©Copyright 2023
Taylor Okonek

Statistical Methods for Official Statistics and Mortality Estimation

Taylor Okonek

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

Doctor of Philosophy

University of Washington

2023

Reading Committee:
Jon Wakefield, Chair
Katherine Wilson
Ken Rice

Program Authorized to Offer Degree:
Department of Biostatistics

University of Washington

Abstract

Statistical Methods for Official Statistics and Mortality Estimation

Taylor Okonek

Chair of the Supervisory Committee:
Jon Wakefield
Department of Statistics

This dissertation develops statistical methodology for small area estimation, the analysis of official statistics, and mortality estimation. All methods are developed with the statistical challenges faced in data from low- and middle-income countries in mind. We first consider a problem in small area estimation known as benchmarking, where subnational estimates must agree in some sense with a national estimate or estimates. We then consider two related problems addressing data challenges specific to low- and middle-income country child mortality data, where data comes from a survey, is interval censored, and may need to be smoothed across time to ensure reasonable and precise estimates in situations with little data. The outline of the dissertation is as follows. In Chapter 2, we propose a benchmarking method for subnational estimates of a proportion in a setting where a national estimate is available with uncertainty. In Chapter 3 we develop a pseudo-likelihood approach to mortality estimation that allows us to obtain continuous mortality curves for children under the age of 5 across time from interval-censored survey data. In Chapter 4 we consider different temporal smoothing models for the direct estimates produced in Chapter 3, and develop a multivariate random walk prior to simultaneously smooth multiple, correlated summaries across time. Finally, we conclude with discussion of future work in Chapter 5.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

B3: Bayesian B-spline bias-reduction

BYM: Besag-York-Mollie

BYM2: Besag-York-Mollie 2

CMR: Child mortality rate

DHS: Demographic and Health Surveys

EM: Expectation-Maximization

ESS: Effective sample size

ETSP: Exponentially-truncated shifted-power

GADM: Database of Global Administrative Areas

GMRF: Gaussian Markov random field

HMD: Human Mortality Database

HSRC: Human Sciences Research Council

HT: Horvitz-Thompson

ICAR: Intrinsic conditional auto-regressive

IGME: Inter-agency Group for child Mortality Estimation

IHME: Institute for Health Metrics and Evaluation

IMR: Infant mortality rate

INLA: Integrated nested Laplace approximation

LMIC: Low- and middle-income country

LSV: Left singular vector

MCMC: Markov chain Monte Carlo

MH: Metropolis-Hastings

MICS: Multiple Indicator Cluster Surveys

MSE: Mean-squared-error

NMR: Neonatal mortality rate

NPLMLE: Nonparametric maximum likelihood estimator

PC: Penalized-complexity

PSU: Primary sampling unit

RMSE: Root-mean-squared-error

SVD: Singular value decomposition

TMB: Template model builder

U5MD: Under-5 Mortality Database

U5MR: Under-5 mortality rate

UN: United Nations

ACKNOWLEDGMENTS

I would like to thank my advisor, Jon Wakefield, for encouraging me and supporting me throughout my Ph.D., and for continually reminding me that my first-year theory grades are inversely related to my ability to do good applied work. Thank you to Noah Simon for being an incredible academic advisor and support system throughout my first few years in the program. Thank you to Jessica Godwin for being the best mentor and collaborator. My interest in demography stems almost entirely from our conversations. Thank you to my TRIO McNair mentors, Janis and Melissa, without whom I never would have made it to graduate school in the first place.

Thank you to my cohort (Si Cheng, Avi Kenny, Pearl Liu, María Valdez, and Charlie Wolock) for providing a welcoming, inclusive space within higher academia. Thank you to my friends from home, especially Emma and Kristen, for reminding me that there is more to life than statistics. Thank you to Becca for keeping me sane throughout the past few years. Thank you to my parents. I know I haven't always followed the path you may have imagined for me, and I appreciate your support in my academic endeavors. Thank you to Angela. You know what you did, and I'm forever grateful.

And finally, thank you to Serge. There aren't enough words to explain how much you, Alice, and Ellie mean to me, so I'll leave it with this: I love you the most.

DEDICATION

To Alice and Ellie.

Chapter 1
INTRODUCTION

Child mortality estimates are important health indicators for countries, and are one metric through which countries may monitor their progress towards the Sustainable Development goals (United Nations General Assembly, 2015). Child mortality estimation in low- and middle-income countries (LMICs), where child mortality is particularly high and may follow dissimilar trends to high income countries, contains specific challenges for modelers related to survey data and sparsity of information. This dissertation aims to address some of the challenges in the particular context of estimating child mortality in LMICs. Chapter 2 addresses the need to reconcile subnational, official statistics, with national estimates that are estimated with uncertainty. In Chapter 3 we develop a pseudo-likelihood approach to child mortality estimation that produces a full survival curve for children under the age of 5. Chapter 4 builds upon this pseudo-likelihood approach, and considers various temporal smoothing models that can produce simultaneously smoothed estimates of the neonatal mortality rate (NMR) and under-5 mortality rate (U5MR).

Estimates of child mortality are often produced in an official statistics context; for example, national and subnational estimates produced by the UN Inter-Agency Group for child Mortality Estimation (UN IGME) (Alkema and New, 2014; Li et al., 2019). In recent times, the data used to obtain estimates of child mortality in LMICs primarily arise from nationally representative surveys, such as the Demographic and Health Surveys (DHS) or Multiple Indicator Cluster Surveys (MICS) rather than vital registration or census data. Since survey data is particularly sparse in small areas, subnational estimates of mortality typically rely on model-based approaches that incorporate spatial and temporal smoothing to obtain estimates in all subnational regions with reasonable precision. National estimates

may be produced with use of additional data, and need not be model-based if sufficient data is present. These circumstances typically lead to a problem specific to official statistics, concerning a lack of internal *consistency* between national and subnational estimates. By consistency, we mean that subnational estimates and national estimates must agree in some sense; typically a population weighted sum of subnational estimates should equal a national estimate. The process of enforcing this agreement is known as *benchmarking*. Traditionally, benchmarking approaches have been developed for situations where national estimates are estimated with no uncertainty, and are unbounded. For estimating prevalences of disease or mortality at a subnational levels in LMICs, we have national estimates with an associated measure of uncertainty and outcomes that lie between 0 and 1. In Chapter 2, we develop a computationally efficient approach to benchmarking that addresses the gap in the benchmarking literature with regards the unique concerns when estimating outcomes that lie between 0 and 1 in LMICs.

Just as benchmarking approaches have been developed primarily in a high-income country setting, so too have continuous parametric survival models for mortality. Continuous, parametric approaches to estimating U5MR provide appealingly simple interpretations of parameters, as well as necessary structure to estimates produced when little data is available. Existing approaches that provide continuous survival curves for children under the age of 5 have been shown to insufficiently capture the pattern of mortality in LMICs (Guillot et al., 2022; Eilerts et al., 2021), with bias towards patterns seen in high-income countries. In Chapter 3, we propose a pseudo-likelihood approach to estimating a continuous, parametric survival curve for children under the age of 5 that addresses key issues specific to LMICs including complex survey design aspects and patterns of mortality that differ from high income countries, such as late or early patterns of mortality (Guillot et al., 2022). Our approach is compatible with a variety of parametric models, and we include assessments for parametric model fit using a survey-weighted non-parametric maximum likelihood estimate (NPMLE).

One aim of Chapter 3 is to assess whether certain parametric assumptions seem rea-

sonable when there are large amounts of survey data available to produce estimates at a national level, so that those assumptions, if reasonable, may be applied in scenarios with less data to obtain more precise estimates. These scenarios include estimates produced at a subnational level, or estimates at a fine temporal scale where survey data is relatively sparse in LMICs. In such cases, smoothing is needed to provide precise estimates across space and time, though the form of smoothing model may differ. In Chapter 4, we propose a flexible type of multivariate Gaussian Markov random field (GMRF) that can smooth multiple, potentially correlated parameters across space or time. We apply our method to NMR, the infant mortaliry rate (IMR) and U5MR from the 27 districts of mainland Malawi over a ten-year period. We compare smoothed parameter estimates from Chapter 3's pseudo-likelihood approach to those from a simple random walk on point estimates. Our proposed approach has the benefit of guaranteeing that smoothed estimates of NMR, IMR, and U5MR obey the logical constraint NMR $<$ IMR $<$ U5MR.

We conclude with a discussion of future work in Chapter 5.

Chapter 2

A COMPUTATIONALLY EFFICIENT APPROACH TO FULLY BAYESIAN BENCHMARKING

2.1 Introduction

In a public health context, small area estimates, where small areas are defined as domains with little to no data, are often produced by official statistics agencies for the purpose of informing targeted public health interventions. In small area estimation, where we have by definition little to no data in certain domains, it may not be possible to obtain direct, survey estimates with reasonable precision. In these settings, model-based methods are often used when estimates at small levels of aggregation are required (Rao and Molina, 2015). Model-based methods that incorporate spatial random effects, in the absence of reliable covariate information, allow for estimates in all small areas to be produced with little to no data, introducing some bias into the estimates in exchange for tighter interval estimates (Knorr-Held, 2000; Riebler et al., 2016; Wakefield et al., 2020). Area-level models, such as the popular Fay-Herriot model (Fay and Herriot, 1979), consider data at the level of the small area, whereas unit-level models consider data at the level of the individual or cluster (Battese et al., 1988). When area-level sample sizes are particularly small, unit-level models may be more desirable. For a review of both area-level and unit-level models, see Chapters 4-7 of Rao and Molina (2015).

It is often required that small area estimates agree with estimates at a higher level of aggregation. For example, subnational estimates may be required to aggregate to a national level estimate. These higher level estimates are referred to as benchmarks, and are frequently considered more reliable than small area estimates since more data are available to inform them, or they are direct (weighted) estimates and therefore less dependent on model

assumptions. Note that benchmarking as we consider it in this chapter is distinct from calibration weighting (Särndal et al., 2003; Si and Zhou, 2020). The latter involves incorporating known, population-level *covariate* information into a model to adjust for potential bias in a subnational model, while the former involves incorporating known, population-level *outcome* information into a model to enforce internal consistency in an official statistics setting.

Benchmarks may be from the same data source that was used to produce small area estimates or from an outside source, referred to as internal benchmarking and external benchmarking, respectively (Bell et al., 2013). In an internal benchmarking setting, the use of the same data twice—once to calculate model-based small area estimates and also to define benchmarks at a higher level of spatial aggregation—may lead to an understating of statistical uncertainty unless additional care is taken. The method of You et al. (2002) can be used to quantify the posterior MSE of an internally benchmarked predictor in a Bayesian context. We also direct the reader to Pfeffermann and Tiller (2006) for an approach to assessing the uncertainty of internally benchmarked predictors. We note that the particular applications we consider in this chapter are in an external benchmarking setting, and that this setting is relatively common in global health applications (Osgood-Zimmerman et al., 2018; Eaton et al., 2021) as well as more recent applications in agriculture (Chen et al., 2022). Internal benchmarking settings and their limitations are outside the scope of this chapter. We direct the reader to Pfeffermann and Tiller (2006) for one approach to assess uncertainty of internally benchmarked estimators.

Many existing benchmarking approaches treat benchmarking as a constraint problem, where estimates at smaller areas are constrained to agree with estimates at a higher level of aggregation. Approaches vary by the ways in which constraints have been incorporated into a modeling framework, and as such, the interpretation of resulting benchmarked estimates varies by approach as well. There are many ways to categorize benchmarking approaches, but one we consider is a one-step versus two-step procedure. In a two-step procedure, estimates and uncertainty are first obtained from a model that is agnostic to the benchmarking constraint and are then adjusted to satisfy the benchmarking constraint. Datta et al. (2011);

Ghosh and Steorts (2013); Steorts et al. (2020); Wang et al. (2008); Patra and Dunson (2018); Patra (2019); Ghosh et al. (2015); Williams and Berg (2013) and Berg and Fuller (2018) all consider a two-step procedure where benchmarked estimates are calculated by minimizing posterior expected loss (for a given loss function) subject to a benchmarking constraint. Unbenchmarked estimates are first obtained, and then projected into a constrained space.

Other two-step procedures include difference benchmarking and ratio benchmarking (henceforth referred to as raking), described in Erciulescu et al. (2019). Again, a model that is agnostic to the benchmarking constraint is first fit, and then estimates are adjusted by a constant, constructed via a population-weighted average, so that the benchmarking constraint is satisfied (Erciulescu et al., 2019, 2018, 2020). You et al. (2002) consider a hierarchical Bayesian model for unbenchmarked estimates, but obtain benchmarked estimates using the raking method, and quantify uncertainty via posterior MSE as opposed to estimating full posterior distributions for benchmarked estimates.

Other benchmarking approaches incorporate the benchmarking constraint into the data likelihood—referred to as an augmented model in Bell et al. (2013), Berg and Fuller (2018), and Stefan and Hidiroglou (2021)—and thus produce automatically benchmarked estimates in a single modeling step. You and Rao (2002) also follow this approach, and refer to this as a "self-benchmarking" property.

Others (Nandram and Sayit, 2011; Janicki and Vesper, 2017; Erciulescu et al., 2019; Nandram et al., 2019; Zhang and Bryant, 2020) propose benchmarking approaches that produce full, benchmarked posterior distributions for small area estimates, making uncertainty quantification straightforward in an external benchmarking setting. Following Zhang and Bryant (2020), we refer to such approaches as fully Bayesian benchmarking approaches. Note that we consider this type of approach to be distinct from fitting a Bayesian model and benchmarking only point estimates. Nandram and Sayit (2011) develop a fully Bayesian benchmarking approach specifically for betabinomial models. Janicki and Vesper (2017) perform fully Bayesian benchmarking by minimizing the KL divergence between a benchmarked and unbenchmarked posterior using moment constraints. Nandram et al. (2019) perform fully

Bayesian benchmarking using a transformation of the benchmarking constraint that corresponds to "deleting" a single small area, but note that benchmarked estimates vary based on which area is deleted. Erciulescu et al. (2019) consider four different benchmarking approaches, one of which is the Bayesian method described in Nandram and Sayit (2011), and three of which involve fitting a hierarchical Bayesian model and benchmarking point estimates using ratio or difference benchmarking. Zhang and Bryant (2020) develop a fully Bayesian benchmarking approach that produces full posterior distributions conditional on either a soft (inexact) or hard (exact) benchmarking constraint.

Benchmarking approaches have also been developed in a time series context, with certain similarities to small area estimation in requiring finer estimates in time to agree with a aggregate estimates accross multiple years (Dagum and Cholette, 2006). Different benchmarking approaches are more or less appealing than others—depending on context—with regards to obtaining measures of uncertainty, computational tractability, and the way in which the benchmarking constraint is enforced.

Benchmarking methods may also differ depending on whether benchmarking must be *exact* or *inexact*. In exact benchmarking, as the name suggests, the benchmarking constraint must hold exactly, whereas in inexact benchmarking the constraint must hold within some margin of error. The latter can be viewed as a soft constraint as opposed to a hard constraint. Exact benchmarking may be appropriate if the benchmarks are unbiased and have little to no uncertainty, which can occur when they come from a census (Trabelsi and Hillmer, 1990). Inexact, external benchmarking may be appropriate if national estimates come from a different survey and contain appreciable sampling error, as discussed in Hillmer and Trabelsi (1987), or if national estimates are model-based with appreciable uncertainty, as in the settings we consider in our application.

In this chapter, we present two novel implementations of the fully Bayesian benchmarking approach described in Zhang and Bryant (2020) that are more flexible and computationally tractable in many settings. Our approaches combine an unbenchmarked model with either a rejection sampler or Metropolis-Hastings algorithm to produce fully Bayesian benchmarked

posteriors, which we describe in subsection 2.3.4. We compare our method to that described in Zhang and Bryant (2020) in the setting of modeling HIV prevalence in South Africa, as well as modeling U5MR in Namibia. These applications were chosen to demonstrate the flexibility of the proposed fully Bayesian approaches in estimating outcomes that lie between 0 and 1 with unique benchmarking constraints, the efficiency of the method in cases where the benchmarks are very consistent or inconsistent with the small area estimates, and the flexibility of the approaches to handle both area-level and unit-level models. To emphasize the novelty of our approaches and computational advantages of the proposed method over the Markov chain Monte Carlo (MCMC) samplers used in Zhang and Bryant (2020) in terms of flexibility, we use integrated nested Laplace approximation (INLA) and Template Model Builder (TMB) as alternative ways to conduct Bayesian inference using Laplace approximations, that are fast and do not require users to code model-specific fitting routines (Rue et al., 2009; Kristensen et al., 2016). We additionally compare run times for our proposed approaches to that of Zhang and Bryant (2020) in a simulation, and show that our proposed approaches provide not only increased flexibility in terms of modeling for Bayesian inference, but computational speed gains as well. All code for fitting the models described in this chapter is available via the R package `stbench`, found at github.com/taylorokonek/stbench.

2.2 Small area models in low- and middle-income countries

The estimation of under-5 mortality rates (U5MR) in LMICs at a subnational level motivates our desire for benchmarking. The UN Inter-agency Group for Child Mortality Estimation (IGME) produces annual, national level estimates of U5MR for all countries using a Bayesian B-spline bias-reduction (B3) method (Alkema and New, 2014). Various data are used to produce B3 estimates, including vital registration, census, and household surveys, and many of these sources cannot be used for producing subnational estimates because geographic information is lacking, or the data type is not amenable to incorporation into a small area model. Subnational estimates of U5MR are of interest in addition to national level estimates, in accordance with the Sustainable Development Goals

(https://sustainabledevelopment.un.org/post2015/transformingourworld).

Model-based small area estimation has a long history in LMICs. While small area estimates that incorporate the survey design directly are preferred, they are often impractical at a small area level, with either too little precision to be practically useful or no data available in some small areas (Lehtonen and Veijanen, 2009; Wakefield et al., 2020). This lack of data primarily comes from a disconnect between the administrative level at which these surveys are designed to produce reliable estimates (often administrative level 1, or state) and the administrative level at which public health interventions are made (often administrative level 2, or county). Model-based methods with spatial smoothing terms allow us to obtain precise estimates in *all* small areas, as required for public health estimates (Datta, 2009; Wakefield et al., 2020).

Due to these smoothing terms and the inability to use data sources without geographic information to produce small area estimates, benchmarking is required to align subnational estimates with national estimates from the B3 model for internal consistency within the production of UN IGME estimates. Subnational estimates are currently produced for a handful of countries using a Beta-binomial model described in Wu et al. (2021), but benchmarking approaches in this context have not yet been rigorously explored. As national estimates are model-based (with uncertainty) in this context, we aim to use a benchmarking approach that incorporates national level uncertainty (i.e., is *inexact*).

Many public health outcomes, including HIV prevalence and U5MR that we consider for our applications, are estimated from Bernoulli or binomial counts. As such, binomial models are a typical choice in an LMIC setting (Eaton et al., 2021; Wu et al., 2021). We aim to use a benchmarking approach that is well-suited to binomial models with estimates that are proportions, and hence lie between 0 and 1. Just as the unbenchmarked estimates will lie in this range, the benchmarked estimates from our chosen benchmarking approach should as well.

Additionally, the approach we use should be fast and computationally flexible. These properties are particularly important in an LMIC setting, as practitioners and national

statistics offices in LMICs may have fewer computational resources than those in high-income countries.

While our motivation for benchmarking primarily comes from subnational estimation of U5MR, the application to HIV prevalence is similarly important in a LMIC context, in that small area estimates of HIV prevalence are produced in an official statistics setting by UNAIDS and require benchmarking for internal consistency (Eaton et al., 2021). Similarly to U5MR, subnational estimates of HIV prevalence inform public health interventions and allow countries to monitor their progress towards the Sustainable Development Goals. The theoretical properties, computational speed, and flexibility of our proposed approaches are relevant to HIV prevalence estimation in LMICs, as (similarly to U5MR) national estimates are produced with uncertainty, and benchmarked estimates should lie between 0 and 1.

2.2.1 Small area models

Below we describe the unbenchmarked models that we consider for the HIV application, but they are generally appropriate for a binary outcome. The unbenchmarked models for the U5MR application are described in Section 2.2 of Appendix A. If there are sufficient data in each target area, then weighted (direct) estimates are reliable. Such estimates include the Horvitz-Thompson (HT) (Horvitz and Thompson, 1952) and Hájek estimators (Hájek, 1971). In situations where the direct estimates can be calculated but have unacceptably high design variance, one may smooth using an area-level model, such as the model of Fay and Herriot (1979). When the data are sparser still, area-level direct estimates may be unusable. With very little data, the variance estimate can be statistically unreliable or may not be calculable at all; for example, if all or no observations in an area contain cases. In these circumstances, the raw data (counts, in the binary context) are modelled (in our HIV and U5MR examples, at the cluster level) to give a unit-level model.

Area-level model: Spatial Fay-Herriot

Let the HT estimator for area i be denoted $\hat{\theta}_i^{HT}$ with design-based variance $\hat{V}_i^{*,HT}$. Under the Fay-Herriot model, a working likelihood may be based on the distribution,

$$y_{1i}^* = \text{logit}(\hat{\theta}_i^{HT}) \sim \text{N}(\eta_i, \hat{V}_i^{HT}) \tag{2.1}$$

where $\hat{V}_i^{HT} = V_i^{*,HT}/(\hat{\theta}_i^{2,HT}(1-\hat{\theta}_i^{2,HT}))^2$ is obtained via the delta method. Treating y_{1i}^* and $\hat{V}_i^{HT}$ as fixed, as in the Fay-Herriot model, we then consider the area-level model with the top level given in Equation (2.1), and mean model

$$\eta_i = \beta_0 + b_i$$

where β_0 is an intercept term, and b_i follows a BYM2 spatial model (Riebler et al., 2016), denoted $b_i \sim \text{BYM2}(\tau_b, \phi)$, where τ_b is the total precision and ϕ is the proportion of the variance that is spatial. The BYM2 model is a reparameterization of the BYM spatial random effect developed by Besag et al. (1991), which consists of an unstructured (iid) random effect and a structured, ICAR spatial random effect (Besag, 1974). A sum-to-zero constraint is placed on the structured component of the BYM2 random effect to ensure identifiability (Besag et al., 1991). The above model is often used in the estimation of U5MR in LMICs as it directly accounts for survey design and incorporates spatial smoothing random effects (Li et al., 2019; Wakefield et al., 2020), and is readily applicable to applications of HIV prevalence as well (Wakefield et al., 2020).

For priors, we set $\beta_0 \sim \text{N}(0, 0.001^{-1})$, $\phi \sim \text{Beta}(0.5, 0.5)$. We use penalized complexity (PC) priors for the precision τ_b with values $U = 1$ and $\alpha = 0.01$, corresponding to a prior belief that the probability $\sigma_b = 1/\sqrt{\tau_b}$ is greater than 1 is 1% (Simpson et al., 2017). Note that we chose noninformative hyperpriors for some of our model parameters. In practice, priors for hyperparameters should be carefully chosen for the given application.

Unit-level model: Binomial

We assume binomial observations of the outcome for each cluster c within area i. Let $y_{1i[c]}$ be the number of cases observed in $n_{i[c]}$ total people in cluster c within area i. We consider the unit-level model

$$\begin{aligned} y_{1i[c]} \mid n_{i[c]}, \theta_{i[c]} &\sim \text{Binomial}(n_{i[c]}, \theta_{i[c]}), \\ \eta_{i[c]} = \text{logit}(\theta_{i[c]}) &= \beta_0 + b_i + e_{i[c]}, \end{aligned}$$

where $\theta_{i[c]}$ is prevalence in cluster c of area i, β_0 is an intercept term, and b_i again follows a BYM2 model (Riebler et al., 2016), denoted $b_i \sim \text{BYM2}(\tau_b, \phi)$, and $e_i[c] \overset{iid}{\sim} \text{N}(0, \sigma_e^2)$ is an iid cluster-level random effect. Constraints and priors are the same as for the area-level model, with the addition of a PC prior on σ_e with values $U = 1$ and $\alpha = 0.01$.

Area-level predictions $\hat{\theta}_i$ are obtained by marginalizing over the cluster random effect to get

$$\hat{\theta}_i = \text{expit}\left(\frac{\beta_0 + b_i}{\sqrt{1 + h^2\sigma_e^2}}\right),$$

where $h^2 = 0.346$. The cluster-level random effect is excluded from predictions, and a correction is done using h and σ_e^2, as described in Dong and Wakefield (2021) and detailed in Section 9.13.1 and Exercise 9.1 of Wakefield (2013). The correction accounts for within-cluster variation that induces cluster-level overdispersion. If we instead believed the cluster-level random effect reflected true *between*-cluster differences, we could obtain predictions such as those obtained in Dong and Wakefield (2021).

2.3 Methods

Let y_2 be a national level estimate of the outcome of interest, possibly with uncertainty given by a confidence interval or standard error, $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_n)^\top$ be small area estimates for areas $i = 1, \ldots, n$, and w_i be weights that do not depend on $\boldsymbol{\theta}$, standardized so that

$\sum_{i=1}^{n} w_i = 1$. Often, these weights are set to $w_i = N_i / \sum_{j=1}^{n} N_j$, where N_j is the population size for the small area j. Note that the population sizes used should correspond to the population under study. For example, when estimating HIV prevalence using survey data, N_j should be population counts for individuals within the age range in the sampling frame. For DHS surveys, this consists of individuals aged 15-49. To our knowledge, incorporating uncertainty about the population counts into benchmarking approaches has not yet been considered in the benchmarking literature.

Throughout the chapter, we consider benchmarking constraints of the most basic form, $\sum_{i=1}^{n} w_i \hat{\theta}_i = y_2$. While more complex benchmarking constraints, or even multiple benchmarking constraints as noted in Zhang and Bryant (2020), may be useful in certain settings, this simple equality constraint is reasonable for the HIV prevalence and U5MR applications we consider, and is the constraint currently used in practice by both UNAIDS and the UN IGME (Eaton et al., 2021; Wu et al., 2021). For examples of benchmarking applications with inequality constraints, see Steorts et al. (2020) or Nandram et al. (2022).

In the following subsections we describe a subset of existing approaches to benchmarking and propose two novel approaches to fully Bayesian benchmarking. The methods we describe in detail were chosen either because they are commonly used benchmarking methods in an official statistics setting, or because they are particularly relevant to our motivating application. One method is not *inherently* preferable, though we argue that some are more appropriate than others in the context of estimating outcomes between 0 and 1 in LMICs. The only existing fully Bayesian approach we describe is that of Zhang and Bryant (2020), as our proposed approach builds directly from their methodology.

2.3.1 Benchmarked Bayes Estimate Approach

The first approach we describe was developed in a Bayesian, decision theoretic framework (Datta et al., 2011; Steorts et al., 2020). This involves minimizing expected posterior MSE loss subject to the benchmarking constraint, which results in a projection of the unbenchmarked estimates into a benchmarked (constrained) space. Since its development, others

have extended this decision theoretic approach with different loss functions (Ghosh et al., 2015; Williams and Berg, 2013; Berg and Fuller, 2018). Methods have also recently been developed to obtain benchmarked uncertainty for these estimates under this decision theoretic framework (Patra and Dunson, 2018; Patra, 2019). As it involves minimizing expected posterior loss subject to a benchmarking consraint, we call this approach the Benchmarked Bayes Estimate approach. Though we will argue that this approach is not appropriate for our motivating application, we detail it here as it is theoretically appealing, fast, and used in many benchmarking applications.

For n small areas, let $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_n)^\top$ be the direct estimators of the small area means $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top$. We are interested in computing the benchmarked Bayes estimator $\hat{\boldsymbol{\theta}}^{BM} = (\hat{\theta}_1^{BM}, \ldots, \hat{\theta}_n^{BM})^\top$ of $\boldsymbol{\theta}$ such that the constraint $\sum_{i=1}^n w_i \hat{\theta}_i^{BM} = y_2$ is satisfied.

As Datta et al. (2011) are interested in an estimate of the benchmarked posterior mean, they consider minimizing the posterior expectation of the weighted squared error loss $\sum_{i=1}^n \phi_i E[(\theta_i - e_i)^2 \mid \mathbf{y}_1]$ under the constraint $\bar{e}_w := \sum_{i=1}^n w_i e_i = y_2$, where ϕ_i are weights not necessarily equal to w_i, $\mathbf{y}_1$ is a vector of area-level observations, and the solution to the minimization problem will be denoted $e_i = \hat{\theta}_i$. They note that the weights ϕ_i could be different for different policy makers, and a simple default is setting $\phi_i = 1$ for all $i = 1, \ldots, n$. The resulting benchmarked Bayes estimate (solution to the constrained minimization of the posterior expectation of the weighted squared error loss) is

$$\hat{\boldsymbol{\theta}}^{BM} = \hat{\boldsymbol{\theta}}^{B} + s^{-1}(y_2 - \bar{\hat{\theta}}_w^B)\mathbf{r}, \tag{2.2}$$

where $\hat{\boldsymbol{\theta}}^B = (\hat{\theta}_1^B, \ldots, \hat{\theta}_n^B)^\top$ is a vector of unbenchmarked posterior means $E[\theta_i \mid \mathbf{y}_1]$ under a given prior, $\bar{\hat{\theta}}_w^B = \sum_{i=1}^n w_i \hat{\theta}_i^B$, $\mathbf{r} = (r_1, \ldots, r_n)^\top$, $r_i = w_i/\phi_i$, and $s = \sum_{i=1}^n w_i^2/\phi_i$. Note that if $\phi_i = 1$ for all $i = 1, \ldots, n$, the benchmarked Bayes estimate becomes $\hat{\boldsymbol{\theta}}^{BM} = \hat{\boldsymbol{\theta}}^B + (\mathbf{w}^\top \mathbf{w})^{-1}(y_2 - \bar{\hat{\theta}}_w^B)\mathbf{w}$. Thus from Equation (2.2) we can see that the benchmarked Bayes estimate is a function of the unbenchmarked Bayes estimate under mean squared error (MSE) loss (posterior means) and user-specified weights. This is computationally appealing, as

obtaining benchmarked estimates with this method is done via a quick post-processing step.

To obtain uncertainty around the benchmarked estimates, posterior samples can be projected into the space defined by the benchmarking constraint (Patra and Dunson, 2018; Patra, 2019). Geometrically, we can interpret this benchmarked Bayes estimate as the point estimate within the space defined by the benchmarking constraint that is as close to the unbenchmarked Bayes estimate as possible, where closeness is measured in terms of expected weighted squared error (Steorts et al., 2020).

The benchmarked estimate from Equation (2.2) is an exactly benchmarked estimate, i.e., the benchmarking constraint holds exactly. Alternatively, in inexact benchmarking, the benchmarking constraint need not hold exactly. If inexact benchmarking is desired in the benchmarked Bayes estimate framework, a penalty term $\lambda > 0$ (pre-specified by the user) is introduced, and a slightly different loss function is considered,

$$L(\boldsymbol{\theta}, \mathbf{e}) = \lambda(y_2 - \bar{e}_w)^2 + \sum_{i=1}^{n} \phi_i(\theta_i - e_i)^2.$$

The Bayes estimate associated with this loss is

$$\hat{\boldsymbol{\theta}}_{\lambda}^{B} = \hat{\boldsymbol{\theta}}^{B} + (s + \lambda^{-1})^{-1}(y_2 - \bar{\hat{\theta}}_{w}^{B})\mathbf{r},$$

where we note that as $\lambda \to \infty$, $\hat{\boldsymbol{\theta}}_{\lambda}^{B}$ approaches the exactly benchmarked Bayes estimate in Equation (2.2). Considering a loss function with the addition of a penalty term for the benchmarking constraint allows the user to incorporate a predetermined level of agreement between the benchmarks and aggregated, unbenchmarked model estimates that must be met, whether that be exact benchmarking (corresponding to $\lambda \to \infty$) or inexact, where the resulting estimate $\hat{\boldsymbol{\theta}}_{\lambda}^{B}$ is a compromise between the exactly benchmarked and unbenchmarked Bayes estimate.

In the context of U5MR and HIV prevalence, the outcome of interest lies between 0 and 1, and may be close to 0. Any estimate that falls outside of $[0, 1]$ would be invalid.

In applications where the outcome of interest lies in a restricted space, the benchmarked Bayes estimate approach described above can return invalid benchmarked estimates. In particular, note that the benchmarked Bayes estimates may possibly lie below zero when $\bar{\hat{\theta}}_w^B > y_2$. Ghosh et al. (2015) note this issue and consider a variant of the Kullback-Leibler loss function rather than weighted MSE loss, which addresses the issue of estimates falling below 0 in the case of positive small area estimates. Williams and Berg (2013) and Berg and Fuller (2018) also note this issue, and the latter propose using a specific form for the weights ϕ_i in Equation (1) to deal with unbenchmarked estimates that are close to the boundary. Their approach does not guarantee, however, that benchmarked estimates will lie within the required restricted space. While they note that in many situations benchmarked estimates that lie outside the restricted space will be rare—and in such cases their approach may be sufficient—when estimating rare disease prevalence or mortality this boundary issue is a concern.

The benchmarked Bayes estimate approach is fast and theoretically justified in settings where estimates fall on the real line, are positive as in Ghosh et al. (2015), or lie well within the boundary of $[0, 1]$. However, with the loss functions considered thus far in the literature, the approach will often fall short for targets that are on a restricted range.

