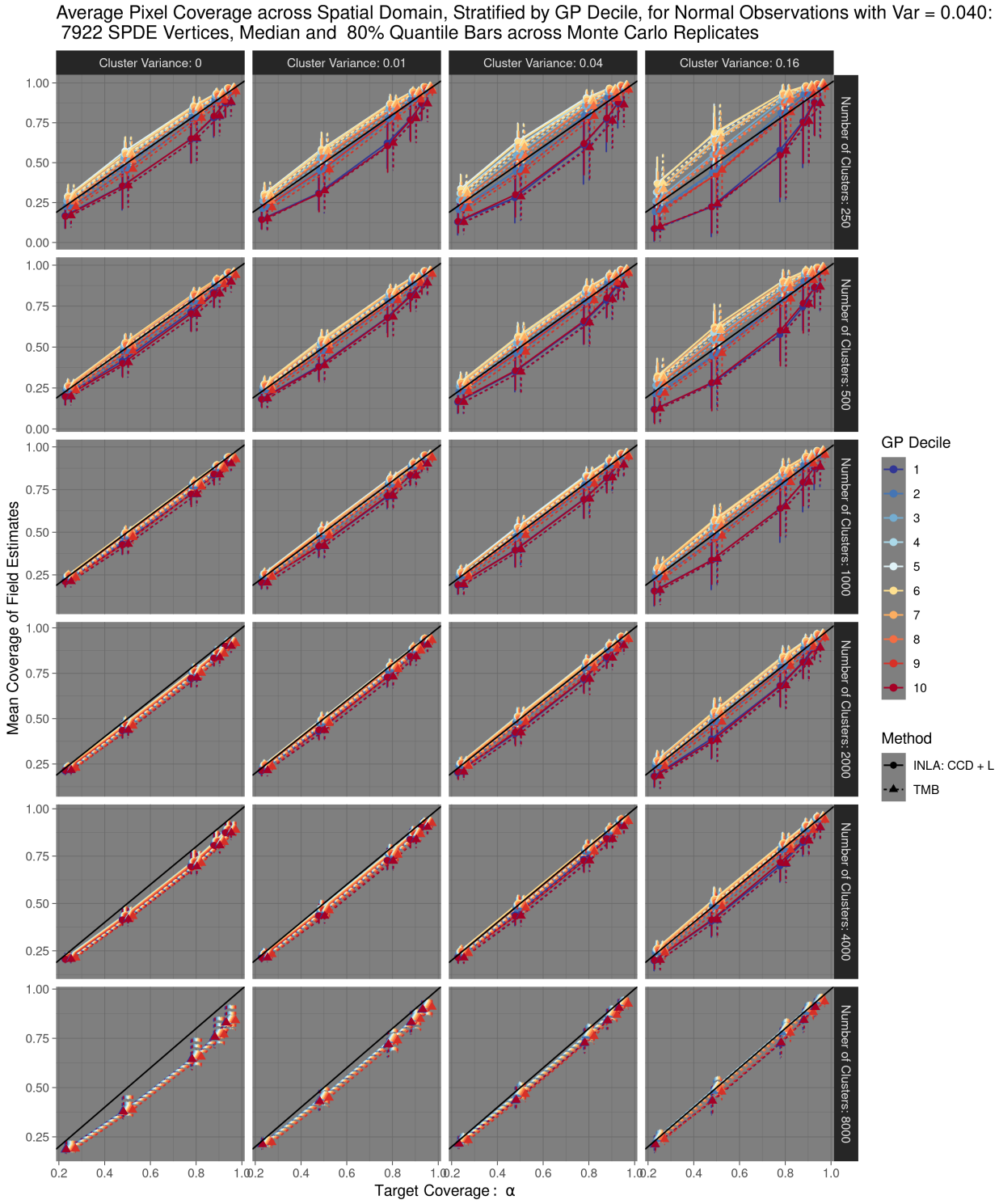


Figure A.2.11: Comparison of the average estimated field coverage of the simulated truth, faceted by cluster (i.i.d. nugget) variance and the number of clusters, from `TMB` (dashed lines) and `R-INLA` using CCD hyperparameter integration and full Laplace approximations (solid lines) plotted against the target nominal coverage, α, for Gaussian observation experiments with $\sigma^2 = 0.04$ and the medium resolution SPDE triangulation. Colors stratify pixels included in the average coverage calculation by the decile of the true GP for the experiment replicate. Each point is the median average coverage of an experiment, calculated across 25 replicates, and the bars represent the middle 80% quantile range of the average coverage across replicates. This figure is shown in the main results section, but is replicated here for easy comparison against the other appendix plots.

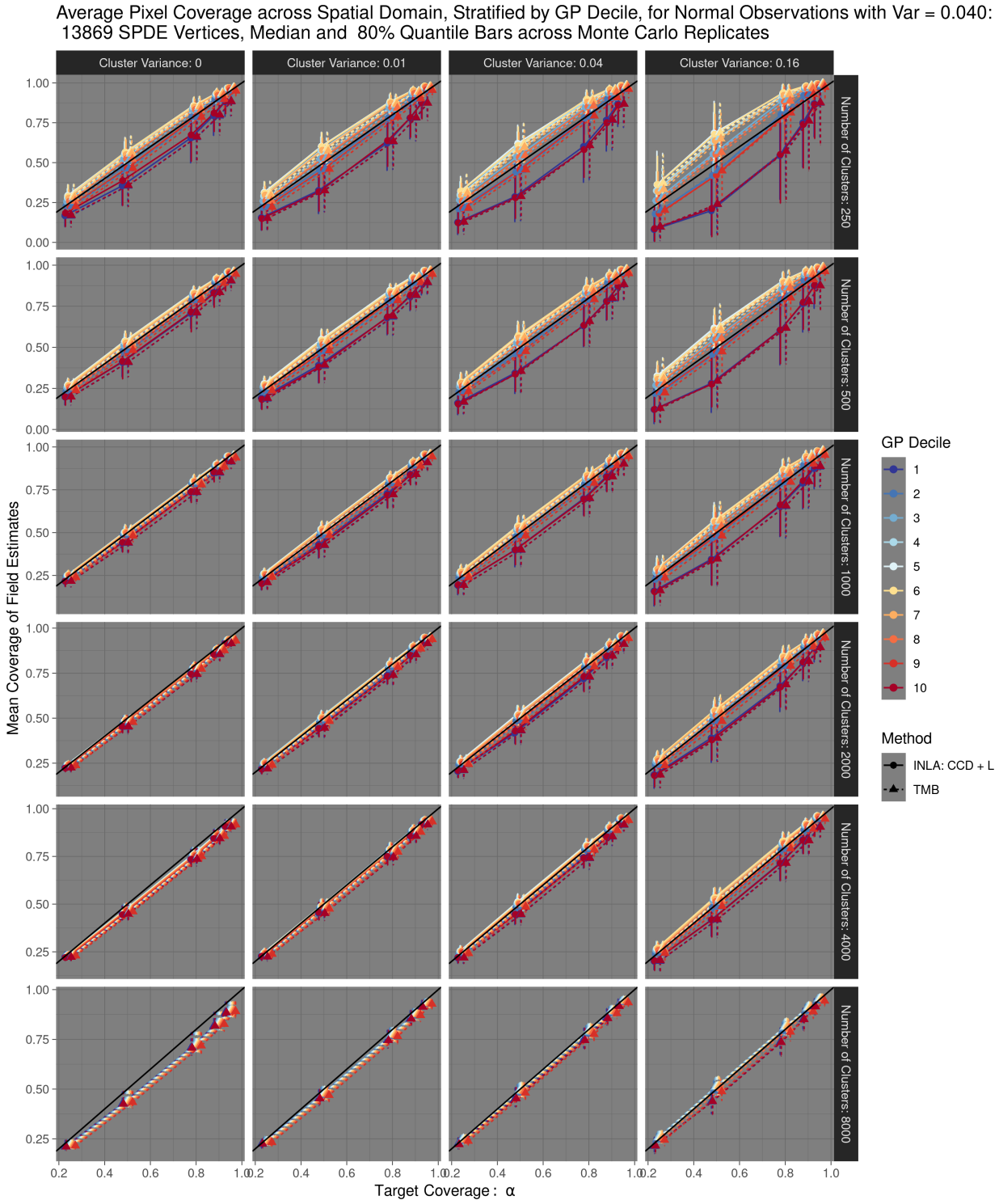


Figure A.2.12: Comparison of the average estimated field coverage of the simulated truth, faceted by cluster (i.i.d. nugget) variance and the number of clusters, from `TMB` (dashed lines) and `R-INLA` using CCD hyperparameter integration and full Laplace approximations (solid lines) plotted against the target nominal coverage, α, for Gaussian observation experiments with $\sigma^2 = 0.04$ and the fine resolution SPDE triangulation. Colors stratify pixels included in the average coverage calculation by the decile of the true GP for the experiment replicate. Each point is the median average coverage of an experiment, calculated across 25 replicates, and the bars represent the middle 80% quantile range of the average coverage across replicates.

A.3 European Breast Cancer Application Details

A.3.1 Data

We work with breast cancer incidence and mortality data from IARC for this report. The current implementation of the methods described in this paper 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.

Figure A.3.1 shows the data type available for the 40 countries from 1990-2010.

Although the data is available across age groups and time, this current project will focus on ages 50-54 (age group 11). While we will use data from all time periods, allowing the country type to vary in time, we assume that the underlying parameters are fixed in time and only estimate spatial variability.

A.3.2 Model

We assume a 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.

First, notation is defined. For a country, c, an age group, a, and a time period (year), t, we use L to denote local registry data and R to denote the remainder of the data. In

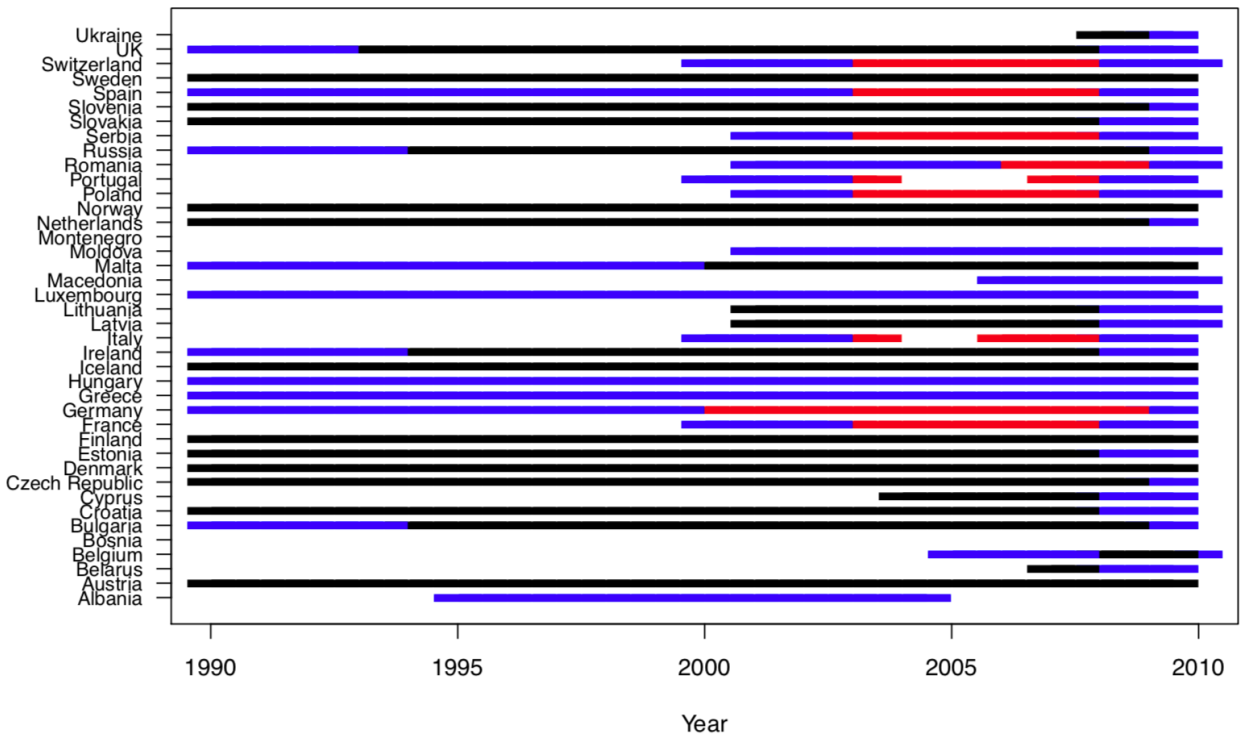


Figure A.3.1: Data type by country from 1990–2010. Type IV (Montenegro) is blank.

countries with no local registry data (L), all the data will fall into the remainder category (R). With terms, we define:

- N_{act}^{L} = Population for age group a in country c at time t covered by the available registry,
- Y_{act}^{L} = Total incident reported cases from all registries for age group a, country c, and time t,
- Z_{act}^{L} = Total reported deaths (mortality) from all registries for age group a in country c at time t,
- N_{act}^{R} = Population for age group a in country c at time t not covered by the available registry,

- Y^R_{act} = Total incident reported cases not covered by registries for age group a in country c at time t,
- Z^R_{act} = Total reported deaths (mortality) not covered by registries for age group a in country c at time t,
- $N_{act} = N^L_{act} + N^R_{act}$ is the total population for age group a in country c at time t,
- $Y_{act} = Y^L_{act} + Y^R_{act}$ is all reported cases for age group a in country c at time t,
- $Z_{act} = Z^L_{act} + Z^R_{act}$ is all reported deaths for age group a in country c at time t,
- $p_{act} = P($ Reported incidence $|a, c, t)$,
- $r_{act} = P($ Reported mortality $|a, c, t)$,
- $q_{act} = P($ Reported mortality $|$ Reported incidence $, a, c, t)$.

For countries that have both national mortality and incidence data (type I) we can assume a Poisson process for cancer incidence, and then conditional on having cancer, we model mortality as a binomial outcome. This also induces a Poisson process for mortality when incidence is unobserved. We suppress age and time notation. Our base model for type I countries is:

$$Y_c|N_c, p_c \sim \text{Poisson}(N_c p_c), \quad p_c = \exp(\alpha^I_c) \tag{A.3.1}$$

$$Z_c|Y_c, r_c \sim \text{Binomial}(Y_c r_c), \quad r_c = \frac{\exp(\alpha^{MI}_c)}{1 + \exp(\alpha^{MI}_c)} \tag{A.3.2}$$

which implies the unconditional mortality model:

$$Z_c|N_c, p_c \sim \text{Poisson}(N_c q_c), \quad q_c = p_c \cdot r_c. \tag{A.3.3}$$

We assume a log- and logit-linear model for incidence and conditional mortality and we assume the following forms:

$$\alpha_c^I = \alpha^I + b_c^I \tag{A.3.4}$$

$$\alpha_c^{MI} = \alpha^{MI} + b_c^{MI} \tag{A.3.5}$$

where α^* are global intercepts, and b_c^* are country random effects that are assumed to have BYM-2 structure comprising of a spatially correlated term as well as an unstructured (iid) country specific term. Specifically, each vector of the country random effects are assumed to independent from one another take the form of the BYM2 effects defined in (3.20) and (3.21).

The α^I, α^{MI}, $\boldsymbol{b}^I$, and $\boldsymbol{b}^{MI}$ parameters are used for across all the country types, but the way that these parameters learn and leverage the information depend on the country data type and thus the way the data enter into the joint likelihood.

For type II countries, those with local incidence and mortality data and national mortality data, we assume the same model used for type I countries for the local registry data and the implied mortality Poisson process for the remaining national mortality data.

That is, for the local data we assume:

$$Y_c^L | N_c^L, p_c \sim \text{Poisson}(N_c^L p_c), \quad p_c = \exp(\alpha_c^I) \tag{A.3.6}$$

$$Z_c^L | Y_c^L, r_c \sim \text{Binomial}(Y_c r_c), \quad r_c = \frac{\exp(\alpha_c^{MI})}{1 + \exp(\alpha_c^{MI})} \tag{A.3.7}$$

where the intercept parameters are of the form shown in 3.25 and 3.26.

Furthermore, for the remaining non-registry mortality data, we assume that the MI ratio is the same in the local registry and national remainder data and we model it as the implied (unconditional) Poisson process:

$$Z_c^R | N_c^R, q_c \sim \text{Poisson}(N_c^R q_c), \quad q_c = \exp(\alpha_c^I) \cdot \frac{\exp(\alpha_c^{MI})}{1 + \exp(\alpha_c^{MI})}. \tag{A.3.8}$$

For type III countries, those with only national level mortality, we use the induced unconditional Poisson model as written in (3.24).

Finally, for type IV countries, those with no data, we rely on the global intercept and the country random effect (both the iid and the spatially correlated and smoothed random components from the BYM-2) from the posterior distribution to estimate their incidence and mortality rates.

To complete the Bayesian specification, we assign the following prior distributions:

- $aI,\ aMI \sim (iid)\ N(0, \sigma^2 = 100)$
- $\varphi_* \sim (iid)$ Beta$(.5, .5)$
- $1/\sqrt{\tau_*} = \sigma_* \sim (iid)\ \mathrm{N}(0, 5^2)\mathbb{1}_{\sigma>0}$.

The model is fit in R using *Template Model Builder* and the nonlinear optimizer, *nlminb*.

A.3.3 Simulation

To assess the feasibility of this model, we first consider a small simulation restricting ourselves to a single year and age group. We set $\alpha^I = -6.5$, $\alpha^{MI} = -1.0$ and the standard deviation of the spatial random effects to be 0.5 (with no iid country effect - effectively setting the mixing term for the BYM-2 to be 1.0). We use the form of the data (country types and populations) from 2008 and age-group 50-54. This resulted in 14, 2, 21, and 3 type I, II, III, and IV countries respectively. Conditional on these true parameters, country data types, and the observed populations, data was simulated from the model outlined in Section A.3.2. Results for the fits are summarized in Figure A.3.3 and indicate that overall the model is performing well, even in the challenging situation with over half of the countries set to type III or IV. Notably, the precision for the MI country random effects has been estimated to be too high and this has resulted in some over-shrinkage of the MI estimates.

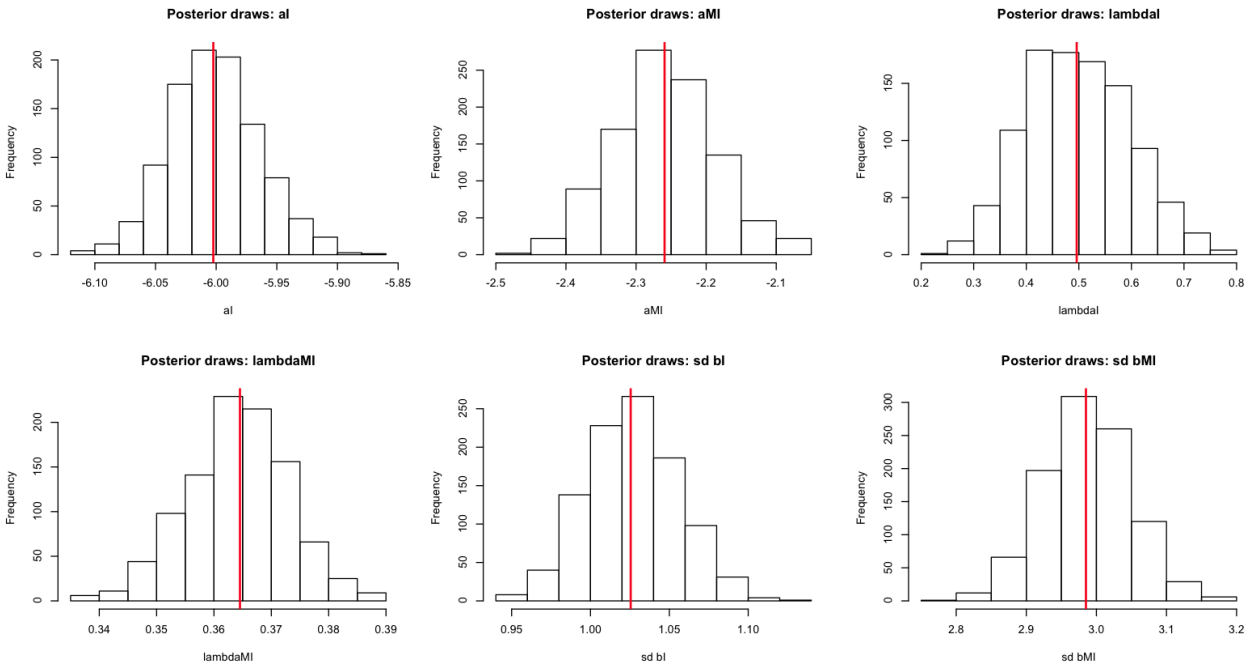


Figure A.3.2: Top row: histograms of the incidence rate intercept aI, mortality-incidence ratio intercept aMI, and the BYM2 incidence mixture parameter λ_I. Top row: histograms of the BYM2 mortality-incidence mixture parameter λ_{MI}, and the standard deviations $(\tau_*^{-1/2})$ of the two BYM2 processes.