2.3.2 Raking approach

The second benchmarking approach we consider is simple and popular: raking, also referred to as the ratio-adjustment method (Datta et al., 2011; Zhang and Bryant, 2020; Ghosh et al., 2015). A version of the raking approach is used by the Institute for Health Metrics and Evaluation (IHME) in a variety of applications (e.g., Osgood-Zimmerman et al. (2018); Local Burden of Disease HIV Collaborators (2021)), in the Naomi model for estimating HIV prevalence and incidence (Eaton et al., 2021), and in the subnational U5MR estimates currently produced by UN IGME (Wu et al., 2021). This approach is commonly applied post-hoc. The key feature of the raking approach is a ratio comparing an unbenchmarked

national estimate to the national level benchmark

$$R = \frac{\hat{\theta}^N}{y_2}, \tag{2.3}$$

where $\hat{\theta}^N$ denotes some unbenchmarked national level estimate, and y_2 again denotes the national level benchmark. The unbenchmarked national estimate could be obtained via the weighted sum of unbenchmarked small area estimates as $\hat{\theta}^N = \sum_{i=1}^n w_i \hat{\theta}_i^M$, where $\hat{\theta}_i^M$ denote the posterior estimates (means or medians) in each area from our unbenchmarked model. Note that the raking approach is considered internal or external benchmarking based on whether y_2 is calculated from the same data as $\hat{\theta}^N$.

In a post-hoc raking approach and with a sampling-based method, the posterior draws of θ_i from an unbenchmarked model are multiplied by $1/R$ so that the constraint $\sum_{i=1}^n w_i \hat{\theta}_i^M = y_2$ holds. Of note, this benchmarking approach will treat the unbenchmarked estimates in every area in the same fashion, regardless of the uncertainty in the unbenchmarked estimates. As such, the ranking of regions based on posterior medians/means will be preserved between unbenchmarked and benchmarked estimates. This behavior follows because of the ad hoc nature of the raking adjustment, as noted by Datta et al. (2011). It may be preferable to instead treat unbenchmarked estimates with more uncertainty differently than those with less.

The raking approach to benchmarking can also be approximated via the inclusion of a log offset term for the ratio comparing an unbenchmarked national estimate to the national level benchmark in logistic models when the outcome of interest is rare, as in under-5 mortality estimation. In the supplement of Wakefield et al. (2019) they show that, for rare outcomes, including a log offset for R in a logistic regression model corresponds approximately to the same multiplicative bias adjustment that would be done in the post-hoc raking approach. This approach is currently used in the UN IGME's subnational U5MR estimates. We note that the form of raking involving the inclusion of a log offset for R does in fact produce fully Bayesian estimates, in the sense that a full posterior distribution for the benchmarked

estimates is produced. However, the approach differs from the fully Bayesian approach described in Section 2.3.3 in that a likelihood is not specified for the benchmarks themselves.

2.3.3 Fully Bayesian benchmarking approach: Zhang and Bryant (2020)

Consider an area-level, Bayesian hierarchical model for small area estimation. For n small areas, let the area-level parameters we wish to estimate be denoted by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^\top$, with a hierarchical structure specified through a model $\pi(\boldsymbol{\theta} \mid \boldsymbol{\phi})$ with a vector of hyperparameters $\boldsymbol{\phi}$. The area-level observations are denoted by $\mathbf{y}_1 = (y_{11}, \ldots, y_{1n})^\top$. For example, the values y_{1i} may be binomial counts of HIV status, with the parameters θ_i corresponding to HIV prevalence in area i. We can write the posterior distribution as

$$\pi(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{y}_1) \propto \pi(\mathbf{y}_1 \mid \boldsymbol{\theta}, \boldsymbol{\phi})\pi(\boldsymbol{\theta} \mid \boldsymbol{\phi})\pi(\boldsymbol{\phi}).$$

Though Zhang and Bryant (2020) consider more complex forms of benchmarking constraints, for simplicity we again consider the constraint $\sum_{i=1}^n w_i\theta_i = y_2$, where w_i and y_2 are defined previously. For simplicity, we will refer to y_2 as benchmarks and $\sum_{i=1}^n w_i\theta_i = y_2$ as a benchmarking constraint in an inexact benchmarking setting as well as an exact setting, despite the fact that in an inexact setting there are no hard constraints. The method we describe is generally applicable to a variety of benchmarking constraints.

To incorporate the benchmarking constraint into their hierarchical model, Zhang and Bryant (2020) define an additional likelihood term for the benchmarks, $\pi(y_2 \mid \boldsymbol{\theta})$. This results in the benchmarked posterior distribution

$$\pi(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{y}_1, y_2) \propto \pi(\mathbf{y}_1 \mid \boldsymbol{\theta}, \boldsymbol{\phi})\pi(y_2 \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta} \mid \boldsymbol{\phi})\pi(\boldsymbol{\phi}), \tag{2.4}$$

with the assumptions that $\mathbf{y}_1$ and y_2 are conditionally independent given $\boldsymbol{\theta}$, and that $\pi(y_2 \mid \boldsymbol{\theta})$ does not depend on hyperparameters $\boldsymbol{\phi}$. Notably, this first assumption does not hold in internal benchmarking settings. The likelihood term for the benchmarks, $\pi(y_2 \mid \boldsymbol{\theta})$, pulls the

likelihood for the area-level observations, $\pi(\mathbf{y}_1 \mid \boldsymbol{\theta}, \boldsymbol{\phi})$, towards the benchmarks. Uncertainty quantification is thus straightforward, as we obtain an entire benchmarked posterior distribution as opposed to simply a benchmarked point estimate.

Under exact benchmarking, we set $\pi(y_2 \mid \boldsymbol{\theta}) = \mathrm{I}\left[\sum_{i=1}^{n} w_i\theta_i = y_2\right]$. Under inexact benchmarking, $\pi(y_2 \mid \boldsymbol{\theta})$ is a non-degenerate distribution specified by the user. One can incorporate a discrepancy parameter λ into this distribution to allow for varying levels of agreement between the benchmarks and the aggregate parameters if desired. For example, in the application we consider setting $\pi(y_2 \mid \boldsymbol{\theta})$ equal to a normal distribution with mean $\sum_{i=1}^{n} w_i\theta_i$ and variance equal to the variance of the national benchmark. This would be equivalent to setting $1/\lambda$ equal to the variance of the national estimate in the Bayes estimate approach described in Section 2.3.1.

A benefit of this benchmarking approach is that it allows for nonlinear benchmarking constraints. This is particularly relevant for logistic models, where the benchmarking constraint may take the form $\sum_{i=1}^{n} w_i\theta_i = \sum_{i=1}^{n} w_i\text{expit}(\eta_i)$ for a mean model η_i. Unlike the estimates from a benchmarked Bayes estimate approach, the fully Bayesian benchmarking approach ensures that benchmarked estimates remain within the boundaries of the parameter space.

The fully Bayesian benchmarking approach can be used with both area-level and unit-level models. Though Zhang and Bryant (2020) describe an extension of their fully Bayesian approach to unit-level models, an implementation of this does not currently exist. Zhang and Bryant (2020) provide code in their R package `demest` for Poisson, binomial, multinomial, and normal models, with optional point mass, Poisson, binomial, and normal distributions for the likelihood for the benchmark, found at `https://github.com/StatisticsNZ/demest`. They provide an outline for a MCMC scheme for a general model. However, implementation of this MCMC scheme will be model-specific, which will likely limit the uptake of the method. Statistical programs such as INLA and TMB provide alternative ways to conduct Bayesian inference which have great computational advantages over MCMC samplers. These computational advantages are described in detail for INLA in Rue et al. (2009) and for TMB in Kristensen et al. (2016), and involve the use of Laplace approximations to perform

posterior inference at a fraction of the time it would take using MCMC algorithms. The methods are particularly suited to space and space-time modeling with Markov random field models—situations in which MCMC can be especially computationally demanding because of the dependence in the posterior (Margossian et al., 2020).

The benchmarking approach we propose in the following section is a more general implementation of the inexact, fully Bayesian approach described by Zhang and Bryant (2020), and allows us to obtain fully Bayesian benchmarked estimates from any method that produces area-level samples from an unbenchmarked model. The approach we propose is readily applicable to both area- and unit-level models, so long as area-level samples can be produced, and can be used in conjunction with statistical programs such as INLA and TMB.

2.3.4 Proposed approach

We propose an external benchmarking approach that combines an unbenchmarked model with either a rejection sampler or Metropolis-Hastings algorithm to produce fully Benchmarked posterior distributions conditional on a benchmarking constraint under *inexact* benchmarking. The key to our proposed approach is that from Equation (2.4) we can write

$$\begin{aligned}\pi(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{y}_1, y_2) &\propto \pi(\mathbf{y}_1 \mid \boldsymbol{\theta}, \boldsymbol{\phi})\pi(y_2 \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta} \mid \boldsymbol{\phi})\pi(\boldsymbol{\phi}), \\ &\propto \pi(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{y}_1)\pi(y_2 \mid \boldsymbol{\theta}).\end{aligned}$$

Intuitively, we can think of $\pi(y_2 \mid \boldsymbol{\theta})$ as a likelihood for the benchmarks and $\pi(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{y}_1)$ as a "prior" that corresponds to the posterior based on area-level observations $\mathbf{y}_1$. For concreteness we consider a normal distribution for $\pi(y_2 \mid \boldsymbol{\theta})$, i.e.,

$$y_2 \mid \boldsymbol{\theta} \sim N\left(\sum_{i=1}^{n} w_i\theta_i, \sigma^2_{y_2}\right),$$

where w_i are population weights that sum to one across all regions, and $\sigma^2_{y_2}$ is treated as the known, national level variance for y_2. One could consider other distributions for the

benchmark likelihood, however we focus on the normal case in the following derivations.

Importantly, $\pi(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{y}_1)$ is the posterior distribution from an unbenchmarked model, i.e., a model that is agnostic to the benchmarking constraint. In practice, this is the small area, subnational model fit without consideration of the national benchmarks. The most appropriate unbenchmarked subnational model will depend on the context of the statistical problem, though we note that if unbenchmarked subnational estimates are particularly far from reliable national benchmarks, this may indicate that the unbenchmarked model is inappropriate for the data. The approach we describe below will produce samples from the posterior conditional on the benchmarking constraint, or "benchmarked posterior," for a given unbenchmarked model.

We note that though we target the same constrained posterior distribution as Zhang and Bryant (2020), our approach is distinct in the implementation in that we do not obtain samples from this posterior distribution in a single step, such as with an MCMC algorithm as implemented in Zhang and Bryant (2020). Rather, our approach allows for first obtaining samples from an unbenchmarked model (not necessarily using MCMC, hence greater flexibility and potential speed gains) and benchmarking in a second step involving either a rejection sampler or a Metropolis-Hastings algorithm.

Rejection sampler

In a rejection sampling framework, we can obtain samples from $\pi(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{y}_1, y_2)$ by filtering samples from $\pi(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{y}_1)$ through the information provided by $\pi(y_2 \mid \boldsymbol{\theta})$. Note that in the following we assume $\pi(y_2 \mid \boldsymbol{\theta})$ is not an indicator function (i.e., we are working under inexact benchmarking). This filtering is done via the following rejection sampler (Smith and Gelfand, 1992):

1. Generate $U \sim \text{Uniform}(0, 1)$ and $(\boldsymbol{\theta}, \boldsymbol{\phi}) \sim \pi(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{y}_1)$ independently.

2. Accept $(\boldsymbol{\theta}, \boldsymbol{\phi})$ if

$$U < \frac{\pi(y_2 \mid \boldsymbol{\theta})}{\sup_{\boldsymbol{\theta}} \pi(y_2 \mid \boldsymbol{\theta})} = \exp\left(-\frac{1}{2\sigma_{y_2}^2}\left(\sum_{i=1}^{m} w_i\theta_i - y_2\right)^2\right),$$

Otherwise, return to Step 1.

This rejection sampling approach targets the same constrained posterior distribution as the fully Bayesian benchmarking approach of Zhang and Bryant (2020) but in a computationally straightforward manner. As the rejection sampling approach only requires posterior draws from an unbenchmarked posterior distribution, this allows practitioners to use a wider array of computation tools to conduct fully Bayesian benchmarking, even when benchmarking constraints are nonlinear. A relevant example of such a computational tool is INLA, which does not allow for non-linear predictors, and therefore INLA cannot directly incorporate the likelihood from a nonlinear benchmark into the model fitting (Rue et al., 2009). Crucially, unbenchmarked posterior samples can be generated from an INLA analysis.

Metropolis-Hastings

The rejection sampling approach to fully Bayesian benchmarking may be inefficient in cases where benchmarks are far from the population aggregated estimates from an unbenchmarked model. One potential way to address this inefficiency is to instead use an independence Metropolis-Hastings (MH) algorithm (Tierney, 1994). Intuitively, we can fit an *adjusted* unbenchmarked distribution that has been shifted towards the national benchmarks and adjust for this shift in the acceptance rate. This serves to increase the proportion of accepted samples relative to using the unbenchmarked distribution as a proposal distribution by moving the proposal distribution closer to where the constrained posterior should be.

Here we present one example of an adjusted unbenchmarked distribution that can be used as a proposal distribution, and show how the shift can be corrected for in the acceptance rate. Note that we assume in this example that the unadjusted, unbenchmarked model we

would want to fit has a fixed intercept β with a flat prior, $\pi(\beta) \propto 1$. This assumption is needed for terms to cancel in the acceptance rate, and is not an unreasonable assumption given that the models fit in our motivating application typically satisfy this assumption (Wu et al., 2021). We note that this assumption can be relaxed if other proposal distributions are considered, which may improve the acceptance rate, but that the acceptance probability may not cancel as conveniently in these cases.

Suppose that area-level parameters are linked to a mean model $g(\boldsymbol{\theta}) = \beta + \boldsymbol{\eta}$ via a link function g, where β is a fixed intercept term with $\pi(\beta) \propto 1$, and $\boldsymbol{\eta}$ denoting a summation of any remaining fixed or random effect terms in the unbenchmarked model (e.g., spatio-temporal smoothing terms, covariates, etc.). Let $\pi^+(\beta) \sim N(g(y_2), \sigma_+^2)$ denote an alternative prior for the intercept, centered at $g(y_2)$ with fixed variance σ_+^2. Then we have

$$\begin{aligned}
\pi(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{y}_1, y_2) &\propto \pi(\mathbf{y}_1 \mid \boldsymbol{\eta}, \beta, \boldsymbol{\phi})\pi(y_2 \mid \boldsymbol{\eta}, \beta)\pi(\boldsymbol{\eta} \mid \boldsymbol{\phi})\pi(\boldsymbol{\phi})\pi(\beta), \\
&= \pi(\mathbf{y}_1 \mid \boldsymbol{\eta}, \beta, \boldsymbol{\phi})\pi(y_2 \mid \boldsymbol{\eta}, \beta)\pi(\boldsymbol{\eta} \mid \boldsymbol{\phi})\pi(\boldsymbol{\phi})\pi(\beta) \left(\frac{\pi^+(\beta)}{\pi^+(\beta)} \right), \\
&\propto \pi^+(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{y}_1) \left(\frac{\pi(y_2 \mid \boldsymbol{\eta}, \beta)}{\pi^+(\beta)} \right),
\end{aligned}$$

where $\pi^+(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{y}_1)$ is an adjusted unbenchmarked distribution with $\pi^+(\beta)$ specified as the prior for the intercept.

In this framework, we can obtain samples from the same benchmarked posterior distribution $\pi(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{y}_1, y_2)$ using independent samples from the adjusted unbenchmarked posterior distribution $\pi^+(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{y}_1)$. The algorithm is executed as follows:

1. Initalize $(\beta^0, \boldsymbol{\eta}^0)$.

2. Sample $(\beta', \boldsymbol{\eta}') \sim \pi^+(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \mathbf{y}_1)$.

3. Compute the acceptance probability

$$A = \min\left(1, \frac{\pi(\boldsymbol{\theta}', \boldsymbol{\phi}' \mid \mathbf{y}_1, y_2)\pi^+(\boldsymbol{\theta}^0, \boldsymbol{\phi}^0 \mid \mathbf{y}_1)}{\pi(\boldsymbol{\theta}^0, \boldsymbol{\phi}^0 \mid \mathbf{y}_1, y_2)\pi^+(\boldsymbol{\theta}', \boldsymbol{\phi}' \mid \mathbf{y}_1)}\right)$$
$$= \min\left(1, \frac{\pi(y_2 \mid \boldsymbol{\eta}', \beta')\pi^+(\beta^0)}{\pi(y_2 \mid \boldsymbol{\eta}^0, \beta^0)\pi^+(\beta')}\right)$$

4. With probability A, accept the proposed value $(\beta', \boldsymbol{\eta}')$ and set $\beta^0, \boldsymbol{\eta}^0) = \beta', \boldsymbol{\eta}')$. Otherwise, set $(\beta^0, \boldsymbol{\eta}^0) = (\beta^0, \boldsymbol{\eta}^0)$.

Just as with the rejection sampling approach, the MH approach targets the same constrained posterior distribution as the fully Bayesian benchmarking approach of Zhang and Bryant (2020) but in a computationally convenient manner.

Convergence diagnostics

While the MH algorithm may have a higher acceptance rate than the rejection sampling approach (and therefore greater computational speed), additional care must be taken to ensure that the algorithm has mixed and converged properly. Though not a convergence diagnostic, one basic check is to aim for an acceptance rate near 23.4%, as suggested in Gelman et al. (1997) as the optimal acceptance rate under normal posteriors for random walk Metropolis algorithms. It may be desirable to vary the prior $\pi^+(\beta)$ in the MH algorithm to obtain an acceptance rate close to this optimal value.

A common convergence diagnostic for MCMC sampling is the potential scale reduction factor $\hat{R}$, introduced by Gelman and Rubin (1992). Intuitively, $\hat{R}$ (or split-$\hat{R}$ as it is extended in Gelman et al. (2013)) compares the variance of each individual Markov chain to the variance of all of the chains combined. If the sampler has mixed appropriately, $\hat{R}$ should be close to 1. More recently, new convergence diagnostics have been developed for MCMC methods that go beyond $\hat{R}$, as there are certain situations where $\hat{R}$ can fail to correctly diagnose poor mixing. In brief, Vehtari et al. (2021) introduce rank-normalized split-$\hat{R}$ and bulk effective sample size (bulk-ESS) as measures that have good asymptotic efficiency

through avoiding a normality assumption in non-rank-normalized measures. For more detail see Vehtari et al. (2021) for an overview of existing convergence diagnostics and details on those they propose.

Following the guidelines in Vehtari et al. (2021), we recommend running at least four chains if using the MH approach to benchmarking that we propose. Aiming for a rank-normalized split-$\hat{R}$ less than 1.01 and a bulk-ESS that exceeds 400 (with 4, 1000-iteration chains after warmup) provide default guidelines. In addition to these numeric checks, we recommend running multiple chains and manually reviewing traceplots to see if the chains provide approximately the same answers, to get a practical handle on whether the difference in chains is important or not.

Limitations

Despite the potential improvements in speed from the MH approach over the rejection sampler approach, both methods may still be inefficient in cases where benchmarks are far from the population aggregated estimates from an unbenchmarked model. It may be necessary in such cases to use an approach such as Zhang and Bryant (2020)'s in order to produce benchmarked estimates. However, we caution that if the proportion of accepted samples is very small this may be indicative of inconsistencies between the two data sources, and model elaboration may be required in this case.

Additionally, the rejection sampling and MH approaches cannot benchmark estimates with zero variance, as the distribution $\pi(y_2 \mid \boldsymbol{\theta})$ would be a point mass. We note this benchmarking scenario could be approximated by setting $\sigma^2_{y_2}$ to be very small, but the rejection sampling approach would likely be very inefficient due to the likelihood for the benchmark being very concentrated. The scenario where benchmarks have positive uncertainty is the most practically relevant in our context, as situations where the national benchmark does not have some degree of uncertainty in an LMIC context is rare. For fully Bayesian benchmarking to benchmarks with zero variance, we refer the reader to the MCMC methods described in Zhang and Bryant (2020).

A final consideration with the MH approach is that using this algorithm with the suggested proposal distribution will not allow for comparison between the resulting benchmarked estimates and the unbenchmarked estimates from the unadjusted, unbenchmarked model, without separately fitting an unadjusted, unbenchmarked model. In some official statistics settings it may be desirable or even required to compare benchmarked estimates to unbenchmarked estimates, and if the time it takes to separately fit an unadjusted, unbenchmarked model is time-consuming, the MH approach may be impractical.

2.4 Simulation

To demonstrate the improved computational speed of our approach compared to Zhang and Bryant (2020), we compare run-times for our approach used with INLA to that of Zhang and Bryant (2020) implemented in Stan; a probabilistic programming language that uses a variant of Hamiltonian Monte Carlo to do full Bayesian inference (Carpenter et al., 2017). We chose not to compare the computational speed of our approach to that of Datta et al. (2011) or the raking approach, as both of those benchmarking approaches target different benchmarked estimates than our proposed method. Note however that both the raking and benchmarked Bayes estimate approach will be the fastest approach to benchmarking in general, as they involve a simple, post-hoc adjustment to unbenchmarked draws/estimates, and do not rely on acceptance rates (as our methods do to achieve a reasonable number of effective samples) or MCMC methods.

We simulate unit-level (cluster) binomial observations, using the nine provinces (small areas) of South Africa as our spatial structure. For each simulation setting, binomial probabilities were given by $p_i = 0.28, 0.29, \ldots, 0.35, 0.36$, with 100 binomial trials in each cluster. Equal probability weights were given to each province. We varied: the number of samples taken in each area, $m = \{5, 10, 100, 1000\}$, the national-level benchmark, $y_2 = \{0.29, 0.3\}$, and the variance of the national-level benchmark, $\sigma^2_{y_2} = \{0.01, 0.0001\}$. For each simulation setting, we generated 10 unique datasets using the given parameters.

The unbenchmarked model we fit to the generated data is the unit-level model used in

our application to South Africa (and described in Section 2.2.1), where we have binomial observations for clusters c within area i. Let $y_{1i[c]}$ be the number of cases observed in $n_{i[c]}$ total observations in cluster c within area i. We consider the unbenchmarked, unit-level model

$$\begin{aligned} y_{1i[c]} \mid n_{i[c]}, \theta_{i[c]} &\sim \text{Binomial}(n_{i[c]}, \theta_{i[c]}), \\ \eta_{i[c]} = \text{logit}(\theta_{i[c]}) &= \beta_0 + b_i + e_{i[c]} \end{aligned}$$

where $\theta_{i[c]}$ is case prevalence in area i and cluster c, β_0 is an intercept term, b_i follows a BYM2 model (Riebler et al., 2016), denoted $b_i \sim \text{BYM2}(\tau_b, \phi)$, and $e_i[c] \overset{iid}{\sim} \text{N}(0, \sigma_e^2)$ is an independent and identically distributed (iid) cluster-level random effect. Note that although the model we fit was not used to generate the data, as we are interested only in comparing run times from each method, and ensuring that the benchmarked distributions are the same from each method, this does not affect the validity of our simulation results. More details on hyperprior specification and simulating the data can be found in Section 3 of the Appendix A.

The modified unbenchmarked model, used for the MH algorithm is the same as above with the exception of the prior for the intercept being $\pi^+(\beta_0) \overset{d}{=} \text{N}(\text{logit}(y_2), \sqrt{0.1})$, where y_2 is the national-level benchmark. The variance of $\sqrt{0.1}$ was chosen in part to reflect how national-level estimates are much more precise than area-level analyses permit. The benchmarked model we fit using the Zhang and Bryant (2020) approach includes the additional likelihood

$$y_2 \mid \boldsymbol{\theta}, \sigma_{y_2}^2 \sim \text{N}\left(\sum_{i=1}^{n} w_i \theta_i, \sigma_{y_2}^2\right),$$

and the proposed rejection sampler and MH algorithm approaches are carried out as described in subsections 2.3.4 and 2.3.4. Code for reproducing the simulation can be found at github.com/taylorokonek/stbench.

To fairly compare run-times, we compare the time it takes to:

1. Fit the fully Benchmarked model per Zhang and Bryant (2020) in Stan, and obtain a bulk-ESS of 1000.

2. Fit the unbenchmarked model in INLA, draw posterior samples, and obtain 1000 accepted samples using the rejection sampler.

3. Fit the modified unbenchmarked model in INLA, draw posterior samples, and obtain a bulk-ESS of 1000 using the MH algorithm.

For both Stan and the MH algorithm, we run four chains with 1000 burn-in samples each, and the appropriate number of samples after to obtain the desired bulk-ESS.

In Figure 2.1 we compare methods under the setting $\{y_2 = 0.29, \sigma_y^2 = 0.01\}$, and in Figure 2.2 we compare methods under the setting $\{y_2 = 0.29, \sigma_y^2 = 0.0001\}$. As expected, our proposed approaches outperform that of Zhang and Bryant (2020) in most simulation settings, as noted in Figures 2.1 and 2.2. This is in large part due to the fact that the Zhang and Bryant (2020) approach relies on MCMC methods, whereas our approach is compatible with Bayesian computational programmes that take advantage of Laplace approximations. In particular, the amount of time needed to obtain 1000 samples from the benchmarked posterior distribution is much lower for both of our proposed approaches than that of Zhang and Bryant (2020) when the number of samples in each area is large. When the national variance is smaller, as in Figure 2.2, the Zhang and Bryant (2020) tends to outperform both the rejection sampler and MH algorithm at low sample sizes (5 and 10 per area), but not at larger sample sizes. Finally, note that, in settings with smaller national variance, the Metropolis-Hastings algorithm outperforms the rejection sampler approach in terms of computational speed, while the reverse is true when national variance is large. Under the setting $\{y_2 = 0.29, \sigma_y^2 = 0.01, m = 5\}$, our proposed methods are 3-5 times faster than that of Zhang and Bryant (2020) on average, and under the setting $\{y_2 = 0.29, \sigma_y^2 = 0.01, m = 1000\}$, our proposed methods are 75-100 times faster that that of Zhang and Bryant (2020) on average. The speed of the MH algorithm could potentially be optimized by modifying prior

for the intercept, $\pi^{+}(\beta_0)$, and we suggest that when possible, multiple priors should be tested if the acceptance rate in the MH algorithm is lower than desired. The included figures and above summary are representative of the simulation results, and additional results from the simulation can be found in Section 3 of Appendix A.

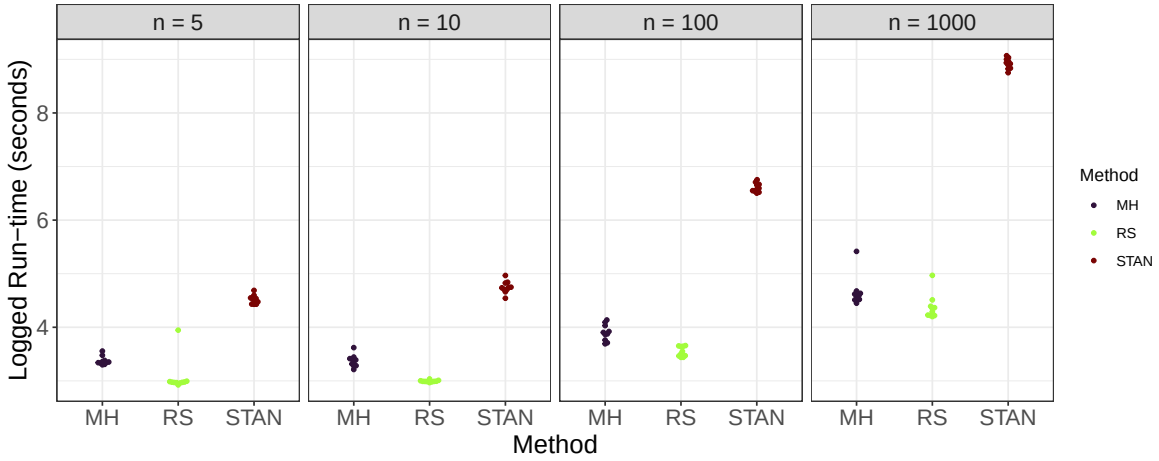


Figure 2.1: Setting: $y_2 = 0.29$, $\sigma^2_{y_2} = 0.01$

Comparative, total run-time (in seconds) needed to obtain 1000 samples from the rejection sampling (RS) approach, or 1000 bulk-ESS from the Metropolis-Hastings (MH) approach or the approach of Zhang and Bryant (2020) (Stan). Each beeswarm plot contains 10 observations, with data generated under the given simulation setting for 10 different seeds.

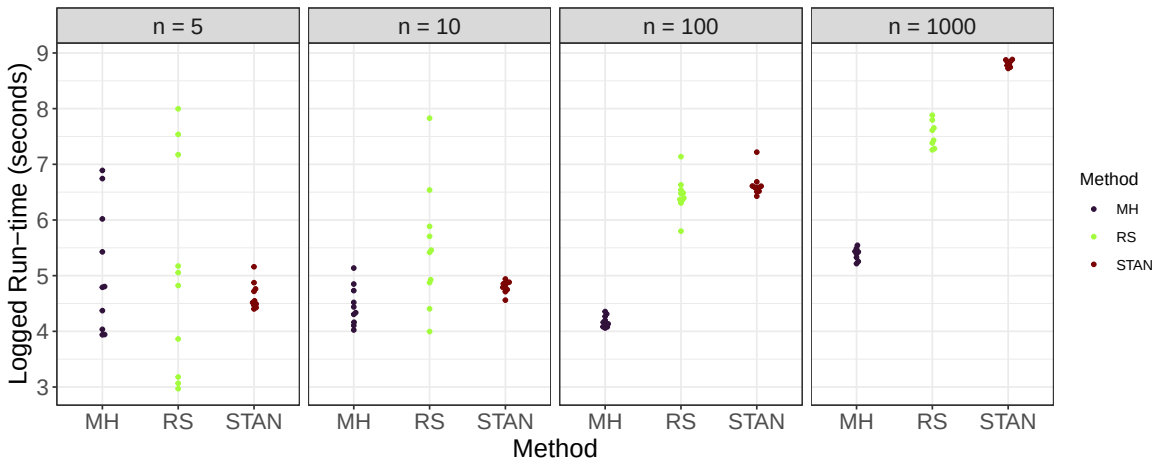


Figure 2.2: Setting: $y_2 = 0.29$, $\sigma^2_{y_2} = 0.0001$

Comparative, total run-time (in seconds) needed to obtain 1000 samples from the rejection sampling (RS) approach, or 1000 bulk-ESS from the Metropolis-Hastings (MH) approach or the approach of Zhang and Bryant (2020) (Stan). Each beeswarm plot contains 10 observations, with data generated under the given simulation setting for 10 different seeds.

2.5 Application

Below we describe the data, models, and software used in the HIV prevalence application. The application to U5MR is in Section 2 of Appendix A.

We fit both area-level and unit-level models to demonstrate the flexibility of our proposed method over existing computational software. As noted in subsection 2.3.3, there is currently no available implementation of the fully Bayesian benchmarking approach described in Zhang and Bryant (2020) to unit-level models. As unit-level models are commonly used in LMICs when estimating public health outcomes in an official statistics setting (Wakefield et al., 2020; Wu et al., 2021), it is important for practitioners to have a method available to conduct fully Bayesian benchmarking with unit-level models without needing to write their own MCMC algorithm. We demonstrate that our proposed method is readily applicable to both area- and unit-level models.

For our unbenchmarked area-level model, we consider the spatial Fay-Herriot model defined in Mercer et al. (2015). The model is detailed in Section 2.2.1. The Fay-Herriot model

is one of the most well-known small area models, that directly incorporates aspects of the survey design using a transformed survey-weighted estimate and an iid random effect to increase precision of the resulting estimates (Fay and Herriot, 1979). The spatial extension incorporates both an iid random effect *and* an additional structured spatial random effect to further capture spatial dependency in the observations (Mercer et al., 2015).

For our unbenchmarked unit-level model, we consider a hierarchical Bayesian model with a spatial random effect term for the application to HIV prevalence, and the model currently used to produce subnational estimates of U5MR for the UN IGME (Wu et al., 2021). The model is detailed in Section 2.2.1. Of note, these unit-level models do not directly account for the survey design like the area-level model. It has been suggested that, when possible, the inclusion of covariates related to the survey design could be included in such a model-based framework to account for the survey design (Wakefield, 2013; Wu et al., 2021). One example of this can be found in Paige et al. (2022), though as noted in Wakefield et al. (2020), in many surveys conducted in LMICs the covariates corresponding to the survey design may be unavailable.

2.5.1 Data

Spatial boundary files for South Africa are obtained from GADM, the Database of Global Administrative Areas (Database of Global Administrative Areas (GADM), 2019).

We estimate HIV prevalence in South Africa from the 2016 South Africa DHS survey. The survey followed a multi-stage, stratified design, and was designed to provide estimates at the Administrative 1 (admin1) level, which consists of nine provinces, which we considered to be our small areas for this application. Each of the nine provinces were stratified by urban/farm/traditional area status, and therefore resulted in 26 strata, as the Western Cape province does not have a traditional area geotype. The sampling frame was established from the 2011 census, and 750 enumeration areas (primary sampling units, PSUs) were selected across strata. The second stage of sampling sampled dwelling units, or households, from the enumeration areas, and every individual within the household (if available) was included in

the survey. Only men and women aged 15-49 were included in the HIV dataset. Households within a given enumeration area are given a single geographic location, and we denote these as clusters from here on. GPS coordinates are displaced by up to 2km for urban clusters and 5km for rural clusters, but are never displaced outside of their area of stratification.

Of note, all nine small areas (provinces) used in model fit for this application contained at least one binomial observation. The number of PSUs in each small area ranged from 56 to 88. The binomial counts within each primary sampling unit ranged from 0 to 11, with totals in each primary sampling unit ranging from 1 to 30. The distribution of observed binomial proportions in each PSU can be seen in Figure 2.3.