A.3.4 Results

The model can be run quite quickly in R and once it has finished fitting, 1000 multivariate normal draws are taken from the joint posterior of all parameters. These draws are then summarized and some relevant quantities are shown in Figures 3.11 and A.3.2.

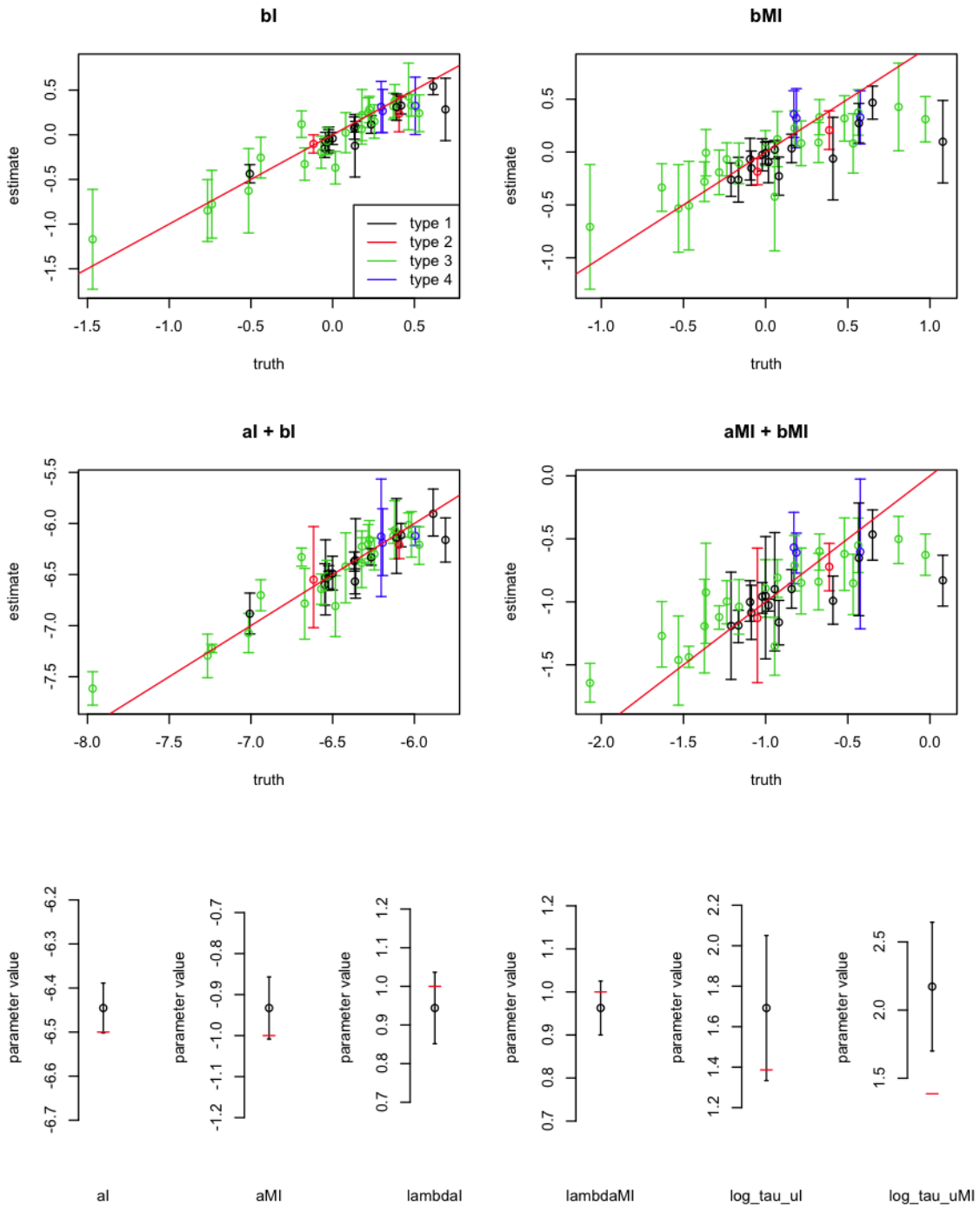


Figure A.3.3: Fitted results from one run of the simulation study. The top row shows the simulated country random effects from the BYM2 specification plotted against the associated fitted median and 95% credible intervals. The second row shows the simulated overall country effect (intercept plus country random effects) plotted against the associated fitted median and 95% credible intervals. In the third row we have the true values for each of the fixed and hyperparameters (shown with the red line) plotted against the associated fitted median and 95% credible intervals.

A.4 Example spatial model code

The complete code used in this study is available online at: https://faculty.washington.edu/jonno/software.html. The following two sections provide succinct yet complete examples of continuous and discrete spatial model inference in both `TMB` and `R-INLA`. We start by simulating data on the unit square, and then use `R-INLA` functions to make the SPDE objects for fitting, some of which are re-used by `TMB`.

A.4.1 Simulating data and generating SPDE objects

```
## install missing packages
pkgs <- c('data.table', 'ggplot2', 'RColorBrewer', 'RandomFields',
          'raster', 'TMB', 'viridis')
new.packages <- pkgs[!(pkgs %in% installed.packages()[,"Package"])]
if(length(new.packages)) install.packages(new.packages)
if(!('INLA' %in% installed.packages()[, 'Package'])){
  ## INLA is not on CRAN
  install.packages("INLA",
                   repos=c(getOption("repos"),
                     INLA="https://inla.r-inla-download.org/R/stable"),
                   dep=TRUE)
}
# load packages
invisible(lapply(c(pkgs, 'INLA'), library, character.only = TRUE))

## setup continuous domain
set.seed(413206)
x <- seq(0, 10, length = 200)
grid.pts <- expand.grid(x, x)

## set up matern params, also set param priors to be used in modeling
sp.alpha <- 2
sp.kappa <- 0.5
sp.var   <- 0.5
gp.int  <- -2
# prior on spde parameters: c(a, b, c, d), where
# P(sp.range < a) = b
# P(sp.sigma > c) = d
```

```
matern.pri <- c(10, .95, 1., .05) ## a, b, c, d
# mean and sd for normal prior on fixed effects (alpha and betas)
alpha.pri <- c(0, 3) ## N(mean, sd)

## sample from matern RF on our grid
model <- RMmatern(nu    = sp.alpha - 1, ## from INLA book
                  scale = sqrt(2 * (sp.alpha - 1)) / sp.kappa,
                  var   = 1)
true.gp <- RFsimulate(model, x = x, y = x, n =1, spConform = FALSE)

## insert into a raster
gp.rast <- raster(nrows=length(x), ncols=length(x),
                  xmn=0, xmx=10, ymn=0, ymx=10,
                  vals=(true.gp + gp.int))

## define cluster locations and sample size at each
n.clust <- 500
clust.mean.ss <- 35
dat <- data.table(x = runif(n.clust, min = min(x), max = max(x)),
                  y = runif(n.clust, min = min(x), max = max(x)),
                  n = rpois(n.clust, clust.mean.ss)
                  )

## extract value of raster at cluster locs and logit transform
##   to binom probs
dat[, latent.truth := raster::extract(x = gp.rast, y = cbind(x, y))]
dat[, p.truth := plogis(latent.truth)]

## sample binomial data
dat[, obs := rbinom(n = .N, size = n, p = p.truth)]

## make SPDE triangulation mesh over our domain
mesh.s <- inla.mesh.2d(loc.domain = grid.pts,
                       max.e = c(0.25, 5))
## check number of vertices
mesh.s[['n']]

## plot true latent field, the observed/empirical binom probs at
##   cluster locs, and the mesh
par(mfrow = c(3, 1))
plot(gp.rast, maxpixels = length(x) ^ 2,
     xlim = range(x), ylim = range(x), main = 'latent truth')
```

```
fields::quilt.plot(dat[, x], dat[, y], dat[, obs] / dat[, n],
                   main = 'empirical binom probs')
plot(mesh.s)
polygon(x = c(0, 0, 10, 10, 0), y = c(0, 10, 10, 0, 0),
col = NA, border = 2, lwd = 5)

## make the SPDE objects (including prec components)
spde <- inla.spde2.pcmatern(mesh = mesh.s, alpha = 2,
                            prior.range = matern.pri[1:2],
                            prior.sigma = matern.pri[3:4])

## make projector matrices to:
## 1) project data to mesh
## 2) project mesh to raster grid
A.proj <- inla.spde.make.A(mesh = mesh.s,
                           loc = dat[, as.matrix(x, y)])
A.pred <- inla.spde.make.A(mesh = mesh.s,
                           loc = as.matrix(grid.pts),
                           group = 1)
```

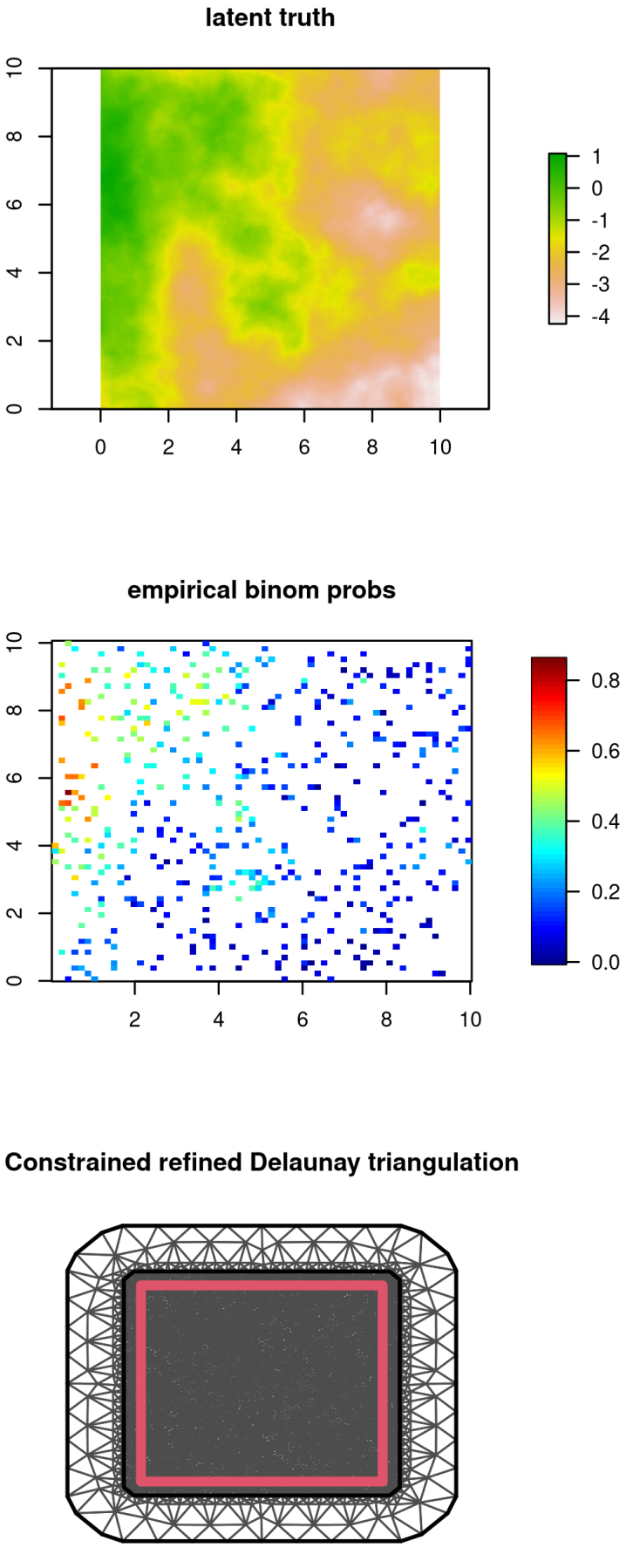


Figure A.4.1: Simulated GP on 10×10 grid, simulated data locations and empirical probabilities, and SPDE mesh from the preparation portion of the code example.

A.4.2 Continuous GP modeling with SPDE in R-INLA

```
## prep inputs for INLA
design_matrix <- data.frame(int = rep(1, nrow(dat)))
stack.obs <- inla.stack(tag='est',
                        data=list(Y = dat$obs, ## response
                                  N = dat$n), ## binom trials
                        A=list(A.proj, ## A.proj for space
                               1),     ## 1 for design.mat
                        effects=list(
                          space = 1:mesh.s[['n']],
                          design_matrix))

## define the INLA model
formula <- formula(Y ~ -1 + int + f(space, model = spde))

## run INLA
i.fit <- inla(formula,
              data = inla.stack.data(stack.obs),
              control.predictor = list(A = inla.stack.A(stack.obs),
                                       compute = FALSE),
              control.fixed = list(expand.factor.strategy = 'inla',
                                   prec = list(default = 1 / alpha.pri[2] ^ 2)),
              control.inla = list(strategy = 'simplified.laplace',
                                  int.strategy = 'ccd'),
              control.compute=list(config = TRUE),
              family = 'binomial',
              Ntrials = N,
              verbose = FALSE,
              keep = FALSE)

## take draws from inla
i.draws <- inla.posterior.sample(n = 500, i.fit,
                                 use.improved.mean = TRUE,
                                 skew.corr = TRUE)

## summarize the draws
par_names <- rownames(i.draws[[1]][['latent']])
s_idx <- grep('^space.*', par_names)
a_idx <- which(!c(1:length(par_names)) %in%
                 grep('^space.*|Predictor|clust.id', par_names))

```

```r
# project from mesh to raster, add intercept
pred_s <- sapply(i.draws, function (x) x[['latent']][s_idx])
pred_inla <- as.matrix(A.pred %*% pred_s)
alpha_inla_draws <- sapply(i.draws, function (x) x[['latent']][a_idx])
pred_inla <- sweep(pred_inla, 2, alpha_inla_draws, '+')


## find the median and sd across draws, as well as 90% intervals
summ_inla <- cbind(median = (apply(pred_inla, 1, median)),
                   sd     = (apply(pred_inla, 1, sd)),
                   lower = (apply(pred_inla, 1, quantile, .05)),
                   upper = (apply(pred_inla, 1, quantile, .95)))

## make summary rasters
ras_med_inla <- ras_sdv_inla <- ras_lower_inla <-
  ras_upper_inla <- ras_inInt_inla <- gp.rast
values(ras_med_inla)   <- summ_inla[, 1]
values(ras_sdv_inla)   <- summ_inla[, 2]
values(ras_lower_inla) <- summ_inla[, 3]
values(ras_upper_inla) <- summ_inla[, 4]
values(ras_inInt_inla) <- 0
ras_inInt_inla[gp.rast < ras_lower_inla | ras_upper_inla < gp.rast] <- 1

## plot truth, pixels falling within/without the 90% interval,
##  post. median, and post sd
# set the range for the truth and median
rast.zrange <- range(c(values(gp.rast), values(ras_med_inla)), na.rm = T)
# plot
par(mfrow = c(2, 2))
plot(gp.rast, main = 'Truth', zlim = rast.zrange, col = (viridis(100)))
points(dat[, .(x, y)])
plot(ras_inInt_inla, main = 'Pixels where 90% CIs did not cover Truth')
points(dat[, .(x, y)])
plot(ras_med_inla, main = 'INLA Posterior Median',
     zlim = rast.zrange, col = (viridis(100)))
points(dat[, .(x, y)])
plot(ras_sdv_inla, main = 'INLA Posterior Standard Deviation')
points(dat[, .(x, y)])
```

A.4.3 Continuous GP modeling with SPDE in TMB

```
## define the TMB model using c++ template code
##  this is usually done in a separate file,
##  but it can be done all from within 1 R script

tmb_spde <-
  "// include libraries
#include <TMB.hpp>
#include <Eigen/Sparse>
#include <vector>
using namespace density;
using Eigen::SparseMatrix;

// helper function for detecting NAs in the data supplied from R
template<class Type>
bool isNA(Type x){
  return R_IsNA(asDouble(x));
}

// helper function to make sparse SPDE precision matrix
// Inputs:
//    logkappa: log(kappa) parameter value
//    logtau: log(tau) parameter value
//    M0, M1, M2: these sparse matrices are output from:
//     R::INLA::inla.spde2.matern()$param.inla$M*
template<class Type>
SparseMatrix<Type> spde_Q(Type logkappa, Type logtau, SparseMatrix<Type> M0,
                          SparseMatrix<Type> M1, SparseMatrix<Type> M2) {
    SparseMatrix<Type> Q;
    Type kappa2 = exp(2. * logkappa);
    Type kappa4 = kappa2*kappa2;
    Q = pow(exp(logtau), 2.)  * (kappa4*M0 + Type(2.0)*kappa2*M1 + M2);
    return Q;
}

// helper function to use the same penalized complexity prior on
//  matern params that is used in INLA

template<class Type>
Type dPCPriSPDE(Type logtau, Type logkappa,
                Type matern_par_a, Type matern_par_b,
```

```
                Type matern_par_c, Type matern_par_d,
                //vector<Type> matern_pri(4),
                int give_log=0)
{

  // matern_pri = c(a, b, c, d): P(range < a) = b; P(sigma > c) = d

  Type penalty; // prior contribution to jnll

  Type d = 2.;  // dimension
  Type lambda1 = -log(matern_par_b) * pow(matern_par_a, d/2.);
  Type lambda2 = -log(matern_par_d) / matern_par_c;
  Type range   = sqrt(8.0) / exp(logkappa);
  Type sigma   = 1.0 / sqrt(4.0 * 3.14159265359 * exp(2.0 * logtau) *
                            exp(2.0 * logkappa));

  penalty = (-d/2. - 1.) * log(range) - lambda1 * pow(range, -d/2.) -
              lambda2 * sigma;
  // Note: (rho, sigma) --> (x=log kappa, y=log tau) -->
  //  transforms: rho = sqrt(8)/e^x & sigma = 1/(sqrt(4pi)*e^x*e^y)
  //  --> Jacobian: |J| propto e^(-y -2x)
  Type jacobian = - logtau - 2.0*logkappa;
  penalty += jacobian;

  if(give_log)return penalty; else return exp(penalty);
}

/////////////////////////////
// the main function     //
// to calculate the jnll //
/////////////////////////////
template<class Type>
Type objective_function<Type>::operator() ()
{

  // ~~~~~~~~~------------------------------------------------------~~
  // FIRST, we define params/values/data that will be passed in from R
  // ~~~~~~~~~~~~------------------------------------------------------

  // normalization flag - used for speed-up
  DATA_INTEGER( flag ); // flag == 0 => no data contribution added to jnll