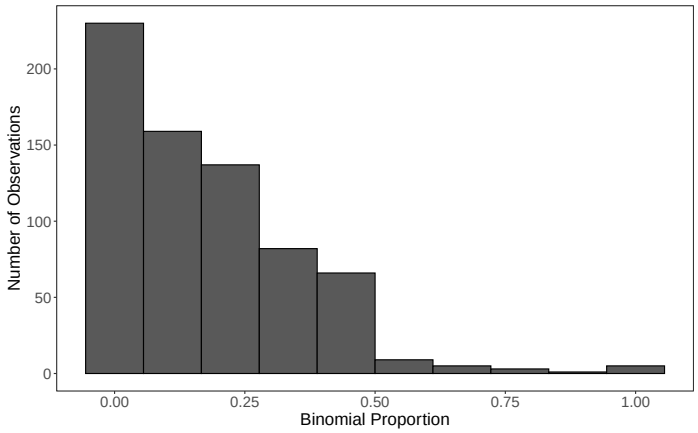


Figure 2.3: Observed binomial proportions in each PSU.

We obtain a national level estimate for 2016 in South Africa from the national Thembisa model, an HIV epidemic projection model that produces the official estimates published by UNAIDS for South Africa (Johnson et al., 2017b; Mahy et al., 2019; Stover et al., 2019). The national level estimate of HIV prevalence for 2016 in South Africa is 17.1%, with a 95% confidence interval of (15.6%, 18.3%). For the benchmark likelihood for the fully Bayesian benchmarking approaches we need a standard error for this benchmarked estimate, which we take to be $(18.3 - 17.1)/1.96 = 0.61$ based on the assumption that the national level estimate is asymptotically normally distributed. The Thembisa model incorporates data from

five Human Sciences Research Council (HSRC) surveys conducted from 2002 to 2017, the 2016 DHS survey, Antenatal Sentinel HIV and Syphilis Surveys, and antiretroviral therapy coverage data. Although the DHS survey is included in the model for the benchmarks, we treat this as an external benchmarking scenario as multiple other data sources are included in the national model as well.

The population count data we use for our HIV application comes from the provincial Thembisa model (Johnson et al., 2017a). We use this specific population data as they are currently used in UNAIDS models of subnational HIV prevalence in South Africa (Eaton et al., 2021). In Figure 2.4, we plot the spatial distribution of individuals aged 15-49 (the population in our HIV example) in 2016 in South Africa, i.e., out of all people aged 15-49 in South Africa, the proportion of that sub-population who live in each region is displayed. The population proportions are also presented in 2.1, along with the number of observations in each province and their direct, survey-weighted estimates.

Table 2.1: Proportion of the 15-49 population living in each province, number of observations in each province, direct survey estimates, and their standard errors provided in parentheses. Provinces are arranged in ascending order of direc estimates.

Province	**Population Proportion**	**Sample Size**	**Direct Estimate (%)**
Limpopo	0.09	815	9.8 (1.19)
Northern Cape	0.02	453	11.4 (2.25)
Western Cape	0.12	397	12.3 (2.96)
Eastern Cape	0.10	1056	16.4 (1.40)
Gauteng	0.28	528	19.6 (2.07)
North West	0.07	917	20.6 (1.52)
Free State	0.05	849	21.3 (1.76)
Mpumalanga	0.08	893	23.9 (1.84)
KwaZulu-Natal	0.19	1004	25.9 (2.08)

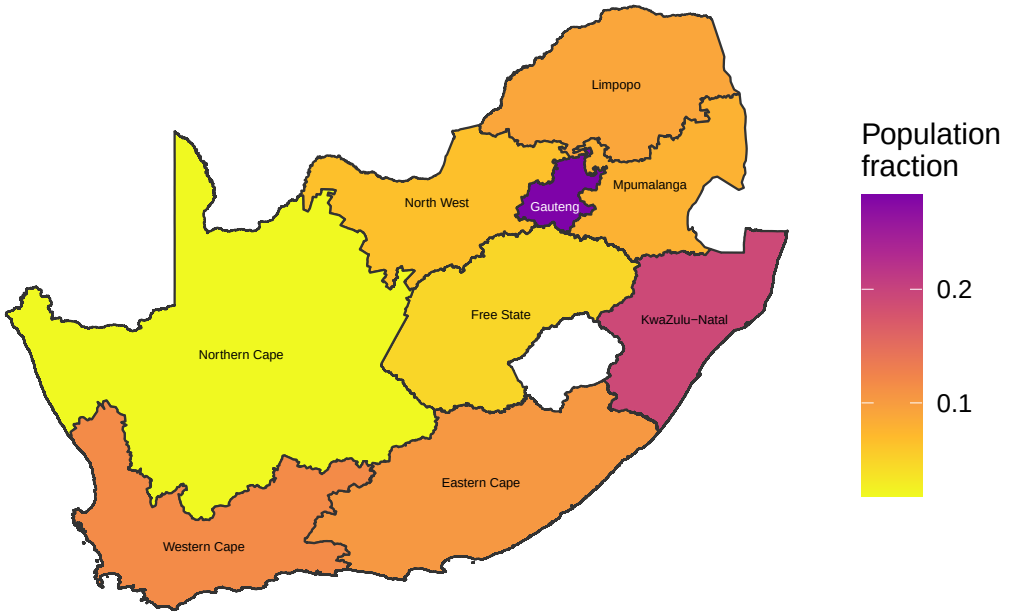


Figure 2.4: Proportion of the 15-49 population living in each province. These proportions are the weights used to aggregate province level estimates to the national level.

2.5.2 Benchmarked models

We compute benchmarked estimates using the Fully Bayesian approach of Zhang and Bryant (2020), the proposed approaches, and the raking approach. This allows us to show: (1) the benchmarked estimates from our proposed approaches are identical to those using the Zhang and Bryant (2020) method, and (2) a comparison of the benchmarked estimates from the proposed approach to what is currently used in practice (raking) by the UN when estimating both HIV Prevalence and U5MR subnationally (Eaton et al., 2021; Wu et al., 2021). We do not compute benchmarked estimates using the benchmarked Bayes estimate approach, as we have noted in Section 2.3.1 that is inappropriate for these applications. We obtain credible intervals for our benchmarked and unbenchmarked estimates using draws from the posterior distributions.

Approaches

Raking:

We obtain posterior medians $\hat{\theta}_i^M$ from an unbenchmarked model for each area i, and compute R from Equation (2.3), using weights w_i and our benchmark y_2, setting $\hat{\theta}^N := \sum_{i=1}^n w_i \hat{\theta}_i^M$. We then adjust unbenchmarked posterior draws $\hat{\theta}_i^{(k)}$, $k = 1, \ldots, K$, to obtain benchmarked posterior draws $\hat{\theta}_i^{R(k)} = \hat{\theta}_i^{(k)}/R$.

Fully Bayesian: Zhang and Bryant (2020)

Using weights w_i and our benchmark y_2 with associated variance $\sigma_{y_2}^2$, we fit the unbenchmarked model with the additional likelihood

$$y_2 \mid \boldsymbol{\theta} \sim \text{Normal}\left(\sum_{i=1}^m w_i \theta_i, \sigma_{y_2}^2\right), \tag{2.5}$$

with $y_2 = 17.1\%$ and $\sigma_{y_2}^2 = 0.37$, and denote benchmarked posterior draws by $\hat{\theta}_i^{FB1(k)}$.

Fully Bayesian: Rejection sampler:

We obtain $k = 1, \ldots, K$ posterior draws $\hat{\theta}_i^{(k)}$ for each area i from the unbenchmarked unit level (or area level) model, and apply the algorithm described in Section 2.3.4 using the benchmark likelihood in Equation (2.5) to obtain benchmarked posterior draws $\hat{\theta}_i^{FB2(k)}$.

Fully Bayesian: Metropolis-Hastings algorithm:

We obtain $k = 1, \ldots, K$ posterior draws $\hat{\theta}_i^{(k)}$ for each area i from the *adjusted* unbenchmarked unit level or area level model, and apply the algorithm described in Section 2.3.4 using the benchmark likelihood in Equation (2.5) to obtain benchmarked posterior draws $\hat{\theta}_i^{FB3(k)}$. The adjusted unbenchmarked model uses the prior $\pi^+(\beta_0) \sim \text{Normal}(\text{logit}(y_2), 0.001^{-1})$.

2.5.3 Model Validation

As we view benchmarking as a necessary adjustment needed to an existing unbenchmarked model in an official statistics setting, we choose to validate the unbenchmarked models we fit *prior* to benchmarking. Note that if benchmarking is instead viewed as a way to reduce potential bias in subnational estimates, model validation may need to be done to the resulting benchmarked models as opposed to the unbenchmarked models.

Wakefield et al. (2020), among others (Wu et al., 2021; Osgood-Zimmerman et al., 2018), suggest that one way to approach model validation in small area models in a survey setting is to check whether the direct estimates in each area lie within an uncertainty interval surrounding the posterior means of the predictive distribution in that area. This also gives one an idea of average coverage across areas: 80% of the direct estimates (left out data) should lie within 80% credible intervals based on the predictive distribution. In a survey setting in LMICs, the direct estimates are considered the gold standard; hence, model-based estimates should not stray too far from the direct estimates.

We take this suggested approach to model validation for both our unbenchmarked area-level and unit-level models. Details and figures containing the results of model validation for the HIV application can be found in Section 1.4 of Appendix A, and Section 2.6 of Appendix A data for the U5MR application.

2.5.4 Computation

As the novelty of our approach is computational, we chose the statistical programmes used in the application to demonstrate the flexibility of our approach compared to the fully Bayesian approach of Zhang and Bryant (2020).

For the approach described in Zhang and Bryant (2020), we implement models in Template Model Builder (TMB) (Kristensen et al., 2016). TMB is a flexible modeling tool that takes advantage of Laplace approximations for computational efficiency, and allows users to specify a wide variety of models in `C++`. For a review of TMB, see Section 4 of Osgood-

Zimmerman and Wakefield (2021). Importantly, TMB allows users to specify nonlinear predictors, unlike INLA, which is needed for implementing the fully Bayesian benchmarking approach of Zhang and Bryant (2020) in settings with binary outcomes. Of note, TMB still requires that users write their own `C++` implementations for models, and therefore does not have as nice of an interface as other programs (such as INLA, which we will describe below).

To show that the results from our proposed method are identical to those from the Zhang and Bryant (2020) method, we also fit the unbenchmarked HIV models in our application using TMB. This ensures that the posterior distributions are directly comparable. Not that our proposed method could have been implemented using INLA to fit the unbenchmarked model as well, but as INLA uses a different Laplace approximation than TMB, the results would not have been exactly, directly comparable.

For our U5MR application, we fit all unbenchmarked models in INLA to show that our proposed method is compatible with the software currently used to produce official estimates of U5MR for the UN IGME. INLA is an appealing, alternative Laplace approximation programme for obtaining posterior distributions for latent Gaussian models (Rue et al., 2009). Unlike TMB, INLA does not allow users to specify nonlinear predictors. Since our proposed approach does not require the use of such nonlinear predictors for the benchmarking constraint by taking advantage of a two-step procedure, we implement the models for our U5MR application using INLA to demonstrate the benefit of the increased flexibility of our approach to accommodate fast Laplace approximation techniques such as INLA.

2.5.5 Results

In Figure 2.5 and Table 2.2, we display national level estimates from Thembisa, the unbenchmarked, and benchmarked results from the unit-level model. Results from the area-level model are very similar to those from the unit-level model, and can be found in Table 2.2 as well, with more detailed results in Section 1.2 of the Supplemental Data and direct comparisons between the unit-level and area-level models found in Section 1.3 of Appendix A. We see that the Fully Bayesian benchmarking approaches produce essentially identical

results, and that they are a compromise between the likelihood given by the national level benchmark and the unbenchmarked posterior estimate. The raking approach enforces exact benchmarking, which is evidenced by the overlap between the density of the Thembisa national estimate and that of the benchmarked method's national aggregated estimate. The fully Bayesian rejection sampler approach accepts 7.6% of unbenchmarked samples using the unit-level model, and 8.6% of unbenchmarked samples using the area-level model. The fully Bayesian Metropolis-Hastings approach accepts 18.8% of unbenchmarked samples using the unit-level model, and 18.9% of unbenchmarked samples using the area-level model.

Table 2.2: Aggregated national level HIV prevalence estimates from Thembisa, unbenchmarked, and benchmarked results from the unit-level and area-level models. 95% credible intervals are given next to posterior medians.

Model	**Unit-Level: Median (%)**	**Area-Level: Median (%)**
Thembisa	17.1 (15.6, 18.3)	17.1 (15.6, 18.3)
Unbenched	19.1 (17.6, 20.6)	19.0 (17.5, 20.7)
FB	18.0 (17.1, 19.0)	17.9 (17.0, 18.8)
FB: Rejection sampler	17.9 (17.0, 18.9)	17.9 (17.0, 18.8)
FB: MH	17.9 (17.0, 18.8)	17.9 (16.9, 18.8)
Raking	17.1 (15.8, 18.5)	17.1 (15.8, 18.7)

Figure 2.6 and Table 2.3 display the subnational breakdown of benchmarked and unbenchmarked estimates from the unit-level model. Of note, the benchmarked posterior medians and credible intervals for the Fully Bayesian approaches look roughly identical (as they should), and the posterior distributions for each province from the raking approach are all slightly lower than those from the Fully Bayesian approaches. This difference is due to the raking approach enforcing the benchmarking constraint exactly, thus pulling the unbenchmarked estimates down farther than the Fully Bayesian approaches, as the national benchmark is lower than the unbenchmarked national estimate. Though more difficult to see, the Fully Bayesian approach treats regions with more uncertainty differently than those with less uncertainty. In particular, note that in the Eastern Cape province the fully Bayesian benchmarked posterior medians are not as different from the unbenchmarked median as

in the Gauteng province, where uncertainty is much higher in the unbenchmarked estimate. This is opposed to the Raking approach, which shifts all estimates by the same multiplicative constant, regardless of uncertainty. Table 2.4 displays the subnational breakdown of benchmarked and unbenchmarked estimates from the area-level model, with additional results in Section 1.2 of Appendix A.

Table 2.3: Admin1 HIV prevalence estimates from unbenchmarked and benchmarked unit-level models. Point estimates provided are medians, with 95% credible intervals. Provinces are arranged in the order of lowest unbenchmarked median to highest.

Province	**Unbenched (%)**	**FB (%)**	**FB: Rejection sampler (%)**	**FB: MH (%)**	**Raking (%)**
Limpopo	11.2 (9.0, 13.9)	11.0 (8.8, 13.6)	10.8 (8.8, 13.4)	10.8 (8.7, 13.4)	10.1 (8.1, 12.5)
Western Cape	11.5 (8.6, 15.3)	10.9 (8.2, 14.4)	10.7 (8.0, 14.0)	10.7 (8.2, 14.0)	10.4 (7.7, 13.8)
Northern Cape	12.0 (9.2, 15.4)	12.0 (9.2, 15.6)	11.8 (9.0, 15.3)	11.7 (8.9, 15.4)	10.8 (8.3, 13.8)
Eastern Cape	17.5 (15.6, 19.4)	16.9 (15.3, 18.7)	16.8 (15.1, 18.5)	16.7 (15.1, 18.5)	15.7 (14.1, 17.4)
Gauteng	18.6 (15.3, 22.5)	16.7 (14.1, 19.6)	16.7 (14.1, 19.5)	16.7 (14.1, 19.4)	16.7 (13.7, 20.2)
Free State	21.0 (17.9, 24.2)	20.6 (17.7, 24.1)	20.5 (17.5, 23.7)	20.5 (17.7, 23.7)	18.8 (16.1, 21.8)
North West	21.1 (18.0, 24.5)	20.7 (17.8, 24.0)	20.5 (17.6, 23.8)	20.5 (17.6, 23.7)	18.9 (16.2, 22.0)
Mpumalanga	23.4 (20.2, 27.0)	22.9 (19.8, 26.3)	22.8 (19.7, 26.2)	22.7 (19.8, 26.1)	21.0 (18.2, 24.3)
KwaZulu-Natal	26.4 (23.1, 30.0)	25.1 (22.1, 28.3)	25.1 (22.2, 28.2)	25.1 (22.2, 28.1)	23.7 (20.7, 26.9)

Table 2.4: Admin1 HIV prevalence estimates from unbenchmarked and benchmarked area-level models. Point estimates provided are medians, with 95% credible intervals. Provinces are arranged in the order of lowest unbenchmarked median to highest.

Province	**Unbenched (%)**	**FB (%)**	**FB: Rejection sampler (%)**	**FB: MH (%)**	**Raking (%)**
Limpopo	10.4 (8.3, 13.1)	10.1 (8.0, 12.8)	10.2 (8.0, 12.8)	10.1 (8.0, 12.8)	9.4 (7.5, 11.8)
Northern Cape	12.5 (8.9, 17.4)	12.2 (8.6, 17.0)	12.3 (8.7, 17.1)	12.4 (8.8, 17.3)	11.3 (8.0, 15.7)
Western Cape	12.6 (8.2, 18.8)	11.1 (7.5, 16.4)	11.0 (7.4, 16.2)	11.1 (7.5, 16.0)	11.3 (7.4, 17.0)
Eastern Cape	16.8 (14.6, 19.4)	16.3 (14.3, 18.5)	16.3 (14.1, 18.7)	16.2 (14.2, 18.6)	15.1 (13.1, 17.4)
Gauteng	19.3 (15.9, 23.6)	17.5 (14.7, 20.6)	17.4 (14.7, 20.5)	17.3 (14.5, 20.3)	17.4 (14.3, 21.2)
North West	20.2 (17.5, 23.3)	19.9 (17.2, 22.8)	19.9 (17.2, 22.9)	20.0 (17.3, 22.9)	18.2 (15.7, 21.0)
Free State	21.0 (17.9, 24.4)	20.6 (17.7, 24.1)	20.7 (17.6, 24.0)	20.7 (17.6, 24.0)	18.9 (16.1, 21.9)
Mpumalanga	23.4 (20.2, 27)	22.9 (19.7, 26.4)	22.9 (19.6, 26.4)	22.8 (19.6, 26.3)	21.1 (18.2, 24.3)
KwaZulu-Natal	25.5 (21.7, 29.4)	24.0 (20.7, 27.6)	24.0 (20.7, 27.4)	24.0 (20.6, 27.5)	22.9 (19.5, 26.5)

The model validation results for the HIV application for both unit- and area-level models

suggest that both model do not indicate gross misspecification. We compare posterior medians of the predictive distribution in each area (having left that area out of model fitting) to the direct estimate in each area, respectively. The 80% credible intervals capture 7/9 and 8/9 direct estimates in the unit- and area-level models, respectively, and the 50% credible intervals capture 3/9 and 4/9 in the unit- and area-level models, respectively.

Additional plots and tables, including those for the area-level model, can be found in Section 1 of Appendix A.

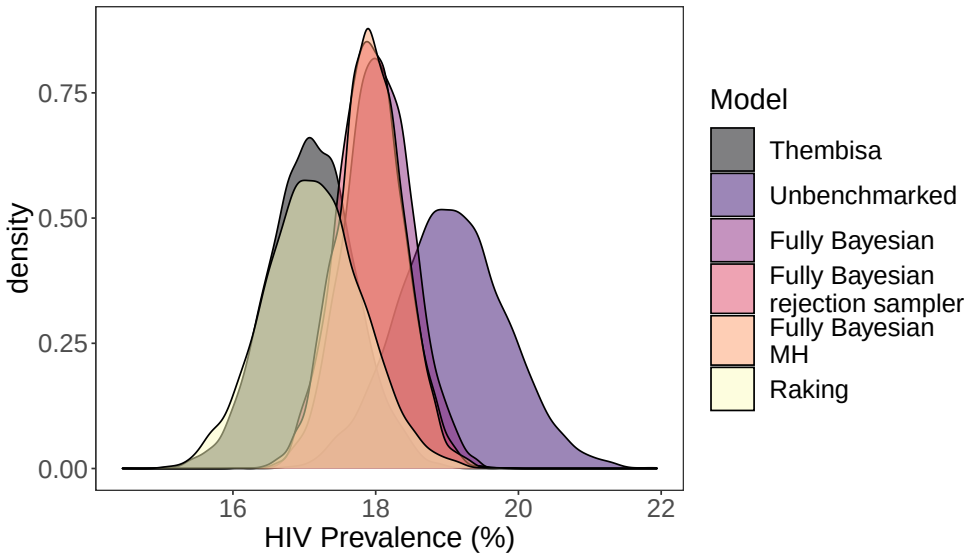


Figure 2.5: Aggregated national level HIV prevalence estimates from Thembisa, unbenchmarked, and benchmarked results from the unit-level model. The Fully Bayesian model refers to the approach described in Zhang and Bryant (2020), and the Fully Bayesian rejection sampler and Fully Bayesian MH (Metropolis-Hastings) refer to the proposed approach. All densities are based on 5000 samples.

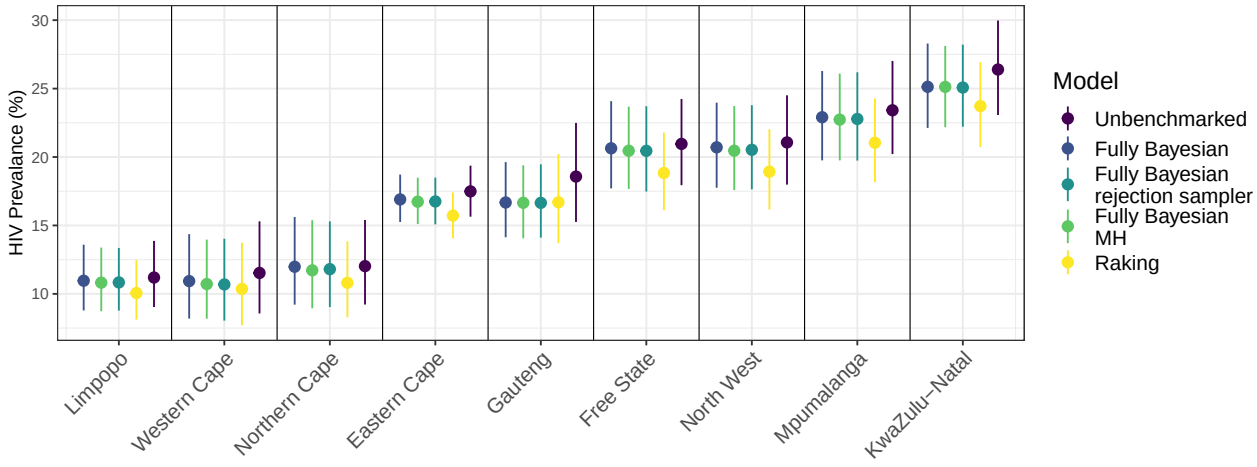


Figure 2.6: Comparison of HIV Prevalence estimates from benchmarked and unbenchmarked unit-level models at a subnational level. Error bars correspond to 95% credible intervals, and provinces are arranged from left to right in order of unbenchmarked median.

2.5.6 U5MR

The results for the U5MR application are in Section 2 of Appendix A.

2.6 Discussion

In this chapter we have summarized existing benchmarking approaches and their benefits and drawbacks with regard to data with rare binary outcomes in LMICs. We consider benchmarking methods that make use of a benchmarking constraint via one-step or two-step approaches, and pay particular attention to the resulting interpretation of the benchmarked estimates, the ease of uncertainty quantification, the acknowledgement of uncertainty in national estimates, the use of non-linear constraints, and computational tractability. We believe that the proposed rejection sampling and MH approaches to fully Bayesian benchmarking provide a desirable balance all of these concerns, and provide alternative computational approaches to the one-step, fully Bayesian benchmarking method developed by Zhang and Bryant (2020), which in many cases may be inflexible with regards to modeling choice and computational tractability, while targeting the same benchmarked, posterior distribution.

We show via an application of various benchmarking approaches to an HIV prevalence example that the proposed two-step approaches to fully Bayesian benchmarking produce the same benchmarked estimates as the one-step Zhang and Bryant (2020) approach, and that the resulting estimates are a compromise between the national level estimate and unbenchmarked estimates. The rejection sampling and MH approaches allow us to take advantage of potentially faster computational programs than traditional MCMC samplers, such as INLA or TMB, as evidenced via our application to U5MR. Inference for spatial models (continuous models especially) using MCMC methods is computationally challenging, and our approach allows users to conduct fully Bayesian benchmarking while relying on Laplace approximation methods for inference that are more suited for such models (Osgood-Zimmerman and Wakefield, 2021). Additionally, the rejection sampling and MH approaches are easily applied to both unit-level and area-level models, as evidenced in both our HIV and U5MR applications. Though not explored in this chapter, additional algorithms could be considered to similarly obtain benchmarked posterior distributions in a two-step fashion, such as the sampling-importance resampling approach (see Section 10.4 of Gelman et al. (2013)).

Though the purpose of this chapter was not to provide an in-depth review of small area models in LMICs, we note that our unit-level models do not directly account for the survey design. As with any application involving survey data, careful consideration should to be given to the survey design even if a model-based approach is required for adequate precision.

There are several limitations with the proposed approaches to fully Bayesian benchmarking. The first and potentially most pressing is the inability to conduct exact benchmarking, which may be required in some settings and is especially relevant if national benchmarks come from a census, though censuses have uncertainty in practice. Scenarios where true exact benchmarking is required are rare in an LMIC context. While we note that exact benchmarking can be approximated via the proposed approaches, it will likely be computationally inefficient. The one-step approach to fully Bayesian benchmarking may be a more useful approach if exact benchmarking is required, or a posterior projection approach using a loss function that respects the bounds on the parameters that are used for the weights.

Additionally, our method does not account for uncertainty in the population count data. The population data used in our applications did not have reported uncertainty. If uncertainty for population count data were available, this would ideally be incorporated into the likelihood for the benchmarking constraint. Worldpop has recently started quantifying uncertainty in population data for select countries in their bottom-up Bayesian models, which may be of interest to include in future applications (Leasure et al., 2020).

Chapter 3

A PSEUDO-LIKELIHOOD APPROACH TO UNDER-5 MORTALITY ESTIMATION

3.1 Introduction

Estimates of mortality rates for specific age groups at a national and subnational level provide important information on the health of a country and inform targeted public health interventions. Historically, estimates of interest have been the neonatal mortality rate (NMR: probability of dying before age 1 month), the infant mortality rate (IMR: probability of dying before age 1 year), the child mortality rate (CMR: probability of dying between ages 1 and 5 years given survival to age 1), and the under-5 mortality rate (U5MR: probability of dying before the age of 5). While these summaries that are grouped across ages give us a rough picture of the pattern of mortality under the age of 5, they do not constitute a *complete* pattern of mortality before the age of 5. As such, producing a full, continuous survival curve for children under the age of 5 is of interest for informing targeted interventions (Verhulst et al., 2022; Guillot et al., 2022), as well as for better quantifying the differences in mortality patterns between countries.

Modern demographic methods for estimating a full mortality schedule for children under the age of 5 have been developed in a high-income country setting that assumes vital registration information is readily available (Guillot et al., 2022; Eilerts et al., 2021; Verhulst et al., 2022). One such method is the log-quad model (Guillot et al., 2022), which makes use of the Human Mortality Database (HMD) (Barbieri et al., 2015) to obtain a continuous curve quantifying the relationship between age and the (log) probability of dying before a given age. This approach makes use of the HMD to obtain tables of parameter values, which may then be plugged into the log-quad model's formula to obtain full, continuous curves.

Guillot et al. (2022) note, however, that because the HMD consists of high-quality vital registration data from high-income countries, the patterns of mortality that are estimated from the model are importantly different from the observed data in low- and middle-income countries (LMICs). Eilerts et al. (2021) and Verhulst et al. (2022) note that sub-Saharan African and south Asian countries typically observe higher levels of CMR for a given IMR when compared to high-income countries. Verhulst et al. (2022) call this a "very late" pattern of under-5 mortality.

Another popular method that makes use of HMD life tables is the Singular Value Decomposition (SVD) approach in Clark (2019). Here, the information from HMD life tables are compressed into three or four principal components that summarise observed full mortality schedules over an entire lifetime. Although used specifically with the HMD in Clark (2019), this approach can be used more generally with any combination of lifetables. For example, Alexander et al. (2017) use annual life tables from the French *Institut National de la Statistique et des Études Économiques* to apply a Bayesian model to French mortality data.

In addition to different patterns of under-5 mortality in LMICs compared to high-income countries, the data sources available in LMICs typically differ from high-income countries. In most high-income countries, vital registration and reliable census information are readily available, hence the individual-level mortality data is more granular and potentially subject to fewer biases than that in LMICs. In countries without vital registration data or reliable census information, we instead need to rely on nationally representative surveys to estimate demographic and health outcomes. In many LMICs, the Demographic and Health Surveys (DHS) are considered the most reliable source of information for such outcomes, and they are conducted with reasonably high frequency (the aim is every 5 years). In order to account for unequal probability of selection in the survey, weighted estimates of health outcomes with variance estimates that account for the survey design are preferred when there is enough data to obtain weighted estimates with high precision. This is the case for estimating U5MR at a national level, where we have enough data and are therefore do not need to use model-based

methods that make additional assumptions about the data structure.

As noted in Hill (1995), Lawn et al. (2008), and Guillot et al. (2022), surveys such as the DHS may be subject to biases in addition to other data quality limitations. One example of bias is age-heaping, where more children are recorded as having died at particular ages than is truly the case. In DHS surveys, this often occurs at age 12 months (see Appendix B.1 for examples). Additionally, rather than exactly observing the deaths of each child for each mother surveyed, observations after a certain age range are interval censored. Specifically, children under the age of one month have [actual] date of death recorded, children between ages one month and 24 months are recorded as having died within a one month interval, and children older than 24 months are recorded as having died within a one year interval. This censoring mechanism combined with the need to appropriately account for survey weights and potential biases from age-heaping form statistical modeling challenges that are unique to surveys in LMICs; to the best of our knowledge all of these challenges have not yet been addressed simultaneously in the literature.

An additional challenge specific to U5MR estimation is distinguishing between cohort- and period-estimates of mortality. When estimating U5MR, we typically want to obtain period-specific estimates rather than cohort-specific estimates, as the most recent, cohort-specific estimates of U5MR we could obtain will always be five years in the past. Period estimates are for "synthetic" children, where we envisage a cohort of children that live their first five years of life in a single time period. This is opposed to a real cohort of children who are born in one time period and move through time (periods) as they age. While synthetic children are of course not real, the concept allows us to provide estimates of demographic indicators such as life expectancy or U5MR for more recent time periods than would otherwise be possible, and provide a reasonable summary of the current state of the mortality pattern. In practice, when estimating a demographic indicator for synthetic people, we consider what a real person would contribute to each period *as though* they were a synthetic person. As detailed in Section 3.2, in a survival analysis framework this corresponds to treating time period as a time-varying covariate. While existing methods have made use of this approach

in a discrete survival setting (Mercer et al., 2015), none to our knowledge have *explicitly* formulated the problem as that of a time-varying covariate in continuous time.

In this paper we aim to (1) reframe the production of period estimates for under-5 mortality rates in LMICs using continuous survival models for mortality with a time-varying covariate representation and potential censoring, and (2) propose a pseudo-likelihood estimate of full mortality schedules for children under the age of 5 in LMICs, that takes full advantage of the granularity of the data available while accounting for both the survey design and potential biases in the surveys. Rather than assume a model based on data from high-income countries, we instead deal with DHS data directly to obtain both a parametric or nonparametric estimate of the survival curve in LMICs at a national level. These methods are easily extended to alternative parametric distributions as well as subnational models.

In Section 3.2, we frame the estimation problem explicitly as a survival analysis problem, and bridge the nomenclature between the demography and statistics literature. In Section 3.3 we describe our proposed nonparametric and parametric, pseudo-likelihood approaches to obtaining a full mortality curve for children under the age of 5. In Section 3.4 we describe three existing approaches to estimating a full survival curve for children under the age of five in greater detail, and describe their potential shortcomings in LMICs. In Section 3.5 we apply our methods and the three approaches described in Section 3.4 to DHS survey data from Burkina Faso, Malawi, Senegal, and Namibia in time periods $[2000, 2005)$ and $[2005, 2010)$. We conclude with a summary of results and discussion. All code to fit the proposed methods is in the `R` package `pssst`, available at `https://github.com/taylorokonek/pssst`.

3.2 Survival framework

To begin, we define some notation that is common in the demography and statistics literature, and is used throughout this paper. Mortality is typically estimated as either a rate or probability. While probabilities have number of persons in the denominator, rates have number of person-years in the denominator.. The under-5 mortality *rate* (U5MR) is in fact the *probability* that a child dies before the age of 5. Three common probabilities that are

often of demographic interest include: U5MR, the probability of dying before the age of 5 years; IMR, the probability of dying before the age of 1 year; and NMR, the probability of dying in the first month of life. Let X be survival time. We denote the probability that a child died between the ages of n and $x+n$, given that they survived until at least age n, as ${}_xq_n = \Pr(X < x+n \mid X > n)$. With age given in months, as we do throughout, we therefore denote U5MR as ${}_{60}q_0$, IMR as ${}_{12}q_0$, and NMR as ${}_1q_0$.

We treat mortality as a time-to-event outcome in a survival framework. In this framework the estimand of interest is typically either: a survival curve $S(x)$, or the probability of surviving to at least age x; a hazard (instantaneous risk of death) $h(x)$; or some measure of differences between multiple survival curves or hazards based on a covariate of interest. In our applications, we are interested in obtaining a full mortality schedule for children under the age of 5 years, and therefore our estimand of interest is $S(x)$, where x denotes the age of a child. We can directly translate quantities ${}_xq_0$ to a survival curve via $S(x) = 1 - {}_xq_0$, and ${}_xq_0$ can also be computed from conditional probabilities.

An important distinction in demography, that again has a survival analysis flavor, is period versus cohort estimates. The age-period-cohort distinction is subtle but well-documented (see Carstensen (2007), for example). Importantly, the subset of data used to estimate cohort and period estimates consequently differs. In Figure 3.1, we illustrate this difference. For simplicity, we assume all children are born on January first of a given year. We see that the data used to obtain a cohort estimate of U5MR for the cohort born in 2000 consists of only children born in the year 2000. Note that we will always be five years behind schedule in terms of estimate production because we need to observe the full, first five years of a cohort before calculating cohort U5MR.