```

```
// Indices
DATA_INTEGER( num_i );   // Number of data points in space
DATA_INTEGER( num_s );   // Number of mesh points in space mesh

// Data (all except for X_ij is a vector of length num_i)
DATA_VECTOR( y_i );   // obs per binomial experiment at point i (clust)
DATA_VECTOR( n_i );   // Trials per cluster
DATA_MATRIX( X_alpha );  // 'design matrix' for just int

// SPDE objects
DATA_SPARSE_MATRIX( M0 );
DATA_SPARSE_MATRIX( M1 );
DATA_SPARSE_MATRIX( M2 );
DATA_SPARSE_MATRIX( Aproj );

// Options
DATA_VECTOR( options );
// options[0] == 1 : use normalization trick
// options[1] == 1 : adreport transformed params

// Prior specifications
DATA_VECTOR( alpha_pri );
DATA_VECTOR( matern_pri);
// matern_pri = c(a, b, c, d): P(range < a) = b; P(sigma > c) = d
Type matern_par_a = matern_pri[0]; // range limit:    rho0
Type matern_par_b = matern_pri[1]; // range prob:     alpha_rho
Type matern_par_c = matern_pri[2]; // field sd limit: sigma0
Type matern_par_d = matern_pri[3]; // field sd prob:  alpha_sigma

// Fixed effects
PARAMETER( alpha ); // Intercept
// Log of INLA tau param (precision of space covariance matrix)
PARAMETER( log_tau );
// Log of INLA kappa (related to spatial correlation and range)
PARAMETER( log_kappa );

// Random effects for each spatial mesh vertex
PARAMETER_VECTOR( Epsilon_s );

// ~~~~~~~~~------------------------------------------------~~
// SECOND, we define all other objects that we need internally
// ~~~~~~~~~------------------------------------------------~~
```

```
125
126 // objective function -- joint negative log-likelihood
127 Type jnll = 0;
128
129 // Make spatial precision matrix
130 SparseMatrix<Type> Q_ss = spde_Q(log_kappa, log_tau, M0, M1, M2);
131
132 // Transform some of our parameters
133 Type sp_range = sqrt(8.0) / exp(log_kappa);
134 Type sp_sigma = 1.0 / sqrt(4.0 * 3.14159265359 *
135                 exp(2.0 * log_tau) * exp(2.0 * log_kappa));
136
137 // Define objects for derived values
138 vector<Type> fe_i(num_i); // main effect: alpha
139 // Logit estimated prob for each cluster i
140 vector<Type> latent_field_i(num_i);
141 // value of gmrf at data points
142 vector<Type> projepsilon_i(num_i);
143
144 // fixed effects is just alpha in this example
145 fe_i = X_alpha * Type(alpha); // initialize
146
147 // Project GP approx from mesh points to data points
148 projepsilon_i = Aproj * Epsilon_s.matrix();
149
150 // ~~~~~~~~~-------------------------------------------------~~-
151 // THIRD, we calculate the contribution to the likelihood from:
152 // 1) priors
153 // 2) GP field
154 // 3) data
155 // ~~~~~~~~~-------------------------------------------------~~-
156
157 /////////
158 // (1) //
159 /////////
160 // the random effects. we do this first so to do the
161 //   normalization outside of every optimization step
162 // NOTE: likelihoods from namespace 'density' already return NEGATIVE
163 //       log-liks so we add other likelihoods return positive log-liks
164 if(options[0] == 1){
165   // then we are not calculating the normalizing constant in the inner opt
166   // that norm constant means taking an expensive determinant of Q_ss
```

```
    jnll += GMRF(Q_ss, false)(Epsilon_s);
    // return without data ll contrib to avoid unneccesary log(det(Q)) calcs
    if (flag == 0) return jnll;
  }else{
    jnll += GMRF(Q_ss)(Epsilon_s);
  }

  /////////
  // (2) //
  /////////
  // Prior contributions to joint likelihood (if options[1]==1)

  // add in priors for spde gp
  jnll -= dPCPriSPDE(log_tau, log_kappa,
                     matern_par_a, matern_par_b, matern_par_c, matern_par_d,
                     true);

  // prior for intercept
  jnll -= dnorm(alpha, alpha_pri[0], alpha_pri[1], true); // N(mean, sd)

  /////////
  // (3) //
  /////////
  // jnll contribution from each datapoint i

  for (int i = 0; i < num_i; i++){

    // latent field estimate at each obs
    latent_field_i(i) = fe_i(i) + projepsilon_i(i);

    // and add data contribution to jnll
    if(!isNA(y_i(i))){

     // Uses the dbinom_robust function, which takes the logit probability
        jnll -= dbinom_robust( y_i(i), n_i(i), latent_field_i(i), true );

    } // !isNA

  } // for( i )


  // ~~~~~~~~~~~
```

```
  // ADREPORT: used to return estimates and cov for transforms?
  // ~~~~~~~~~~~~
  if(options[1]==1){
    ADREPORT(sp_range);
    ADREPORT(sp_sigma);
  }

  return jnll;

}"

## write model to file, compile, and load it into R
dir.create('TMB_spde_example')
write(tmb_spde,file="TMB_spde_example/tmb_spde.cpp")
compile( "TMB_spde_example/tmb_spde.cpp")
dyn.load( dynlib("TMB_spde_example/tmb_spde") )

## prep inputs for TMB
data_full <- list(num_i = nrow(dat),  # Total number of observations
                  num_s = mesh.s[['n']], # num. of vertices in SPDE mesh
                  y_i   = dat[, obs],# num. of pos. obs in the cluster
                  n_i   = dat[, n],  # num. of exposures in the cluster
                  X_alpha  = matrix(1, nrow = nrow(dat), ncol = 1),# des.mat
                  M0    = spde[['param.inla']][['M0']], # SPDE sparse matrix
                  M1    = spde[['param.inla']][['M1']], # SPDE sparse matrix
                  M2    = spde[['param.inla']][['M2']], # SPDE sparse matrix
                  Aproj = A.proj,             # Projection matrix
                  options = c(1, ## if 1, use normalization trick
                              1), ## if 1, run adreport
                  # normalization flag.
                  flag = 1,
                  alpha_pri = alpha.pri, ## normal
                  matern_pri = matern.pri
                  )

## Specify starting values for TMB params
tmb_params <- list(alpha = 0.0, # intercept
                   log_tau = 0, # Log inverse of tau (Epsilon)
                   log_kappa = 0, # Matern range parameter
                   Epsilon_s = rep(0, mesh.s[['n']]) # RE on mesh vertices
                   )

```

```
## make a list of things that are random effects
rand_effs <- c('Epsilon_s')

## make the autodiff generated liklihood func & gradient
obj <- MakeADFun(data=data_full,
                 parameters=tmb_params,
                 random=rand_effs,
                 hessian=TRUE,
                 DLL='tmb_spde')

## we can normalize the GMRF outside of the nested optimization,
## avoiding unnecessary and expensive cholesky operations.
obj <- normalize(obj, flag="flag", value = 0)

## run TMB
opt0 <- nlminb(start       =    obj[['par']],
               objective   =    obj[['fn']],
               gradient    =    obj[['gr']],
               lower = rep(-10, length(obj[['par']])),
               upper = rep( 10, length(obj[['par']])),
               control     =    list(trace=1))

## Get standard errors
SD0 <- TMB::sdreport(obj, getJointPrecision=TRUE,
                     bias.correct = TRUE,
                     bias.correct.control = list(sd = TRUE))
## summary(SD0, 'report')

## take samples from fitted model
mu <- c(SD0$par.fixed,SD0$par.random)

## simulate draws
rmvnorm_prec <- function(mu, chol_prec, n.sims) {
  z <- matrix(rnorm(length(mu) * n.sims), ncol=n.sims)
  L <- chol_prec #Cholesky(prec, super=TRUE)
  z <- Matrix::solve(L, z, system = "Lt") ## z = Lt^-1 %*% z
  z <- Matrix::solve(L, z, system = "Pt") ## z = Pt    %*% z
  z <- as.matrix(z)
  mu + z
}

L <- Cholesky(SD0[['jointPrecision']], super = T)
```

```
t.draws <- rmvnorm_prec(mu = mu , chol_prec = L, n.sims = 500)

## summarize the draws
parnames <- c(names(SD0[['par.fixed']]), names(SD0[['par.random']]))
epsilon_tmb_draws  <- t.draws[parnames == 'Epsilon_s',]
alpha_tmb_draws    <- matrix(t.draws[parnames == 'alpha',], nrow = 1)

# project from mesh to raster, add intercept
pred_tmb <- as.matrix(A.pred %*% epsilon_tmb_draws)
pred_tmb <- sweep(pred_tmb, 2, alpha_tmb_draws, '+')

## find the median and sd across draws, as well as 90% intervals
summ_tmb <- cbind(median = (apply(pred_tmb, 1, median)),
                  sd     = (apply(pred_tmb, 1, sd)),
                  lower = (apply(pred_tmb, 1, quantile, .05)),
                  upper = (apply(pred_tmb, 1, quantile, .95)))

## make summary rasters
ras_med_tmb <- ras_sdv_tmb <- ras_lower_tmb <-
  ras_upper_tmb <- ras_inInt_tmb <- gp.rast
values(ras_med_tmb)   <- summ_tmb[, 1]
values(ras_sdv_tmb)   <- summ_tmb[, 2]
values(ras_lower_tmb) <- summ_tmb[, 3]
values(ras_upper_tmb) <- summ_tmb[, 4]
values(ras_inInt_tmb) <- 0
ras_inInt_tmb[gp.rast < ras_lower_tmb | ras_upper_tmb < gp.rast] <- 1

## plot truth, pixels falling within/without the 90% interval,
##  post. median, and post sd

# set the range for the truth and median
rast.zrange <- range(c(values(gp.rast), values(ras_med_tmb)), na.rm = T)

# plot
par(mfrow = c(2, 2))
plot(gp.rast, main = 'Truth', zlim = rast.zrange, col = (viridis(100)))
points(dat[, .(x, y)])
plot(ras_inInt_tmb, main = 'Pixels where 90% CIs did not cover Truth')
points(dat[, .(x, y)])
plot(ras_med_tmb, main = 'TMB Posterior Median',
     zlim = rast.zrange, col = (viridis(100)))
points(dat[, .(x, y)])
```

```
plot(ras_sdv_tmb, main='TMB Posterior Standard Deviation')
points(dat[, .(x, y)])


## compare INLA and TMB meds and stdevs

med.zrange <- range(c(values(ras_med_tmb), values(ras_med_inla)), na.rm = T)
sdv.zrange <- range(c(values(ras_sdv_tmb), values(ras_sdv_inla)), na.rm = T)

par(mfrow = c(2, 2))
plot(ras_med_inla, main = 'INLA Posterior Median',
     zlim = med.zrange, col = (viridis(100)))
points(dat[, .(x, y)])
plot(ras_sdv_inla, main = 'INLA Posterior Standard Deviation',
     zlim = sdv.zrange)
points(dat[, .(x, y)])
plot(ras_med_tmb, main = 'TMB Posterior Median',
     zlim = med.zrange, col = (viridis(100)))
points(dat[, .(x, y)])
plot(ras_sdv_tmb, main = 'TMB Posterior Standard Deviation',
     zlim = sdv.zrange)
points(dat[, .(x, y)])
```

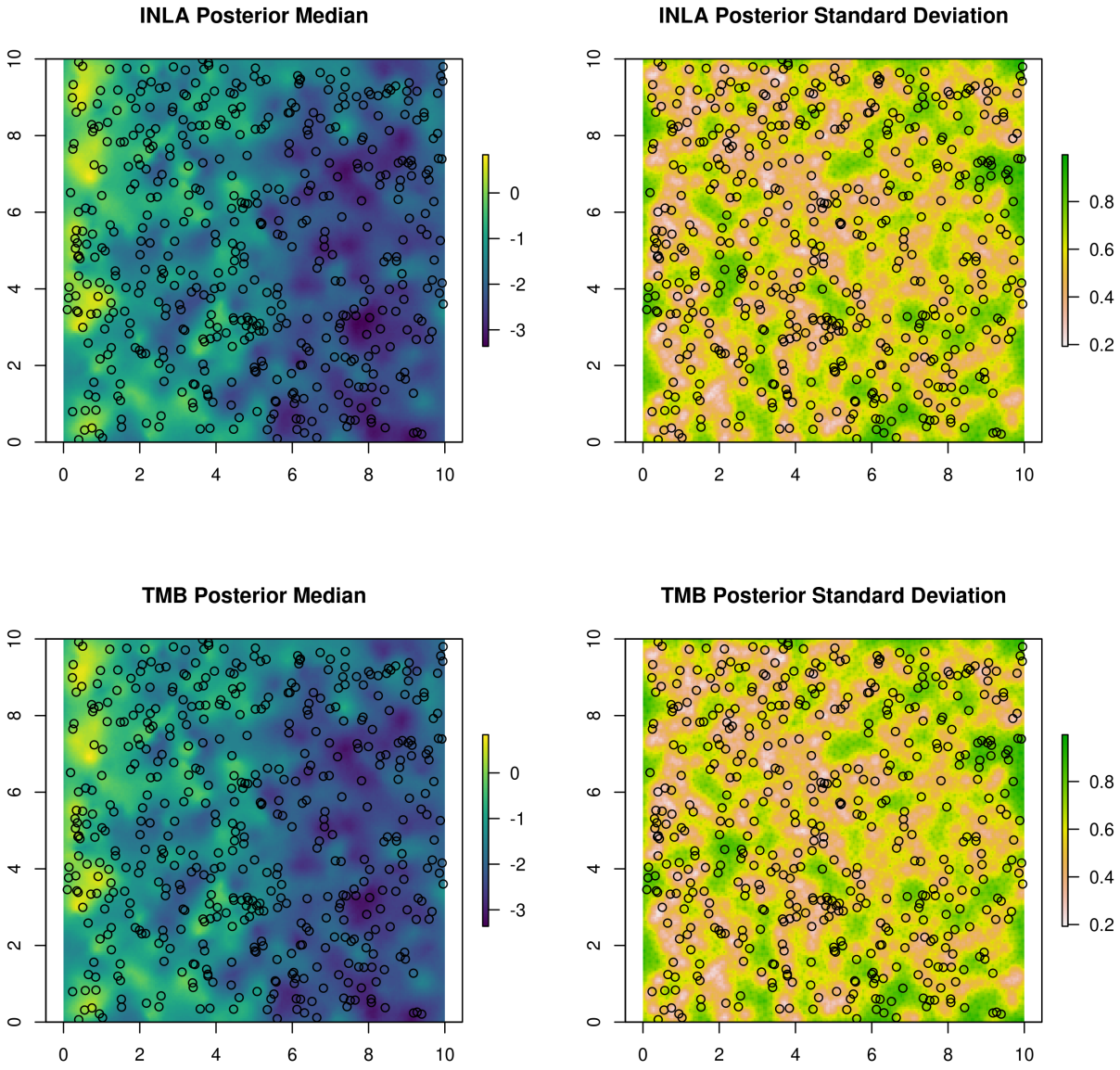


Figure A.4.2: Comparison of the posterior median and standard deviations from the example R-INLA and TMB code. Points indicate cluster locations.

A.5 Software and hardware details

The large scale continuous simulations of Section 3.6.1 were run on a large computing cluster containing a variety of hardware types. The jobs were randomly placed onto different machines as space became available and the effects of the various hardware were averaged over in producing all results, including the serial timing results in Figure 3.7. A singularity image was used to ensure consistent software versions across the nodes. Specifically, these simulations were run on:

- `R` 3.6.1,
- `R-INLA` 20.01.29.9000,
- `TMB` 1.7.16.

All other work was performed on a laptop with an Intel Core i7-8550U CPU (4 cores, 8 threads @ 1.8GHz) and 16Gb of RAM. This machine used the following software versions:

- `R` 4.0.4,
- `R-INLA` 21.02.23, with PARDISO solver enabled,
- `TMB 1.7.18`, with METIS reordering enabled.

Appendix B

APPENDIX FOR CHAPTER 4

B.1 Data

B.1.1 Data Sources and Processing

The breast cancer morality and incidence counts and the population data used for this project were provided by IARC. National mortality and population data was extracted from the WHO mortality database and this was merged with the local registry data aggregated by IARC. Only incidence, mortality, and population counts from females were used.

The data procured from IARC, with few exceptions, came as individual files for each national or subnational location. Subnational data from local registries were merged on to the corresponding national datasets by age group and year. Where necessary, data was subset to observations occurring in 2000 or later. For each age-year with both national and subnational incidence counts or national and subnational mortality counts, the subnational counts and populations were subtracted from the national ones to avoid any duplicate counting. De-duplicated datasets across all countries were merged together, to create a master dataset where each row consisted of a unique age group, year and location observation, and where the locations could either represent national counts, subnational counts, or national remainders. Both cervical and breast cancer counts were provided. The cervical cancer data was only used to help fill in the occasional missing population from the breast cancer dataset and to help distinguish between missing data and unrecorded 0s.

There were a few particular locations which had observations binned to non-standard age groups, such as the Republika of Srpska in Bosnia what a maximum age group of 75+ instead of the standard 85+ used by most locations and used in this study. In these cases, national population proportions were used to partition the non-standard bins into the standard age

groups. A number of Ukrainian subnationals were processed in the same way. There were certain years of national observations from Slovenia which were partitioned using the national population from Slovenia in the nearest year which reported standard age bins.

The UK was processed individually due to multiple (conflicting) sources of available information. IARC was able to provide complete registry coverage for Scotland, England, Wales, and North Ireland, which when summed, provide complete UK national counts. The UK national data extracted from the WHO mortality database differed in substantial ways from the sum of the combined registry datasets. We proceeded by dropping the national observations and have thus treated UK as a type V country for all years (only subnational data) even though there is complete subnational coverage of the country and an second complete national dataset.

Finally, to help differentiate between true missing observations and unrecorded 0 counts (*e.g.* within a young age group), any country or sub-region with no cases and no deaths reported for all ages within a year for breast cancer were recoded from 0s to "missing".

Table B.1.1 provides summaries of the total amount of available data by country after initial processing, including subtracting subnational from national counts, and Figure B.1.1 provide a summary of the national and subnational data available for each country across time.

Country	N I Yrs	S I Yrs	N M Yrs	S M Yrs	N Pop	S Pop	N I Ct	S I Ct	N M Ct	S M Ct	Matched I Ct	Matched M Ct
Austria	13	0	16	0	127446510	0	72293	0	28279	0	61381	19464
Belarus	9	0	9	0	77421902	0	38130	0	14314	0	38130	14314
Belgium	9	0	16	0	136442096	0	95969	0	40818	0	95969	23204
Bosnia	0	5	3	5	6987658	6053886	0	3360	1614	1177	3360	1177
Bulgaria	13	0	10	0	99632540	0	61642	0	16996	0	38873	13532
Croatia	13	0	18	0	63894692	0	36051	0	19400	0	36051	13644
Cyprus	13	0	13	0	9146840	0	5830	0	1472	0	4406	991
Czechia	13	0	17	0	146838334	0	89886	0	37146	0	89886	29055
Denmark	13	0	16	0	69254764	0	61184	0	21827	0	61184	18138
Estonia	13	0	17	0	18883942	0	10530	0	5389	0	10530	4190
Finland	10	0	16	0	66130200	0	39686	0	14360	0	39686	8852
France	0	13	17	5	791723724	93015132	0	103859	209633	11155	47440	11155
Germany	0	13	17	5	981382136	430913130	0	474773	270449	65234	262409	65234
Greece	0	0	2	0	17643592	0	0	0	4513	0	0	0
Hungary	0	0	18	0	147183560	0	0	0	48033	0	0	0
Iceland	13	0	18	0	3372094	0	2620	0	778	0	2620	508
Ireland	13	0	8	0	48630146	0	34962	0	6197	0	18424	4630
Italy	0	13	14	6	623744722	206939630	0	221772	155585	25443	98992	25443
Latvia	13	0	16	0	30172122	0	16125	0	9134	0	16125	7438
Lithuania	13	0	18	0	47557804	0	24781	0	14084	0	24781	10445
Luxembourg	0	0	17	0	4891308	0	0	0	1502	0	0	0
Macedonia	0	0	8	0	10774164	0	0	0	2607	0	0	0
Malta	13	0	17	0	4872736	0	3486	0	1403	0	3486	1075
Moldova	0	0	18	0	47718810	0	0	0	11796	0	0	0
Montenegro	0	0	10	0	3696968	0	0	0	878	0	0	0
Netherlands	13	0	18	0	239016392	0	175463	0	63164	0	175463	46264
Norway	13	0	17	0	61040164	0	40451	0	13202	0	40451	10250
Poland	0	13	18	5	511577038	119743864	0	70225	122501	10349	28915	10349
Portugal	0	12	13	4	111429546	1931298	0	1458	25655	161	550	161
Romania	3	0	18	0	320866634	0	20820	0	92270	0	20820	15294
Russia	0	13	16	5	2451881776	61251490	0	35165	461832	507	2854	507
Serbia	0	0	18	0	107287474	0	0	0	38777	0	0	0
Slovakia	11	0	14	0	61431640	0	30650	0	14234	0	30650	10733
Slovenia	13	0	18	0	29271222	0	16897	0	8050	0	16897	5719
Spain	0	13	17	5	578044636	107569640	0	77453	108488	7925	31233	7925
Sweden	12	0	18	0	128421270	0	82233	0	30287	0	82233	20464
Switzerland	0	13	14	5	59661826	44424212	0	45883	15840	4562	18961	4562
UK	0	13	0	17	0	695894728	0	617429	0	78944	278227	73139
Ukraine	0	11	16	0	792979274	444192801	0	239123	158694	0	0	0

Table B.1.1: Summary by country of the number of years with (N)ational and (S)ubnational (I)ncidence and (M)ortality data, the national and subnational populations represented across all ages and years, the total number of national and subnational incidence and mortality observations across all ages and years, and the total number of incidence and mortality observations which could be matched at the location- age- and year-level (the total counts that could be used to estimate MI ratios).