The data used to obtain a period estimate of U5MR for the year 2004 contains data from five distinct cohorts: cohorts 2000, 2001, 2002, 2003, and 2004. Period estimate are more widely used in practice as health indicators for a country and in forecasting, as we are able to obtain period estimates of demographic rates of interest (such as U5MR) up to the most recent year we have available data, as seen on the right-hand side of Figure 3.1. Of

note, when obtaining cohort estimates, both age and time are synonymous, whereas when obtaining period estimates, age and time are distinct. This is because the age of a synthetic child is not directly tied to time as we observe it. As such, it becomes necessary to index ${}_xq_{n,p}$ by period p in addition to the age range $[n, x+n)$. Note that age and time similarly define a cohort of children. As an aside, this leads to the classic identifiability problem that is encountered in age-period-cohort modeling (Carstensen, 2007). For example, we may write the probability a child dies between the ages of 1 and 2 in the year 2001 as ${}_{12}q_{12,2001}$, where the deaths that inform this estimate must come from the cohort of children born in 2000 who survive until at least age 1.

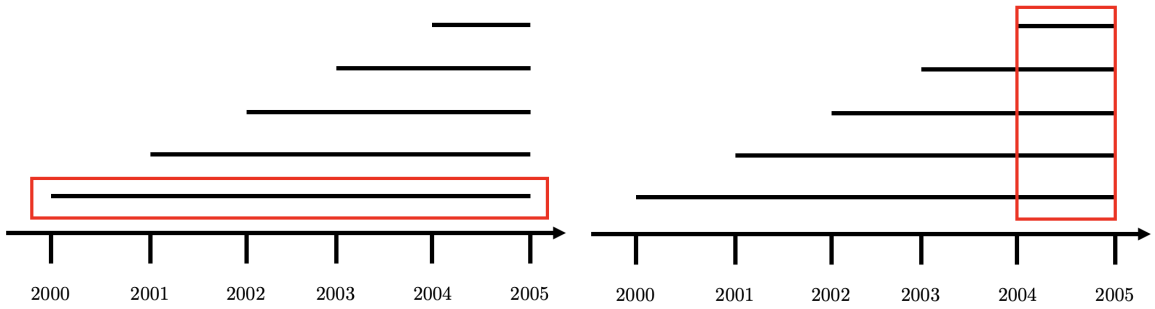


Figure 3.1: Left-hand side: Potential lifespans of observations used to obtain a cohort estimate of U5MR for the cohort born in 2000. Right-hand side: Potential lifespans of observations used to obtain a period estimate of U5MR for the year 2004. All children are assumed to be born on January first of a given year. Horizontal lines indicate the potential lifespans of children up to January 1st, 2005.

From a survival standpoint, it is important to note that when computing period estimates, some of the data will be subject to left-truncation. In general, left-truncation, also known as late entry, occurs if survival time is less than left-truncation time and thus no information is available on the subject. If not dealt with, left-truncation can induce selection bias. In our applications, left-truncation occurs when an individual is not at risk of dying in a specific age band, within a specific time period because they died before the period begins. Truncation accounts for the potential bias that would be introduced into our estimate from the individuals who were born in the earlier cohorts, yet died before our time period of

interest. As an example, in Figure 3.1 (right panel), all individuals born at the beginning of the year 2000 would be left-truncated at age 4 when computing their contribution to the period U5MR estimate for 2004. If we compute the estimate of ${}_{60}q_0$ using five, discrete values, ${}_{12}q_{48,2004}$, ${}_{12}q_{36,2004}$, ${}_{12}q_{24,2004}$, ${}_{12}q_{12,2004}$, ${}_{12}q_{0,2004}$, for each cohort born in 2000 through 2004, respectively, left-truncation is dealt with implicitly through the conditional probability structure of ${}_nq_x$, as we will see in the discrete hazards approach (Allison, 2014; Mercer et al., 2015) that we describe in Section 3.4.2.

If we are interested in obtaining estimates or U5MR (or NMR, IMR) for multiple periods across time, this truncation structure can be incorporated into a model by treating period as a time-varying covariate. This is again done implicitly in a discrete hazards approach, but can be used explicitly in the pseudo-likelihood approach we propose that allows us to use continuous survival models for age. It is odd in a survival setting to treat time itself (discretely categorized) as a time-varying covariate, but perhaps makes more sense when distinguishing between real and synthetic children. The discretely categorized variable period is treated as a covariate that changes *through* time, and is simply an indicator for which synthetic cohort we are considering.

A final piece of the survival framework we set up is how we deal with censored observations. Of course, not all children we observe will die before the 5 years of age for U5MR, 1 year of age for IMR, etc. These children will be right-censored, and therefore will not contribute any time at risk after the age of 5 (or 1, etc.) to a statistical model. Children who die in a time period later than the period in which they were born also contribute right-censored observations to those earlier time periods. A survival framework also allows us to deal with interval censoring, where we know only that an event has occurred for an individual between two ages. The ability to handle interval censoring is particularly relevant for child mortality estimation in LMICs, where the data primarily comes from surveys. DHS surveys in particular, contain daily observed death dates for children who died before the age of 1 month, monthly, interval-censored observations for children who died between 1 month and 24 months (e.g., we may observe that a child that dies between the ages of 2 and

3 months), and yearly, interval-censored observations for children who died after 2 years of age, with some exceptions. Interval censoring can be appropriately addressed by discretely categorizing observations, as is done in all of the existing approaches described in Section 3.4, but can also be addressed in a continuous survival framework as we propose in Section 3.3.

3.3 Methods

The methods we propose for estimating the whole survival curve for children under the age of 5 explicitly consider survival approaches as well as the complex design aspects of survey data, as we are interested in developing methods that can be used with household survey data from LMICs. Historically, survival methods have been extended to the survey setting, in the context of the Cox proportional hazards model (Binder, 1992; Lin, 2000; Breslow and Wellner, 2007, 2008). Such methods address the survey aspect of the data via a pseudo-likelihood approach to estimation (Binder, 1983), in which we weight each individual's likelihood contribution by their sampling weight, and maximize the pseudo-likelihood to give weighted (pseudo) MLEs. The variance of the estimates is approximated via sandwich estimation.

Consider a vector of superpopulation parameters $\boldsymbol{\theta}$ and observations $\mathbf{y}$ that can be written as the solution to the score equations $\mathcal{S}(\boldsymbol{\theta}, \mathbf{y}) = 0$. In a finite population setting, we are interested in the finite population parameters $\boldsymbol{\theta}'$, obtained by solving $\sum_{i=1}^{N} \mathcal{S}(\boldsymbol{\theta}', \mathbf{y}) = 0$, where $i = 1, \ldots, N$ denote all individuals in the finite population. Rather than observe all N individuals in the population, we instead take a probability sample of $j = 1, \ldots, U \leq N$ individuals, with weights w_j equal to the inverse of their inclusion probabilities in the sample. We obtain survey-weighted estimates of the finite population parameters $\boldsymbol{\theta}'$ via the finite population score equations

$$\sum_{j=1}^{U} w_j \times \mathcal{S}(\boldsymbol{\theta}', \mathbf{y}_j) = 0.$$

This framework, developed in Binder (1983), works for sample designs and populations that admit asymptotically normal estimators, with certain regularity conditions (see Binder (1983) for details, as well as a more general case of score equations). As an example, for linear regression we would set $\mathcal{S}(\boldsymbol{\theta}, \mathbf{y}) = -(y_i - \mathbf{x}_i^\top \boldsymbol{\theta})\mathbf{x}_i$. The variance of $\boldsymbol{\theta}'$ then has a sandwich form obtained via a Taylor expansion of the score equations at $\boldsymbol{\theta}' = \boldsymbol{\theta}$.

Bootstrap and jackknife procedures have been developed for various complex survey designs, including the two-stage, stratified cluster design common to DHS surveys. To obtain bootstrapped variance estimates, $n_h - 1$ clusters are sampled with replacement within strata h, where n_h is the number of clusters in strata h (Rao and Wu, 1988). Pointwise confidence intervals may then be constructed using percentiles of the bootstrap samples. A jackknife procedure for the same setting is described in Pedersen and Liu (2012).

For the remainder of this section, we describe two approaches for estimating continuous survival curves for synthetic children across multiple time periods. Both methods are based on pseudo-likelihood; one parametric, and one nonparametric.

3.3.1 Nonparametric approach

The classic and most popular nonparametric estimate of a survival curve is the Kaplan-Meier estimator (Kaplan and Meier, 1958). Let t_i be a time when at least one event (death) occurred, d_i be the number of events that occurred at time t_i, and n_i be the number of children who have not had an event or been censored up to time t_i. Then the Kaplan-Meier estimator of the survival curve at time t is

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right).$$

Under noninformative censoring, the Kaplan-Meier estimator is the nonparametric maximum likelihood estimator (NPMLE) of the survival curve. However, the Kaplan-Meier estimator, in it's simplest form, is unsuitable for interval censored data.

A generalization of the Kaplan-Meier estimator to arbitrarily truncated and censored

observations is the Turnbull estimator (Turnbull, 1976). Below we introduce the notation used in this paper to describe the estimator in its full generality, alongside an example for our motivating application.

Suppose we are interested in estimating a full mortality schedule (survival curve) from ages 0 to 5 at a national level in time periods $[2000, 2005)$ and $[2005, 2010)$. For simplicity, we consider age in full years rather than months in this example. Let $[t_0, t_1]$ denote the interval in which an observation is censored, with the convention $[t_0, \text{inf})$ for right-censored observations. Our data is recorded as follows:

Individual	Year born	t_0	t_1
1	2002	2	4
2	1998	5	∞
3	2007	1	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Individual	Year born	B_i	A_i	Period
1	2002	$(-\infty, \infty)$	[2, 3]	[2000, 2005)
1	2002	[3, ∞)	[3, 4]	[2005, 2010)
2	1998	[2, ∞)	[5, ∞)	[2000, 2005)
3	2007	$(-\infty, \infty)$	[1, 1]	[2005, 2010)
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Table 3.1: (Left) Example data for children alive between the ages of 0 and 5 years in the time periods $[2000, 2005)$ and $[2005, 2010)$. (Right) Example data, separated into (truncated) observations for each time period.

Individual 1 is interval censored to the age range $[2, 4]$, individual 2 is right censored at age 5, individual 3 is observed to die at exactly 1 year, and so forth. To account for time period as a time-varying covariate in our model, we rearrange our data by splitting each observation into multiple observations, one for each time period in which they contribute to the risk set. The observations are left-truncated at the beginning of each time period by $\max(0, \text{age at which they enter the time period})$, where we denote this truncation in set notation as B_i. If $B_i = (-\infty, \infty)$, this indicates that no truncation has occurred, which in this context means the individual was born in the given time period and not before it. Individuals are censored according to the sets $A_i = [L_i, R_i]$, where if $L_i = R_i$ the observation is exactly observed (uncensored), and if $A_i = [L_i, \infty)$ the observation is right censored at L_i. The number of observations in our expanded dataset is $N \equiv \sum_{i=1}^{n} p_i$, where n is the

number of individuals in our dataset and p_i is the number of time periods in which individual i contributes to the risk set.

In the above example, for $i = 1, \ldots, N$ observations, let A_i denote an individual's censoring set, and let B_i denote an individuals truncation set such that the likelihood contribution for an individual can be denoted $P(X_i \in A_i \mid X_i \in B_i)$, where X_i is the random variable corresponding to the death of child i. The data is thus in the form of pairs $(A_1, B_1), \ldots, (A_n, B_n)$.

The Turnbull estimator is the NPLME of the cumulative distribution function $\hat{F}$ of F, and therefore, also produces the NPMLE of the survival curve $\hat{S} = 1 - \hat{F}$. We write the likelihood in a convenient way that allows us to optimize the proportions of the CDF that lie within given intervals subject to the constraint that the proportions sum to 1 and are greater than 0. This allows us to view the likelihood maximization as a constrained optimization problem that can be solved using an Expectation-Maximization (EM) algorithm (Dempster et al., 1977), which we now describe.

Assume that each A_i can be written as a finite union of disjoint, closed intervals, with a single point $A_i = X_i$ being written as the closed interval $[X_i, X_i]$. Then for each censoring set A_i we can write,

$$A_i = \bigcup_{j=1}^{k_i} [L_{ij}, R_{ij}],$$

for $i = 1, \ldots, n$. Then the likelihood for all observations can be written as

$$\text{Likelihood} = \prod_{i=1}^{n} \frac{\left[\sum_{j=1}^{k_i} F(R_{ij}) - F(L_{ij})\right]}{P(B_i)}. \tag{3.1}$$

Let $[q_1, r_1], \ldots, [q_m, r_m]$ denote all unique intervals defined by $[L_{ij}, R_{ij}]$, and $s_j = F(r_j) - F(q_j)$. These s_j define the proportions of the CDF that lie within an interval $[q_j, r_j]$.

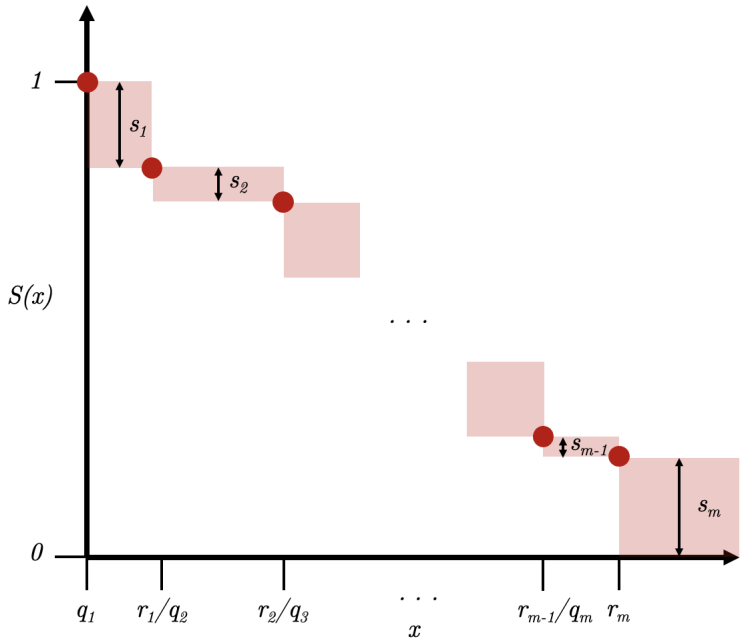


Figure 3.2: A visual representation of the values s_j that make up the Turnbull estimator.

We can rewrite the likelihood defined in Equation (3.1) as

$$\text{Likelihood} = \prod_{i=1}^{n} \left(\frac{\sum_{j=1}^{m} I\{[q_j, r_j] \in A_i\} s_j}{\sum_{j=1}^{m} I\{[q_j, r_j] \in B_i\} s_j} \right).$$

Maximizing the above subject to the constraints $s_j \geq 0$, $\sum_{j=1}^{m} s_j = 1$ then corresponds to maximizing the likelihood for arbitrarily censored and truncated observations, and provides us with a nonparametric estimate of the MLE. A visual description of the Turnbull estimator is provided in Figure 3.2.

Turnbull (1976) suggests the following procedure for obtaining the MLE:

1. Obtain initial values for $\mathbf{s}^0 = s_1^0, \ldots, s_m^0$ subject to $\sum_{j=1}^{m} s_j^0 = 1$, $s_j^0 \geq 0$

2. Compute

$$\mu_{ij}(\mathbf{s}) = \frac{I\{[q_j, r_j] \in A_i\} s_j}{\sum_{k=1}^{m}, I\{[q_k, r_k] \in A_i\} s_k}$$
$$\nu_{ij}(\mathbf{s}) = \frac{(1 - I\{[q_j, r_j] \in B_i\}) s_j}{\sum_{k=1}^{m} I\{[q_k, r_k] \in B_i\} s_k}$$
$$\pi_j(\mathbf{s}) = \frac{\sum_{i=1}^{n} (\mu_{ij}(\mathbf{s}) + \nu_{ij}(\mathbf{s}))}{\sum_{i=1}^{n} \sum_{j=1}^{m} (\mu_{ij}(\mathbf{s}) + \nu_{ij}(\mathbf{s}))}$$

3. Set $s_j^1 = \pi_j(\mathbf{s}^0)$.

4. Return to Step 1.

The procedure exits once some predetermined, required tolerance is achieved.

Groeneboom and Wellner (1992) note that, compared to the Kaplan-Meier estimator, the Turnbull estimator has less appealing asymptotics. The estimator converges pointwise (i.e. at a fixed value t) at a rate of $n^{1/3}$ to a non-Gaussian distribution. The question of obtaining valid confidence bands for the Turnbull estimator remains an open statistical question. Though some have recommended using a bootstrap procedure for variance estimation (see Sun (2001), for example), the coverage of these procedures is not well justified (and therefore not necessarily correct) due to the rate of convergence and non-Gaussian asymptotics.

Incorporating survey weights into the Turnbull estimator is straightforward, as the likelihood contribution for each individual is simply multiplied by their survey weight. This corresponds to altering Step 2 in the algorithm described, replacing $\pi_j(\mathbf{s})$ with $\tilde{\pi}_j(\mathbf{s})$, where $\tilde{\pi}_j(\mathbf{s})$ is

$$\tilde{\pi}_j(\mathbf{s}) = \frac{\sum_{i=1}^{n} w_i (\mu_{ij}(\mathbf{s}) + \nu_{ij}(\mathbf{s}))}{\sum_{i=1}^{n} w_i \sum_{j=1}^{m} (\mu_{ij}(\mathbf{s}) + \nu_{ij}(\mathbf{s}))},$$

and w_i is the survey weight for a given individual. The use of $\tilde{\pi}_j(\mathbf{s})$ in the algorithm produces a pseudo-NPMLE for arbitrarily truncated and censored data with survey weights. This modification makes the Turnbull estimator applicable to scenarios where the data come from

a survey with a complex design, such as those in LMICs.

Although the bootstrap is not well-justified for the Turnbull estimator, we do use a bootstrap procedure appropriate for a two-stage, stratified sampling design from Rao and Wu (1988) to assist with model comparison in our application. The procedure is described in Section 3.3. In our model comparison approach, we treat the Turnbull estimator as a baseline estimate of the survival curve, and aim to determine whether a given parametric model is "reasonably" close to the Turnbull estimator. Obtaining some measure of uncertainty for the Turnbull estimator facilitates this comparison.

As a well-justified variance estimator is not available for the Turnbull estimator, we do not recommend using the Turnbull estimator for *official* estimates of full mortality schedules for children under the age of 5 in LMICs. It is especially important in scenarios where the data does not come from a census or other vital registration source to accurately quantify the uncertainty of estimates. The Turnbull estimator does, however, describe the survival function, and provide a useful reference when assessing how well a parametric distribution summarizes the pattern of U5MR in LMICs, as its point estimates do not rely on parametric assumptions.

3.3.2 Parametric approach

Suppose we have individuals $i = 1, \ldots, n$. Let,

- $p = 1, \ldots, P$: consecutive time periods, which may be single years or combinations of years
- l_p: length of period p, measured in the same units as age of child
- y_p: date at the start of time period p
- b_i: child's date of birth
- t_i: child's age at right-censoring or age at death

- I_i: an indicator that child i is interval-censored. If $I_i = 1$, individual i is interval-censored. If $I_i = 0$, individual i is right-censored or has an exact death time
- t_{0i}: child's age at beginning of interval censoring, if child is interval censored
- t_{1i}: child's age at end of interval censoring, if child is interval censored
- $a_{pi} = y_p - b_i$: the age the child would be at y_p
- E_i: an indicator that child i's death is exactly observed. If $E_i = 1$, then $I_i = 0$, and if $E_i = 0$, then I_i could be 0 or 1
- $\tilde{p}_i$: if $E_i = 1$, the period in which that child died
- $U_{x_i}(p) = \{p : a_{pi} > -l_p, a_{pi} < x_i\}$. $U_{x_i}(p)$ is the set of periods for which child i is alive and at risk of dying

Let F denote the CDF for the specified parametric distribution, and H the corresponding cumulative hazard function. The likelihood for all individuals in our dataset across all time periods can be written as

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1}^{n} L_i(\boldsymbol{\theta}) \\ &= \prod_{i=1}^{n} \left[1 - F_{\boldsymbol{\theta},i}(t_i)\right]^{1-I_i} \left[F_{\boldsymbol{\theta},i}(t_{1i}) - F_{\boldsymbol{\theta},i}(t_{0i})\right]^{I_i} \left[f_{\boldsymbol{\theta},i}(t_i)\right]^{E_i}, \\ &= \prod_{i=1}^{n} \underbrace{\left[\exp(-H_{\boldsymbol{\theta},i}(t_i))\right]^{1-I_i}}_{right-censored} \underbrace{\left[\exp(-H_{\boldsymbol{\theta},i}(t_{0i})) - \exp(-H_{\boldsymbol{\theta},i}(t_{1i}))\right]^{I_i}}_{interval-censored} \underbrace{\left[\exp(-H_{\boldsymbol{\theta},i}(t_i)) h_{\boldsymbol{\theta},\tilde{p}_i}(t_i)\right]^{E_i}}_{exact}, \end{aligned}$$

where

$$H_{\boldsymbol{\theta},i}(x_i) = \sum_{U_{x_i}(p)} \int_{\max\{a_{p_i},0\}}^{\min\{x_i, a_{p_i}+l_p\}} h_{\boldsymbol{\theta},p}(u)du,$$

and $h_{\boldsymbol{\theta},p}(u)$ is a period-specific hazard function for a specified distribution.

To obtain survey-weighted estimates, we obtain pseudo-MLEs (Binder, 1983) of the distribution-specific parameters by maximizing the sum of log likelihood contributions for each individual observation multiplied by their respective survey weights. To obtain finite population variance estimates, we use a trick where we treat our estimator as a weighted total, for which we can easily obtain finite population variance estimates using `R`'s `survey` package.

For a generic parametric model (in a non-survey context) with $j = 1, \ldots, J$ parameters, let $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_J)$ denote the MLE. We can write $\hat{\theta}_j$ as asymptotically linear, meaning

$$\hat{\theta}_j - \theta_j = \frac{1}{n} \sum_{i=1}^{n} \Delta_i + o_p(n^{-1/2}),$$

with influence functions Δ_i given by

$$\Delta_i = \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log\Big(L_i(\hat{\boldsymbol{\theta}}) \Big) \right] \left[\mathbf{H}_{\log(L_i)} \right],$$

where $\frac{\partial}{\partial \boldsymbol{\theta}} \log\Big(L_i(\hat{\boldsymbol{\theta}}) \Big)$ is an $n \times J$ dimensional matrix of score functions for each individual $i = 1, \ldots n$ with respect to parameters $j = 1, \ldots J$, and $\mathbf{H}_{\log(L_i)}$ denotes the Hessian of the log likelihood. Then for an influence function of a pseudo-MLE with weights w_i, we can write

$$\hat{\theta}_j - \theta_j = \frac{1}{n} \sum_{i=1}^{n} w_i \Delta_i + o_p(n^{-1/2}).$$

To estimate the finite population variance of $\hat{\theta}_j$, calculating the finite population variance of $\sum_{i=1}^{n} w_i \Delta_i$ corresponds to the Taylor-linearization method described in Binder (1983). The convenience here lies in that $\sum_{i=1}^{n} w_i \Delta_i$ is a survey total, and the influence functions are simple to obtain since we have a parametric model. In fact, only the gradient of the log likelihood evaluated at the pseudo-MLE and the Hessian obtained during optimization of

the weighted log likelihood are needed to calculate the finite population variance. Details of this calculation are given in Appendix B.2.

Many common parametric distributions used in survival analysis are currently implemented in the `R` package `pssst`, including: exponential, Weibull, piecewise exponential with arbitrary cutpoints, generalized Gamma, lognormal, Gompertz, and the distribution characterized by the exponentially-truncated power shifted family of hazards defined in Schöley (2019). The proposed parametric method may also be used in conjunction with any parametric form of period-specific hazard, and so can be readily extended.

3.4 Literature review

As our proposed methodology focuses on providing a continuous, age-specific mortality curve for children under the age of 5, we focus on three existing methods that can provide this, modulo a few assumptions: the log-quad model (Guillot et al., 2022), the discrete hazards model (Li et al., 2019; Wu et al., 2021), and the SVD model (Clark, 2019). The latter two require the assumption that the discrete hazards ${}_nq_x$ estimated for each x are constant within the interval $[x, x+n)$ in order to obtain a full survival curve. We believe that the paucity of literature focused on providing estimates for mortality schedules by continuous age is perhaps due to demography's historical focus on estimates grouped by five- or one-year period and age.

Of note, Schöley (2019) recently proposed a continuous, parametric approach model for infant mortality. There are similarities between it and our proposed approach, notably the use of a continuous hazard to assist in defining a survival curve for children. The method differs, however, in its focus on the pattern of infant mortality as opposed to U5MR, the use of daily observed deaths from a high-income country which removes the need to account for interval-censored observations, and the use of data that does not come from a survey and therefore does not need to account for the survey design. The methods described in Schöley (2019) serve as high income country analogues to our proposed methods, and we consider one family of hazards they deem best fitting to US data in our proposed methodology.

3.4.1 Log-quad model

The log-quad model described in Guillot et al. (2022) builds on the log-quad approach in Wilmoth et al. (2012), and can provide an estimate of a continuous survival curve from ages 0 to 5 using only an observed or previously estimated ${}_{60}\hat{q}_0$. Other optional inputs to the log-quad model include values ${}_xq_0$ for different ages x. Following Clark (2019), we call the model "empirical" because the coefficients input to the model are not estimated during the modeling process, but instead are computed ahead of time using data from the Under-5 Mortality Database (U5MD) (Guillot et al., 2022). The model specifies

$$\log({}_xq_0) = a_x + b_x \log({}_{60}\hat{q}_0) + c_x \log({}_{60}\hat{q}_0)^2 + v_x k,$$

where x takes on one of the 22 values $\{7d, 14d, 21d, 28d, 2m, 3m, 4m, 5m, 6m, 7m, 8m, 9m, 10m, 11m, 12m, 15m, 18m, 21m, 2y, 3y, 4y, 5y\}$. The age-specific coefficients $\{a_x, b_x, c_x, v_x\}$ are provided in the U5MD (and are estimated using lifetables in the HMD, as described in more detail in Guillot et al. (2022)), ${}_{60}\hat{q}_0$ is input to the model as a fixed covariate, and the parameter k is an optional parameter describing whether the age pattern of mortality is "early" or "late." By early, we mean that NMR and IMR are higher than typically observed when compared to U5MR, and by late we mean that NMR and IMR are lower than typically observed when compared to U5MR. By typically observed, we mean the patterns of mortality before the age of 5 in countries with highly reliable child mortality data, such as those included in the U5MD. As a consequence, mortality patterns in high-income countries are often referred to as "typical" in the demography literature, and those in LMICs are referred to as atypical.

The age-specific coefficients are estimated for the 22 ages $\{7d, \ldots, 5y\}$, and therefore predictions are made for $\log({}_xq_0)$ at those values of x. The output of the log-quad model provides a continuous survival curve under the age of 5 under the assumption that ${}_xq_0$ is constant within age intervals defined by the 22 values of x.

The only parameter that is estimated in the modeling step when predicting the U5MR

pattern for a new country is k, unless the average pattern of mortality observed across data in the HMD is desired, in which case k is set to 0. (Of note, Guillot et al. (2022) consider ${}_{60}\hat{q}_0$ to be an additional parameter for the log-quad model. Here we consider ${}_{60}\hat{q}_0$ to be data rather than a parameter, as ${}_{60}\hat{q}_0$ is input to the model as a fixed value rather than estimated during the modeling step.) The parameter k is estimated in one of two ways:

- **Option 1**: If only a single value x is supplied for ${}_xq_0$ to the model as a fixed constant, k is estimated as

$$\hat{k} = \frac{e(x)}{v_x},$$

  where $e(x)$ is the difference between the predicted and observed values of ${}_xq_0$ when the the model is fit with $k = 0$.

- **Option 2**: If more than one x is supplied for ${}_xq_0$ to the model as data, k is estimated as

$$\hat{k}^* = \frac{\sum_x w(x)e(x)v_x}{\sum_x w(x)v_x^2},$$

  where $\hat{k}^*$ is the value of k that minimizes the root-mean-square error (RMSE) of predicted values of ${}_xq_0$ to observed values of ${}_xq_0$ across all values of x supplied, and $w(x)$ is the weight corresponding to the length of the previous age interval ending at age x (i.e. $w(1) = 7d$, $w(2) = 7d$, $\ldots$, $w(22) = 1y$).

When all 22 possible values for x are supplied to the model, Guillot et al. (2022) propose an uncertainty band around the estimated survival curve that can be computed as

$$\hat{k}^* \pm 1.96 \times \sqrt{Var(\hat{k}^*)},$$
$$Var(\hat{k}^*) = \frac{22}{21}\left(\frac{\sum_x w(x)e(x)^2}{\sum_x w(x)v_x^2} - (\hat{k}^*)^2\right).$$

They propose a separate uncertainty band when only one value for x is supplied to the model, but we exclude it in our summary as that scenario is of little importance in our applications. In the derivation of this variance estimator, Guillot et al. (2022) assume that the errors $e(x)$ are independent across values x, that the weighted errors $w(x)e(x)$ are homoskedastic, and that $\hat{k}^*$ is approximately normally distributed.

Additionally, they note that almost all data used to estimate the age-specific coefficients $\{a_x, b_x, c_x, v_x\}$ in the U5MD are estimated with values of k that fall in $(-1, 1)$. Due to this observed range of values, they state that values of k that are estimated outside the range $(-1.1270, 1.5047)$ (the exact range of all observed values) have no "empirical basis" and may in fact produce estimates of ${}_xq_0$ that progress nonmonotically for children under the age of 5, whereas actual survival curves must be monotonically nonincreasing. Therefore, they suggest a rule of thumb that the estimates should only be used when k is estimated in the range $(-1.1, 1.5)$.

Multiple follow-up papers (e.g. Eilerts et al. (2021); Verhulst et al. (2022)), as well as Guillot et al. (2022) itself, note that the log-quad model is likely unsuitable for use in LMICs, or in countries with (broadly) early or late patterns of child mortality. This is perhaps unsurprising given that the coefficients in the U5MD are estimated from high-income countries which likely have differing health care systems, and structural and programmatic support for decreasing child mortality. Guillot et al. (2022) also note that there are known biases in the data sources available in LMICs. One of these issues, age-heaping, can be addressed using the log-quad model. In DHS surveys especially, this typically occurs at age 12 months. Rather than input all 22 possible age groups into the model for estimating the k parameter, the user may instead leave out a range of ages (Guillot et al. (2022) suggest 9 to 21 months) that they believe covers the ages where data is heaped. Note that this is distinct from treating deaths between the ages of 9 and 21 months as interval censored, but rather is akin to removing these deaths entirely from the data. The rationale for this approach is that in removing those deaths, the estimated curve will essentially smooth over any age heaping that occurs. Guillot et al. (2022) demonstrate this approach to addressing age-heaping using

data from Bolivia's 1989 DHS. A downside to this approach is that it involves throwing away information about the pattern of U5MR.

While the log-quad model can address age-heaping, it has additional characteristics that may be unsuitable in LMICs. Due to its formulation, the log-quad model's prediction of U5MR is identical to the value of U5MR input as a covariate (when $x = 5y$, the age-specific coefficients from the model are estimated as $\{a_x, b_x, c_x, v_x\} = \{0, 1, 0, 0\}$). In settings with reliable data, this may be a reasonable (even desirable) property. However, in LMICs where U5MR is estimated with considerable uncertainty, we do not necessarily want our predicted value of U5MR to align perfectly with a point estimate, but rather to lie within a range of reasonable values defined by the confidence interval for U5MR.

3.4.2 Discrete hazards approach

The discrete hazards approach described in Allison (2014) (as well as Mercer et al. (2015); Li et al. (2019); Wu et al. (2021)) formulates child mortality data in a more explicit survival framework than some other existing demographic methods. Note that this framework is currently used by both the UN and DHS for estimating subnational U5MR in LMICs. Suppose we are interested in estimating U5MR, or ${}_{60}q_0$, where age is in months. The discrete hazards model splits the time before the age of 60 months into J discrete intervals $[x_1, x_2), [x_2, x_3), \ldots, [x_J, x_{J+1})$ where $x_{j+1} = x_j + n_j$, $x_1 = 0$. Then U5MR can be computed as

$$ {}_{60}q_0 = 1 - \prod_{j=1}^{J} (1 - {}_{n_j}q_{x_j}). \tag{3.2}$$

As examples, the DHS (USAID, 2014) use seven age intervals, while Mercer et al. (2015) divide the first 60 months of life for individuals into six intervals, $J = 6$: $[0, 1)$, $[1, 12)$, $[12, 24)$, $[24, 36)$, $[36, 48)$, $[48, 60)$, where $(x_1, \ldots, x_6) = (0, 1, 12, 24, 36, 48)$, $(n_1, \ldots, n_6) = (1, 11, 12, 12, 12, 12)$. Data is tabulated into binomial counts indexed by age group j, and potentially indexed by time period p as well, where the number of observations y_{jp} in a given

bin corresponds to the number of deaths observed in that age group and time period, and the number at risk N_{jp} in a given bin corresponds to the number of children alive in that age group and time period. Note that by construction of the age intervals, we can also estimate NMR and IMR from this model.

Mercer et al. (2015) then fit a logistic regression model,

$$
\begin{aligned}
y_{jp} \mid N_{jp}, \eta_{jp} &\sim \text{Binomial}(N_{jp}, {}_{n_j}q_{x_j,p}), \\
\text{logit}({}_{n_j}q_{x_j,p}) &= \beta_{jp},
\end{aligned}
$$

where β_{jp} is an age-period specific intercept. Pseudo-MLEs of β_{jp} are obtained by fitting this model in `R`'s `survey` package, using the `svyglm()` function. By writing the likelihood as a product of binomial likelihoods, we can use the pseudo-MLEs estimaed from the logistic regression model to construct estimates of ${}_{60}\hat{q}_0$ (or NMR, or IMR) using Equation (3.2). We stress that although the binomial likelihood does not reflect the exact data generating mechanism, many sampling schemes in LMICs (including that used by the DHS) allow data to be aggregated to binomial counts by cluster. Having well-established, pre-written code to obtain the pseudo-MLEs is particularly appealing.

Inherently, the discrete hazards approach assumes a constant hazard within the specified age groups. Therefore, while we can estimate a full survival curve for children under the age of 5, we know its shape will not be realistic, as the probability of survival should change smoothly with age rather than make discrete jumps. To obtain a more continuous survival curve, we could have 60 age groups for each 1-month breakdown in the discrete hazards approach. There is a balance here between flexibility and parsimony: the model fitted with more age groups better reflects the underlying smooth changes in hazard, but each hazard estimate is less precise than we might get fitting a more parsimonious model.

As with the log-quad model, age-heaping can be handled in the discrete hazards model by construction of the age intervals. For example, one could consider age intervals ($J = 7$): $[0, 1)$, $[1, 9)$, $[9, 21)$, $[21, 24)$, $[24, 36)$, $[36, 48)$, $[48, 60)$, where we group deaths recorded between the

ages of 9 and 21 months into a single age group.

We make two further notes on using the discrete hazards model in conjunction with DHS surveys. First, in DHS surveys deaths are recorded at exact days between ages 0 and 1 month, monthly until 24 months, and yearly onwards. The discrete hazards model does not take advantage of the fine-scale, daily data available prior to 1 month, and instead groups those deaths together to form a neonatal age group. If NMR is the smallest demographic rate we wish to estimate, this grouping is not inherently an issue. However, daily recorded deaths may be informative of the overall *pattern* of mortality before the age of 5, so if we instead want to estimate an accurate survival curve over the entire age range from 0 to 5, grouping all deaths within the first month of life together will not capture the expected sharp decline in survival in the first week of life, or even the first two weeks of life.

A second benefit of aggregating our data across age groups is that, especially at small levels of spatial aggregation, we may have very little data available on the hazard in some age groups. Consequently, if data is sparse we may prefer to use fewer age groups in the discrete hazards model, as if no deaths were present in a certain age-period group (which is common for small regions in small area estimation, or fine-scale time periods), their estimated hazard will be exactly zero. Hazards of exactly zero are undesirable, as they are both implausible and it is difficult to get inference around such an estimate.

3.4.3 SVD approach

A third, popular approach for estimating a full mortality schedule across an entire lifetime is Singular Value Decomposition (SVD), as described in Clark (2019, 2015), and Alexander et al. (2017) among others. Note that this approach is intended to provide full mortality schedules separately for binarized sex (male/female) rather than a full mortality schedule for both binarized sexes combined. We point the reader to Clark (2019) for an in-depth review of older demographic methods using SVDs and the related method of principal components analysis.

The general idea of the SVD approach to modeling full mortality schedules is to incorpo-

rate external summaries of observed demographic patterns. These summaries are typically left singular values (LSVs) from a singular value decomposition of historic life tables from the HMD. Clark (2019) suggest that four LSVs are typically enough to accurately capture a full mortality schedule across all ages for a variety of countries. To maintain consistency with the demographic literature, we will hereafter refer to models that take advantage of SVDs, such as that in Clark (2019), as an SVD model.

Of note, as the life tables available in the HMD are calculated for 1 year period by 1 year age groups at the finest level, the SVD models in Clark (2019) only predict mortality schedules at that 1×1 level. As a consequence, these models may not accurately estimate NMR, or any metric that is not defined on a yearly scale, since to estimate summary measures at a finer scale than yearly requires the assumption of constant hazards between years.

Below, we state the modeling steps described in Clark (2019) that are relevant to our applications in LMICs, where only an estimate ${}_{60}q_0$ is available for each sex rather than both ${}_{60}q_0$ and ${}_{540}q_{180}$. Note that the latter is important in the SVD approach because the model was designed to produce full mortality curves for ages well beyond 60 months. With coefficients estimated from various models involving the HMD (see Clark (2019) for details), rhe framework for obtaining estimates of ${}_{12}q_x$ for $x = 0, 12, \ldots, 48$ from the SVD model in Clark (2019) is as follows:

1. Estimate ${}_{60}\hat{q}_{0,F}$ and ${}_{60}\hat{q}_{0,M}$, and predict $\text{logit}({}_{540}\hat{q}_{180,F})$ and $\text{logit}({}_{540}\hat{q}_{180,M})$ using the input value for ${}_{60}\hat{q}_0$ as

$$\begin{aligned}
\text{logit}({}_{540}\hat{q}_{180,F}) &= -10.797\,{}_{60}\hat{q}_{0,F} + 4.005\text{logit}({}_{60}\hat{q}_0)_F + 0.695 * \text{logit}({}_{60}\hat{q}_0)_F^2 \\
&\quad + 0.046 * \text{logit}({}_{60}\hat{q}_0)_F^3 + 5.921 \\
\text{logit}({}_{540}\hat{q}_{180,M}) &= 2.883\,{}_{60}\hat{q}_{0,M} + 0.359\text{logit}({}_{60}\hat{q}_0)_M + 0.104 * \text{logit}({}_{60}\hat{q}_0)_M^2 \\
&\quad + 0.016 * \text{logit}({}_{60}\hat{q}_0)_M^3 - 0.703
\end{aligned}$$

where the coefficients for each term come from Table D3 of the online appendix for

Clark (2019).

2. Using $\text{logit}({}_{60}\hat{q}_{0,F})$, $\text{logit}({}_{60}\hat{q}_{0,M})$, $\text{logit}({}_{540}\hat{q}_{180,F})$, and $\text{logit}({}_{540}\hat{q}_{180,M})$, compute $\hat{w}_{zi}$ (where $z = \{F, M\}$) as

$$\begin{aligned}\hat{w}_{zi} =& c_{zi} + \beta_{z1i} \times {}_{60}\hat{q}_{0,z} + \beta_{z5i} \times {}_{540}\hat{q}_{180,z} \\ &+ \beta_{z2i} \times \text{logit}({}_{60}\hat{q}_{0,z}) \\ &+ \beta_{z2i} \times \text{logit}({}_{60}\hat{q}_{0,z})^2 + \beta_{z6i} \times \text{logit}({}_{540}\hat{q}_{180,z})^2 \\ &+ \beta_{z3i} \times \text{logit}({}_{60}\hat{q}_{0,z})^3 + \beta_{z7i} \times \text{logit}({}_{540}\hat{q}_{180,z})^3 \\ &+ \beta_{z8i} \times \text{logit}({}_{60}\hat{q}_{0,z}) \times \text{logit}({}_{540}\hat{q}_{180,z})\end{aligned}$$

where the coefficients $\{c_{zi}, \beta_{z1i}, \ldots, \beta_{z8i}\}$ are available for $z = \{F, M\}, i = 1, \ldots, 4$ in Tables D1 and D2 of the online appendix for Clark (2019). For completeness, we include the values in Appendix B.3 as well.

3. Obtain a full, yearly mortality schedule ${}_{12}\hat{q}_x$ for each $z = \{F, M\}$ as

$$\text{logit}({}_{12}\hat{q}_{x,z}) = \sum_{i=1}^{4} \hat{w}_{zi} s_{zi} \mathbf{u}_{zi}$$

where $\mathbf{u}_{zi}$ and s_{zi} are the left singular vectors (LSVs) and singular vectors (SVs), respectively, of an $A \times L$ matrix, $\mathbf{Q}_z$ containing mortality schedules for each sex, where A indexes age groups (and rows) in $\mathbf{Q}_z$, and L indexes life tables (and columns) in $\mathbf{Q}_z$. The life tables and mortality schedules included in $\mathbf{Q}_z$ in Clark (2019) are part of the HMD, where each mortality schedule is offset by -10 to ensure all age groups have approximately the same influence on the resulting LSVs. This is the step for which the SVD model is named.

4. Alter the predictions $\text{logit}({}_{12}\hat{q}_{0,z})$ using the original estimates of ${}_{60}\hat{q}_{0,z}$ from Step 1,

using the formula

$$\text{logit}({}_{12}\hat{q}_{0,F}) = -0.951 + 0.659\text{logit}({}_{60}\hat{q}_{0,F}) - 0.038\text{logit}({}_{60}\hat{q}_{0,F})^2,$$
$$\text{logit}({}_{12}\hat{q}_{0,M}) = -0.828 + 0.689\text{logit}({}_{60}\hat{q}_{0,M}) - 0.037\text{logit}({}_{60}\hat{q}_{0,M})^2,$$

Clark (2019) note that this alteration is because as ${}_{60}\hat{q}_0$ predicts ${}_{12}\hat{q}_{12}$ and ${}_{12}\hat{q}_{48}$ well, but not ${}_{12}\hat{q}_0$, where the latter is better predicted using only ${}_{60}\hat{q}_0$ separately.

5. Obtain final estimates ${}_{12}\hat{q}_x$ via

$${}_{12}\hat{q}_x = \text{expit}({}_{12}\hat{q}_x + 10)$$

where the addition of 10 accounts for the offset used when calculating the SVDs of the HMD mortality schedules.

Note that in the above framework, uncertainty of the estimate of ${}_{60}q_0$ is not considered. This is true of the log-quad model as well.

One approach to fitting the SVD model for estimating U5MR in LMICs would be to follow the steps above, using a survey-weighted estimate for ${}_{60}q_{0,z}$, then to average the resulting curves over sex using population-level sex weights to obtain an estimate for both sexes combined.

A second approach is to use the general structure of the SVD method in a model similar to the discrete hazards approach, and this is what we propose in the next subsection. Of note, the approach described in Clark (2019) requires separate model fits for males and females, whereas the (pseudo-likelihood) approach we propose allows for straightforward uncertainty quantification and can provide an estimate for males and females combined, without a secondary aggregation step.

Proposed SVD approach

For each yearly age group $x = \{0, 12, 24, 36, 48\}$, let y_{xp} be binomial counts and N_{xp} be totals in each time period p. We fit the logistic regression model

$$y_{xp} \mid N_{xp}, \eta_{xp} \sim \text{Binomial}(N_{xp}, \eta_{xp}),$$
$$\text{logit}(\eta_{xp}) = \boldsymbol{\beta}\mathbf{X},$$

where $\mathbf{X} = (\mathbf{1}^\top, \mathbf{u}_1, \ldots, \mathbf{u}_4)$ is a covariate matrix containing four LSVs $\mathbf{u}_1, \ldots, \mathbf{u}_4$ and an intercept term. The LSVs are computed from a SVD of a subset of the life tables contained in the HMD, but restricted to ages 0 to 5 years. The logistic regression model above is fitted using `svyglm()` to account for the survey design. We then obtain our final estimates ${}_{12}\hat{q}_x$ in the same manner as for the discrete hazards approach, in Equation (3.2). Details of the variance calculation for ${}_{12}\hat{q}_x$ are in Appendix B.3.1.

Clark (2019)'s analysis used only 4610 of HMD life tables, omitting those with implausible mortality at older ages, or incomplete data. We also remove the life tables for Belgium from 1914-1918 as they contain no data. Of the remaining life tables, 82 contained 0 estimates for ${}_{12}q_x$ for at least one age group, and hence those tables were also removed from the analysis. In total we used 4842 life tables in our analysis. The HMD data used in this paper were downloaded on April 23, 2022 from the HMD website (`https://www.mortality.org/File/GetDocument/hmd.v6/zip/by_statistic/lt_both.zip`). The resulting singular values are listed in Table 3.2 and visualized in Figure 3.3. As noted in Clark (2019), the first LSV is entirely negative and captures the underlying average shape of mortality across age. From 0 to 60 months, this corresponds to a stronger decrease at earlier ages that levels off slightly for later months. The reamining LSVs reflect age-specific deviations from the average pattern of mortality.

Age (months)	u_1	u_2	u_3	u_4
12	-0.28	-0.67	0.67	-0.13
24	-0.44	-0.45	-0.51	0.40
36	-0.47	-0.03	-0.33	-0.12
48	-0.50	0.27	-0.04	-0.72
60	-0.51	0.53	0.42	0.54

Table 3.2: LSVs obtained from 4842 life tables in the HMD subset to include ages 0 to 60 months.

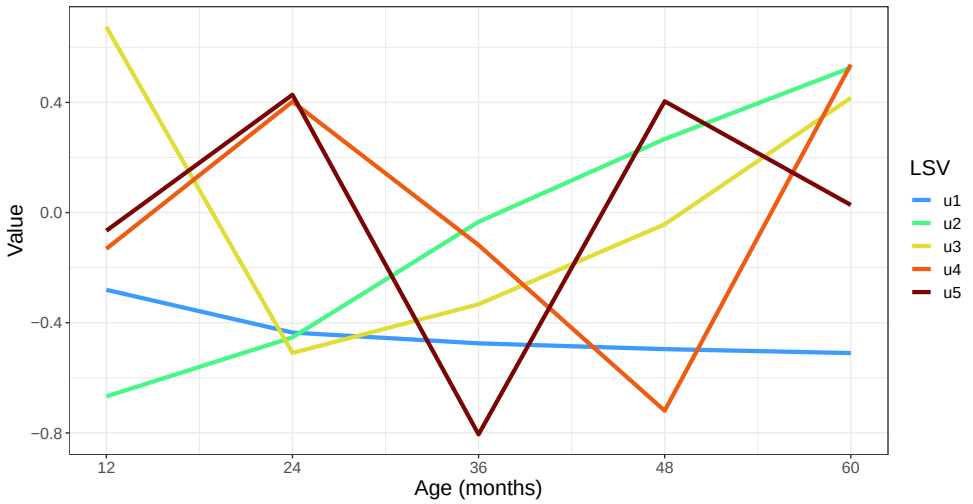


Figure 3.3: LSVs obtained from 4842 life tables in the HMD subset to include ages 0 to 60 months.

3.5 Application

We apply our proposed, parametric pseudo-likelihood approach to child mortality data from Burkina Faso, Malawi, Senegal, and Namibia. We chose single DHS surveys from each of these countries, and used the proposed approach to obtain continuous survival curves for the time periods $[2000, 2005)$ and $[2005, 2010)$ to demonstrate the ability of our approach to produce period estimates throughout time. The data used in the application is described in detail in Section 3.5.1, and all parametric models considered are noted in Section 3.5.2. We additionally fit a survey-weighted version of the Turnbull estimator, with boostrapped

confidence bands, to validate the parametric approaches, as described in Section 3.5.3.

For further comparison, we compare our approach to estimates from the log-quad model using all 22 age inputs (calculated from the Turnbull estimate), the proposed SVD approach described in Section 3.4.3, and the discrete hazards model from Mercer et al. (2015).

For all parametric approaches, we estimate the survival curves in each time period, uncertainty bands surrounding each survival curve (95% confidence bands based on finite population variances for all approaches other than log-quad, and the derived uncertainty band from Guillot et al. (2022) for the log-quad approach), and estimates of NMR, IMR, and U5MR from these survival curves.

3.5.1 Data

All data used in our application comes from the Demographic and Health Surveys (DHS) programme. The DHS programme is one of the largest producers of surveys in LMICs, and includes many health indicators including child mortality. Child death data is collected via interviewing mothers, and asking them the birth and death dates of all children they have had. As previously noted, child death dates are recorded daily up to 1 month of life, at monthly intervals up to 24 months, and at yearly intervals onward. Therefore, we treat deaths prior to one month as exact, and interval-censored afterwards with the interval given as a single month or a single year depending on when the child died.

It has previously been noted that DHS surveys are subject to potential biases that may negatively impact the resulting estimates of child mortality (Hill, 1995; Lawn et al., 2008; Guillot et al., 2022). The main concern for estimates of mortality under the age of 5 years is age-heaping at age 12 months, where more children are recorded as having died at 12 months than would otherwise be expected, especially compared to deaths recorded at surrounding months. Lawn et al. (2008) additionally note that age-heaping in DHS surveys may occur at 7 days, 14 days, and 1 month.

In our application, we address age-heaping at 12 months by interval-censoring all observations recorded as having died between 6 and 18 months for that entire 12 month period

$[6, 18)$. We chose this 12-month window to capture a wide range of potential age-heaping surrounding 12 months, but a smaller or larger (or multiple) windows could instead be chosen, depending on the assumption one wants to make about when age-heaping occurred. Note that in aggregating our data over these 12 months, we will lose some precision in our estimate of the survival curve but should decrease bias due to age-heaping. We emphasize that the benefit of this straightforward approach to addressing age-heaping is that the assumptions involved are made clear, in this case, that the only age-heaping in our data occurs between 6 and 18 months. Incorporating additional assumptions about where age-heaping occurs is straightforward; we include additional intervals surrounding the dates where age-heaping is thought to occur (for example, ages 3-10 days for age-heaping at 7 days).

All DHS surveys used in our application follow a two-stage, stratified cluster design, and were designed to provide accurate estimates at the Administrative 1 (admin1) subnational level. Strata are defined by admin1 region and urban/rural status. Each sampling frame is established from a previous census. Primary sampling units (PSUs), or clusters, are selected across strata, and the second stage of sampling consists of households within PSUs. GPS coordinates are displaced by up to 2km for urban clusters and 5km for rural clusters, and are not displaced outside of their strata. Information related to the sampling design for the surveys used in our application is given in Table 3.3.

Country	Survey Year	Census	Admin1 Regions	PSUs (U/R)	Households (U/R)
Burkina Faso	2010	2006	13	574 (176/398)	14924 (4576/10348)
Malawi	2016	2008	28*	850 (173/677)	27531 (5190/22341)
Senegal	2010	2002	14	392 (147/245)	8232 (3087/5145)
Namibia	2013	2011	13	554 (269/285)	11080 (5380/5700)

Table 3.3: Sampling information from DHS surveys. Census year is the year of the census upon which the sampling frame for the survey is based upon. PSUs and Households listed are the number of PSUs and Households in the sample, not the sampling frame, and counts are additionally disaggregated by Urban/Rural (U/R). *At the time of survey, Malawi's 28 districts were considered Admin2 regions, with Northern, Central, and Southern regions being Admin1. Some shapefiles now consider the 28 districts to be Admin1, with a finer grid as 243 Admin2 subregions.

3.5.2 Parametric Models

The parametric distributions considered for our proposed approach are listed in Table 3.4. Note that the exponentially-truncated shifted-power (ETSP) family of hazards we consider is slightly different than that considered in Schöley (2019), as we set $c = 0$ as opposed to estimating it via profile likelihood. The generalized Gamma distribution is parametrized as in the `flexsurv` package in `R`, as it is more numerically stable than the original parameterization (Prentice, 1974).

Distribution	Characterization	Parameters
Exponential	$f(x) = \beta e^{-\beta x}$	β
Piecewise Exponential	$f(x) = \beta_0 e^{-\beta_0 x} I[x < 1] + \beta_1 e^{-\beta_1 x} I[1 \leq x < 12] + \beta_2 e^{-\beta_2 x} I[x \geq 12]$	$\beta_0, \beta_1, \beta_2$
Weibull	$f(x) = \beta k (\beta x)^{k-1} e^{-(\beta x)^k}$	β, k
Generalized Gamma	$f(x) = \frac{\lvert Q \rvert (Q^{-2})^{Q^{-2}}}{\sigma x \Gamma(Q^{-2})} e^{Q^{-2}(Q\omega - e^{Q\omega})}$	Q, σ, ω
Lognormal	$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(\log(x)-\mu)^2}$	σ, μ
Gompertz	$f(x) = \beta k e^{k + \beta x - k e^{\beta x}}$	β, k
Exponentially-truncated shifted power (ETSP)*	$h(x) = a(x + c)^{-p} e^{-bx}$	a, b, c, p

Table 3.4: Parametric distributions considered and their characterizations in terms of a probability density function $f(x)$ or hazard $h(x)$, with relationship $h(x) = f(x)/(1 - F(x))$. Note that the ETSP hazard as described in Schöley (2019) contains four parameters, but in our applications we set $c = 0$.

3.5.3 Model Validation

To assist with model validation, we fit a survey-weighted version of the Turnbull estimator, with bootstrapped confidence bands, to provide a guideline for how well each of the parametric distributions is able to capture the underlying survival curve in each time period. Here, the Turnbull estimator of the survival curve is treated as a reasonable reference point for the underlying survival curve as it is free of parametric assumptions. However, despite our use of bootstrapped CIs we recall (from Section 3.3.1) that there are no well-justified variance

estimates for the Turnbull estimator, making our comparisons to the Turnbull estimator only crudely calibrated.

Let a sample k from the bootstrapped distribution of the Turnbull estimate at age x be denoted $\tilde{\theta}_x^{(k)}$, and a sample k from the asymptotic distribution of the parametric survival curve at age x be denoted $\hat{\theta}_x^{(k)}$. We obtain $k = 1, \ldots, 500$ samples, and compute $\hat{\theta}_x^{(k)} - \tilde{\theta}_x^{(k)}$ to obtain samples from the empirical distribution of the difference between the Turnbull and parametric distribution at a given age x.

We then produce visual comparisons of the range of values $\hat{\theta}_x^{(k)} - \tilde{\theta}_x^{(k)}$ at each age x, and observe whether or not these ranges contain 0, as would be expected if the Turnbull estimate and parametric estimate came from the same underlying distribution. We additionally calculate the proportion of uncertainty intervals derived from $\hat{\theta}_x^{(k)} - \tilde{\theta}_x^{(k)}$ at ages x that contain 0 as a rough estimate of how closely each parametric model aligns with the Turnbull estimate. Note that *this is not a formal hypothesis test*, but rather a means of assessing how close the parametric estimate is to the Turnbull estimate while accounting for uncertainty in *both* estimates.

3.6 Results

In this section, we display a subset of results from the application of the seven parametric models, log-quad model, proposed SVD approach, and discrete hazards model to DHS data from Burkina Faso, Malawi, Senegal, and Namibia. Additional results can be found in Appendix B.4, with comparisons to models where the data is not adjusted for age-heaping in Appendix B.5.

In the top row of Figure 3.4 we display the fitted survival curves for the Weibull model in both time periods for Malawi, and compare them to the Turnbull estimator, log-quad model, proposed SVD approach, and discrete hazards model. Note that compared to the Turnbull estimator the Weibull model tends to estimate higher survivorship at younger ages, and lower survivorship at older ages. In the bottom row of Figure 3.4, we display the same comparison for Malawi but for the lognormal model. It appears the lognormal model captures

the sharp increase in mortality within the first 12 months of life more accurately than the Weibull model. Showing just summary measures of mortality (NMR, IMR, U5MR), we see the same patterns in Figure 3.5. The Weibull model in each time periods underestimates NMR, and overestimates U5MR, relative to the Turnbull estimator, particularly for the period $[2000, 2005)$. In contrast, the lognormal model confidence intervals cover NMR, IMR, and U5MR in both time periods, with the exception of IMR in $[2005, 2010)$ where only the Weibull model captures the Turnbull estimate. We emphasize that these figures should all be interpreted with care, as the boostrapped confidence intervals for the Turnbull estimator are not well-justified. Therefore, small deviations from the Turnbull estimator may not necessarily imply a bad parametric fit. As seen in Table 3.5, the differences between the Weibull estimates and Turnbull estimates capture zero for 41.7% and 60.7% of ages where the Turnbull estimate is defined, prior to age 60 months, for $[2000, 2005)$ and $[2005, 2010)$, respectively. In contrast, the differences between the lognormal estimates and Turnbull estimates capture zero for 91.7% and 77.4% of ages. These percentages align with the visualizations to suggest that the lognormal model is a better parametric fit for the mortality curve for children under the age of 5 than the Weibull model.

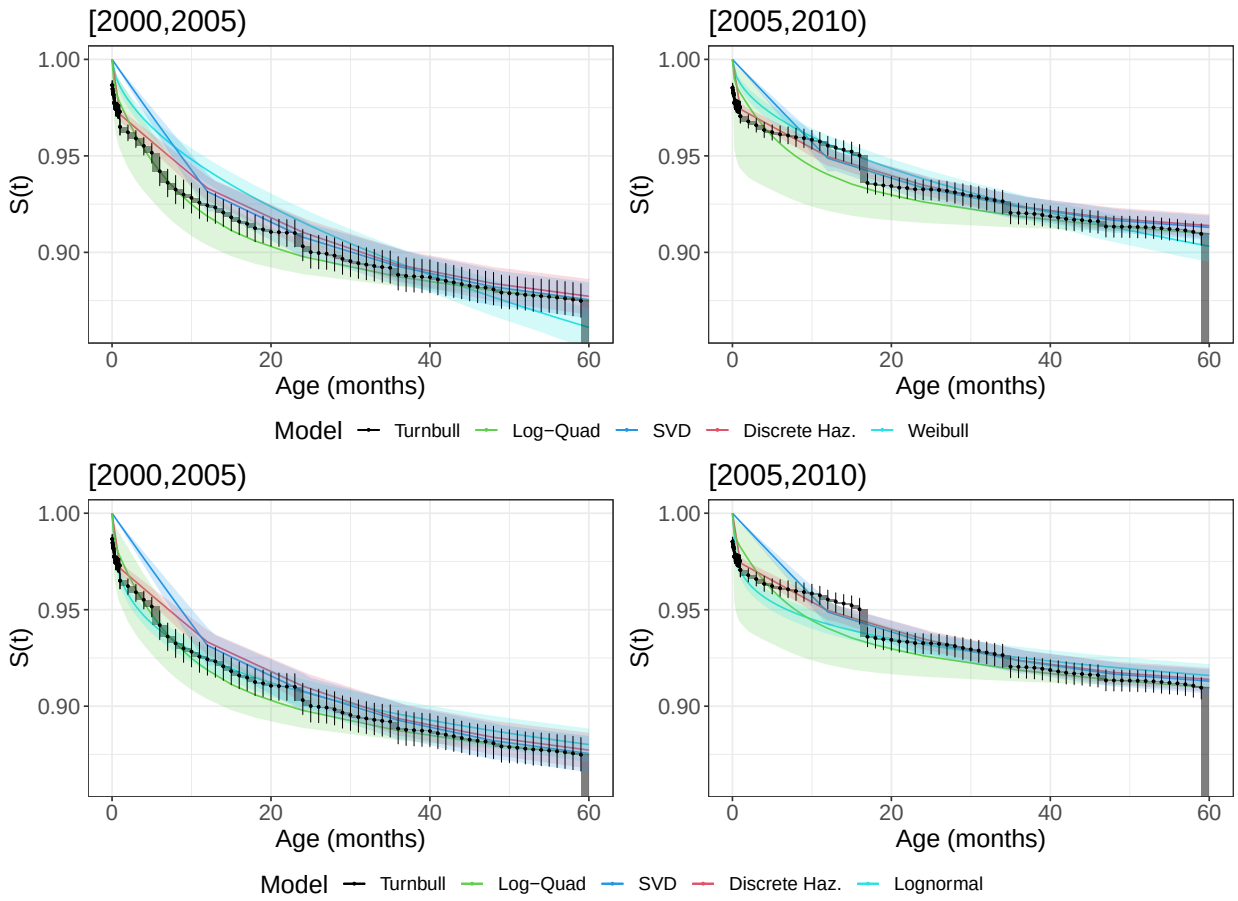


Figure 3.4: Top: Estimated Weibull survival curves for time periods $[2000, 2005)$ (left) and $[2005, 2010)$ (right) for Malawi. Bottom: Estimated lognormal survival curves for time periods $[2000, 2005)$ (left) and $[2005, 2010)$ (right) for Malawi.

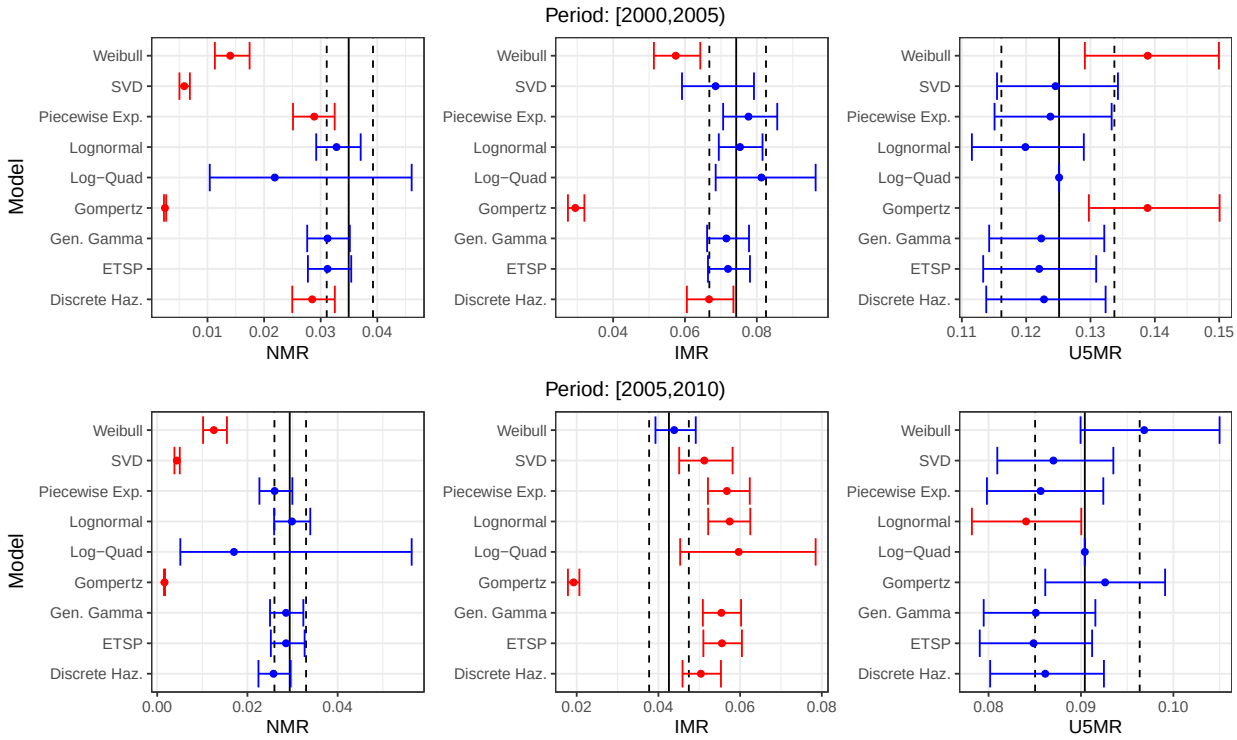


Figure 3.5: Estimates of NMR, IMR, and U5MR for Malawi in periods $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom). Turnbull point estimates are denoted by vertical black lines, with dashed vertical black lines denoting the 95% uncertainty interval estimated from the bootstrap samples. Horizontal error bars are blue if the interval captures the Turnbull point estimate, or red if the interval does not capture the Turnbull point estimate. All 95% confidence intervals are based on finite population variances, with the exception of the log-quad model where uncertainty is calculated as in Guillot et al. (2022).

We obtain similar results for Burkina Faso and Senegal, that suggest the Piecewise Exponential and Lognormal models are most similar to the survival curve given by the Turnbull estimator. For Namibia, a larger collection of parametric models may be appropriate, as the only parametric models we consider that do not align with the Turnbull estimator are the Weibull and Gompertz models.

The proposed SVD approach performs adequately in all countries in terms of capturing IMR and U5MR, but underestimates NMR in all time periods and countries. This makes sense, as the proposed SVD approach assumes a constant hazard between ages 0 and 12

months. Hence, unless mortality is decreasing linearly in the first year of life, the SVD approach will always underestimate NMR; if not linear, the survival curve will be convex. If finer scale life tables were available (for example, monthly life tables), this might be ameliorated.

Country	Period	Weibull	Piecewise Exponential	Generalized Gamma	Lognormal	Gompertz	ETSP	Discrete Hazards
Burkina Faso	[2000, 2005)	17	37	**93**	54	12	34	48
	[2005, 2010)	14	39	**71**	66	8	60	60
Malawi	[2000, 2005)	42	**76**	**95**	**92**	18	**94**	**73**
	[2005, 2010)	61	63	**82**	**77**	12	**80**	70
Senegal	[2000, 2005)	30	**73**	**85**	**83**	17	**84**	**72**
	[2005, 2010)	36	**73**	64	**85**	12	**91**	**72**
Namibia	[2000, 2005)	56	**86**	**100**	**100**	23	**99**	**86**
	[2005, 2010)	53	**86**	**100**	**100**	27	**99**	**85**

Table 3.5: Model validation results. Percentage of samples (out of 500) from $\hat{\theta} - \tilde{\theta}$ that contain 0 for all parametric models, countries, and periods. Results that contain more than 70% of samples noted in bold.