B.1.2 IARC Data Scoring Systems

Tables B.1.2 and B.1.3 provide summaries of the IARC data scoring systems for incidence and mortality. Incidence scores are determined by the availability and coverage of registries within the country. The number of registries per country and the population covered by any given registry, in either total numbers or population percentages, varies quite a bit within and between countries. Some countries only provide national or regional rates. Across all of these data types it can be quite difficult to assess the reliability of these numbers. For the sake of this project, we focus on providing estimates of rates like those reported by countries and do not attempt to correct for any reporting error or undercoverage.

Table B.1.2: 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 B.1.3: 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

B.1.3 Data Types

The current implementation of the methods described in this paper 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 six types of data; (I) national incidence and mortality, (II) subnational incidence and mortality (from registries) and national mortality, (III) only national mortality, (IV) only national incidence, (V) only subnational data, and (VI) no available data. The relationship between the IARC scores and our categories are shown in Table B.1.4. The available data type for each country and year is shown in the top of Figure B.1.1.

To protect patient confidentiality we do not have information about individual registries beyond cases, deaths, and catchment population by standard five-year age groups. For the purposes of analyses discussed in this article, each different subnational registry constitutes a data observation that enters the model. 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 B.1.4: Comparison of the six data types (I-VI) used in our methodology to the mortality and incidence source and quality scores defined by IARC in Tables B.1.2 and B.1.3. The comparison between our data types and the scores used by IARC is not one-to-one because, for example, high-quality national or regional incidence data coverage (incidence score A) could feature in any of our data type categories which have national incidence data (data types I, II, IV, and V). For this project, we do not use any incomplete or sample vital registration data resulting in the empty mapping of mortality score 4.

IARC Mortality Data Scores	IARC Incidence Data Scores						
	A	B	C	D	E	F	G
1	I/II	II	II	I	II	–	III
2	I/II	II	II	I	II	–	III
3	I/II	II	II	I	II	–	III
4	–	–	–	–	–	–	–
5	IV/V	V	V	IV	V	–	–
6	IV/V	V	V	IV	V	–	VI

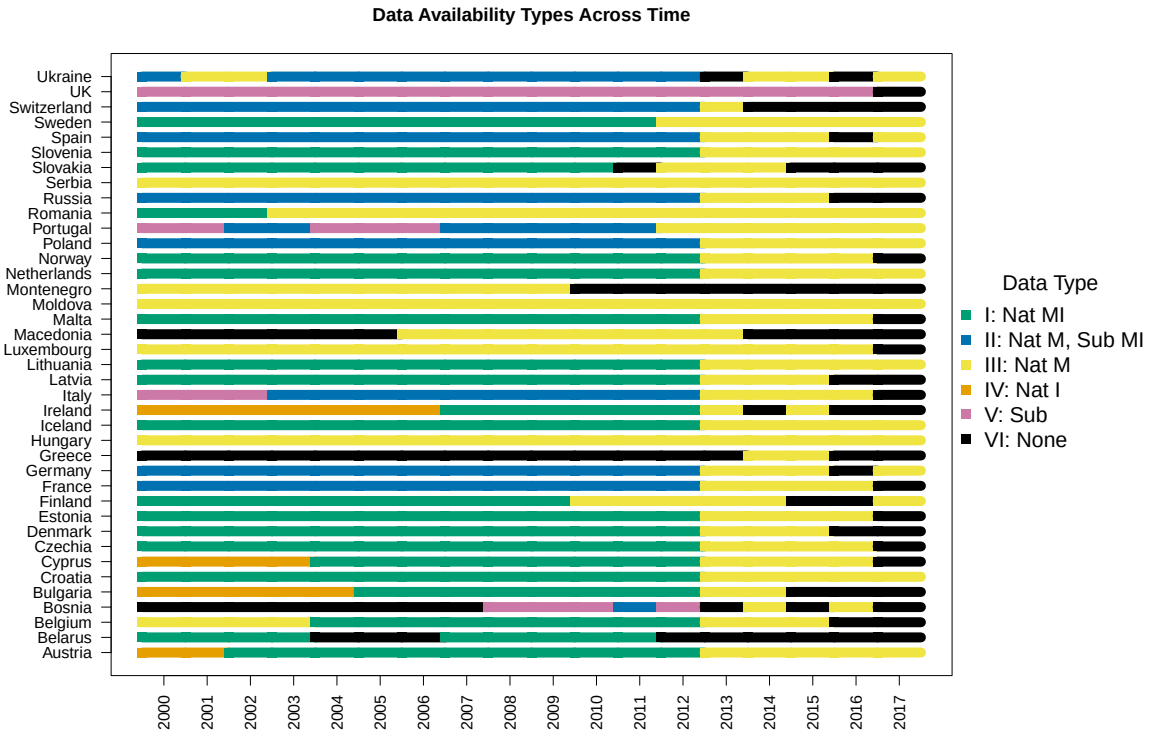


Figure B.1.1: Data type by country from 2000–2017.

B.2 Factors associated with breast cancer incidence and mortality rates

The purpose of this project is to estimate the number of cases that would appear in the health care system. This is neither etiology nor estimation of the true underlying incidence, but a descriptive and surveillance-related effort. The estimates we produce will be akin to the numbers planners are used to receiving from cancer registries. There are many factors that may influence the rates of reporting of breast cancer in each country and even within the populations covered by each local registry. As shown in Appendix B.1 the data quality and sources vary between countries and across times. Furthermore, the location of registries may not be representative of the country as a whole which has the potential to introduce bias. Additionally we would expect routine screening and diagnostic methods, which vary

across ages, locations and times and are associated with more developed health care systems, to increase incidence rates. Lastly, risk factors and population differences will affect the true underlying rates and will vary by country and potentially across time. With country-level data and limited data on risk factors, it is difficult to disaggregate the impact of the many factors that may affect reported incidence and mortality rates (*e.g.* registry reporting rates versus differences in diagnostic tools), but in this section we will briefly discuss some of the factors and their likely effects.

There are many potential sources of bias in both the registry and national cancer data. Systematic underreporting of incident cases is likely if the registry does not have access to all possible sources of information. In the case of breast cancer, poor access to screening can be associated with higher than expected mortality, because cases are identified at more advanced stages after symptoms have presented, and access to treatment may be more limited in low-screening settings. This could lead to higher than expected mortality rates given the, likely, lower than expected reported incidence rates. Registries data may also biased due to the fact that they do not represent a random sample from the population. Further exacerbating this, one could imagine that registries are more often established 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 particular to each location, as these could increase or decrease the true rates. CDC (2022b) reports the primary risk factors for breast cancer as:

- age (getting older),
- genetic mutations (such as BRCA1 and BRCA2),
- early menstrual periods before the age of 12
- later menopausal transition after age 55,

- higher breast density,
- personal history of breast cancer or non-cancer breast diseases,
- family history of breast or ovarian cancer,
- previous radiation therapy treatment,
- exposure to diethylstilbestrol (DES),
- not being physically active,
- overweight/obese after menopause,
- use of combination hormone therapy,
- oral contraceptives,
- first pregnancy after 30, not breastfeeding, never having a full-term pregnancy, and
- drinking alcohol.

Age is straight forward to incorporate in a modeling strategy and how we included it is discussed in Appendix B.3, but most of the other risk factors are difficult to find reliable data on at a country level, let alone across age and time.

In summary, there are many factors related to the reported rates of breast cancer. Our goal is to provide short term backcasts of the reported number of cases, as these are directly related to health care utilization, are relevant for planning and resource allocation, and are easier to obtain given the reported data.

B.3 Joint Model of Mortality and MI Ratio

In this section we detail the complete Bayesian hierarchical model (BHM) including the priors and hyperpriors, the methods we use to sample from the approximate posterior distribution, and the metrics used to perform model selection and validation.

B.3.1 Hierarchical Model Description

Berliner (1996), in the context of time series modeling, introduced a general three-layer description of BHMs which is widely applicable and which we find useful for describing this work. The three layers represent the data sampling, the latent process, and the hyperpriors and were discussed in Chapter 2.1. We now describe each of the three stages of the BHM as used within this project.

B.3.2 Data Stage

The approach we describe rely on probabilistic models for incidence, mortality, and mortality given incidence. For many countries an alternative would be to rely only on separate unconditional models for incidence and mortality. The MI modeling approach facilitates estimating national incidence in countries without national or local incidence data by providing an explicit link between mortality and incidence. The data stage of the BHM, representing the conditional layer shown in (2.2), consists of the product of likelihoods for each data observation. The observations fall into one of five data types (I-V, VI being no data), and the likelihoods for the five data types are described in detail in the following sections after notation is defined.

Notation

We begin by establishing notation with a indexing 5-year age-groups, c indexing country, t indexing time in years, L indicating local (registry) data from the L^{th} subnational registry in country c at time t, and R denoting the remainder data (not covered by local registries).

The number of subnational registries varies by countries and times but we suppress that notation and allow $L \in \{1, \ldots, n_{ct}^L\}$ as appropriate for any country-time index ct.

For age group a, country c, and time t we define:

- N_{act}^L = population in time t, age group a, and local area L in country c (population covered by registry L),
- Y_{act}^L = reported cases (incidence) in time t, age group a, and local area L in country c (cases reported at registry L)
- Z_{act}^L = reported deaths (mortality) in time t, age group a, and local area L in country c (deaths reported at registry L),
- N_{act}^R = population remainder in time t, age group a, and country c (population not covered by registries),
- Y_{act}^R = reported cases (incidence) remainder in time t, age group a, and country c among the remainder population (cases not covered by registries),
- Z_{act}^R = reported deaths (mortality) remainder in time t, age group a, and country c among the remainder population (deaths not covered by registries),
- $Y_{act} = \sum_L Y_{act}^L + Y_{act}^R$ = all reported cases in time t, age group a and country c,
- $Z_{act} = \sum_L Z_{act}^L + Z_{act}^R$ = all reported deaths in time t, age group a and country c,
- $N_{act} = \sum_L N_{act}^L + N_{act}^R$ = total population in time t, age group a and country c,
- $p_{act} = P(\text{incident case}|a, c, t)$ = reported incidence risk in time t, age group a and country c,

- $q_{act} = P(\text{ death } | a, c, t) =$ reported unconditional mortality risk in time t, age group a and country c, and

- $r_{act} = \Pr(\text{death } | \text{incident case}, a, c, t) =$ the MI ratio in time t, age group a and country c.

Country-times with national mortality and incidence

We first consider the model specification for country-times with both national mortality and incidence data (type I data). In the absence of local registry data, the national data is still denoted with a remainder superscript R to indicate we are denoting all non-registry data (even though no local registry data has been subtracted from it). Our approach assumes the following two conceptual models for incidence and mortality given incidence, respectively:

$$Y^R_{act} | N^R_{act}, p_{act} \sim \text{Poisson}(N_{act} \times p_{act}), \qquad p_{act} = \frac{q_{act}}{r_{act}} \tag{B.3.1}$$

$$Z^R_{act} | Y^R_{act}, r_{act} \sim \text{Binomial}(Y_{act}, r_{act}), \qquad r_{act} = \frac{\exp(\eta^{MI}_{act})}{1 + \exp(\eta^{MI}_{act})}. \tag{B.3.2}$$

The log linear model in (B.3.1) acknowledges the rare disease assumption. This pair of models imply the (unconditional) mortality model:

$$Z^R_{act} | N^R_{act}, q_{act} \sim \text{Poisson}(N^R_{act} \times q_{act}), \qquad q_{act} = p_{act} \times r_{act} = \exp(\eta^M_{act}). \tag{B.3.3}$$

We model the linear predictor parameters as:

$$\eta^M_{act} = \alpha^M + (\beta^M_a a + b^M_a) + b^M_c + (\beta^M_t t + b^M_t) + \dots \tag{B.3.4}$$

$$\eta^{MI}_{act} = \alpha^{MI} + (\beta^{MI}_a a + b^{MI}_a) + b^{MI}_c + (\beta^{MI}_t t + b^{MI}_t) + \dots \tag{B.3.5}$$

where the BYM2 model of Section B.3.3 was selected to model the random effects $b^{\bullet}_{\bullet}$ across 5-year age-groups, countries, and years. Linear terms for age and time were also included.

Different combinations of higher-order random effects interactions were assessed to select a final model from a suite of candidate models, as described in 4.3.4.

The parameters $\alpha^{\bullet}, \beta^{\bullet}_{\bullet}$ and the hyperparameters associated with each random effect, $b^{\bullet}_{\bullet}$, are shared across all ages, countries, and times, regardless of the available data. Let ω_1 denote the set of indices for country-times with both national incidence and mortality data, and then let

- $\boldsymbol{N}^{(1)R} = \{y_{act} : \{c,t\} \in \omega_1\}$ national population in age a in type I country-times,
- $\boldsymbol{y}^{(1)R} = \{y_{act} : \{c,t\} \in \omega_1\}$ national incidence data in age a in type I country-times,
- $\boldsymbol{z}^{(1)R} = \{z_{act} : \{c,t\} \in \omega_1\}$ national mortality data in age a in type I country-times,
- $\boldsymbol{b}^{(1)M} = \{b^{M}_{\bullet} : \{c,t\} \in \omega_1\}$ age-country-time-specific random effects for mortality in type I country-times, and
- $\boldsymbol{b}^{(1)MI} = \{b^{MI}_{\bullet} : \{c,t\} \in \omega_1\}$ age-country-time-specific random effects for the MI ratio in type I country-times.

The likelihood of the data from type I countries is

$$p(\boldsymbol{y}^{(1)R}, \boldsymbol{z}^{(1)} | N^R_{act}, \alpha^M, \beta^M_a, \beta^M_t, \boldsymbol{b}^{(1)M}, \alpha^{MI}, \beta^{MI}_a, \beta^{MI}_t, \boldsymbol{b}^{(1)MI}) =$$
$$\underbrace{p(\boldsymbol{y}^{(1)R} | \boldsymbol{N}^{(1)R}, \alpha^M, \beta^M_a, \beta^M_t, \boldsymbol{b}^{(1)M}, \alpha^{MI}, \beta^{MI}_a, \beta^{MI}_t, \boldsymbol{b}^{(1)MI})}_{\prod_{\{c,t\}\in\omega_1} \prod_a \text{Poisson}(N^R_{act} \times p_{act})} \times$$
$$\underbrace{p(\boldsymbol{z}^{(1)R} | \boldsymbol{y}^{(1)R}, \alpha^{MI}, \beta^{MI}_a, \beta^{MI}_t, \boldsymbol{b}^{(1)MI})}_{\prod_{\{c,t\}\in\omega_1} \prod_a \text{Binomial}(Y^R_{act},\ r_{act})}.$$

Country-times with subnational mortality and incidence and national mortality

Next we consider a modeling approach for the country-times with local incidence and mortality data from registries and national mortality data (type II countries). For these country-

times our goal is to use the observed local MI ratio applied to the remainder national mortality to estimate total national incidence. Our approach relies on jointly modeling the local incidence and mortality in addition to the national mortality counts that are not contained in the registries. Define $Z^R_{act} = Z_{act} - \sum_L Z^L_{act}$ and $N^R_{act} = N_{act} - \sum_L N^L_{act}$. Our MI ratio assumption states that all local registry and national MI ratios are equivalent (that is, $\forall l: \ r^L_{act} = r^R_{act} = r_{act}$). This yields the following model for the remainder mortality:

$$Z^R_{act}|N^R_{act}, q_{act} \sim \text{Poisson}(N^R_{act} \times q_{act}), \quad q_{act} = \exp(\eta^M_{act}) \tag{B.3.6}$$

For the local data, each registry, L, is included as a data observation:

$$Y^L_{act}|N^L_{act}, p^L_{act} \sim \text{Poisson}(N^L_{act} \times p_{act}), \qquad p_{act} = \frac{q_{act}}{r_{act}} = \frac{\exp(\eta^M_{act})(1+\exp(\eta^{MI}_{act}))}{\exp(\eta^{MI}_{act})} \tag{B.3.7}$$

$$Z^L_{act}|Y^L_{act}, r^L_{act} \sim \text{Binomial}(Y^L_{act}, \ r_{act}), \qquad r_{act} = \frac{\exp(\eta^{MI}_{act})}{1+\exp(\eta^{MI}_{act})} \tag{B.3.8}$$

where the linear predictor parameters are assumed to be of the form (B.3.4) and (B.3.5).

Our model formulation models the observed incidence, using the quotient of two parameters, q_{act} and r_{act}. Both parameters are influenced by all the data: the mortality parameter, q_{act} enters directly into the likelihood for the national mortality data but it also features in the likelihood for the local incidence data. Likewise, the MI parameter is directly influenced by the local incidence and mortality data but the joint formulation, made explicit by the nonlinear relationship used to define the incidence parameter, p_{act}, means the national mortality remainder also contributes to the MI parameters estimates. By combining the modeled MI ratio, estimated from the local registries, with the remainder mortality, the model allows inference on the unobserved national incidence rate. Depending on the registry coverage for a country-time, as $\sum_L N^L/N$ approaches one the parameters will be more heavily influenced by data from the registries, and when $\sum_L N^L/N$ is small the parameters will be more heavily influenced by the national mortality data. Let ω_2 index country-times with subnational incidence and mortality data from registries and national mortality data, and then define:

- $\boldsymbol{N}^{(2)L} = \{N^L_{act} : \{c,t\} \in \omega_2\}$ local population served by registry L in age a in type II country-times,
- $\boldsymbol{N}^{(2)R} = \{N^R_{act} : \{c,t\} \in \omega_2\}$ national remainder population not covered by any local registry in age a in type II country-times,
- $\boldsymbol{y}^{(2)L} = \{y^L_{act} : \{c,t\} \in \omega_2\}$ local incidence data from registry L in age a in type II country-times,
- $\boldsymbol{z}^{(2)L} = \{z^L_{act} : \{c,t\} \in \omega_2\}$ local mortality data from registry L in age a in type II country-times,
- $\boldsymbol{z}^{(2)R} = \{z^R_{act} : \{c,t\} \in \omega_2\}$ national mortality remainder data in age a in type II country-times,
- $\boldsymbol{b}^{(2)M} = \{b^M_\bullet : \{c,t\} \in \omega_2\}$ age-country-time-specific random effects for mortality in type II countries, and
- $\boldsymbol{b}^{(2)MI} = \{b^{MI}_\bullet : \{c,t\} \in \omega_2\}$ age-country-time-specific random effects for MI in type II countries.

The likelihood of the data from type II countries is

$$
\begin{aligned}
p(\boldsymbol{y}^{(2)L}, \boldsymbol{z}^{(2)L}, \boldsymbol{z}^{(2)R} | \boldsymbol{N}^{(2)L}, \boldsymbol{N}^{(2)R}, \alpha^I, \beta^M_a, \beta^M_t, \boldsymbol{b}^{(2)I}, \alpha^{MI}, \beta^{MI}_a, \beta^{MI}_t, \boldsymbol{b}^{(2)MI}) = \\
\underbrace{p(\boldsymbol{y}^{(2)L} | \boldsymbol{N}^{(2)L}, \alpha^I, \beta^M_a, \beta^M_t, \boldsymbol{b}^{(2)M}, \alpha^{MI}, \beta^{MI}_a, \beta^{MI}_t, \boldsymbol{b}^{(2)MI})}_{\prod_l \prod_{\{c,t\} \in \omega_2} \prod_a \text{Poisson}(N^{(2)L}_{act} \times p_{act})} \times \\
\underbrace{p(\boldsymbol{z}^{(2)L} | \boldsymbol{y}^{(2)L}, \alpha^{MI}, \beta^{MI}_a, \beta^{MI}_t, \boldsymbol{b}^{(2)MI})}_{\prod_l \prod_{\{c,t\} \in \omega_2} \prod_a \text{Binomial}(Y^L_{act},\ r_{act})} \times \\
\underbrace{p(\boldsymbol{z}^{(2)R} | \boldsymbol{N}^{(2)R}, \alpha^I, \beta^M_a, \beta^M_t, \boldsymbol{b}^{(2)M})}_{\prod_{\{c,t\} \in \omega_2} \prod_a \text{Poisson}(N^{(2)R}_{act} \times q_{act})}.
\end{aligned}
$$

In the event that the registry L only reports incidence data, that Binomial likelihood term for conditional mortality is left out.

Country-times with only national mortality

There are a number of country-times for which only national mortality data is available (type III countries). This is particularly common in recent years where the registry data reporting lags behind the national mortality reporting. For these observations, we fit the unconditional mortality model (B.3.3) and then impute the national incidence estimates by borrowing information about the mortality-incidence ratio, p_{act}, from nearby age-country-time locations. We learn about the national incidence rate p_{act} by modeling the national mortality rate q_{act} via the model:

$$Z^R_{act}|N^R_{act}, q_{act} \sim \text{Poisson}(N^R_{act} \times q_{act}), \qquad q_{act} = \exp(\eta^M_{act}). \tag{B.3.9}$$

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

- $\boldsymbol{N}^{(3)R} = \{N^R_{act} : \{c,t\} \in \omega_3\}$ national population in age a in type III country-times,
- $\boldsymbol{z}^{(3)R} = \{z^R_{act} : \{c,t\} \in \omega_3\}$ national mortality data for age a in type III country-times, and
- $\boldsymbol{b}^{(3)M} = \{b^M_\bullet : \{c,t\} \in \omega_3\}$ age-country-time-specific random effects for mortality in type III country-times.

The likelihood of the data for the model shown in (B.3.9) is:

$$p(\boldsymbol{z}^{(3)R}|\boldsymbol{N}^{(3)R}, \alpha^M, \beta^M_a, \beta^M_t, \boldsymbol{b}^{(3)M}) = \prod_{\{c,t\}\in\omega_3} \prod_a \text{Poisson}(N^R_{act} \times q_{act}).$$

Country-times with only national incidence

There are a few country-times for which only national incidence data is available (type IV countries). This is happens most often earlier in the time series. These observations let us directly estimate the national incidence, which in turn provides information for both the MI ratio and mortality parameters. In the absence of other data, incidence alone is would not allow the MI ratio and mortality to be identifiable in our model, but the smoothing random effects allow us to borrow information from incidence-only country-years to improve our estimates. We model the national incidence rate p_{act} as:

$$\begin{aligned} Y^R_{act} | N^R_{act}, p_{act} &\sim \text{Poisson}(N^R_{act} \times p_{act}), \\ p_{act} = \frac{q_{act}}{r_{act}} &= \frac{\exp(\eta^M_{act})(1+\exp(\eta^{MI}_{act}))}{\exp(\eta^{MI}_{act})}. \end{aligned} \tag{B.3.10}$$

Let ω_4 denote the set of indices for country-times with national incidence data and no mortality nor subnational data, and then let

- $\boldsymbol{N}^{(4)R} = \{N^R_{act} : \{c,t\} \in \omega_4\}$ national population in age a in type IV country-times,
- $\boldsymbol{y}^{(4)R} = \{y^R_{act} : \{c,t\} \in \omega_4\}$ national incidence data for age a in type IV country-times,
- $\boldsymbol{b}^{(4)M} = \{b^M_\bullet : \{c,t\} \in \omega_4\}$ age-country-time-specific random effects for mortality in type IV country-times, and
- $\boldsymbol{b}^{(4)MI} = \{b^{MI}_\bullet : \{c,t\} \in \omega_4\}$ age-country-time-specific random effects for the MI ratio in type IV country-times.

The likelihood of the data for the model shown in (B.3.10) is:

$$p(\boldsymbol{y}^{(4)R} | \boldsymbol{N}^{(4)R}, \alpha^M, \beta^M_a, \beta^M_t, \boldsymbol{b}^{(4)M}, \alpha^{MI}, \beta^{MI}_a, \beta^{MI}_t, \boldsymbol{b}^{(4)MI}) = \prod_{\{c,t\} \in \omega_4} \prod_a \text{Poisson}(N^R_{act} \times p_{act}).$$

Country-times with only subnational data

Some country-times only have local registry data (type V countries). Due to the assumption that local rates are equal to national rates, these observations are modeled identically to the registry data in type II observations as shown in (B.3.7) and (B.3.8).

Let ω_5 denote the set of indices for country-times with national mortality data and no incidence data, and then let

- $\boldsymbol{N}^{(5)L} = \{N^L_{act} : \{c,t\} \in \omega_5\}$ local population served by registry L in age a in type V country-times,
- $\boldsymbol{y}^{(5)L} = \{y^L_{act} : \{c,t\} \in \omega_5\}$ local incidence data from registry L in age a in type V country-times,
- $\boldsymbol{z}^{(5)L} = \{z^L_{act} : \{c,t\} \in \omega_5\}$ local mortality data from registry L in age a in type V country-times,
- $\boldsymbol{b}^{(5)M} = \{b^M_\bullet : \{c,t\} \in \omega_5\}$ age-country-time-specific random effects for mortality in type V countries, and
- $\boldsymbol{b}^{(5)MI} = \{b^{MI}_\bullet : \{c,t\} \in \omega_5\}$ age-country-time-specific random effects for MI in type V countries.

The likelihood of the data from type V countries is

$$p(\boldsymbol{y}^{(5)L}, \boldsymbol{z}^{(5)L} | \boldsymbol{N}^{(5)L}, \boldsymbol{N}^{(5)R}, \alpha^M, \beta^M_a, \beta^M_t, \boldsymbol{b}^{(5)M}, \alpha^{MI}, \beta^{MI}_a, \beta^{MI}_t, \boldsymbol{b}^{(5)MI}) =$$
$$\underbrace{p(\boldsymbol{y}^{(5)L} | \boldsymbol{N}^{(5)L}, \alpha^M, \beta^M_a, \beta^M_t, \boldsymbol{b}^{(5)M}, \boldsymbol{b}^{(5)M}, \alpha^{MI}, \beta^{MI}_a, \beta^{MI}_t, \boldsymbol{b}^{(5)MI})}_{\prod_l \prod_{\{c,t\} \in \omega_5} \prod_a \text{Poisson}(N^{(5)L}_{act} \times p_{act})} \times$$
$$\underbrace{p(\boldsymbol{z}^{(5)L} | \boldsymbol{y}^{(5)L}, \alpha^{MI}, \beta^{MI}_a, \beta^{MI}_t, \boldsymbol{b}^{(5)MI})}_{\prod_l \prod_{\{c,t\} \in \omega_5} \prod_a \text{Binomial}(Y^L_{act},\ r_{act})}.$$

In the event that the registry L only reports incidence data, the Binomial mortality term in the likelihood is left out.

Countries with No Data

During time periods where countries have no data, age effects, spatial effects, and temporal effects will be generated from the prior which borrows information from nearby age-country-times. In these country-years the base mortality model is

$$q_{act} = \exp\left(\alpha^M + (\beta_a^M a + b_a^{M\star}) + b_c^{M\star} + (\beta_t^M t + b_t^{M\star})\right)$$

where $b_a^{M\star}, b_c^{M\star}$, and $b_t^{M\star}$ will be influenced by data in neighboring ages, countries, times, and the rest of the observations through the random effects priors. Similarly, we will infer the conditional probability of mortality given incidence as:

$$r_{act} = \frac{\exp\left(\alpha^{MI} + (\beta_a^{MI} a + b_a^{MI\star}) + b_c^{MI\star} + (\beta_t^{MI} t + b_t^{MI\star})\right)}{1 + \exp\left(\alpha^{MI} + (\beta_a^{MI} a + b_a^{MI\star}) + b_c^{MI\star} + (\beta_t^{MI} t + b_t^{MI\star})\right)}$$

where $b_a^{MI\star}, b_c^{MI\star}$, and $b_t^{MI\star}$ are simulated from the prior. Incidence rates are then imputed as $p_{act} = q_{act}/r_{act}$.

B.3.3 The Process Stage: BYM2 Model

To model the random effects across space, age, and time, we implement and use the BYM2 Gaussian Markov random field (GMRF), a modern formulation of the classic Besag-York-Mollie model developed by Riebler *et al.* (2016). The random effects prior takes the form:

$$\begin{aligned} \boldsymbol{b} &= \frac{1}{\sqrt{\tau}}\left(\sqrt{1-\varphi}\boldsymbol{v} + \sqrt{\varphi}\boldsymbol{u}_\star\right) \\ \boldsymbol{v} &\sim N(\mathbf{0}, \boldsymbol{I}) \\ \boldsymbol{u}_\star &\sim N\left(\mathbf{0}, \boldsymbol{Q}_\star^{-1}\right), \; s.t. \; \sum_{i=1}^{37} u_{\star_i} = \mathbf{0}, \end{aligned} \tag{B.3.11}$$

for BYM2 GMRF, $\boldsymbol{b}$, with total variance τ^{-1}, mixing parameter φ controlling the contribution of $\boldsymbol{v}$, the unstructured i.i.d. portion of the BYM2 field, and $\boldsymbol{u}_\star$, the scaled spatially structured component of the BYM2. The structured portion of the BYM2 is specified with precision $\boldsymbol{Q}_\star$, a scaled version of the precision from the classic BYM ICAR model, and is constrained to sum to zero. The model, including the hard constraint was implemented in Template Model Builder (TMB, Kristensen *et al.*, 2016) by appropriately transforming the unconstrained optimization parameters into the appropriate constrained field (Gelfand *et al.*, 2010, Section 12.1.7.4). Each draw from the posterior underwent the same constraint correction procedure.

The discrete BYM2 spatial model was implemented over the 39 countries in our European region. Likewise, BYM2 models were implemented over the 18 age groups and the 18 years included in the model. The neighborhood adjacency structure for each dimension is shown in Figure B.3.1.

B.3.4 The Parameter Stage: Hyperpriors

To complete the Bayesian model specification, priors were assigned to all remaining parameters: the global intercepts for mortality and MI, $\alpha^\bullet$, the linear coefficients for age and time for both mortality and MI, $\beta^\bullet_\bullet$, and the precisions and mixing parameters of all included BYM2 random effects, $\tau^\bullet_\bullet$, and $\phi^\bullet_\bullet$. Priors were independently assigned across mortality, MI, and the age, space, and time dimensions:

$$\begin{aligned}
\alpha^\bullet &\sim N(0, \sigma^2 = 10) \\
\beta^\bullet_\bullet &\sim N(0, \sigma^2 = 10) \\
\phi^\bullet_\bullet &\sim \text{logitBeta}(a = .5, b = .5) \\
\tau^\bullet_\bullet &\sim \text{logTGaussian}(\mu = 0, s = \sqrt{5}).
\end{aligned} \tag{B.3.12}$$

The priors for the intercepts and linear coefficients were taken to be moderately wide mean zero Gaussian distributions. The logit-Beta and log-truncated-Gaussian priors for the BYM2 parameters were implemented in TMB and selected to align with those used by the R-INLA

package (Rue *et al.*, 2009; Martins *et al.*, 2013), which uses similar inferential approximations similar to those implemented in TMB.

Logit Beta Prior

This is a prior for a mixing parameter, $\varphi \in (0,1)$ which represented internally on the logit scale by θ:

$$\theta = g^{-1}(\varphi) = \log\frac{\varphi}{1-\varphi} = \text{logit}(\varphi)$$

with density defined on θ such that the φ has a Beta(a,b) distribution:

$$\pi(\varphi) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\varphi^{a-1}(1-\varphi)^{b-1}$$

and the density for θ can be found using standard change of variable methods for $\varphi = g(\theta) = \frac{\exp(\theta)}{1+\exp(\theta)}$:

$$\begin{aligned}
\pi(\theta) = f_\Theta(\theta) &= f_P(g(\theta)) \times |g^{'}(\theta)| \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\left(g(\theta)\right)^{a-1}\left(1-g(\theta)\right)^{b-1} \times \frac{\exp(\theta)}{\left(1+\exp(\theta)\right)^2} \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\left(\frac{\exp(\theta)}{1+\exp(\theta)}\right)^{a-1}\left(1-\frac{\exp(\theta)}{1+\exp(\theta)}\right)^{b-1} \times \frac{\exp(\theta)}{\left(1+\exp(\theta)\right)^2}.
\end{aligned}$$

Log Truncated Gaussian Prior

For the logTGaussian prior, INLA specifies the distribution for the standard deviation, $\sigma = \sqrt{1/\tau}$, but internally represents the distribution on a transformed version of this parameter to improve symmetry and approximate Gaussianity. Here, INLA internally utilizes the log precision, $\psi = \log(\tau) = \log(\frac{1}{\sigma^2})$ scale. To align TMB with this INLA prior, we again determine the density for the internally used parameter given the specification on the oft-used parameter, σ. For $\sigma = g(\psi) = \exp(\frac{-\psi}{2})$. For $\sigma \sim \text{N}(\mu, s^2)\mathbb{1}\{\sigma > 0\}$, we find the distribution of ψ to be:

$$\begin{aligned}\pi(\psi) = f_\psi(\psi) &= f_\sigma(g(\psi)) \times |g^{'}(\psi)| \\ &= \frac{1}{\sqrt{2\pi s^2}}\exp\left[\frac{-1}{2s^2}\left(g(\psi)-\mu\right)^2\right] \times \left|\frac{-\exp(\frac{-\psi}{2})}{2}\right| \\ &= \frac{1}{\sqrt{8\pi s^2}}\exp\left[\frac{-1}{2s^2}\left(\exp(\frac{-\psi}{2})-\mu\right)^2 - \frac{\psi}{2}\right].\end{aligned}$$

Note: since the truncation ensures $\sigma > 0$, this relaxes to optimizing over $-\infty < \psi < \infty$ which, no doubt, is one of the reasons the INLA developers use this transformation.

B.3.5 Model Fitting

Model fitting is performed with Template Model Builder (TMB, Kristensen *et al.*, 2016), a flexible random effects modeling package. The model is coded in a `C++` template, and random effects parameters are specified (all BYM2 field parameters, $\boldsymbol{b}^{\bullet}_{\bullet}$, of (B.3.11)). TMB fitting proceeds via maximization of the marginal joint distribution of the data, the process, and the priors shown in (2.5) having first integrated over all parameters defined to be random effects. TMB returns a vector of estimates for all parameters, and a corresponding joint precision matrix. From these, 1000 posterior samples are simulated from which all derivative quantities (age-time-country specific mortality rates, MI ratios, incidence rates, ASRs, etc) are calculated per draw before being summarized. BYM2 constraints are applied within each step of the optimization and for each sample from the posterior.

B.3.6 Model Selection

Model selection, evaluation, and assessment was performed using two likelihood metrics, the deviance information criteria (DIC) and the log pseudo-marginal likelihood (LPML), as well as metrics comparing the fitted model to the observed data such as the bias and root mean square error (RMSE) of the posterior median and the posterior predictive coverage for both

the observed incidence and mortality ASRs. The bias, RMSE, and coverage metrics were calculated across space-time, space, time, and collapsed to a single comprehensive summary metric.

The DIC is calculated as initially proposed in Spiegelhalter *et al.* (2002):

$$\text{DIC} = D(\bar{\theta}) + 2p_d,$$

where $D(\theta) = -2\log(p(\text{data}|\theta))$ is the deviance, θ is the vector of all model parameters, $\bar{\theta}$ is the posterior mean of the parameters, and $p_d = \overline{D(\theta)} - D(\bar{\theta})$ is the difference between the expected deviance and the deviance of the expectation of the posterior which approximates the effective number of parameters in the model. The DIC provides a metric of how well the likelihood fits the observed data while accounting for the tendency of more complex models to allow for closer fits to data via penalization proportional to the number of parameters. Smaller DIC values are preferable.

The LPML is defined to be the sum of the log conditional predictive ordinates, $\text{LPML} = \sum_i \log(\text{CPO}_i)$. The CPO is based on the concept of leave-one-out validation and is defined as:

$$CPO_i = \sum_i p(data_i|data_{-i}) = \left(\int [p(data_i|\boldsymbol{\theta}]^{-1} \times p(\boldsymbol{\theta}|data) \mathrm{d}\theta \right)^{-1} \approx \left(\sum_s \left[p(data_i|\boldsymbol{\theta}^{(s)} \right]^{-1} \right)^{-1}$$

where $data_{-i}$ is all observations except the i^{th}, and s indexes over the S posterior draws used to empirically approximate the posterior distribution. Values closer to zero are preferable.

The bias is in country c at time t, is calculated as the difference between the observed ASR, ASR_{ct} and the estimated ASR, $\widehat{ASR}_{ct}$: $bias_{ct} = \widehat{ASR}_{ct} - ASR_{ct}$. This is then averaged over countries, times, or both:

$$\text{Bias} = \frac{1}{C} \sum_{c=1}^{C} \left[\frac{1}{T} \sum_{t=1}^{T} \left(\widehat{ASR}_{ct} - ASR_{ct} \right) \right],$$

where the estimated ASR is calculated by averaging over $s \in \{1, \ldots, S\}$ posterior draws: $\widehat{ASR}_{ct} = \frac{1}{S}\sum_{s=1}^{S} \widehat{ASR}_{ct}^{(s)}$.

The variance is calculated as

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

and the RMSE=$\sqrt{\text{Bias}^2 + \text{Variance}}$. In general, there is a trade-off between models which reduce bias at the cost of increased variability when additional parameters are added to a model but which do not contribute to the fit, and RMSE is a metric which balances these competing outcomes. For this particular application, bias is more complex due to the two interacting outcomes.

Lastly, we calculate the model's average posterior predictive coverage (PPC). This assesses if new datasets simulated from the fitted model are similar to the observed data. From S posterior predictive draws $\widetilde{ASR}_{ct}^{(s)}$ is sampled from $p(ASR_{ct}|\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{N})$, and for a particular probability, α, the lower $l_\alpha = \alpha/2$ and upper $u_\alpha = 1 - \alpha/2$ quantiles are approximated from the draws, yielding the $(100-\alpha)\times 100$ posterior predictive credible interval: $(\widetilde{ASR}_{ct}^{l_\alpha}, \widetilde{ASR}_{ct}^{u_\alpha})$. The coverage for this country-time is a binary outcome indicating if an observation is contained within the corresponding interval, $\mathbb{1}_{\{\widetilde{ASR}_{ct}^{l_\alpha} \leq ASR_{ct} \leq \widetilde{ASR}_{ct}^{u_\alpha}\}}$. The PPC^α for each country, time, or for the entire model is found by averaging over one or both of these dimensions:

$$\text{PPC}^\alpha = \frac{1}{C}\sum_{c=1}^{C}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbb{1}_{\{\widetilde{ASR}_{ct}^{l_\alpha} \leq ASR_{ct} \leq \widetilde{ASR}_{ct}^{u_\alpha}\}}\right],$$

Figure 4.4 displays the DIC, effective number of parameters, and the LPML across the ten models. Table 4.3 contains the bias, standard deviation, RMSE, and $\text{PPC}^{.9}$ of the ASRs for incidence and mortality. Based on the model selection criteria, Model IX was selected. It demonstrated DIC, LPML, and posterior predictive coverage comparable those of model X, the most complicated model and the one which technically performs best in these metrics. Generally, Model IX performed well across all metrics, exhibiting the lowest mortality ASR

bias and RMSE. While model IX did not perform best in any of the incidence metrics, it has the second best and reasonable PPC[.9], indicating that it adequately describes the variability in observed incidence ASR, Other models that performed better in incidence ASR bias, standard deviation, and RMSE have more severe PPC undercoverage. As the final deciding factor, we note that the effective number of parameters for model IX is approximately 140 less than that of model X, and that model IX has a notably smaller bias, standard deviation and RMSE for the incidence ASR as compared to model X. The rest of this study focuses on the fit of model IX to the European breast cancer data from 2000–2017.

B.4 IARC and IHME GBD Methods

B.4.1 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 2018 estimates have been published (Ferlay *et al.*, 2018) and the statistical methods were discussed thoroughly in GLOBOCAN 2018, 2012 and GLOBOCAN 2008 (Ferlay *et al.*, 2019, 2013, 2010a).

The GLOBOCAN estimation and projection of country-specific rates can be divided into three categories based on available data. The NORDPRED method, an age-period-cohort model described in Møller *et al.* (2003) and which requires at least 15 consecutive years of data, is used to project the most recently available country-specific rates from national or subnational levels to 2018. For countries with 6-10 years of data available, the DEPPRED method, a linear time prediction model developed by IARC and based on Dyba and Hakulinen (2000), was used to generate estimates. Cancer- and sex-specific predictive models were only fit when the country had at least 50 all-age cancer cases or deaths recorded per year. If none of these conditions were met, 2018 rates were taken to be the annual average rates recorded in the last 3-5 year period with at least 20 all-age cases or deaths

were recorded. For countries with no historical data, the rates were taken to be the most recent available.