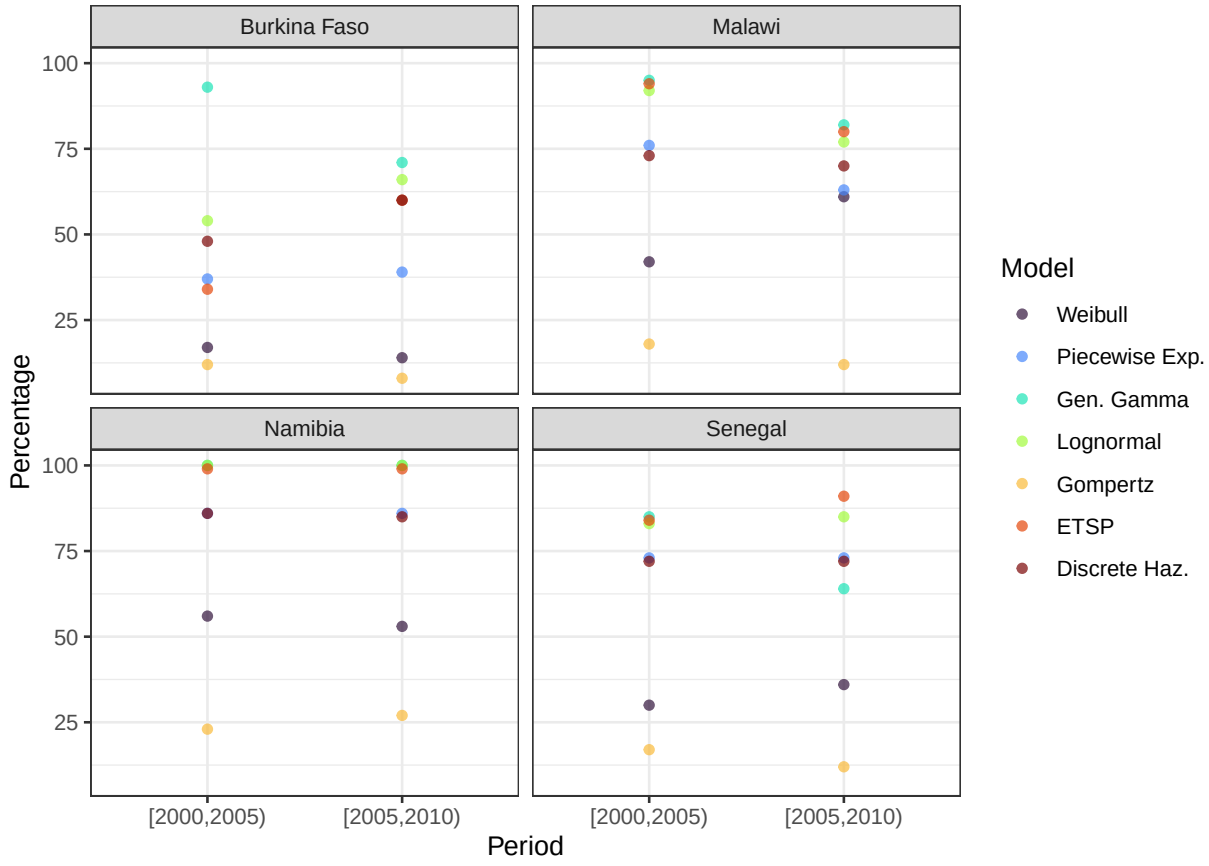


Figure 3.6: Model validation results. Percentage of samples (out of 500) from $\hat{\theta} - \tilde{\theta}$ that contain 0 for all parametric models, countries, and periods.

The log-quad approach performs adequately in general as well, with a few key caveats. The first is that U5MR is assumed to be estimated with no uncertainty, again. This is not a desirable property of this approach in LMICs, since our estimates of U5MR that are input to the log-quad model are themselves estimated with uncertainty. Second, we note that the uncertainty bands around the log-quad point estimates are in general much wider than the confidence bands for the parametric models. The confidence bands surrounding the parametric models may be interpreted at each age x with a 95% confidence interval interpretation based resampling observations from the finite population, whereas the uncertainty surrounding the log-quad model does not have as straightforward of an interpretation. Fur-

thermore, out of all of the analyses conducted, only the log-quad models for Namibia (both time periods) provided estimates and confidence bands that would be considered reasonable by Guillot et al. (2022). All other countries had estimated values for k outside the suggested range $(-1.1, 1.5)$, or an increasing hazard by age in the uncertainty interval computed, which is unrealistic.

In general, the discrete hazards approach performed well, though perhaps not sufficiently better than some of the proposed parametric models (such as lognormal or piecewise exponential) to justify the need for six parameters (coefficients for the discrete hazards) in estimating the survival curve. Futhermore, assuming a constant hazard over certain age intervals is not necessarily an assumption we wish to make, as it is unrealistic even at a fine scale of age groups.

One final thing to note is that both the ETSP and generalized Gamma models performed reasonably well, but again perhaps not significantly better than some of the two-parameter models. This can be seen for Namibia in Table 3.5 in particular, where both differences between the models' respective parametric estimates and the Turnbull estimates capture zero for 99% (ETSP) and 100% (generalized Gamma) of ages where the Turnbull estimate is defined prior to 60 months. However, the lognormal model performs just as well in Namibia according to this metric. Furthermore, fitting both models (generalized Gamma and ETSP) results in computational complexities that may make them less desirable than other parametric options. The ETSP model, for example, does not have a closed form cumulative hazard, and therefore requires numerical integration at every step of the likelihood optimization. Therefore, this model takes more time to fit and is potentially less numerically stable than others with closed form cumulative hazards. The generalized Gamma distribution, on the other hand, occasionally produces unreasonably wide confidence bands. In some countries and time periods, Burkina Faso $[2000, 2005)$, for example, the shape parameter Q is estimated with a large variance relative to the other parameters in the model. This results in the confidence bands produced being highly asymmetric, and so wide as to be unusable in practice. As such, we believe that in certain cases there may not be enough data under the

age of 5 to reliably estimate all three parameters that define the generalized Gamma survival curve.

3.7 Discussion

In this paper, we have proposed a parametric, pseudo-likelihood approach to estimation of a full mortality schedule for children under 5 across multiple time periods. The novelty of our approach lies in the explicit set-up of our methodology in a survival context with a time-varying covariate representation to allow period effects, the ability to work directly with survey data and straightforwardly conduct design-based inference, and the particular focus of our methodology on low- and middle-income countries. We clarify the relationship between existing demographic concepts (e.g. period vs. cohort estimates) and concepts in survival analysis. We describe three existing approaches to estimating full survival curves for children under the age of 5, and describe their potential shortcomings in terms of statistical properties (such as variance calculations), demographic properties (constant hazards between age groups), and applicability to LMICs. We propose a flexible parametric approach that we believe is suited to survey data from LMICs, and easily extended to additional parametric models. We additionally describe a nonparametric pseudo-likelihood approach, the Turnbull estimator, and propose that it may be used informally for model validation of parametric approaches. Finally, we apply our proposed methods to DHS surveys from Burkina Faso, Malawi, Senegal, and Namibia.

Our application suggests that there are potentially very large differences in model fit between parametric distributions, with the Weibull distributions and Gompertz distributions generally providing the worst fit compared to the Turnbull estimator, in terms of capturing the survival curve under the age of 5. In general, the lognormal model seems to fit the countries in our application reasonably well. We note that two of the three parameter models we compared, the piecewise exponential and ETSP model, also adequately captured the survival curve provided by the Turnbull estimator, though the piecewise exponential model has the undesirable property of assuming constant hazards within prespecified age

groups and the ETSP model is computationally challenging to fit. We conclude that for our application, the lognormal model outperforms other parametric models in terms of the ability to capture the point estimate provided by the Turnbull estimator while only requiring two parameters to define the survival curve.

The benefits of a parametric approach to under-5 mortality estimation, and in particular to estimating the full survival curve for children under the age of 5, are many. As laid out in Schöley (2019), we note that correctly specified parametric assumptions about the shape of mortality may greatly assist estimation of the survival curve under the age of 5 in situations with little data. This becomes particularly relevant in a small area setting, where often little to no data is available at small administrative regions (Wakefield et al., 2020). As such, the methods proposed in this paper may serve as a guideline, or perhaps even as prior information in a Bayesian setting, for small area estimation problems of child mortality when a full survival curve is desired. Further benefits of a continuous, parametric approach involve interpretability and parsimony. Perhaps the most well-known parametric, demographic model for mortality estimation, the Heligman-Pollard model (Heligman and Pollard, 1980), provides informative interpretations of the parameters involved in the model, and the same is true of the models we propose. Of course, we rely on the assumption that the parametric distribution we specify is *correctly* specified, which likely is not the case. In fact, it is likely that there is *no* parametric distribution that can perfectly capture the age trend in mortality for every country. However, especially in scenarios with little data, *reasonable* parametric assumptions may still be useful. Hence it is important to observe and test these parametric models in settings with relatively more data, such as the national setting we use in our application. Note that a meaningful question is: Is the fit of a continuous parametric model better than the 6-parameter discrete hazards model currently used by the UN IGME and DHS? When comparing to the Turbull estimator, the lognormal model does outperform the 6-parameter discrete hazards model in terms of our model performance metric (see Table 3.5).

Limitations of our proposed approach to extend the Turnbull estimator include the lack

of a well-justified variance estimate. As previously noted, a variance estimate is not readily available due to the non-Gaussian, cubed-root asymptotics, and a bootstrap estimate of the variance is not applicable for similar reasons. More work needs to be done to provide an estimate of the asymptotic variance of the Turnbull estimator in a superpopulation setting, and then additionally to a finite population setting, before comparisons between the non-parametric and parametric approaches (and model validation procedures) can be made with some degree of calibration.

In conclusion, we have provided a method for obtaining a complete, continuous survival curve for children under the age of 5 using assumed parametric models. Our method enables estimations using interval-censored, left-truncated observations, as is required for period estimates of mortality from DHS data. Furthermore, aspects of survey design, which are particularly relevant in LMICs, may be directly incorporated into our modeling framework to provide design-consistent estimates of mortality with finite population variances. All software for fitting the models proposed in this paper is available in the `R` package `pssst`, available at `https://github.com/taylorokonek/pssst`.

Chapter 4

CORRELATED TEMPORAL SMOOTHING MODELS FOR MORTALITY ESTIMATION

4.1 Introduction

As noted in previous chapters, data available for estimating U5MR in LMICs becomes sparse at small levels of spatial aggregation and finer temporal scales. As a consequence, obtaining direct, survey-weighted estimates that are precise in all regions and time points becomes impossible; particularly when there is no relevant data available in a region. Nevertheless, accurate estimates of U5MR are needed in every region, particularly at the Administrative 2 (admin2) level where public health interventions primarily take place. Model-based approaches to estimating U5MR incorporate spatio-temporal smoothing random effects and are often used at a subnational and fine temporal scale to overcome the lack of relevant data available. These models introduce some bias into the estimates due to shrinkage, in exchange for potentially large gains in precision.

In this chapter, we will discuss existing temporal smoothing terms, and propose a multivariate random walk that allows multiple, potentially-correlated observations to co-vary in time. This approach directly extends the work in Chapter 3, providing a parametric model analogue to the smoothed Fay-Heriot approach developed in Mercer et al. (2015), with the added benefit of producing estimates of NMR, IMR, and U5MR that follow known relationships in mortality; for example, NMR $<$ IMR $<$ U5MR.

In Section 4.2, we detail existing random walk models, and show how a particular derivation of the form for the random walk prior allows us to extend the prior to allow for two correlated random walks in time. We compare survey-weighted estimates produced from parametric models considered in Chapter 3 in 27 districts of Malawi on a yearly scale from

2000 to 2009, and smooth these estimates using the existing and proposed random walks in Section 4.3. We conclude with a discussion of hyperprior choice and possible extensions.

4.2 Methods

We obtain survey-weighted, direct estimates of parameters from the pseudolikelihood approach described in Chapter 3, using the Weibull, lognormal, piecewise exponential (with breakpoint at 1 month), and Gompertz distribution. We determine which of these two-parameter models most closely captures the mortality curve for children under the age of 5 in each area estimated from the survey-weighted Turnbull estimator.

For the thusly determined parametric model, we compare the weighted estimates with these three smoothing approaches:

1. RW2: Smooth NMR, IMR, U5MR directly

   Obtain survey-weighted estimates of NMR and U5MR, then smooth each over time separately using RW2 models

2. Uncorrelated MVRW2: Smooth parameters, then calculate NMR, IMR, U5MR

3. Correlated MVRW2: Smooth parameters, then calculate NMR, IMR, U5MR

In the following subsections, we describe the proposed first-order multivariate random-walk (MVRW1), provide a derivation of the model akin to the derivation of the first-order random walk, and detail the models we use in our application that implement the three smoothing approaches. The extension to second-order random walks used in the application is provided in Appendix C.

4.2.1 Existing temporal GMRFs

Two of the most common temporal smoothing terms are the autoregressive (AR) process and the random walk. Both are specific forms of Gaussian Markov random fields (GMRFs),

which are distinguished by (1) having a Gaussian precision matrix, and (2) containing the Markov property of conditional independence between certain observations given all other observations (Rue and Held, 2005). In a temporal setting, this conditional independence manifests in independent observations for discrete temporal observations that are far apart (more than one time point away for first-order random effects, two time points for second-order random effects, and so on), and results in 0's in the off-diagonal entries of the precision matrix for the GMRF. The 0's in the off-diagonals provide sparsity to the specification of the GMRFs that enable use of fast, "sparse matrix" techniques.

Autoregressive processes are proper random effects with invertible precision matrices. By proper we mean that the random effects have a well-defined density function with precision matrices of full rank. A first-order autoregressive process (AR1) on the vector $x_1, \ldots, x_n$ can be written in its conditional form as

$$\begin{aligned} x_1 &\sim N(0, \tau^{-1}/(1-\phi^2)), \\ x_n \mid x_1, \ldots, x_{n-1} &\sim N(\phi x_{n-1}, \tau^{-1}), \end{aligned}$$

where $|\phi| < 1$, and its joint form as

$$\pi(\mathbf{x}) \sim N(\mathbf{0}_n, \tau^{-1}\mathbf{Q}^{-1}),$$

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

Details on autoregressive processes, as well as other GMRFs, can be found in Rue and Held (2005).

Taking $\phi \to 1$ in the AR1 produces the first-order random walk (RW1), known as an

intrinsic GMRF (IGMRF). Random walks are improper random effects, where by improper we mean the precision matrix of the random effect is not full rank. To see this, note that a RW prior can be written as,

$$\pi(\mathbf{x}) \sim N(\mathbf{0}_n, \tau^{-1}\mathbf{Q}^-),$$
$$\mathbf{Q} = \begin{pmatrix} 1 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 1 \end{pmatrix}, \tag{4.1}$$

and note that multiplying $\mathbf{Q}$ by a column vector of 1's produces a column vector of 0's. In the case of an RW1, this means that the level (intercept) of a temporal trend cannot be estimated from a model that includes only an RW1 random effect. In practice, a model that includes both an RW1 random effect and intercept term can estimate the level of a temporal trend (via the intercept term) so long as the RW1 is constrained so that the intercept is identifiable. The constraints are given by the matrix of eigenvectors that correspond to the zero eigenvalues of the precision matrix; in this case, a $1 \times n$ dimensional matrix of 1's.

We have derived the random walk by beginning with an AR1 and taking $\phi \to 1$. We could instead begin by placing Gaussian priors on the differences:

$$\begin{aligned} x_2 - x_1 &\sim N(0, \tau^{-1}), \\ x_3 - x_2 &\sim N(0, \tau^{-1}), \\ &\vdots \\ x_n - x_{n-1} &\sim N(0, \tau^{-1}). \end{aligned}$$

Assuming the differences are iid (a Markov property) with an improper prior on x_1, then we

can write the joint "density" for $\mathbf{x} = (x_1, \ldots, x_n)$ as

$$\begin{aligned}\pi(x_1, \ldots, x_n \mid \tau) &= \tau^{(n-1)/2} \exp\left(-\frac{\tau}{2} \sum_{i=1}^{n-1} (x_{i+1} - x_i)^2 \right), \\ &= \tau^{(n-1)/2} \exp\left(-\frac{\tau}{2} \mathbf{x}^\top \mathbf{Q} \mathbf{x} \right)\end{aligned}$$

where $\mathbf{Q}$ is defined in Equation (4.1). This derivation will be convenient in the following development.

Note also that we can write

$$\begin{pmatrix} -1 & 1 & & & & \\ & -1 & 1 & & & \\ & & \ddots & \ddots & & \\ & & & -1 & 1 & \\ & & & & -1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_2 - x_1 \\ x_3 - x_2 \\ \vdots \\ x_n - x_{n-1} \end{pmatrix}.$$

Denote the lefthand matrix $\mathbf{R}$ (an $(n-1) \times n$ dimensional matrix), and note that $\mathbf{R}^\top$ is a projection matrix that takes the vector $\mathbf{x}^\top$ of dimension n to the vector of differences $(\mathbf{x}_i - \mathbf{x}_{i-1})^\top = (x_2 - x_1, x_3 - x_2, \ldots, x_n - x_{n-1})^\top$ of dimension $n-1$. As we previously noted, the vector differences are iid $N(0, \tau)$, so we can write $\pi((\mathbf{x}_i - \mathbf{x}_{i-1})) \sim N(\mathbf{0}_{n-1}, \tau^{-1}\mathbf{I}_{n-1})$. To obtain the precision matrix for $\mathbf{x}$ then, we can project our vector using $\mathbf{R}^\top$ to a space where the density is easily defined, and then project back using $\mathbf{R}$ to obtain our final joint density. This corresponds to

$$\pi(\mathbf{x}) \sim N(\mathbf{0}_n, (\mathbf{R}^\top (\tau^{-1}\mathbf{I}_{n-1})^{-1} \mathbf{R})^-) \tag{4.2}$$

and finally, note that $\mathbf{R}^\top (\tau^{-1}\mathbf{I}_{n-1})^{-1}\mathbf{R} = \tau\mathbf{Q}$. Therefore, this derivation, summarized in Equation (4.2), arrives at the same joint distribution for $\mathbf{x}$ as the previous two.

Note that the precision matrix in Equation (4.2) involves a generalized inverse. In prac-

tice, we handle generalized inverses computationally in the same way as INLA (Rue et al., 2009). A small constant (1×10^{-6}, by default) is added to the diagonal of the precision matrix, and the random effect is constrained using the *conditioning by kriging* equations, found in Subsection 2.3.3 of Rue and Held (2005).

4.2.2 Proposed temporal GMRF

Suppose we now have observations $(x_1, \ldots, x_n, w_1, \ldots, w_n)$. Rather than assume that $(x_j - x_{j-1}) \perp\!\!\!\perp (w_j - w_{j-1})$ (as would be the case if we put an RW1 prior on the whole $2n$-dimensional vector), we instead assume

$$\begin{pmatrix} x_j - x_{j-1} \\ w_j - w_{j-1} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{\Sigma} \right), \qquad \mathbf{\Sigma} = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_w \\ \rho\sigma_x\sigma_w & \sigma_w^2 \end{pmatrix} \tag{4.3}$$

where parameter σ_x^2 is the variance of the differences between adjacent x_j indices, σ_w^2 is the variance of the differences between w_j indices, and ρ is the correlation between $(x_j - x_{j-1})$ and $(w_j - w_{j-1})$.

Similarly to the RW1 case, we can construct a projection matrix $\mathbf{R}^\top$ as $\mathbf{R}^\top = (\mathbf{R}_1, \mathbf{R}_2)^\top$,

where

$$\mathbf{R}_1 = \begin{pmatrix} \mathbf{M}_{1,*} \\ \mathbf{0}_t^\top \\ \mathbf{M}_{2,*} \\ \mathbf{0}_t^\top \\ \vdots \\ \mathbf{M}_{n-1,*} \\ \mathbf{0}_n^\top \end{pmatrix}, \qquad \mathbf{R}_2 = \begin{pmatrix} \mathbf{0}_t^\top \\ \mathbf{M}_{1,*} \\ \mathbf{0}_t^\top \\ \mathbf{M}_{2,*} \\ \vdots \\ \mathbf{0}_n^\top \\ \mathbf{M}_{n-1,*} \end{pmatrix},$$

$$\mathbf{M} = \begin{pmatrix} -1 & 1 & & & & \\ & -1 & 1 & & & \\ & & \ddots & \ddots & & \\ & & & -1 & 1 & \\ & & & & -1 & 1 \end{pmatrix},$$

where $\mathbf{M}$ is a $(n-1) \times n$ matrix, and $\mathbf{M}_{i,*}$ denotes the i'th row of $\mathbf{M}$.

In its most general form, let $\mathbf{\Sigma}$ be defined as in Equation (4.3). Then our proposed random walk where the differences between adjacent x_j and w_j are correlated with potentially different precisions can be written as

$$\begin{pmatrix} \mathbf{x}^\top \\ \mathbf{w}^\top \end{pmatrix} \sim N(\mathbf{0}_{2n}, (\mathbf{R}^\top (\mathbf{I}_{n-1} \otimes \mathbf{\Sigma}^{-1}) \mathbf{R})^-),$$

where $\mathbf{A} \otimes \mathbf{B}$ denotes the kronecker product between matrices $\mathbf{A}$ and $\mathbf{B}$. This general case may be simplified. For example, ρ could be set to zero to allow for the differences between adjacent x_j and w_j to have different precisions but enforce independence between x_j and w_j in time, a priori. Note also that $\mathbf{R}^\top \mathbf{R} = \mathbf{I}_2 \otimes \mathbf{Q}$, where $\mathbf{Q}$ is the structure matrix for a RW1. Interestingly, if we have only a single precision parameter in our prior for both $\mathbf{x}$ and

$\mathbf{w}$ with $\rho = 0$, this is exactly a Type-II Knorr-Held interaction between the vectors $\mathbf{x}$ and $\mathbf{w}$ (Knorr-Held, 2000).

Of note, just as the RW1 is improper, so too is the proposed random walk prior. Specifically, the level of the random effect for each vector $\mathbf{x}$ and $\mathbf{w}$ cannot be determined from this prior alone. To make the prior proper, the proposed GMRF requires two constraints: a sum-to-zero constraint across $\sum_{i=1}^{n} x_i = 0$ and $\sum_{i=1}^{n} w_i = 0$. As usual, these constraints are only needed for identifiability if a fixed effect for $\mathbf{x}$ and $\mathbf{w}$ is included in the model, which corresponds to unique intercept terms for $\mathbf{x}$ and $\mathbf{w}$.

The proposed prior can be straightforwardly extended to random walks of higher order, spatial smoothing priors such as the intrinsic conditional autoregressive (ICAR) model, and to vectors of higher dimension (more than simply two vectors defined in time $\mathbf{x}$ and $\mathbf{w}$, for example). The only differences will be in the projection matrix $\mathbf{R}^{\top}$, and the covariance matrix $\mathbf{\Sigma}$, which defines the covariance in the projected space. Constructions for $\mathbf{R}$ for the two vector case for RW2 and ICAR priors are given in Appendix C.1. Examples of what samples from the proposed MVRW2 may look like for different hyperparameter values can be seen in Figure 4.1.

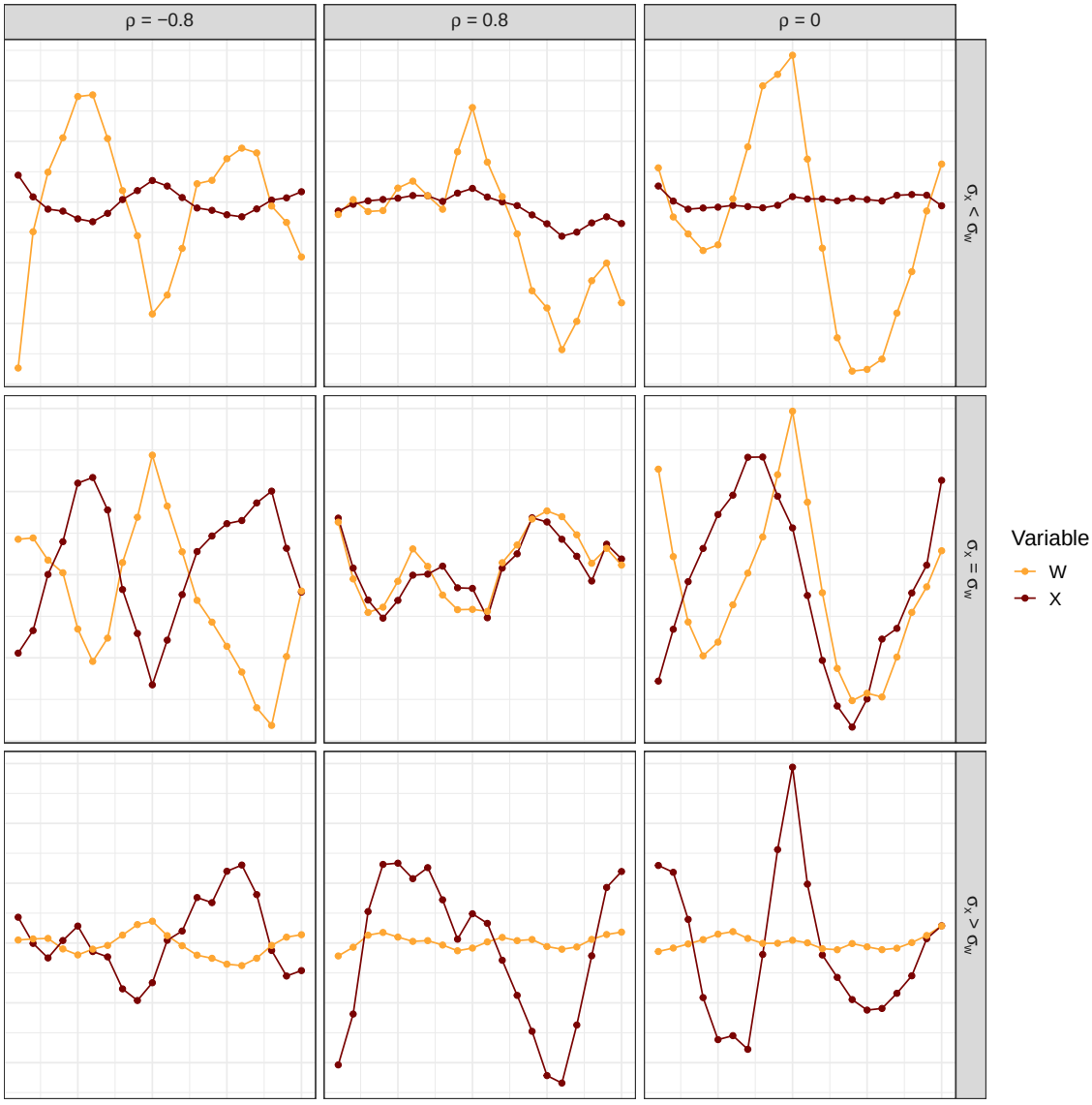


Figure 4.1: Examples of MVRWs generated from the proposed model, with different hyparparameter values (denoted in grey labels). Each time series consists of 20 time points.

4.2.3 Implentation

We implement all models for this chapter in INLA, a fast Bayesian computational programme described in greater detail in Chapter 2, Section 2.5.4. Implementing the proposed models in INLA is nontrivial as the likelihood at the highest level has a nondiagonal covariance

structure, and the random effect has multiple hyperparameters that do not easily fit into existing INLA structures for random effects.

To address the nondiagonal covariance structure in the likelihood, we note that the following models are equivalent as $\sigma \to 0$.

$$\begin{aligned} \mathbf{y} \mid \boldsymbol{\eta} &\sim \text{MultivariateNormal}(\boldsymbol{\eta}, \mathbf{I}\sigma) \\ \boldsymbol{\eta} &= \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \cdots + \boldsymbol{\phi} \end{aligned}$$

and

$$\begin{aligned} \mathbf{y} \mid \boldsymbol{\eta} &\sim \text{MultivariateNormal}(\boldsymbol{\eta}, \mathbf{V}) \\ \boldsymbol{\eta} &= \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \ldots \end{aligned}$$

where $\boldsymbol{\phi} \sim \text{MultivariateNormal}(\mathbf{0}, \mathbf{V})$. While the former model cannot be fit in INLA, the latter can be specified. We include pseudo-code in Listing 4.1 for fitting this model using INLA.

```
formula ← y ~ x1 + ... +
   f(idx, model = "generic0",
     Cmatrix = solve(V),
     hyper = list(prec = list(initial = 0, fixed = TRUE)))
result ← inla(formula,
              data = inla_data,
              family = "gaussian",
              control.family = list(hyper =
              list(prec = list(initial = 15, fixed = TRUE))))
```

Listing 4.1: INLA pseudo-code for MVN likelihood

To specify the correlated multivariate random walk in INLA, we use INLA's `rgeneric` capability to freely specify a random effect. Pseudo-code for each of these models in `rgeneric`

is given in Appendix C, Section C.2.

4.2.4 Model Statements

There are different classes of models we could consider, which are summarized in Figure 4.2. Note that in all cases, weighted estimates are first obtained via a pseudo-likelihood approach, and models that induce smoothing via random effects are then applied to (possible summaries of) the weighted estimates. The first distinction is between obtaining direct estimates via the parametric, pseudo-likelihood approach described in Chapter 4, or obtaining direct estimates view the discrete hazards approach of Mercer et al. (2015). Note that the latter is the approach currently used to estimate summaries of child mortality by both UN IGME and the DHS (Li et al., 2019; Wu et al., 2021). From either of these models, we can then choose to work with the survey-weighted parameters directly, or compute survey-weighted summaries of these parameters. For the discrete hazards model, the parameters are the six, discrete hazards, and for the two-parameter models these are the shape and scale. Summaries computed are typically NMR, IMR, and U5MR, though we note that for the parametric models other summaries of the survival curve could be straightforwardly computed.

Currently, the standard smoothing model is the Smoothed Direct model of summaries of the discrete hazards (Mercer et al., 2015; Li et al., 2019); specifically, smoothing NMR, IMR, and U5MR separately. Note that we could similarly obtain Smoothed Direct estimates from summaries of the parametric direct estimates. Additional smoothing approaches we consider involve smooth survey-weighted estimates from the parametric models.

We consider two transformations of the survey-weighted parameters, and two potential smoothing approaches. The first transformation is simply the identity, where we smooth the parameters as they are estimated from the pseudolikelihood approach. The second computes ${}_1q_0$ (NMR) and ${}_1q_{59}$ from the survey-weighted shape and scale parameters, and smooths those estimates across time. Note that this transformation will allow us to enforce the relationship NMR $<$ U5MR. For each of these transformations, we then consider two smoothing models: a correlated multivariate random walk, and an uncorrelated multivariate walk with separate

precisions for each of the two estimates across time.

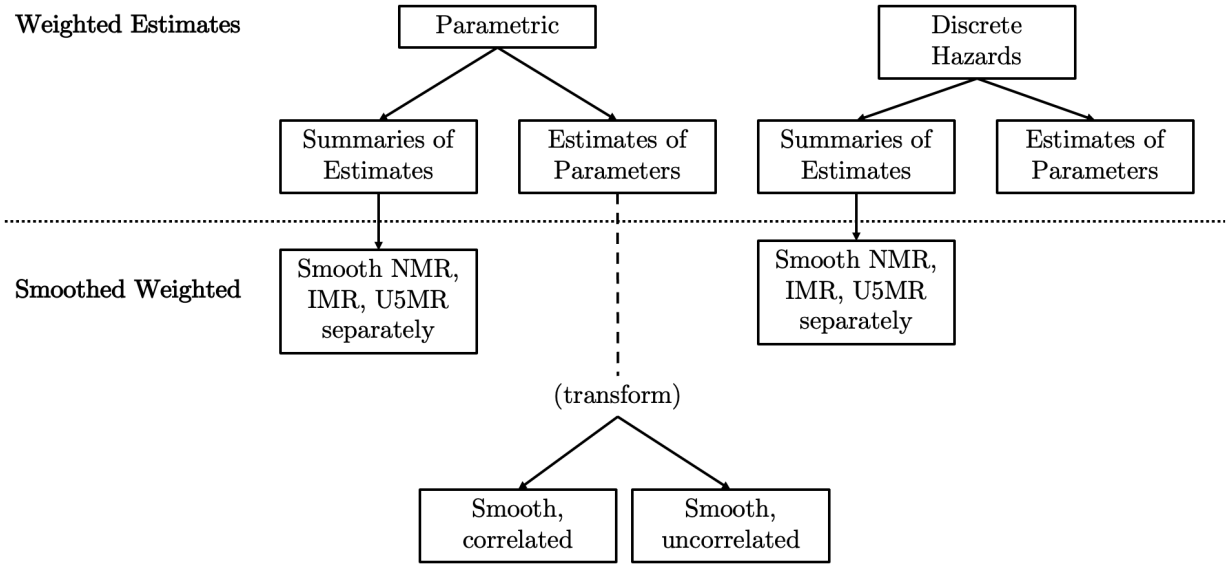


Figure 4.2: Estimates and smoothing models we consider. Direct estimates may come from either the parametric approach in Chapter 3, or the discrete hazards approach currently used by UN IGME and DHS. We then consider smoothing either summaries of the direct estimates (NMR, IMR, U5MR) or the underlying survey-weighted estimates of parameters (potentially transformed).

We now state the smoothing models we will consider. Let $i = 1, \ldots, n$ denote yearly, discrete time points and ${}_x\hat{q}_{0,i}$ the direct estimate of U5MR ($x = 60$), IMR ($x = 12$) or NMR ($x = 1$) at time i, with associated design variance $\hat{V}_{xi}$. Let the direct estimates from the two-parameter parametric model be denoted $\hat{\theta}_{1i}$ and $\hat{\theta}_{2i}$, where both parameters are estimated on

the whole of the real line (for example, by log transforming the parameters) with associated design covariance $\hat{C}$.

We consider the following smoothing models separately for each district:

1. Smooth NMR, IMR, U5MR separately:

$$\text{logit}({}_x\hat{q}_{0,i}) \mid \eta \sim \text{Normal}(\eta_i, \hat{V}_{xi})$$
$$\eta_i = \beta + \phi_i$$

   where $\phi_i \sim \text{RW1}(\tau_\phi)$, $\tau_\phi \sim \text{PC}(U = 1, \alpha = 0.01)$, $\beta \sim N(0, 0.001^{-1})$, for $x = 1$, $x = 12$, and $x = 60$.

2. Smooth, uncorrelated:

$$(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) \mid \boldsymbol{\eta} \sim \text{MultivariateNormal}(\boldsymbol{\eta}, \hat{C})$$
$$\boldsymbol{\eta} = \mathbf{X}^\top \boldsymbol{\beta} + \psi(\tau_{\theta_1}, \tau_{\theta_2})$$

   where $\psi(\tau_{\theta_1}, \tau_{\theta_2}) \sim \text{MVRW2}(\tau_{\theta_1}, \tau_{\theta_2}, \rho \equiv 0)$, $\tau_{\theta_1} \sim \text{PC}(U = 1, \alpha = 0.01)$, $\tau_{\theta_2} \sim \text{PC}(U = 1, \alpha = 0.01)$, $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top \overset{iid}{\sim} N(0, 0.001^{-1})$, and $\mathbf{X} = \begin{pmatrix} \mathbf{1}_n & \mathbf{0}_n \\ \mathbf{0}_n & \mathbf{1}_n \end{pmatrix}$.