For the 40 European countries discussed by Ferlay *et al.* (2018), all had historical national cancer mortality data available through the WHO mortality database for at least the 2004-2016 period, allowing use of the NORDPRED or DEPRED method. Populations were extracted from the same database, unless they were not available in which case they were taken from the UN population

Twenty-four of the countries had national and local incidence data to use to estimate 2018 incidence. Twenty-three of the countries, twenty-two with national incidence data, had sufficient reporting to project recorded incidence rates to 2018. The most recent data from Slovakia, recorded in 2010, were not recent enough to use for projections and were the rates from 2010 were used as a stand-in proxy for 2018. When national incidence estimates were not available, the approach for estimating country-specific incidence rates relied on using sex-, site-, and age-specific incidence to mortality (IM) ratios (Y^L_{ac}/Z^L_{ac}) and the 2018 national mortality estimates to estimate the 2018 national incidence rates:

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

where the IM ratios were calculated by aggregating age- and country-specific local registry incidence data. Aggregation was performed by weighting incidence and mortality by the square-root of corresponding population. Poisson regression models with terms for sex and age were then used to estimate IM ratios, $\widehat{Y^L_{ac}/Z^L_{ac}}$. Seven countries received country-specific models using the most recent available registry data, usually from a 5-year period centered on 2010. By assuming that the IM ratio would be relatively constant in time, this same rate was used to estimate 2018 national incidence. In some cases, these results were deemed unrealistic and the most recent incidence rates from the registries were used as a proxy for the 2018 national incidence rate. For the nine remaining countries with either no incidence data or incidence data lacking sufficient quality, regional models were fit using IM ratios

generated by aggregating cancer cancer registry data from neighboring countries. Ferlay *et al.* (2018) provide summarization of the total estimated cases and deaths as well as the European age standardized rates. The incidence ASR for country c in year t is calculated as:

$$ASR_{ct} = \sum_{a=A}^{A} p_{act} \times N_a^s \tag{B.4.1}$$

where $A = 18$ for the eighteen age groups (0-4,..., 80-84, 85+) and N_a^s are the European age standard populations are shown in Figure B.4.1.

Starting with their 2018 estimates, IARC introduced a method to calculate uncertainty intervals for their estimates. The intervals are created for all estimated sex- and site-specific new cancer cases and cancer deaths for all ages. They calculate standard errors for the observed incidence and mortality rates that are used in estimation on the log scale and then construct 95% uncertainty intervals on the arithmetic scale using the following formulas Ferlay *et al.* (2019) :

$$\text{UI}_{lower} = \exp(\log(CR2018_c - 1.96 \times se) \times P2018_c/100000 \tag{B.4.2}$$

$$\text{UI}_{upper} = \exp(\log(CR2018_c + 1.96 \times se) \times P2018_c/100000, \tag{B.4.3}$$

where $\tilde{se}$ is a bias-corrected version of the standard error (se) for the estimated crude incidence or mortality rate per 100,000 in 2018, $CR2018_c$ for country c and $P2018_c$ is the population of country c in 2018. These intervals are calculated for each cancer and sex for which they produce estimates. The se is corrected for three causes of bias: population coverage, lag time to available data if data is not available from 2018 for this calculation, and the quality of the data. Each of these biases are assumed to be equally important and each country is given a categorical score ranging from 0 (high) to 10 (low) for these three sources

of bias. Using these bias scores, the se is adjusted to:

$$\tilde{se} = se \times \frac{100}{100 - c} \times \frac{100}{100 - t} \times \frac{100}{100 - q},$$

where c is the categorical coverage score, t is the categorical time lag score, and q is the categorical quality score.

B.4.2 GBD Methods

The article 'Breast and cervical cancer in 187 countries between 1980 and 2010: a systematic analysis' by Forouzanfar *et al.* (2011) and the associated web appendix provide details into GBD's methods to generate reliable estimates of worldwide breast and cervical cancer incidence and mortality. They approach the modeling of incidence and mortality using a similar underlying generative model to the one we have proposed, relying on the MI ratio to infer unobserved mortality and incidence. Since that study, they have published a number of cancer papers which estimate mortality, incidence, and other metrics for a wider variety of cancer types and which use similar but updated methods. Unlike our approach, their procedure involves numerous independent fitting steps which are combined together with through a stages of ad-hoc and post-hoc manipulation.

Their most recent article details global estimates for 5 metrics across 29 cancer groups and 10 years, 'Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life Years for 29 Cancer Groups From 2010 to 2019: A Systematic Analysis for the Global Burden of Disease Study 2019,' and was published along with a 226 page online supplemental (Kocarnik *et al.*, 2021). One very important distinction between their earlier work and more recent publications is that cancer incidence and mortality, including estimates for breast and cervical cancer, have been integrated into the Global Burden of Disease (GBD) study. The GBD study and its regular updates are a huge and laudable endeavor, providing much needed estimates of many causes of death and risk factors across age, sex, time, and location. Unfortunately, their methods are often opaque with unknown

statistical properties.

Including cancer (or any cause of death) in the GBD framework has has many repercussions for the estimates, including the forced alignment of mortality counts from all causes to sum to the global all-cause annual mortality envelope that is estimated at an early stage of every GBD update. This constraint necessarily introduces negative correlation between the mortality estimates across all causes of death and makes the work effectively impossible to reproduce. The following sections provide an overview of GBD's most recent cancer incidence and mortality estimation procedure as it is described by Kocarnik *et al.* (2021), in their main text, online supplemental content, and numerous other GBD papers that these authors reference. We attempt to describe the process in some detail since it can be difficult to piece together their modeling approach from the numerous and long appendices that they self-reference.

Cancer Modeling

Kocarnik *et al.* (2021) describe 5 ways in which the cancer estimation for GBD 2019 changed since GBD 2017. These include adding new data sources, improving data processing, particularly for liver cancer, changes in the definition of the youngest age group estimated to align with cancer registry age patterns, updating modeling parameters to "perform additonal smoothing of mortality-to-incidence (MIR) estimates across age and time, reducing improbable variaton from sparse data," and updating the cancer survival methods to estimate age-specific survival curves and to "improve uncertainty estimates". The updates to the modeling parameters is of particular relevance to this comparison and is discussed further below.

GBD 2019 used 929,193 cancer-, location-, and year-specific sources of data in their cancer analysis. Of these 767,514 (>80%) were from vital registration systems, 155,542 (∼15%) from cancer registries, with the remainder (<1%) coming from verbal autopsy reports. Among these breast cancer was included in 23,378 vital registration observations, 5,333 cancer registry observations, and 515 verbal autopsy reports.

The GBD 2019 cancer mortality estimation procedure consisted of two primary steps: estimating MI ratios, followed by a separate modeling step to estimate cancer mortality. Incidence was then estimated in a third modeling step.

MI Ratio Modeling

In their first step, the MI ratio, r, is estimated. This step uses only matched incidence and mortality data. Their approach uses what GBD calls a space-time Gaussian process regression (ST-GPR) approach with the GBD-estimated health care access and quality index (HAQI) as a single covariate and a logit link function:

$$\text{logit}(r_{c,a,s,t}) = \alpha + \beta_1 \text{HAQI}_{c,t} + \sigma_a^A \beta_2 \mathbb{1}_a + \beta_3 \mathbb{1}_s + \epsilon_{c,a,s,t} \tag{B.4.4}$$

for country (or subnational) c, age a, sex s, and time t, where $\mathbb{1}$ represents binary indicators and $\epsilon_{c,a,s,t}$ is the "error term." The ST-GPR has three main parameters that "control for smoothing across time, age, and geography." In addition to presenting the above equation and referencing another paper which describes the ST-GPR method (Vos *et al.*, 2020), Kocarnik *et al.* (2021), in their web supplemental, only supply the following information about this estimation step:

> Predictions were made without the random effects. The ST-GPR model has three main hyper-parameters that control for smoothing across time, age, and geography. These hyper-parameters were adjusted for GBD 2019 in order to improve model performance in locations with sparse data. The time adjustment parameter lambda (λ) aims to borrow strength from neighboring time points (i.e., the value in this year is highly correlated with the value in the previous year but less so further back in time). For GBD 2019, lambda was lowered from 2 to 0.05, increasing the weight of more distant years. The age adjustment parameter omega (ω) borrows strength from data in neighboring age groups and was lowered from 1.0 to 0.5, increasing the weight of more distant age groups. The space adjustment parameter zeta (ζ) aims to borrow strength across the hierarchy of geographical locations. Zeta was lowered from 0.95 to 0.01, reducing the weight of more distant geographical data at the region or super region level. For the remaining parameters in the Gaussian process regression, we lowered the amplitude from 2 to 1 (reducing fluctuation from the mean function) and reduced the scale value

> from 15 to 10 (reducing the time distance over which points are correlated). Compared to GBD 2017 models, these model specification changes generally led to more smoothing of the MI ratio estimates across age and time, and less geographic smoothing at the region or super region level.
>
> Data-cleaning steps for MIR estimation were similar to those for GBD 2017. For each cancer, MIRs from locations in HAQ Index quintiles 1-4 were dropped if they were below the median of MIRs from locations in HAQ Index quintile 5. We also dropped MIRs from locations in HAQ Index quintiles 1-4 if the MIRs were above an outlier threshold calculated as the third quartile + 1.5 * IQR (inter-quartile range). We dropped all MIR data that were based on fewer than 15 incident cases to avoid excessive variation in the ratio due to small numbers (this threshold was 25 cases in GBD 2017, but was lowered in GBD 2019 in order to include additional data). An exception to this threshold was made for mesothelioma and acute myeloid leukemia, where instead we dropped MIRs that were based on fewer than ten cases because of lower data availability for these two cancers. For the lower end of the age spectrum where cancers are generally rarer, we also aggregated incidence and mortality to the youngest five-year age bin where SEER reported at least 50 cases from 1990 to 2015, to avoid unstable MIR predictions in young age groups because of too few data. The MIR estimates in this SEER-based minimum age-bin were then copied down to all younger GBD age groups estimated for that cancer.

Based on this, we note that they are dropping outliers and that they are selecting (not fitting) the smoothing parameters.

The GBD ST-GPR method is used across most of the hundreds of diseases and injuries for which they produce estimates. The main ST-GPR reference paper cited by Kocarnik *et al.* (2021), Vos *et al.* (2020), only has one sentence about ST-GPR in the main methods section, noting "ST-GPR is a set of regression methods that borrow strength between locations and over time for single metrics of interest, such as risk factor exposure or mortality rates." In section 4.3.3 on page 440 of the 1813 appendix 1, Vos *et al.* (2020) describe the ST-GPR process in more detail. While the description is detailed and explicit, it is an ad-hoc multi-stage modeling process whose statistical properties are not obviously understood nor, to the best of our knowledge, have they been assessed.

Before the ST-GPR details, we provide an overview of the multiple regressions and calculations that comprise the ST-GPR process to orient the reader. First, a linear model is fit to the logit of the MI ratios, and the residuals from this fit are calculated. A LOESS smoothing

curve is then applied to the residuals, and the sum of the linear fit and LOESS is taken as a preliminary approximation to the logit MI ratios and will also be used as the mean function of the final Gaussian Process (GP) used to predict logit MI ratios. The variance of the error term for each data point is then calculated, and finally the hyperparameters of the GP covariance function are calculated. Each of these steps is fit independently, but conditionally on the previous steps, and there does not appear to be any uncertainty propagation across these steps.

In the cancer setting, the ST-GPR is used to estimate a smooth prediction over time for the logit of the MI ratios:

$$\widehat{\text{logit}(r_{c,a,s,t})} = g_{c,a,s}(t) + \epsilon_{c,a,s,t} \tag{B.4.5}$$

$$\epsilon_{c,a,s,t} \sim N(0, \sigma^2_{\epsilon_{c,a,s,t}}) \tag{B.4.6}$$

$$g_{c,a,s}(t) = GP(\alpha + \beta_1 \text{HAQI}_{c,t} + \sigma_a^A \beta_2 \mathbb{1}_a + \beta_3 \mathbb{1}_s + h(\delta_{c,a,s,t}), \Sigma) \tag{B.4.7}$$

where $\delta_{c,a,s,t}$ are the residuals calculated from the linear regression in (B.4.4), which comprises one component of the mean of the Gaussian Process (GP). A LOESS locally weighted polynomial regression smoothing function, $h(\delta_{c,a,s,t})$, is fit to the residuals in an attempt to "to systematically estimate this residual variability by borrowing strength across time, age, and space patterns (the spatiotemporal component of ST-GPR)" (Vos *et al.*, 2020, Appendix 1 p. 441). This LOESS curve is a function of the three smoothing parameters mentioned earlier, λ, which smooths time, ω, which smooths age, and ζ, which smooths across the hierarchy of spatial locations (which is not the same as smoothing across spatial distances). These parameters determine the contribution of each data observation in the weighted polynomial regression. Σ is the covariance of the GP.

(Vos *et al.*, 2020, Appendix 1 p. 443) state that the variance of the error term in (B.4.6) is found using either measures of uncertainty from the data observation, or were imputed using

$$\tilde{\sigma}^2_{\epsilon_{c,a,s,t}} = \text{logit}^{-1}(m_{c,a,s,t}) \times (1 - \text{logit}^{-1}(m_{c,a,s,t}))/n_{c,a,s,t}, \tag{B.4.8}$$

where $m_{c,a,s,t}$ is the mean function of the GP in (B.4.7) and is being used as an initial proxy for the MI ratio. This variance was then "transformed into logit-space by using a delta method approximation":

$$\sigma^2_{\epsilon_{c,a,s,t}} = \tilde{\sigma}^2_{\epsilon_{c,a,s,t}} \times \left[\text{logit}^{-1}(m_{c,a,s,t}) \times (1 - \text{logit}^{-1}(m_{c,a,s,t}))\right]^{-2}. \tag{B.4.9}$$

The authors then note that prior to the actual GPR, an approximation of non-sampling error was added to this variance. This was calculated as "the variance of inverse-variance weighted residuals from the space-time estimate at a given location-level hierarchy. If there were < 10 data points at a given level of the location hierarchy, the non-sampling variance was replaced with that of the next highest geography level with > 10 data points."

Lastly, the covariance of the GP needs to be defined. The covariance of is assumed to be Matérn with smoothness parameter $\nu = 2$. The marginal variance, σ^2_m was approximated by taking "the normalised median absolute deviation $MADN(\delta'_c)$ of the difference, which is the normalised absolute deviation of the difference of the first-stage linear regression estimate from the second-stage spatiotemporal smoothing step for each country. We then took the mean of these country-level MADN estimates for all countries with 10+ country-years of data to ensure that differences between first- and second-stage estimates had sufficient data to truly convey meaningful information on model uncertainty."

After all these steps are completed and all parameters of (B.4.5)-(B.4.7) have been sequentially approximated, dense predictions of the logit of the MI ratios across location, age, sex and time are sampled from

$$\widehat{\text{logit}(r_{c,a,s}(t))} \sim N(\alpha + \beta_1 \text{HAQI}_{c,t} + \sigma^A_a \beta_2 \mathbb{1}_a + \beta_3 \mathbb{1}_s + h(\delta_{c,a,s,t}), \sigma_{\epsilon_{c,a,s,t}} \boldsymbol{I} + \Sigma). \tag{B.4.10}$$

1000 samples are taken from this distribution, and the final estimates are taken to be the mean of the samples with the 2.5% and 97.5% quantiles taken to be the lower and upper bounds for the uncertainty estimates.

Estimation of Trends in Mortality by Age

A large amount of vital registration data is processed for use in mortality estimation. In addition, mortality estimates from registry incidence are generated by combining the recently estimated MI ratios and the registry incidence: $\widehat{\text{MI}} \times \text{incidence}_{\text{registry}} = \widehat{\text{mortality}}_{\text{CR input}}$. These mortality estimate are then smoothed using what Kocarnik *et al.* (2021) a "Bayesian noise-reduction algorithm," detailed in (Vos *et al.*, 2020, Appendix 1 Section 2.14), to "deal with zero counts." These smoothed mortality-registry estimates are then combined with vital registration and verbal autopsy mortality data and mortality modeling proceeds via the general GBD Cause of Death Ensemble model (CODEm) process.

Kocarnik *et al.* (2021) provide a summary of CODEm:

> In brief, the CODEm approach is based on several principles: that all types of available data should be used, even if data quality varies; that a diverse set of plausible models with different combinations of covariates should be evaluated; that both individual models and the overall ensemble models should be tested for their predictive validity; and that the best model or sets of models should be chosen based on the out–of-sample predictive validity
>
> Covariates are provided for potential use in the ensemble based on a possible predictive relationship between the covariate and the specific cancer mortality, with an expected level and direction of association. Generally, Level 1 covariates have a proven strong relationship with the outcome, such as etiological or biological roles. Level 2 covariates have a strong relationship but not a known direct biological link. Level 3 covariates have a relationship that may be more distal in the causal chain, or are mediated through Level 1 or 2 covariates. The covariates provided to CODEm, as well as their level and direction, differ by cause and sex. ...
>
> To generate an ensemble model, CODEm generates submodels that evaluate all plausible relationships between covariates and the response variable. Three additive components of data variance are used in CODEm: sampling variance, non-sampling variance, and garbage code redistribution variance. Model performance of all models is evaluated through out-of-sample predictive validity tests. Ensemble models are constructed from the individual models, with the contribution of individual models to the ensemble

> weighted by the basis of their predictive validity ranking. The final ensemble contains 1000 draws from these individual component models, from which a mean estimate and a 95% uncertainty interval are calculated. The 95% uncertainty interval represents the 0.025 and 0.975 quantiles of the draws.

Once the ensemble model estimates for mortality have been generated, they undergo one additional processing step. As previously mentioned, each GBD update depends on an all-cause global mortality estimate. The CODEm estimation procedure for each cause are not constrained at the time of fitting to sum to the global envelope – indeed, the CODEm ensembles for different causes are fit independently and are fit independently from the all-cause mortality estimate. To ensure that the sum of all mortality equals the global estimate, and that all child-causes sum to their respective parent causes, GBD employs their cause of death correction process, CoDCorrect. The core of the correction algorithm has remained the same since GBD 2013. In Appendix 1 section 3.3.2.2 Vos *et al.* (2020) describe the correction equation as:

$$CD_{lyasjd} = D_{lyasjd} \frac{PD_{lyasjd}}{\sum_{j=1}^{j=k} D_{lyasjd}} \tag{B.4.11}$$

where CD is the corrected deaths, D is the uncorrected deaths from the CODEm ensemble, and PD is the deaths in the parent cause of death, all indexed over location l, year y, age a, sex s, cause j, and sample draw d. The CoDCorrect algorithm first rescales the Level 1 causes to match the all cause mortality (taken to be PD at this stage). Then it proceeds to the next level, rescaling Level 2 causes to their already corrected Level 1 parent causes. This continues until the complete cause hierarchy has been compressed into the all cause mortality envelope. This process is done at the draw level, for different draw objects that were independently generated but which should be correlated. It is unclear what downstream effects this may have on the estimates and their uncertainty.

Estimation of Trends in Incidence by Age

The last relevant step for this study is the generation of incidence estimates. The corrected mortality estimates that result from the CODCorrect algorithm are transformed to incidence estimates using the already estimated MI ratios. To do this, 1000 corrected mortality estimate draws are divided by 1000 estimated MI ratio draws to generate 1000 estimated incidence draws. Kocarnik *et al.* (2021) notes that this step assumes that the uncertainty in the MI ratio draws is independent from the uncertainty in corrected mortality estimates.

Problems with the GBD estimation procedure

In an ideal world we would be able to write down the conditional independence assumptions implied by the methods descriptions and flowcharts from Kocarnik *et al.* (2021) and Vos *et al.* (2020) papers and appendices. Unfortunately, their methods require multiple sequential fits with later stages conditioning on estimated quantities and at times their estimation appears circular (for example, using the MI ratio and incidence data to estimate mortality, and then using the MI ratio and mortality to estimate incidence). Although they claim all uncertainty is propagated between steps in the modeling process, it is clear that uncertainty from every modeled quantity is not passed forward every time that quantity is used. Even if draws were always used to propagate uncertainty at every stage for every estimated quantity, the draws were simulated independently for quantities that are often known to be correlated. The GBD method is complicated, unreprodicble, and ad-hoc and as such has unknown statistical properties. GBD fails to demonstrate the validity of their uncertainty estimates which may be more dangerous than not publishing them at all.

Methods Discussion

Ferlay *et al.* (2018) 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 and they provide 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 and that most countries are fit independently from others, missing an opportunity so pool information.

The methods described in Kocarnik *et al.* (2021) and Vos *et al.* (2020) contain many complex steps. The GBD studies use a vast amount of data sources for the incidence and mortality and incorporate many country-level covariates, which may add precision to the estimates. Many of the covariates are also modeled at IHME, and it is unclear if uncertainty from the covariates is propagated to later modeling steps. Generally, GBD uncertainty intervals are generated based on simulated draws from an approximate posterior distribution. An extensive appendix is provided with details related to the modeling procedure, however there are many complex steps, and while they claim that they use draw-level calculations to propagate uncertainty between steps, but from their description it does not seem that this happens every time a modeled quantity is used in a subsequent modeling stage. Some parameters are fixed (or selected from by assessing a small candidate pool of hand-selected options) and appear to be selected to align the estimates with the expectations of the GBD researchers. It is unlikely that the results for a single outcome could be externally reproduced, even with access to the same data sources. The CoDCorrect step applied to all causes of death guarantees this. While they do provide uncertainty intervals for their estimates, the complicated procedure begs the question of whether the uncertainty intervals actually represent what they claim and if they have any real meaning at all.

The model implemented in this study overcomes many of the limitations of both IARC and GBD methods, providing a reproducible, straight-forward, unified modeling process with proper uncertainty intervals whose statistical properties have been assessed and verified.

B.5 Additional Results

Comparison against IARC and GBD estimates

Figure B.5.1 provides a comparison between our 2017 incidence and mortality ASR estimates and 95% credible intervals and the 2018 estimates IARC has published. Using the same data as IARC we see very similar estimates across most countries. Only three of IARC's incidence ASR estimates for Romania, Germany, and Iceland fall outside our 95% credible intervals, and the IARC estimates for Iceland and Germany do not fall far outside our credible intervals. Relatively speaking, the estimates for Romania are the furthest off, but Romania has low registry coverage and IARC pools data from select countries' cancer registries (Bulgaria, Romania, Slovakia and Poland) to estimate the incidence/mortality ratios in Romania. Since estimating incidence is the primary objective of this project, it is promising that the new methods presented here, which are quite different than those used by IARC, yield very similar incidence estimates. Interestingly, there are larger differences between our mortality estimates, even though there is much more recent mortality data and the mortality data influences the well-aligned incidence estimates. Most notably, we see the largest mortality discrepancy in Montenegro, which is missing mortality data after 2009, and which IARC estimates to be quite high. Generally, we see our estimates appear to smooth more than the IARC estimates, which is to be expected.

Figure B.5.2 provides a similar comparison, contrasting our estimates and 95% credible intervals against the GBD 2019 estimates and 95% uncertainty intervals, which uses both different methods and different data. In contrast to the IARC comparison, our estimates for mortality mostly align well with the GBD mortality estimates. Furthermore, the uncertainty intervals from both methods overlap in the vast majority of countries. On the other hand, GBD's estimates for incidence are regularly below our estimates and very often have tighter uncertainty intervals. In addition, this also implies that GBD must be generally producing higher MI ratios than we produce. GBD's relatively tight incidence uncertainty is surprising because the incidence estimates that GBD produces are generated in a later step than the

mortality estimates – and we would expect propagated uncertainty to grow as you proceed further down in the algorithm.

There are large differences between the data used in this project and by GBD, which makes it difficult to compare the results against each other. Incidence estimates are harder to compare than mortality estimates since the quality and coverage of registries and incidence data vary more often than mortality data, which is frequently available at the national level. Figure B.5.3 attempts to qualitatively assess the uncertainty intervals from both methods by plotting the width of the mortality uncertainty intervals against the national population. Knowing that many countries have vital registration systems that provide national mortality counts, it is not unreasonable to expect a negative relationship between population size and mortality interval widths. We see on the right side of this figure that our method generates estimates where smaller countries tend to have more uncertain estimates. The notable exception to this is Greece. Although relatively populous, we only have a few years of data from Greece, a far smaller than the typical number of available years of data compared to most countries, which explains this discrepancy. The left side of this figure shows the analogous plot for GBD estimates, and we see a more varied pattern where the population seems less correlated with the width of the mortality uncertainty intervals. Although this is unexpected, the GBD model is quite complex and there could be valid reasons for this unusual pattern, such as varying power of the predictive covariates used to estimate mortality. The problem is, with a complex unreproducible model that exhibits unintuitive behavior, (this pattern in mortality intervals and the small incidence uncertainty intervals are both not easily explained) the burden is on the estimators to demonstrate the statistical properties of their estimates which the GBD project has failed to do.

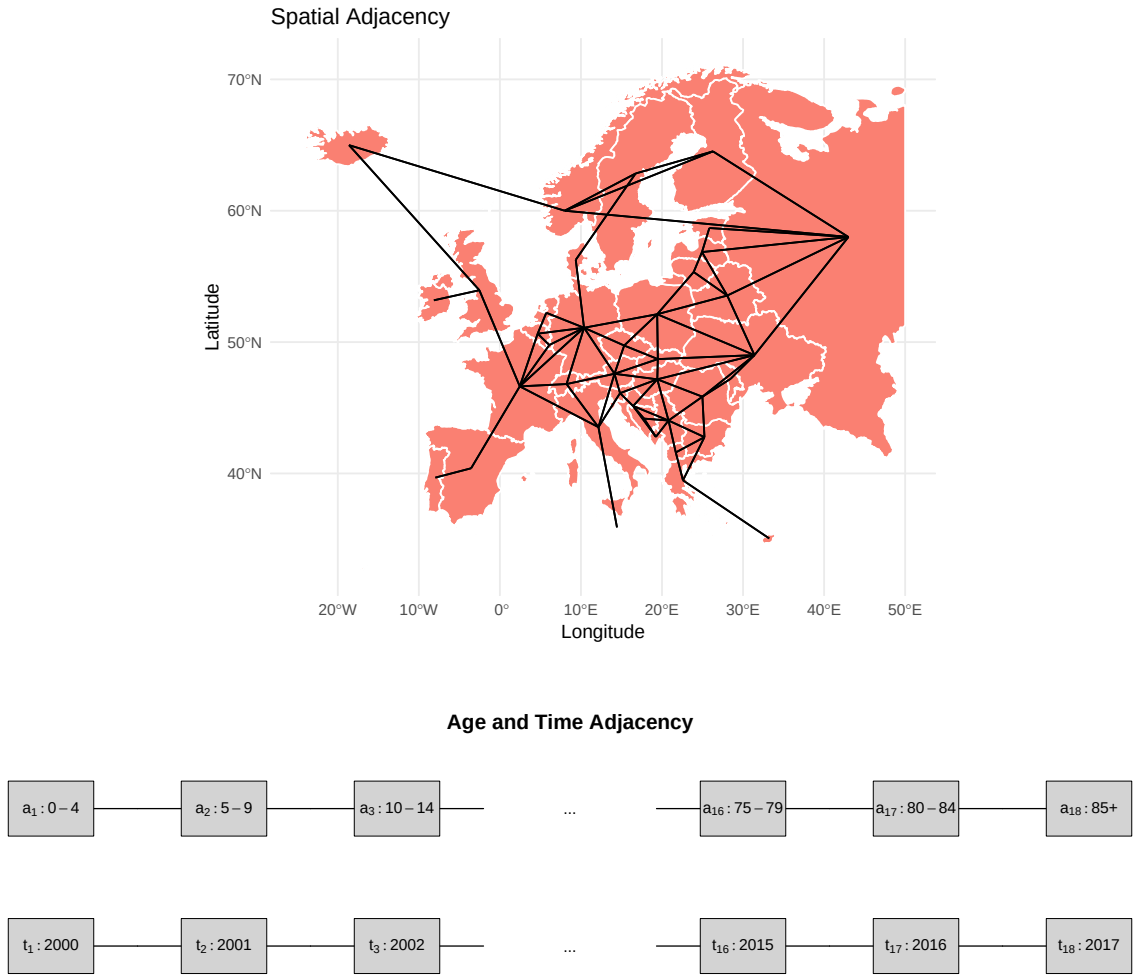


Figure B.3.1: Adjacency structures used in the spatial, age, and temporal BYM2 random effects models.

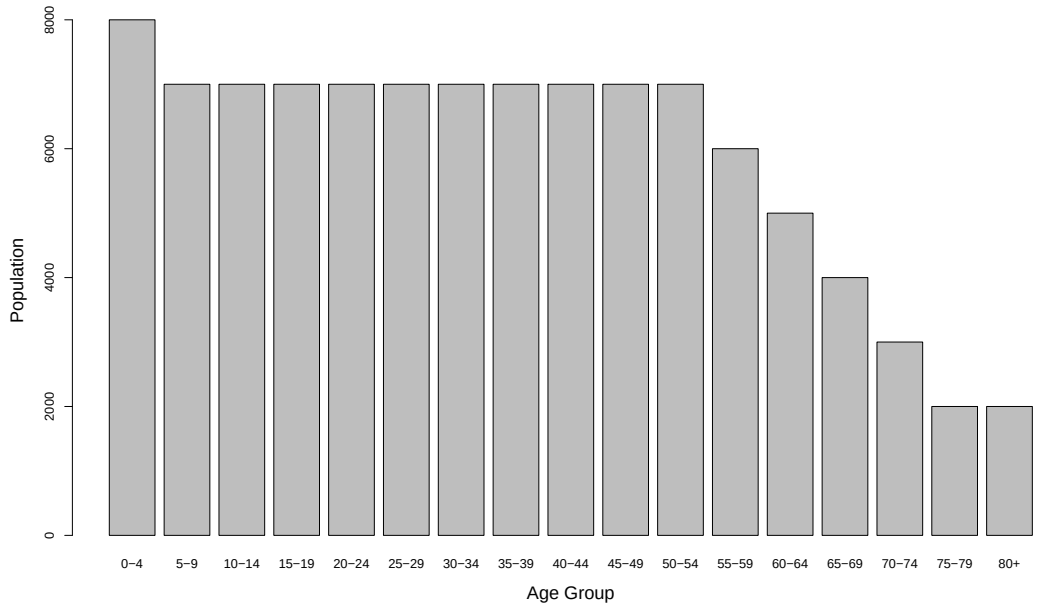


Figure B.4.1: European age standardized population.

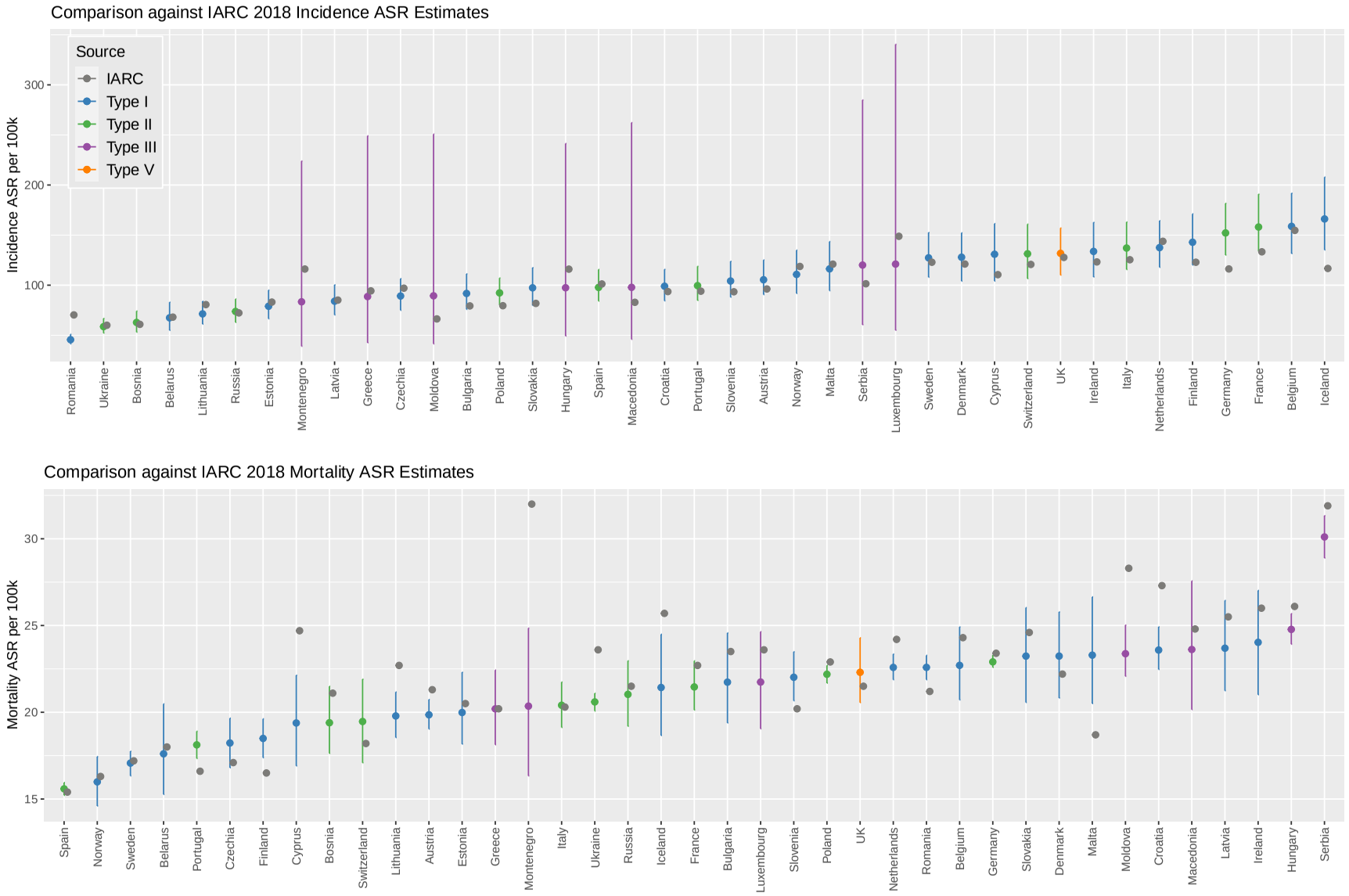


Figure B.5.1: Comparison to published breast cancer incidence (top) and mortality ASRs from IARC. Points in black represent published estimates for 2018 by IARC. Points and intervals in blue, green, purple, and orange represent our estimates with colors corresponding to the best data type available for each country between 2000 and 2017, as shown in Figure B.1.1.

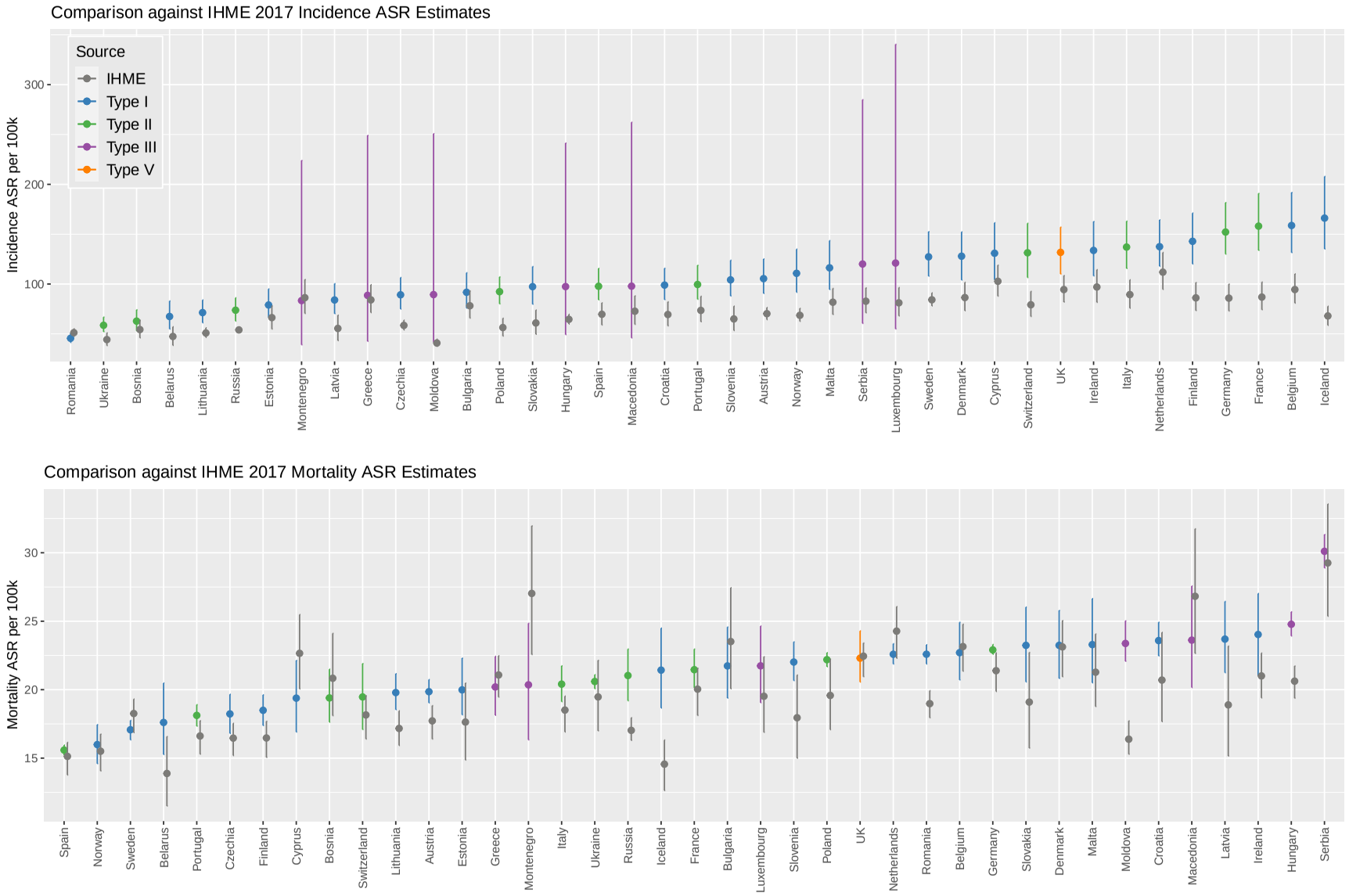


Figure B.5.2: Comparison to published breast cancer incidence (top) and mortality ASRs from IHME. Points in black represent published GBD estimates for 2017 as extracted from the GHDx data base. Points and intervals in blue, green, purple, and orange represent our estimates with colors corresponding to the best data type available for each country between 2000 and 2017, as shown in Figure B.1.1.