   (a) $(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) = (\log(1/\hat{\boldsymbol{\mu}}), \log(\hat{\boldsymbol{\sigma}}))$

   (b) $(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) = (\text{logit}({}_1\hat{\mathbf{q}}_0), \text{logit}({}_{59}\hat{\mathbf{q}}_1))$

3. Smooth, correlated:

$$(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) \mid \boldsymbol{\eta} \sim \text{MultivariateNormal}(\boldsymbol{\eta}, \hat{C})$$
$$\boldsymbol{\eta} = \mathbf{X}^\top \boldsymbol{\beta} + \psi(\tau_{\theta_1}, \tau_{\theta_2}, \rho)$$

   where $\psi(\tau_{\theta_1}, \tau_{\theta_2}) \sim \text{MVRW2}(\tau_{\theta_1}, \tau_{\theta_2}, \rho)$, $\tau_{\theta_1} \sim \text{PC}(U = 1, \alpha = 0.01)$, $\tau_{\theta_2} \sim \text{PC}(U =$

$1, \alpha = 0.01)$, $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top \overset{iid}{\sim} N(0, 0.001^{-1})$, $\mathbf{X} = \begin{pmatrix} \mathbf{1}_n & \mathbf{0}_n \\ \mathbf{0}_n & \mathbf{1}_n \end{pmatrix}$, and $\text{logit}\left(\frac{\rho+1}{2}\right) \sim$ Normal$(0, 1)$.

(a) $(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) = (\log(1/\hat{\boldsymbol{\mu}}), \log(\hat{\boldsymbol{\sigma}}))$

(b) $(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) = (\text{logit}({}_1\hat{\mathbf{q}}_0), \text{logit}({}_{59}\hat{\mathbf{q}}_1))$

4.3 Application

We apply our three smoothing approaches to yearly, survey-weighted estimates for 27 districts of Malawi from 2000-2010, from the 2016 Malawi DHS. The survey followed a two-stage, stratified cluster design, with strata given by district crossed with urban rural and clusters as primary sampling units. There were 12,558 clusters in the sampling frame with 827 sampled clusters. Additional survey design information for the 2016 Malawi DHS can be found in Chapter 3, Subsection 3.5.1. Of note, there are 28 districts in Malawi. Our analyses exclude the island of Likoma, as it is spatially distinct from the other 27 districts of Malawi and has a small population. We first obtain survey-weighted estimates from Weibull, lognormal, piecewise exponential (with breakpoint at 1 month), and Gompertz distributions, as well as the discrete hazards model of Mercer et al. (2015). Additionally, we obtain survey-weighted Turnbull estimates to use as a comparator for assessing fit of our parametric models.

For the parametric model that is deemed best fitting, direct estimates of U5MR, IMR, and NMR are calculated using samples from the finite-population, asymptotic normal distribution for the survey-weighted parameters and the CDF for that distribution. We then fit the three smoothing models described in Section 4.2 as well as the discrete hazards approach of Mercer et al. (2015) with six age-specific hazards, and compare the resulting smoothed estimates. For models 2 and 3, we consider smoothing direct estimates of the shape and scale parameter for the parametric distribution, as well as a transformation of those direct estimates to ${}_1q_0$ and ${}_{59}q_1$, resulting in five total smoothing models considered for the pseudolikelihood estimates.

4.3.1 A note on the Turnbull estimator

When an individual is interval-censored across the boundary of a time period, that individual cannot be split into separate observations that are left truncated at the beginning of a given time period. Intuitively, this would imply that a single individual could contribute more than one death to the risk set, which biases estimates of mortality upwards, and also overstates the amount of information present. When working on a yearly scale, many individuals are interval censored across the boundary of a one-year time period. For 2000-2009 data from Malawi, roughly 3% of all individuals in each district are interval-censored across a time period boundary, which constitutes approximately 40-50% of all observed deaths in each district. Table 4.1 gives the exact breakdown of these percentages by district.

Table 4.1: Percentages of individuals who are interval-censored across time period boundaries out of all individuals at risk, and percentages of individuals who are interval-censored across time period boundaries out of all observed deaths by district in Malawi for 2000-2009.

District	Percent across boundary out of all individuals	Percent across boundary out of all deaths
Balaka	4	42
Blantyre	3	48
Chikwawa	2	41
Chiradzulu	4	51
Chitipa	2	39
Dedza	3	42
Dowa	3	56
Karonga	3	44
Kasungu	3	41
Lilongwe	4	46
Machinga	3	42
Mangochi	3	40
Mchinji	4	40
Mulanje	5	47
Mwanza	3	43
Mzimba	2	36
Neno	5	51
Nkhata Bay	2	33
Nkhotakota	3	47
Nsanje	3	48
Ntcheu	3	44
Ntchisi	3	41
Phalombe	3	40
Rumphi	2	30
Salima	3	51
Thyolo	3	37
Zomba	3	44

To account for an individual who is interval-censored across time period boundaries in the Turnbull estimator, we split the individual into separate, left truncated observations at the beginning of the time period, where the last two observations for this individual will each contain an interval censored observation. We then down-weight these last two observations

by the proportion of the length of the original individual's interval that is included in that time period. Note that this assumes that the age distribution of deaths in each of these two time periods is the same. Though we know this assumption will not hold (due to cohort effects of conflicts, for example), the resulting survey-weighted Turnbull estimator is still a useful comparator, as it is a nonparametric estimator that estimates rates more robustly than a parametric estimator.

4.4 Results

Survey-weighted survival curves from the parametric models, the Turnbull estimator, and the discrete hazards model were fitted to the Malawi data. The resulting curves for Lilongwe–the district that contains the capital city of Malawi–are displayed in Figure 4.3, and the curves for the remaining districts are in Appendix C. We note that the Gompertz and Weibull distributions appear furthest from the Turnbull estimator in nearly every time period, with the lognormal distribution appearing the closest of the parametric distributions. The discrete hazards model appears to follow the Turnbull estimator closely as well, notably in 2008 when the lognormal survival curve estimates a lower U5MR than both the Turnbull and discrete hazards estimator. Similar patterns hold true for the remaining 26 districts. Based on these visual comparisons, and keeping in mind that the survey-weighted Turnbull estimator is imperfect due to the assumption we made about individuals interval-censored across time period boundaries, we chose the lognormal distribution as the best fitting parametric model.

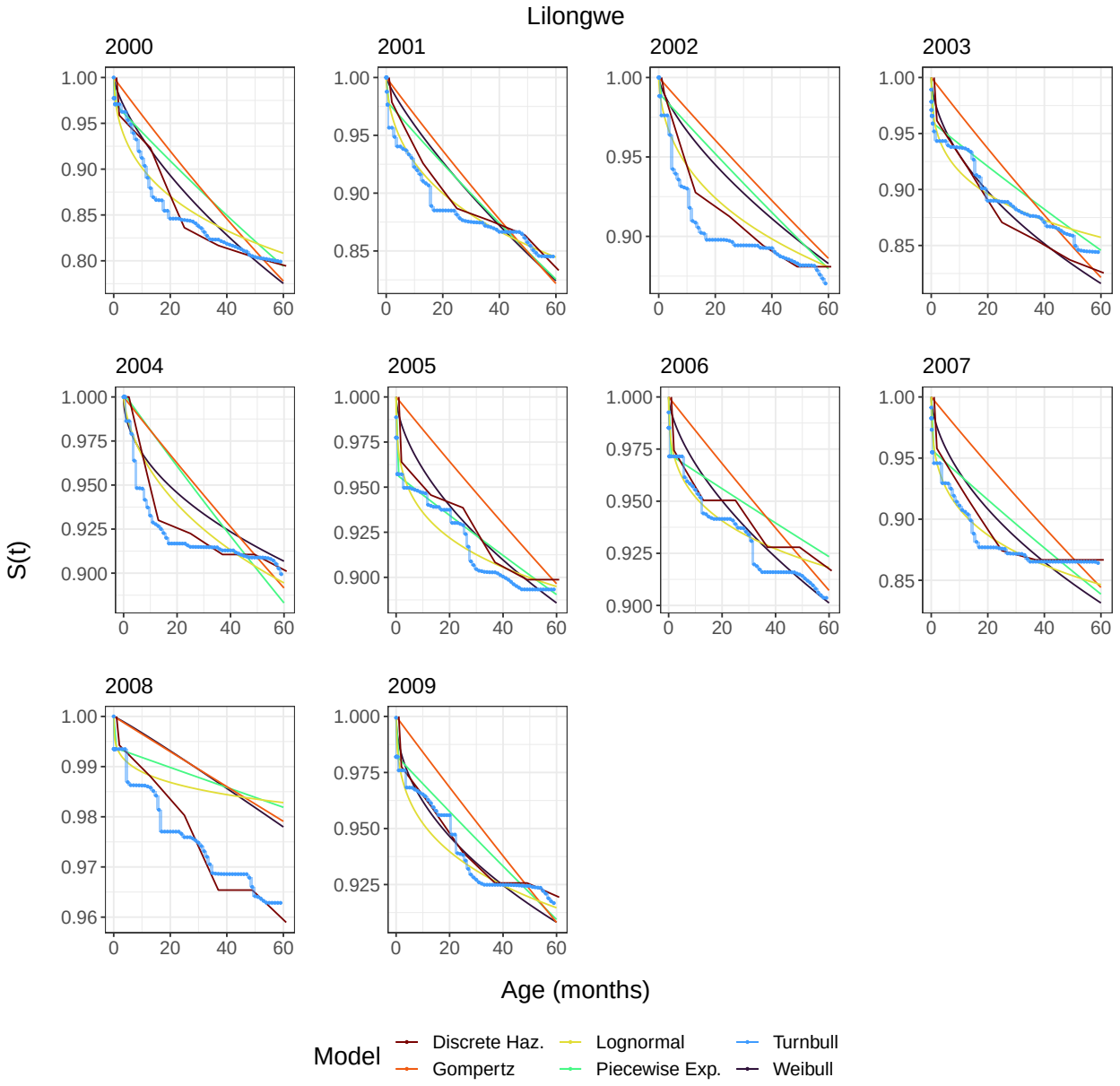


Figure 4.3: Survey-weighted survival curves for the Lilongwe district of Malawi from 2000-2009.

In Figures 4.4, 4.5, 4.6, and 4.7, we map the weighted (transformed) estimates from the lognormal model across space and time. There is no obvious spatial pattern to the estimates of $\log(1/\mu)$ and $\log(\sigma)$ across space, though $\text{logit}({}_1q_0)$ and $\text{logit}({}_{59}q_1)$ may be slightly higher

in more northern districts.

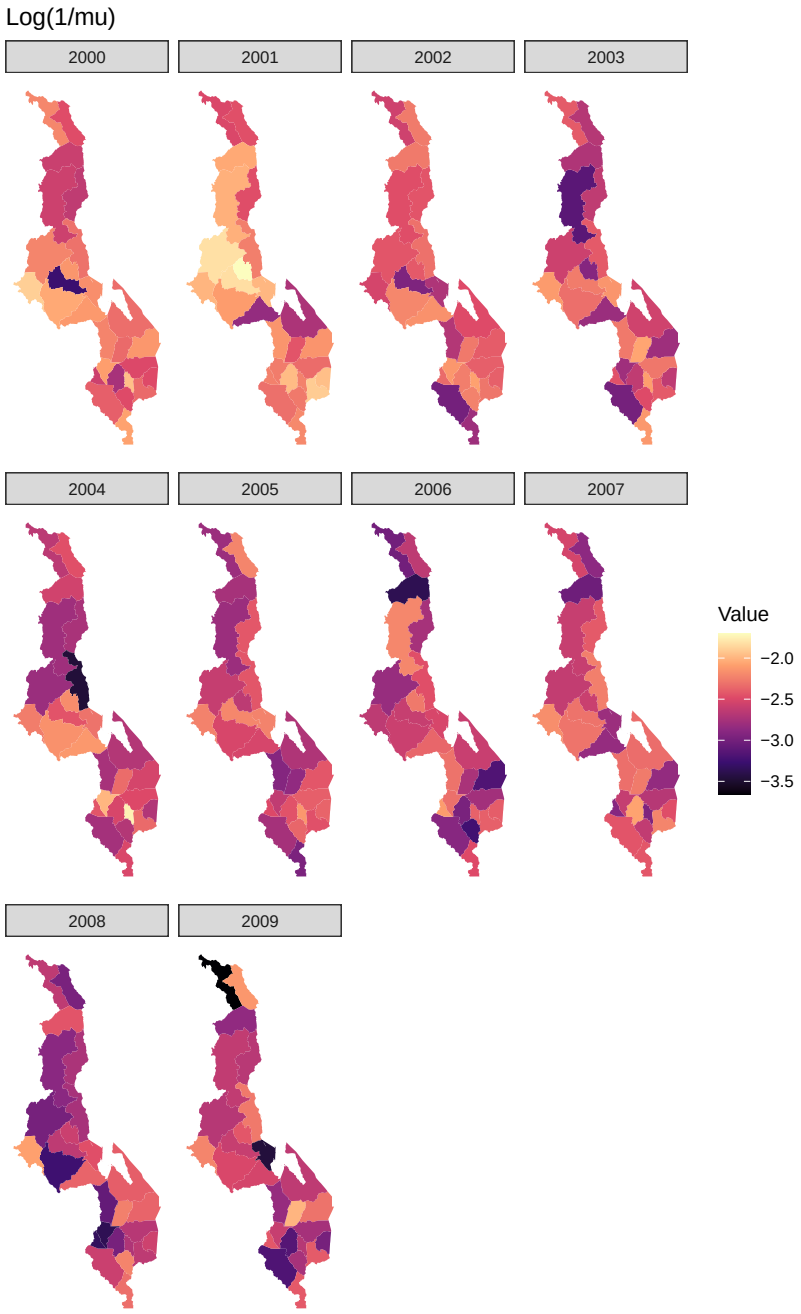


Figure 4.4: Weighted estimates of $\log(1/\mu)$ across space and time.

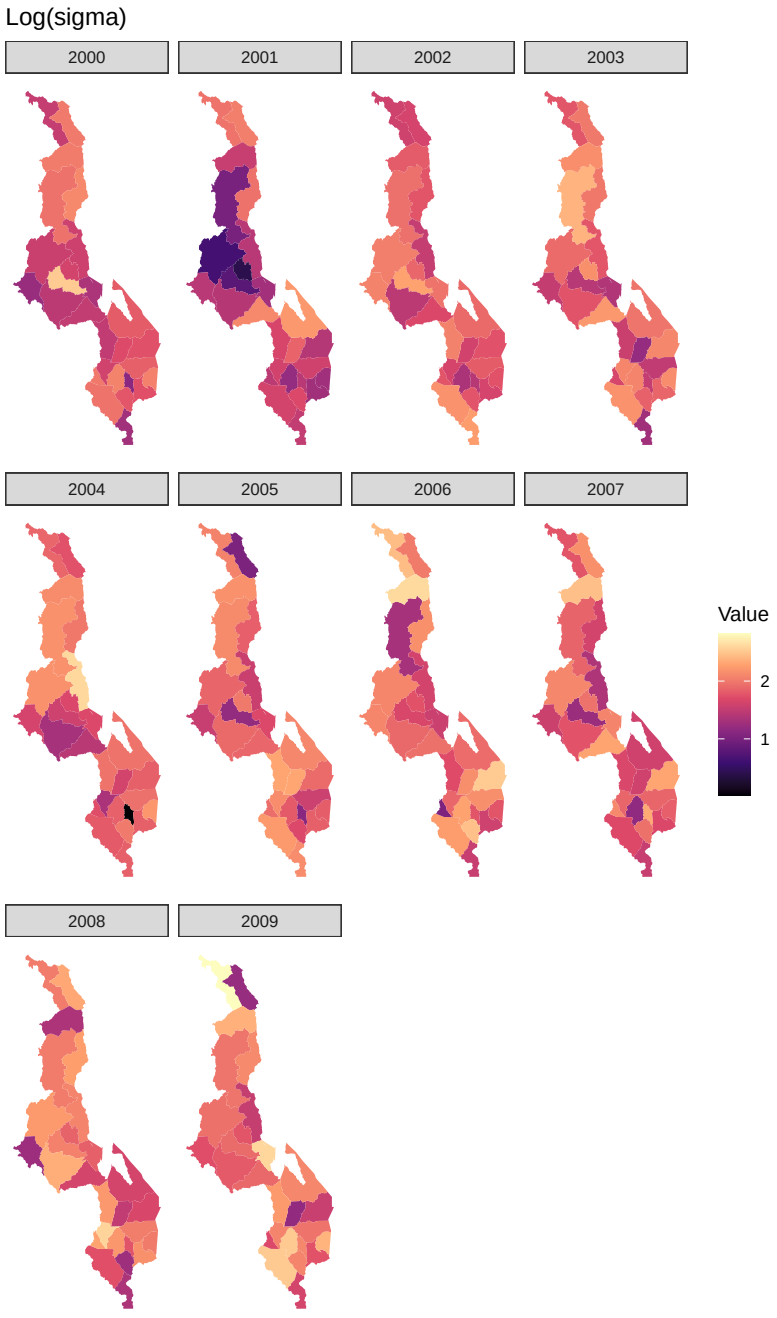


Figure 4.5: Weighted estimates of $\log(\sigma)$ across space and time.

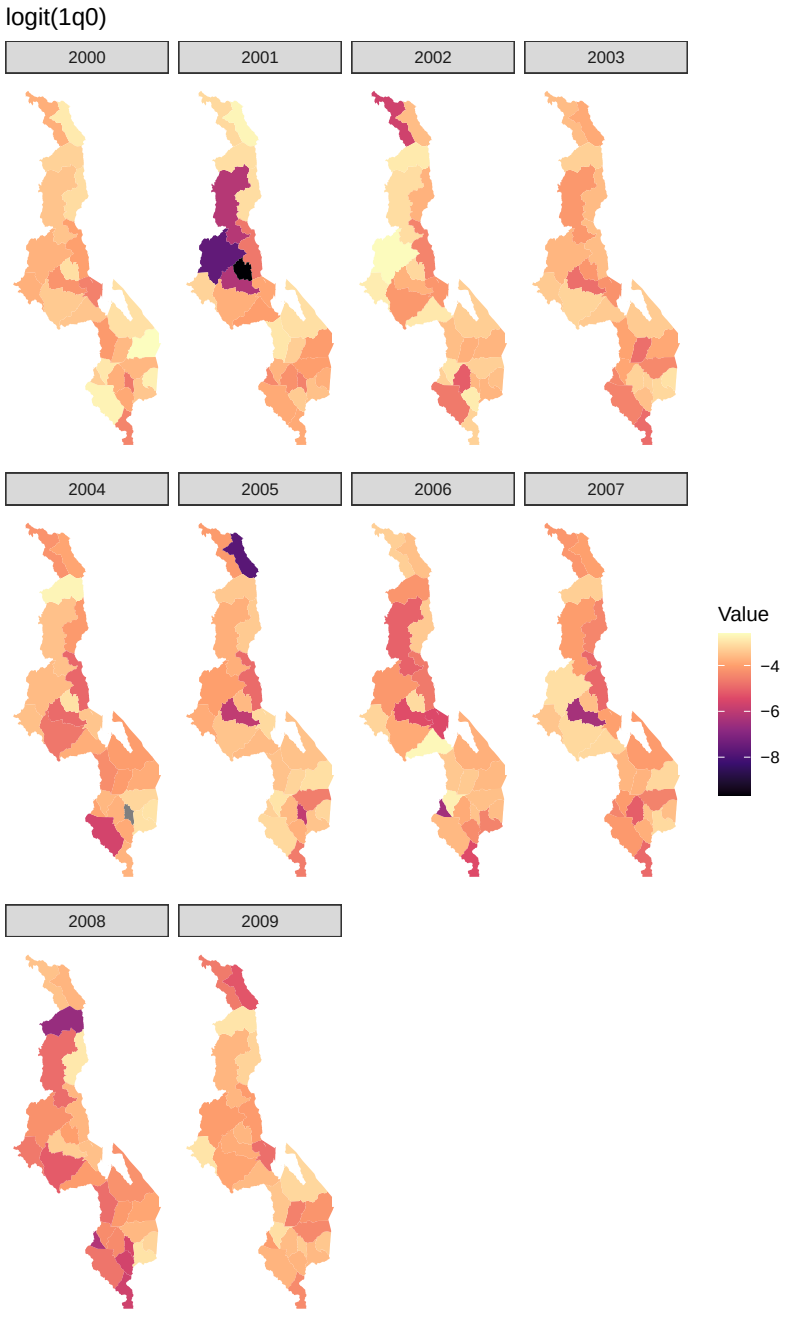


Figure 4.6: Weighted estimates of logit($_1q_0$) across space and time.

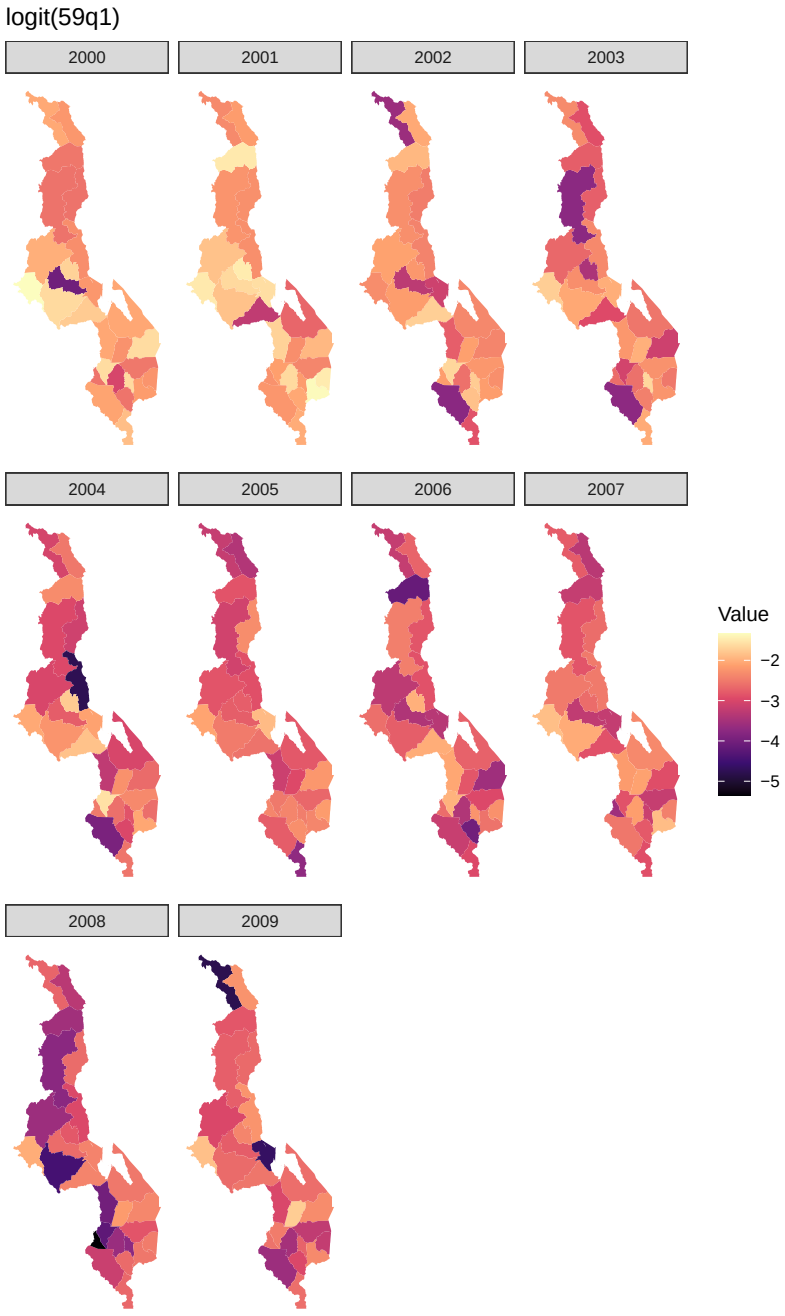


Figure 4.7: Weighted estimates of logit$({}_{59}q_1)$ across space and time.

Below we display the results from the different smoothing models for the Lilongwe region of Malawi, with additional results in Appendix C. In Figure 4.8 we can see that both corre-

lated and uncorrealted multivariate random walk models are smoothing the direct estimates for both the untransformed shape and scale parameter, and the transformed weighted ${}_1q_0$ and ${}_{59}q_1$. Both smoothing models for the transformed parameters decrease the variance of the resulting estimates compared to the direct estimates. For the untransformed scale and shape parameter, however, the variance is increased slightly (see Top of Figure 4.8). This is likely because the direct estimates are very certain and also very non-smooth; enforcing smoothness in this case leads to increased variance compared to the raw, weighted estimates due to lack of fit.

For the shape and scale parameters, we note that the correlated multivariate random walk model is slightly less smooth than the uncorrelated version. This is likely due to the greater flexibility in the correlated model; the additional parameter in the random walk, ρ, allows the model to more closely fit the data and therefore supports less smooth posterior estimates. If the correlation structure of the empirical data is very different from the assumed structure in the correlated multivariate random walk model, we expect to estimate $\hat{\rho} = 0$, and therefore the correlated and uncorrelated smoothing approaches should provide similar posterior estimates. If $\hat{\rho}$ is far from zero, we expect the correlated model to be more dissimilar from the uncorrelated model.

We generally observe that the correlated and uncorrelated smoothing models provide similar smoothness when the posterior density for the correlation parameter in the multivariate random walk is concentrated around -1 or 1 and the temporal patterns are similar, or when the correlation parameter for the multivariate random walk is concentrated around zero. Table 4.2 provides the posterior means for the correlation parameters for the correlated multivariate random walk models that we fit, with the posterior mean for Lilongwe being 0.02 for the model smoothing transformed weighted estimates (${}_1q_0$, ${}_{59}q_1$), and 0.01 for the model smoothing untransformed weighted estimates ($\log(1/\mu)$, $\log(\sigma)$).

We emphasize that the posterior estimates for the correlation parameter should be interpreted with caution, particularly since there is little data available in these models to estimate three hyperparameters for the MVRW (1 observation per parameter per time point). In sce-

narios where more data are available, reasonable values of ρ may be estimated. We provide a simulation study to compare the performance of the two MVRW models in a scenario with more data in Appendix C.3.

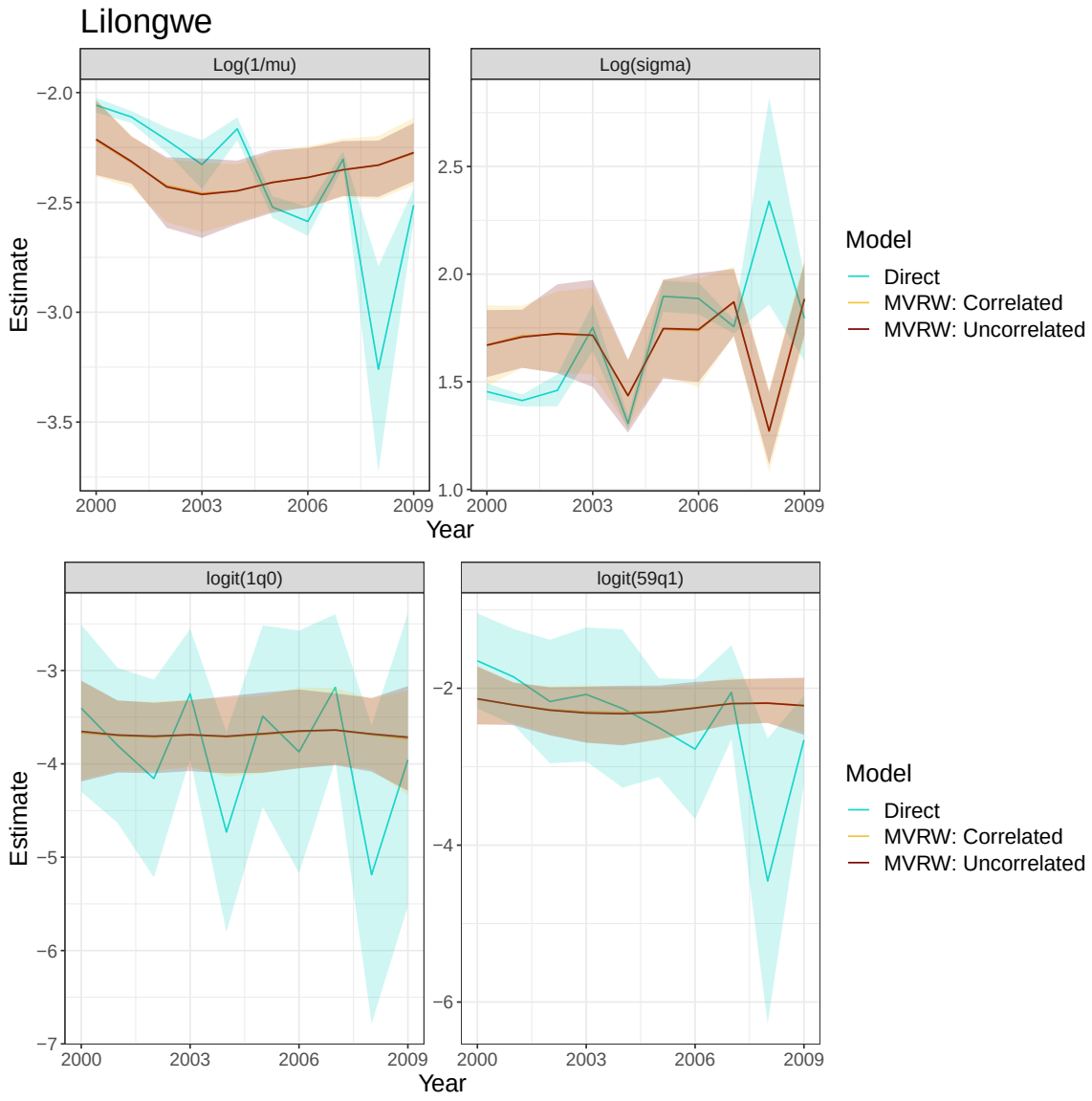


Figure 4.8: Smoothed lognormal parameters (Top) and smoothed, transformed lognormal parameters (Bottom) for the Lilongwe district of Malawi from 2000-2009.

District	Transformed	Untransformed
Balaka	-0.07	-0.49
Blantyre	-0.10	-0.76
Chikwawa	0.20	-0.30
Chiradzulu	0.03	-0.53
Chitipa	0.02	-0.15
Dedza	0.02	-0.12
Dowa	0.03	-0.81
Karonga	0.10	-0.10
Kasungu	-0.01	-0.74
Lilongwe	0.02	0.01
Machinga	0.13	-0.40
Mangochi	0.05	-0.23
Mchinji	-0.03	-0.22
Mulanje	-0.02	-0.17
Mwanza	0.02	-0.61
Mzimba	0.00	-0.45
Neno	0.03	-0.22
Nkhata Bay	0.00	0.06
Nkhotakota	0.04	-0.56
Nsanje	0.00	-0.66
Ntcheu	0.03	-0.12
Ntchisi	0.03	-0.45
Phalombe	-0.02	-0.44
Rumphi	0.07	0.02
Salima	0.09	0.06
Thyolo	0.06	-0.03
Zomba	0.05	0.07

Table 4.2: Posterior means of the correlation parameter ρ in the multivariate random walk models for the models using transformed ($\text{logit}({}_1q_0)$, $\text{logit}({}_{59}q_1)$) and untransformed ($\log(1/\mu), \log(\sigma)$) weighted estimates.

For each of the smoothing models considered, we obtain smoothed estimates of NMR, IMR, and U5MR. The results for Lilongwe are in Figure 4.9. Note that the parametric, smoothed weighted estimates of the NMR, IMR, and U5MR summaries appear the smoothest. This is to be expected, as the smoothing is taking place on the same scale as the

output. The multivariate random walk models on the transformed parameters are similarly smooth. For NMR, this is unsurprising as half of the weighted estimates that are smoothed are exactly the parametric, weighted estimates for NMR. For U5MR, the smoothness likely comes from the fact that the transformation needed to get from the smoothed parameters themselves to U5MR involves only simple multiplication ($_{60}q_0 = 1 - (1 - {}_1q_0)(1 - {}_{59}q_1)$); if the hazards are smooth, survival probabilities should also be smooth.

The estimates of NMR, IMR, and U5MR from the multivariate random walk models on the untransformed ($\log(1/\mu)$ and $\log(\sigma)$) parameters are notably less smooth than the other models. Intuitively, smoothing the parameters on one scale does not necessarily imply that some transformation of those parameters will also be smooth. The transformation matters for maintaining smoothness (such as with the transformation from $_1q_0$ and $_{59}q_1$ to U5MR), though smoothing on either a transformed or untransformed scale increases the precision of our estimates relative to the unsmoothed estimates; discrete hazards or parametric, weighted.