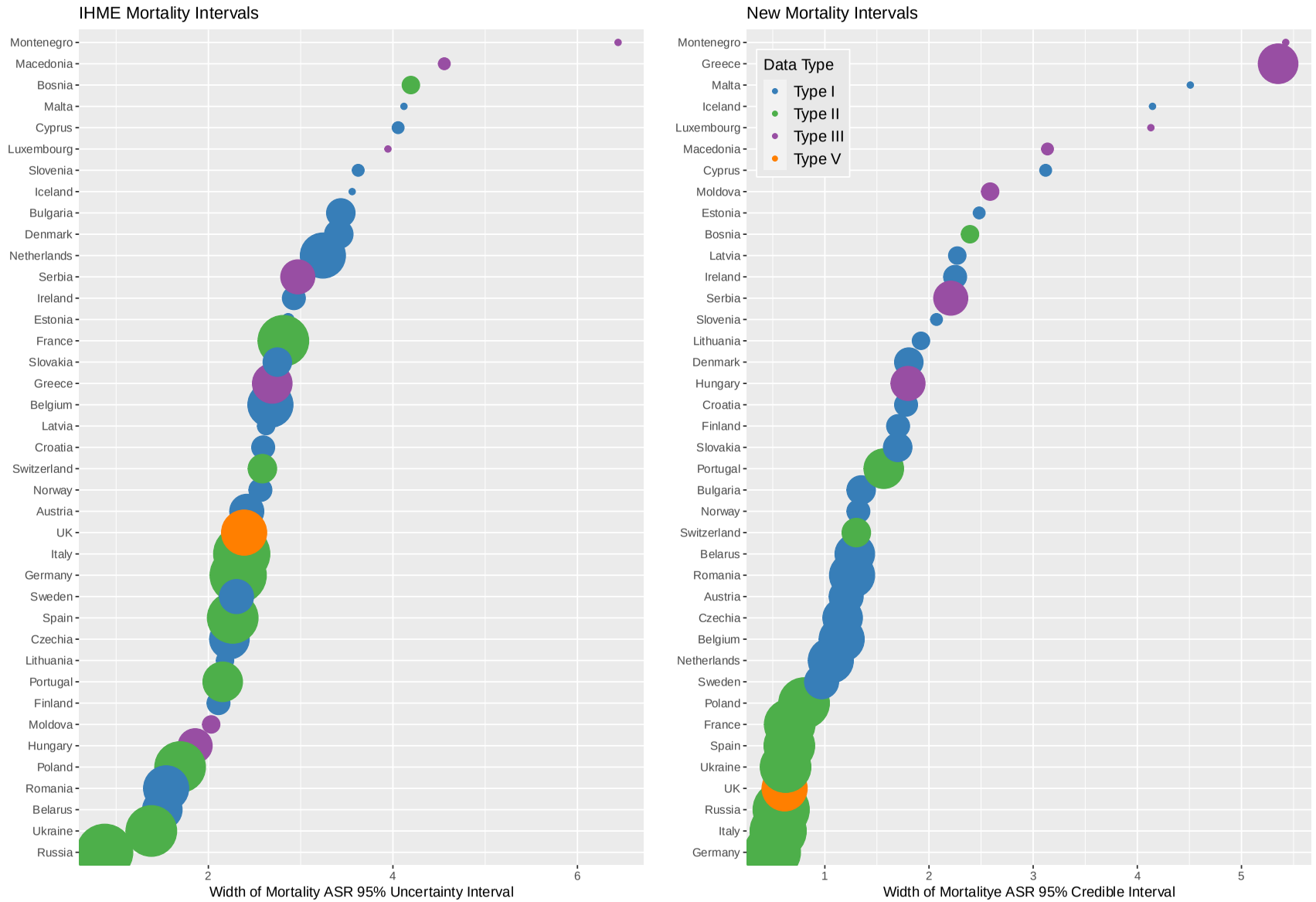


Figure B.5.3: Comparison to published breast cancer incidence (top) and mortality ASRs from IHME. Points in black represent published GBD estimates for 2017 as extracted from the GHDx data base. Points and intervals in blue, green, purple, and orange represent our estimates with colors corresponding to the best data type available for each country between 2000 and 2017, as shown in Figure B.1.1.

Residual plots

Figures B.5.4 and B.5.5 show Pearson residuals for incidence and mortality for our selected model (model IX), plotted against age groups 1-18, with colors for different years of data. Loess smoothing curves have been overlaid to help differentiate patterns across years, though they also tend to make it look as if stronger patterns exist than the individual residuals would suggest. Overall, the magnitude of the residuals is usually less than 2, suggesting few outliers that our model cannot explain. Figures B.5.6 and B.5.7 show the same residual plots for the base model (model I) to help illustrate the need for the interactions and how much structured variability remained before they were included. In the base model, we frequently see residuals of magnitudes exceeding 5 and there are clearly country varying age-time structures in these residuals.

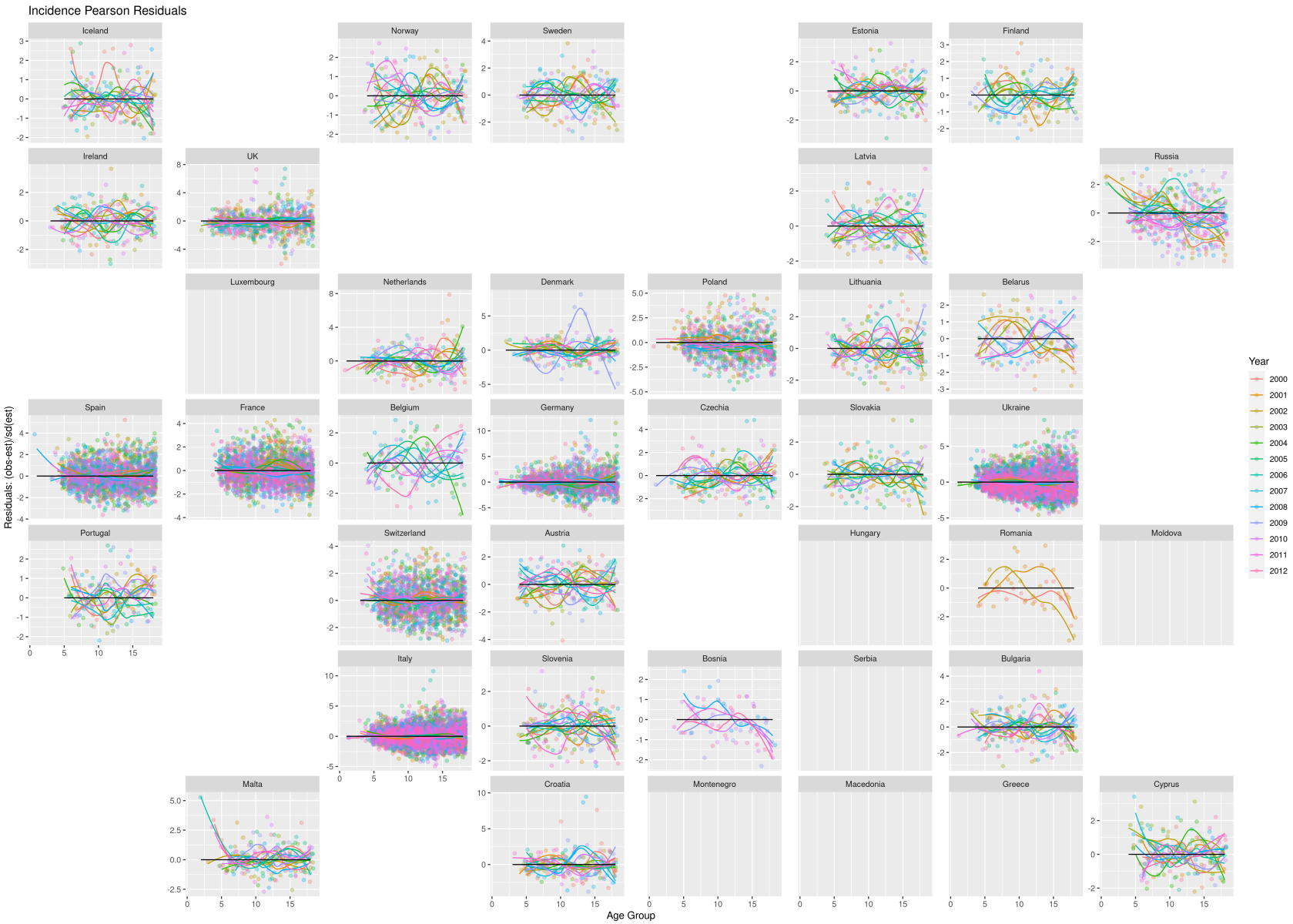


Figure B.5.4: Pearson residuals from model IX (final selected model) for each age, location, and time incidence observation. The residuals are faceted by country, plotted against age groups 1-18 (0-5, ..., 80-84, 85+), and colors and smooth loess curves showing residuals from different years of data.

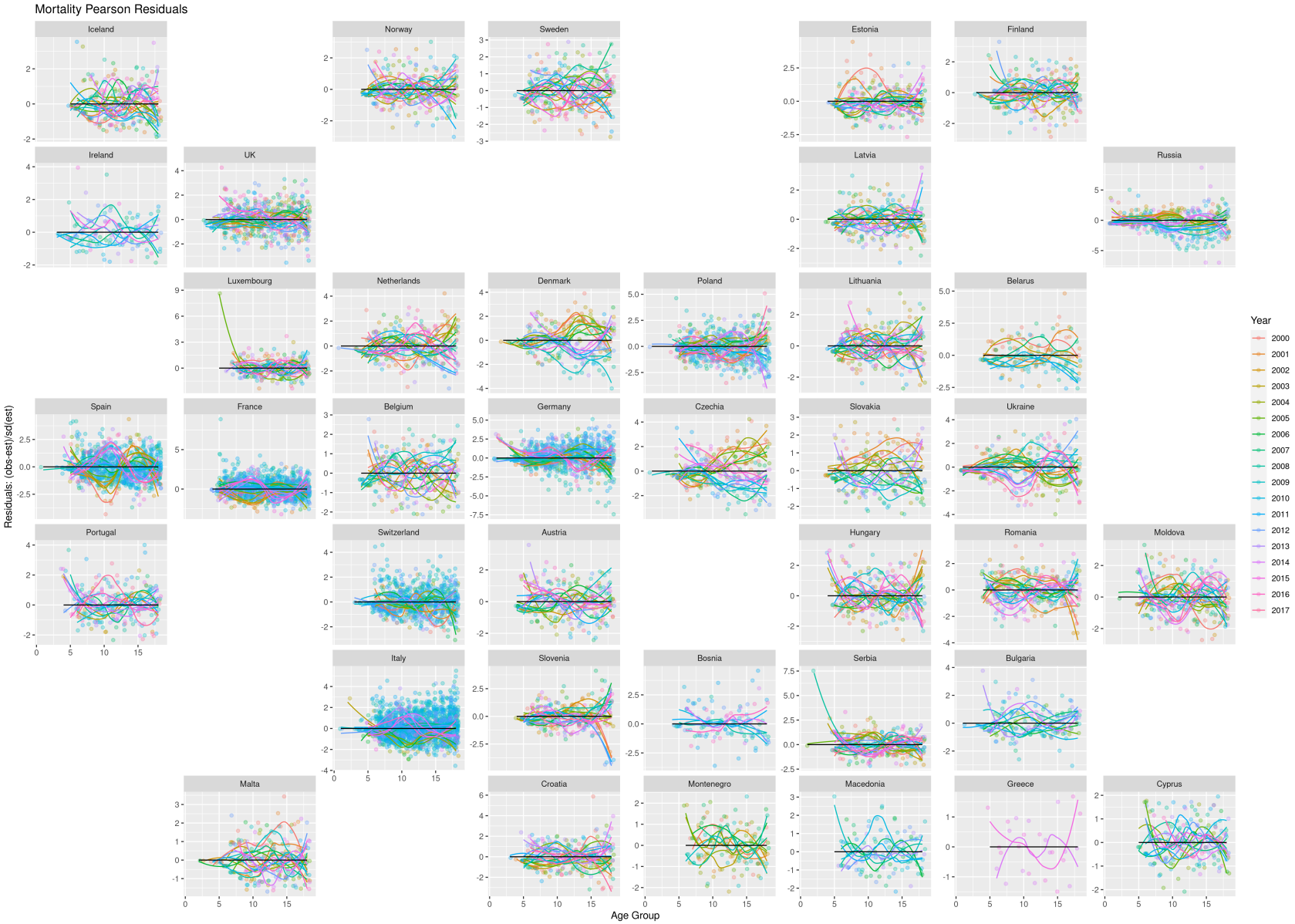


Figure B.5.5: Pearson residuals from model IX (final selected model) for each age, location, and time mortality observation. The residuals are faceted by country, plotted against age groups 1-18 (0-5, ..., 80-84, 85+), and colors and smooth loess curves showing residuals from different years of data.

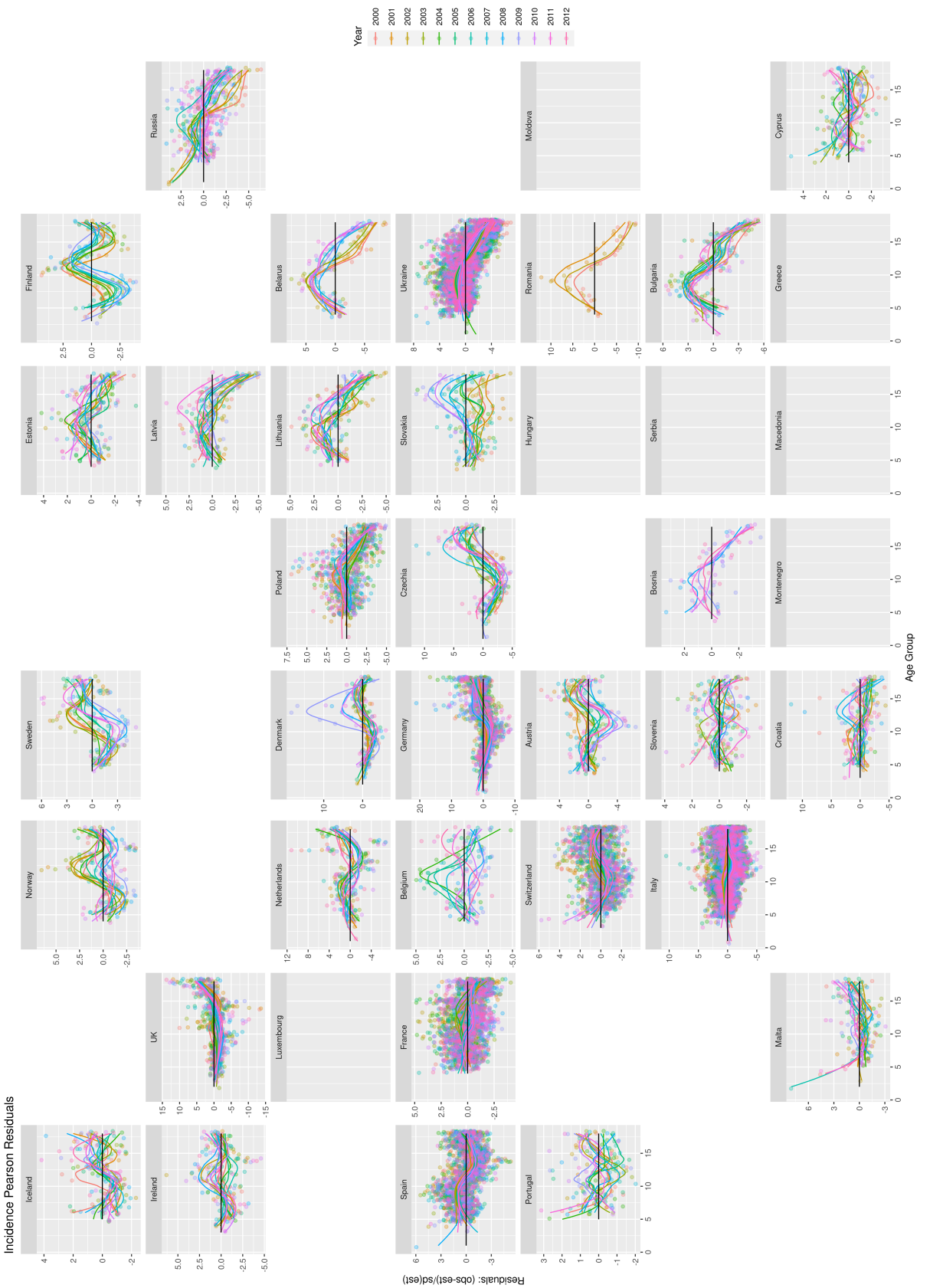


Figure B.5.6: Pearson residuals from model I (base model) for each age, location, and time incidence observation. The residuals are faceted by country, plotted against age groups 1-18 (0-5, ..., 80-84, 85+), and colors and smooth loess curves showing residuals from different years of data.

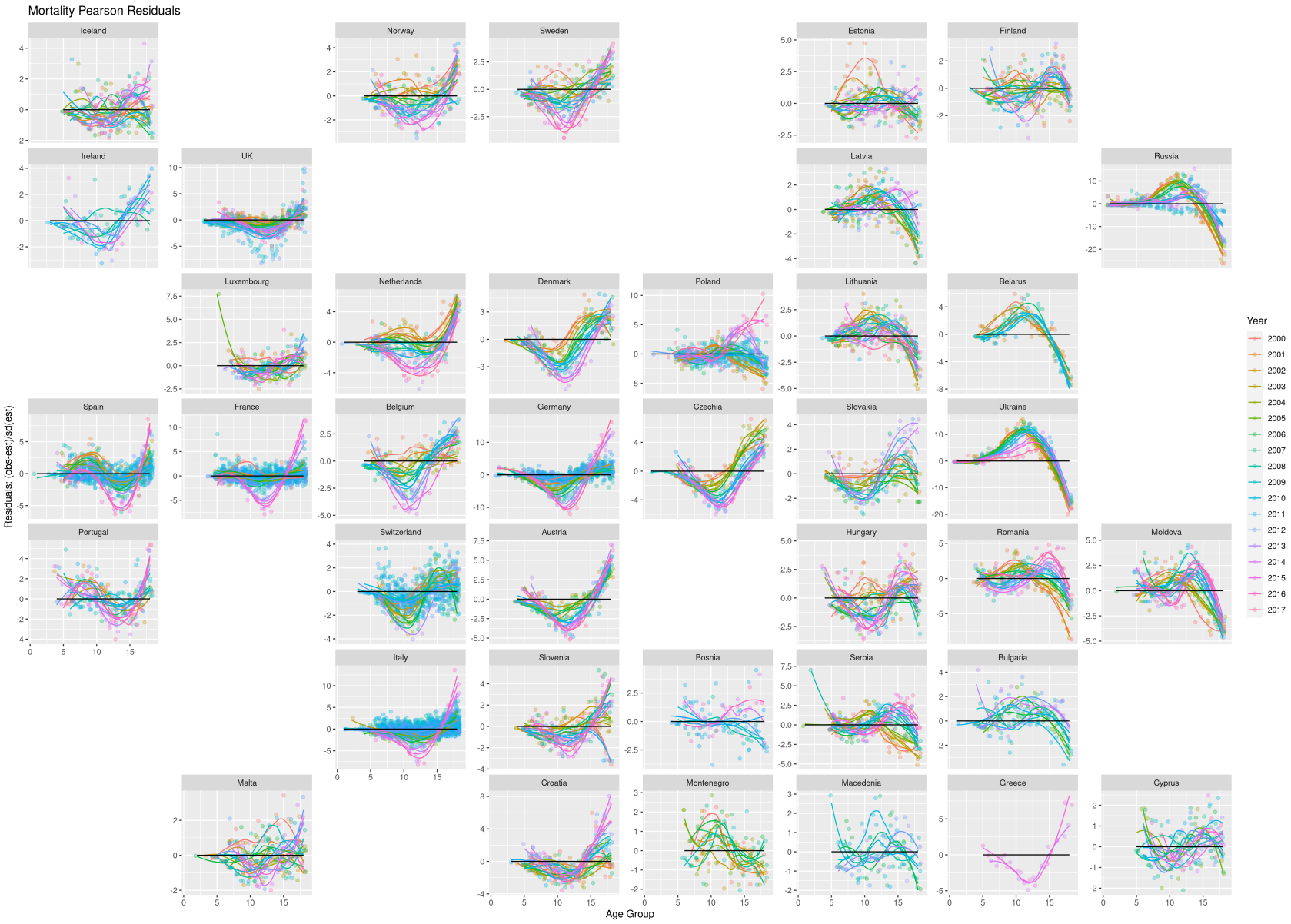


Figure B.5.7: Pearson residuals from model I (base model) for each age, location, and time mortality observation. The residuals are faceted by country, plotted against age groups 1-18 (0-5, ..., 80-84, 85+), and colors and smooth loess curves showing residuals from different years of data.

VITA

Aaron Osgood-Zimmerman was born 'merely' a Zimmerman to Robyn and George Zimmerman in Northampton, Massachusetts in the United States in 1989. He earned a B.A. in Mathematics and B.S. in Engineering from Swarthmore College in 2011, after which he moved to Seattle to pursue his advanced statistics education at the University of Washington. In 2015, he received his M.S in Statistics from the University of Washington under the mentorship of Peter Guttorp. For the next 5 years he worked as a Geostatistics Researcher at the Institute for Health Metrics and Evaluation on the Local Burden of Disease team where he led the development of the statistical models and software used by LBD to make $5 \times 5\text{km}^2$ pixel estimates of children-under-5 mortality and the leading causes of that mortality. In 2016 he married his wife, Logan Osgood-Jacobs, and in 2021 their first child, Asa Osgood-Zimmerman, was born. In 2022, he earned his Ph.D. in Statistics from the University of Washington under the supervision of Dr. Jon Wakefield. Later that year he joined the Math Department at Bucknell University as an Assistant Professor of Statistics.