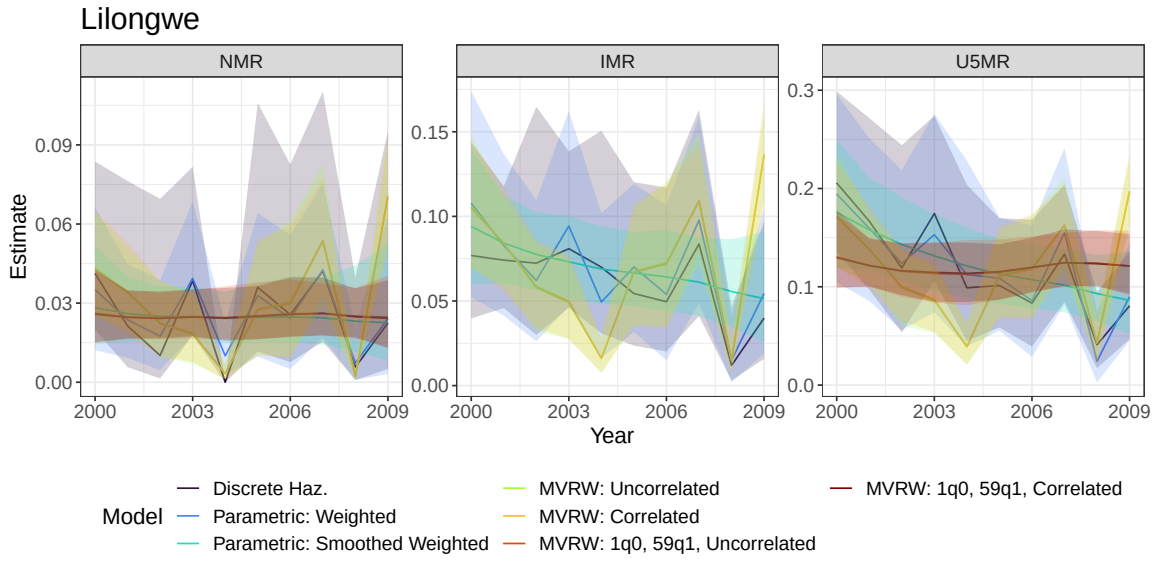


Figure 4.9: Estimates of NMR, IMR, U5MR from direct estimates and smoothing models across time for the Ntchisi district of Malawi from 2000-2009.

Estimates of U5MR across time for the districts of Malawi are shown in Figure 4.10 for

different smoothing models. Additionally, we plot differences in posterior standard deviations for U5MR from the smoothing approaches considered and the discrete hazards approach in Figure 4.11. Figures for NMR and IMR are given in Appendix C.

Note that the discrete hazards model and the parametric weighted model produce similar estimates (see Figure 4.12). As indicated in Figure 4.11, however, the standard deviations of estimates from the parametric, weighted model are larger in general than those from the discrete hazards model. The parametric, smoothed weighted model and uncorrelated multivariate random walk on transformed parameters produce the smoothest estimates of U5MR across time (by visual comparison), and both models see the smallest posterior standard deviations when compared to the discrete hazards model.

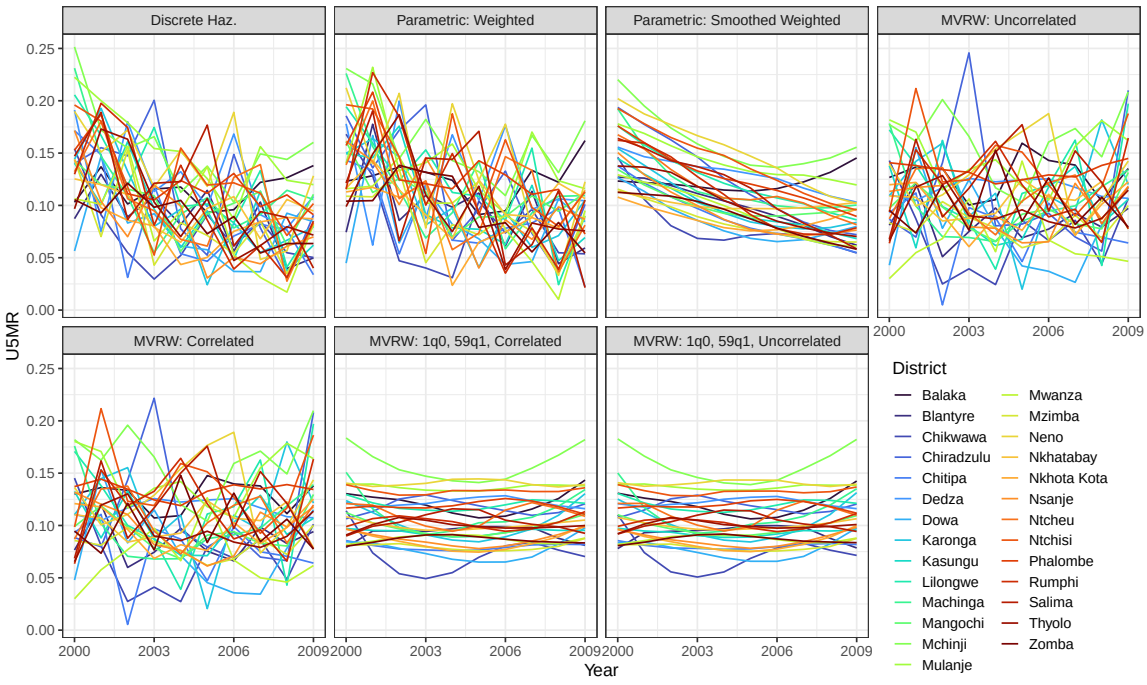


Figure 4.10: Posterior mean estimates of U5MR from 2000-2009 for 27 districts of Malawi from different smoothing models considered.

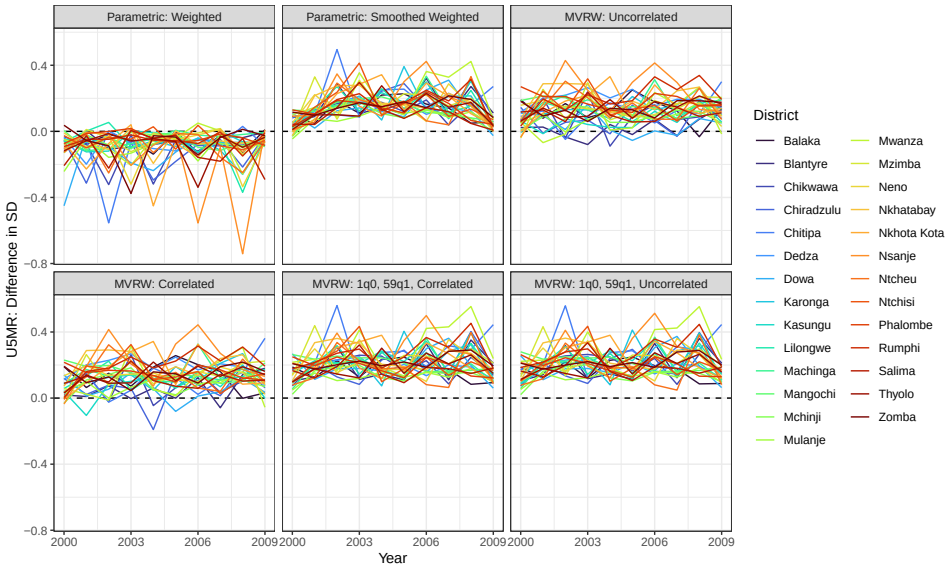


Figure 4.11: Differences in posterior standard deviation for U5MR from the indicated smoothing models and the discrete hazards approach for 2000-2009 for 27 districts of Malawi. A positive difference indicates that the standard deviation from the discrete hazards estimate is higher than the corresponding smoothing model.

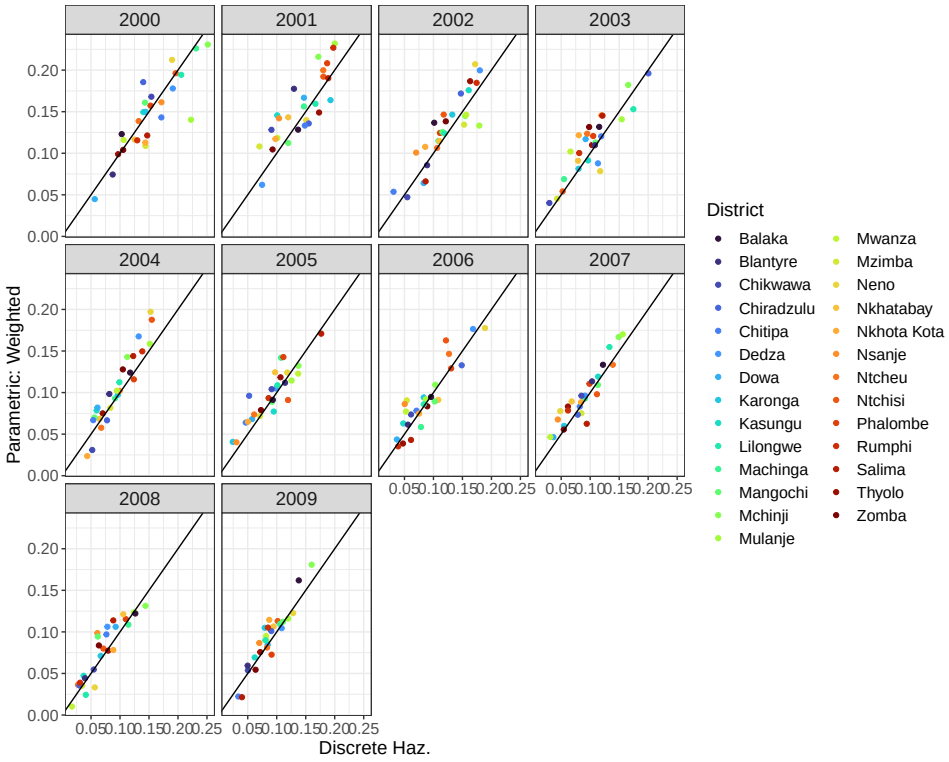


Figure 4.12: Comparison of estimates of U5MR from the discrete hazards model to estimates from the parametric, weighted model across space and time.

Though the differences in standard deviations are small, we do see the smallest posterior standard deviations relative to the discrete hazards model from the uncorrelated multivariate random walk model on transformed parameters. In Figure 4.13, we compare the standard deviations from the two models (parametric, smoothed weighted, and uncorrelated MVRW on transformed parameters) that produced the smoothest estimates and largest precision gains, and note that the proposed uncorrelated MVRW gives the smallest posterior standard devation compared to the parametric, smoothed weighted approach in over 50% of districts across all years for estimates of both NMR and U5MR.

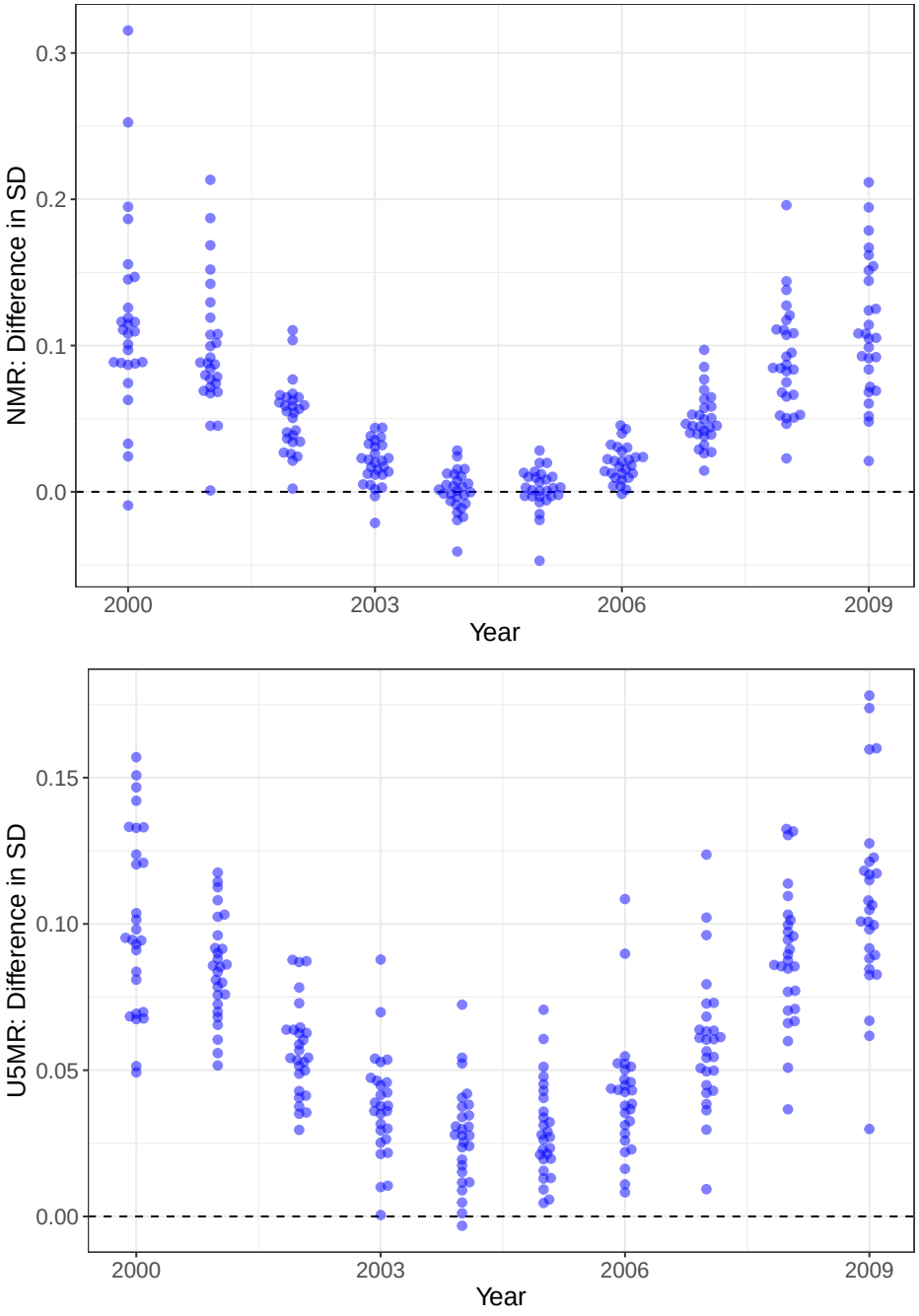


Figure 4.13: Differences in standard deviation between estimates from the parametric, smoothed weighted model and uncorrelated multivariate random walk on transformed parameters (Top: NMR, Bottom: U5MR). A positive difference indicates that the standard deviation from the parametric, smoothed weighted is higher than the uncorrelated multivariate random walk on transformed parameters. Each beeswarm consists of 27 data points; one from each district.

4.5 Discussion

In this chapter, we proposed multivariate random walk models for smoothing weighted estimates from mortality models across time. Existing approaches as currently used by the UN IGME and DHS, smooth summaries of NMR, IMR, and U5MR across time. While this approach smooths estimates across time and produces smaller posterior standard deviations than the raw, weighted estimates, it also fails to enforce known relationships between NMR, IMR, and U5MR; namely, that they must be in increasing order. The multivariate random walk models we propose on weighted parameters (transformed and untransformed) from continuous, parametric models smooth estimates and produce lower posterior standard deviations, with the added benefit of enforcing known survival relationships.

We applied our methods to 27 districts of Malawi and compared the performance of proposed and existing smoothing models across the years 2000 to 2009. All smoothing models had increased precision compared to the discrete hazards approach, with the uncorrelated multivariate random walk on transformed parameters having the highest gains in precision among the models considered.

The only transformation of parameters we considered in this chapter was from the weighted estimates of mean and variance parameters from the lognormal model (log scale) to logit estimates of ${}_1q_0$ and ${}_{59}q_1$. We chose this transformation specifically to ensure that (1) estimates that were smoothed could lie on the whole of the real line, and (2) estimates could be easily combined to produce inference on NMR and U5MR. Other transformations could be considered (computing the NMR, IMR, and U5MR from the parameters used to specify the model is itself a transformation), though not all transformations will guarantee NMR $<$ IMR $<$ U5MR.

Future extensions of the model include extensions to more than 2 parameters. In particular, it may be of interest to smooth the six, discrete hazards from the Mercer et al. (2015) model across time, with multiple correlation parameters in the random effect. Additional precautions may need to be taken when extending this model beyond two dimensions to en-

sure identifiability of parameters. Riebler et al. (2012) note, for example, that the inclusion of more than one correlation parameter in this type of multivariate random effect may require the general Fisher's z-tranformation (Fisher (1958), page 219) on correlation parameters to ensure the precision matrix of the random effect is positive definite (modulo typical random walk constraints). We included extensions of the model to ICAR random effects in Appendix C, and further extensions could be made to space-time random effects. In particular, the four Knorr-Held type random effects distributions could be considered (Knorr-Held, 2000).

Ultimately, the scale on which parameters are smoothed affects (1) how "smooth" resulting estimates are, and (2) posterior standard deviations of resulting estimates. In our application, we noted that weighted estimates that had been transformed to a scale very close to the final, desired estimates (in this case, NMR and U5MR) produced the smoothest estimates across time. Additionally, the MVRW models without the additional correlation parameter had smaller posterior standard deviations than their correlated counterparts, relative to the estimates from the discrete hazards model. One thing to keep in mind is that while we used PC priors on precision hyperparameters in the MVRW, we used a Normal$(0, 1)$ prior on a transformation of the correlation parameter. An additional extension to this model would be the development of a PC prior for the correlation parameter, which would allow use of (more) informative priors, and may influence smoothed estimates in scenarios with little data. Additional simulation studies should be conducted to assess model performance in scenarios with little data. As with the pseudo-likelihood approach developed in Chapter 3, model validation and comparison procedures require further consideration.

Chapter 5
DISCUSSION AND FUTURE WORK

This dissertation has focused on developing statistical methods for estimating child mortality in an official statistics setting, with a focus on outcomes in LMICs. Many existing demographic methods for producing child mortality estimates have been developed in a high-income country setting, and LMICs have unique data challenges that make these existing methods less applicable. The methods in all chapters of this dissertation were applied to DHS survey data from Malawi (2015-16), South Africa (2016), Namibia (2013), Senegal (2010), and Burkina Faso (2010).

In Chapter 2 we addressed computationally efficient benchmarking of small area estimates to national estimates with uncertainty, as opposed to national estimates from a census. Building on the approach of Zhang and Bryant (2020), we develop two fully Bayesian benchmarking approaches that target posterior estimates conditional on a soft benchmarking constraint in a two-step fashion. This allows for increased flexibility with regards to software during the modeling stage with benchmarking done as a second step. The proposed benchmarking approaches are appropriate in an LMIC setting, where national estimates are typically model-based with uncertainty, and computational efficiency is important. A `R` package, `stbench` implementing the methods described in this chapter can be found at `https://github.com/taylorokonek/stbench`. A paper is available as Okonek and Wakefield (2023).

In Chapter 3 we developed a method for producing full survival curves for children under the age of 5 that accounts for complex survey designs and period estimates of mortality. We reframed the demographic concepts of cohort and period estimates in a survival analysis framework, noting that period estimates take advantage of discrete time as a time-varying

covariate. We then proposed a pseudo-likelihood approach to obtaining period-specific estimates of full survival curves for children under the age of 5. The approach produces design-consistent estimators, and may be more appropriate for LMICs than existing demographic approaches to obtaining full survival curves for children under the age of 5. A `R` package, `pssst` implementing the methods can be found at `https://github.com/taylorokonek/pssst`.

Finally, in Chapter 4 we proposed a smoothing model that allows us to obtain precise estimates of mortality at a subnational level, while maintaining known demographic relationships between NMR, IMR, and U5MR. This work extended the methods developed in Chapter 3 to a smoothed direct framework, where survey-weighted estimates with high uncertainty may be smoothed using mixed effects models to obtain more precise estimates. The approach we proposed was a correlated, multivariate, first-order random walk on two parameters, which allows us to smooth any of the direct estimates from the two-parameter models considered in Chapter 3. We additionally provided extensions for how this framework can be extended to correlated, multivariate ICAR models and second-order random walks.

Future directions for research include extending methods to further subnational disaggregations and model validation approaches. A limitation of the benchmarking approach we propose in Chapter 2 is that uncertainty of the resulting benchmarked estimates will be misstated if national benchmarks and subnational estimates are produced from the same data. This setting, known as internal benchmarking, has yet to be satisfactorily addressed in the fully Bayesian benchmarking literature. Additionally, it is potentially of interest to benchmark estimates at multiple levels of aggregation (i.e., admin2 to admin1, admin1 to national). Simiarly, the pseudolikelihood approach proposed in Chapter 3 could be extended to spatial disaggregations. One could then consider correlated, spatio-temporal random effects as an extension of the correlated, multivariate random walk and ICAR models developed in Chapter 4. Software for these correlated MVGMRFs needs to be developed.

Model validation approaches can be improved upon from what was used in Chapter 3. The pseudolikelihood approach for parameter distributions was compared to the NPMLE provided by the Turnbull estimator, though the Turnbull estimator does not have a readily

derived variance estimator. Providing a justified variance estimator for the survey-weighted Turnbull would allow us to better determine parametric model fit for the methods developed in Chapter 3.

BIBLIOGRAPHY

Alexander, M., Zagheni, E., and Barbieri, M. (2017). A flexible Bayesian model for estimating subnational mortality. *Demography*, 54(6):2025–2041.

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

Allison, P. D. (2014). *Event history and survival analysis: Regression for longitudinal event data*, volume 46. SAGE publications.

Bachl, F. E., Lindgren, F., Borchers, D. L., and Illian, J. B. (2019). inlabru: an R package for Bayesian spatial modelling from ecological survey data. *Methods in Ecology and Evolution*, 10(6):760–766.

Barbieri, M., Wilmoth, J. R., Shkolnikov, V. M., Glei, D., Jasilionis, D., Jdanov, D., Boe, C., Riffe, T., Grigoriev, P., and Winant, C. (2015). Data resource profile: the Human Mortality Database (HMD). *International journal of epidemiology*, 44(5):1549–1556.

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83(401):28–36.

Bell, W. R., Datta, G. S., and Ghosh, M. (2013). Benchmarking small area estimators. *Biometrika*, 100(1):189–202.

Berg, E. and Fuller, W. A. (2018). Benchmarked small area prediction. *Canadian Journal of Statistics*, 46(3):482–500.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225.

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.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review/Revue Internationale de Statistique*, pages 279–292.

Binder, D. A. (1992). Fitting Cox's proportional hazards models from survey data. *Biometrika*, 79(1):139–147.

Breslow, N. E. and Wellner, J. A. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to cox regression. *Scandinavian Journal of Statistics*, 34(1):86–102.

Breslow, N. E. and Wellner, J. A. (2008). A Z-theorem with estimated nuisance parameters and correction note for 'Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression'. *Scandinavian Journal of Statistics*, 35(1):186–192.

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).

Carstensen, B. (2007). Age–period–cohort models for the lexis diagram. *Statistics in Medicine*, 26(15):3018–3045.

Chen, L., Nandram, B., and Cruze, N. B. (2022). Hierarchical Bayesian Model with Inequality Constraints for US County Estimates. *Journal of Official Statistics (JOS)*, 38(3).

Clark, S. J. (2015). A singular value decomposition-based factorization and parsimonious component model of demographic quantities correlated by age: Predicting complete demographic age schedules with few parameters. *arXiv preprint arXiv:1504.02057*.

Clark, S. J. (2019). A general age-specific mortality model with an example indexed by child mortality or both child and adult mortality. *Demography*, 56(3):1131–1159.

Dagum, E. B. and Cholette, P. A. (2006). Benchmarking, temporal distribution, and reconciliation methods for time series.

Database of Global Administrative Areas (GADM) (2019). Global administrative areas [shapefiles]. Available at: `https://www.gadm.org`. Downloaded January 2020.

Datta, G., Ghosh, M., Steorts, R., and Maples, J. (2011). Bayesian benchmarking with applications to small area estimation. *TEST*, 20(3):574–588.

Datta, G. S. (2009). Model-based approach to small area estimation. *Handbook of statistics*, 29:251–288.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22.

Dong, T. Q. and Wakefield, J. (2021). Modeling and presentation of vaccination coverage estimates using data from household surveys. *Vaccine*, 39(18):2584–2594.

Eaton, J. W., Dwyer-Lindgren, L., Gutreuter, S., O'Driscoll, M., Stevens, O., Bajaj, S., Ashton, R., Hill, A., Russell, E., Esra, R., et al. (2021). Naomi: a new modelling tool for estimating HIV epidemic indicators at the district level in sub-Saharan Africa. *Journal of the International AIDS Society*, 24:e25788.

Eilerts, H., Prieto, J. R., Eaton, J. W., and Reniers, G. (2021). Age patterns of under-5 mortality in sub-saharan africa during 1990–2018: A comparison of estimates from

demographic surveillance with full birth histories and the historic record. *Demographic research*, 44:415.

Erciulescu, A. L., Cruze, N. B., and Nandram, B. (2018). Benchmarking a triplet of official estimates. *Environmental and Ecological Statistics*, 25(4):523–547.

Erciulescu, A. L., Cruze, N. B., and Nandram, B. (2019). Model-based county level crop estimates incorporating auxiliary sources of information. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(1):283–303.

Erciulescu, A. L., Cruze, N. B., and Nandram, B. (2020). Statistical challenges in combining survey and auxiliary data to produce official statistics. *Journal of official statistics*, 36(1):63–88.

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.

Fisher, R. A. (1958). *Statistical Methods for research workers*. Oliver and Boyd, 13th (rev.) ed edition.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). Bayesian data analysis.

Gelman, A., Gilks, W. R., and Roberts, G. O. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *The annals of applied probability*, 7(1):110–120.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472.

Ghosh, M., Kubokawa, T., and Kawakubo, Y. (2015). Benchmarked empirical Bayes methods in multiplicative area-level models with risk evaluation. *Biometrika*, 102(3):647–659.

Ghosh, M. and Steorts, R. C. (2013). Two-stage benchmarking as applied to small area estimation. *TEST*, 22(4):670–687.

Groeneboom, P. and Wellner, J. A. (1992). *Information bounds and nonparametric maximum likelihood estimation*, volume 19. Springer Science & Business Media.

Guillot, M., Romero Prieto, J., Verhulst, A., and Gerland, P. (2022). Modeling age patterns of under-5 mortality: Results from a log-quadratic model applied to high-quality vital registration data. *Demography*, 59(1):321–347.

Hájek, J. (1971). Discussion of 'An essay on the logical foundations of survey sampling, Part I', by D. Basu. *Foundations of statistical inference*, page 326.

Heligman, L. and Pollard, J. H. (1980). The age pattern of mortality. *Journal of the Institute of Actuaries*, 107(1):49–80.

Hill, K. (1995). Age patterns of child mortality in the developing world. *Population bulletin of the United Nations*, (39):112–132.

Hillmer, S. C. and Trabelsi, A. (1987). Benchmarking of economic time series. *Journal of the American Statistical Association*, 82(400):1064–1071.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685.

Inter-agency Group for Child Mortality Estimation (2020). Levels and Trends in Child Mortality: Estimates. Available at `https://childmortality.org`.

Janicki, R. and Vesper, A. (2017). Benchmarking techniques for reconciling Bayesian small area models at distinct geographic levels. *Statistical Methods and Applications*, 26(4):557–581.

Johnson, L. F., Dorrington, R. E., and Moolla, H. (2017a). Progress towards the 2020 targets for HIV diagnosis and antiretroviral treatment in South Africa. *Southern African Journal of HIV Medicine*, 18(1):1–8.

Johnson, L. F., May, M. T., Dorrington, R. E., Cornell, M., Boulle, A., Egger, M., and Davies, M.-A. (2017b). Estimating the impact of antiretroviral treatment on adult mortality trends in South Africa: A mathematical modelling study. *PLoS Medicine*, 14(12):e1002468.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

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

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

Lawn, J. E., Osrin, D., Adler, A., and Cousens, S. (2008). Four million neonatal deaths: counting and attribution of cause of death. *Paediatric and perinatal epidemiology*, 22(5):410–416.

Leasure, D. R., Jochem, W. C., Weber, E. M., Seaman, V., and Tatem, A. J. (2020). National population mapping from sparse survey data: A hierarchical Bayesian modeling framework to account for uncertainty. *Proceedings of the National Academy of Sciences*, 117(39):24173–24179.

Lehtonen, R. and Veijanen, A. (2009). Design-based methods of estimation for domains and small areas. In *Handbook of statistics*, volume 29, pages 219–249. Elsevier.

Li, Z., Hsiao, Y., Godwin, J., Martin, B. D., Wakefield, J., Clark, S. J., with support from the United Nations Inter-agency Group for Child Mortality Estimation, and its technical advisory group (2019). Changes in the spatial distribution of the under-five mortality

rate: Small-area analysis of 122 DHS surveys in 262 subregions of 35 countries in Africa. *PloS One*, 14(1):e0210645.

Li, Z. R., Martin, B. D., Dong, T. Q., Fuglstad, G.-A., Paige, J., Riebler, A., Clark, S., and Wakefield, J. (2020). Space-Time Smoothing of Demographic and Health Indicators using the R Package SUMMER. *arXiv:2007.05117*.

Lin, D. (2000). On fitting cox's proportional hazards models to survey data. *Biometrika*, 87(1):37–47.

Lindgren, F. and Rue, H. (2008). On the second-order random walk model for irregular locations. *Scandinavian journal of statistics*, 35(4):691–700.

Local Burden of Disease HIV Collaborators (2021). Mapping subnational HIV mortality in six Latin American countries with incomplete vital registration systems. *BMC Medicine*, 19:1–25.

Mahy, M., Marsh, K., Sabin, K., Wanyeki, I., Daher, J., and Ghys, P. D. (2019). HIV estimates through 2018: data for decision-making. *AIDS (London, England)*, 33(Suppl 3):S203.

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*, 33:9086–9097. Available at: `https://proceedings.neurips.cc/paper/2020/file/673de96b04fa3adcae1aacda704217ef-Paper.pdf` (accessed January 2022).

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

Nandram, B., Cruze, N. B., Erciulescu, A. L., Chen, L., et al. (2022). Bayesian Small Area Models under Inequality Constraints with Benchmarking and Double Shrinkage. Technical

report, United States Department of Agriculture, National Agricultural Statistics Service. Available at: https://www.nass.usda.gov/Education_and_Outreach/Reports,_Presentations_and_Conferences/reports/ResearchReport_constraintmodel.pdf (accessed January 2023).

Nandram, B., Erciulescu, A. L., and Cruze, N. B. (2019). Bayesian benchmarking of the Fay-Herriot model using random deletion. *Survey Methodology*, 45(2):365–391.

Nandram, B. and Sayit, H. (2011). A Bayesian analysis of small area probabilities under a constraint. *Survey Methodology*, 37(2):137–152.

Okonek, T. and Wakefield, J. (2023). A Computationally Efficient Approach to Fully Bayesian Benchmarking.

Osgood-Zimmerman, A., Millear, A. I., Stubbs, R. W., Shields, C., Pickering, B. V., Earl, L., Graetz, N., Kinyoki, D. K., Ray, S. E., Bhatt, S., et al. (2018). Mapping child growth failure in Africa between 2000 and 2015. *Nature*, 555(7694):41–47.

Osgood-Zimmerman, A. and Wakefield, J. (2021). A Statistical Introduction to Template Model Builder: A Flexible Tool for Spatial Modeling. *arXiv:2103.09929v1*.

Paige, J., Fuglstad, G.-A., Riebler, A., and Wakefield, J. (2022). Design-and model-based approaches to small-area estimation in a low-and middle-income country context: comparisons and recommendations. *Journal of Survey Statistics and Methodology*, 10(1):50–80.

Patra, S. (2019). *Constrained Bayesian Inference through Posterior Projection with Applications*. PhD thesis, Duke University.

Patra, S. and Dunson, D. B. (2018). Constrained Bayesian inference through posterior projections. *arXiv:1812.05741v3*.

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

Pfeffermann, D. and Tiller, R. (2006). Small-area estimation with state–space models subject to benchmark constraints. *Journal of the American Statistical Association*, 101(476):1387–1397.

Prentice, R. L. (1974). A log gamma model and its maximum likelihood estimation. *Biometrika*, 61(3):539–544.

Rao, J. N. and Molina, I. (2015). *Small Area Estimation*. John Wiley & Sons.

Rao, J. N. and Wu, C. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83(401):231–241.

Riebler, A., Held, L., and Rue, H. (2012). Estimation and extrapolation of time trends in registry data—borrowing strength from related populations. *The Annals of Applied Statistics*, pages 304–333.

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

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*. CRC press.

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

Särndal, C.-E., Swensson, B., and Wretman, J. (2003). *Model assisted survey sampling*. Springer Science & Business Media.

Schöley, J. (2019). The age-tra jec tory of infant mor tal ity in the united states: Parametric mod els and gen er a tive mech a nisms. In *annual meeting of the Population Association of America, Austin, TX*.

Si, Y. and Zhou, P. (2020). Bayes-Raking: Bayesian Finite Population Inference with Known Margins. *Journal of Survey Statistics and Methodology*, 9(4):833–855.

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

Smith, A. F. and Gelfand, A. E. (1992). Bayesian statistics without tears: a sampling–resampling perspective. *The American Statistician*, 46(2):84–88.

Stefan, M. and Hidiroglou, M. A. (2021). Small area benchmarked estimation under the basic unit level model when the sampling rates are non-negligible. *Survey Methodology*, 47(1):123–150.

Steorts, R. C., Schmid, T., and Tzavidis, N. (2020). Smoothing and Benchmarking for Small Area Estimation. *International Statistical Review*.

Stevens, F. R., Gaughan, A. E., Linard, C., and Tatem, A. J. (2015). Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PloS One*, 10(2):e0107042.

Stover, J., Glaubius, R., Mofenson, L., Dugdale, C. M., Davies, M.-A., Patten, G., and Yiannoutsos, C. (2019). Updates to the Spectrum/AIM model for estimating key HIV indicators at national and subnational levels. *AIDS (London, England)*, 33:S227.

Sun, J. (2001). Variance estimation of a survival function for interval-censored survival data. *Statistics in Medicine*, 20(8):1249–1257.

Tatem, A. J. (2017). WorldPop, open data for spatial demography. *Scientific data*, 4(1):1–4.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *The Annals of Statistics*, pages 1701–1728.

Trabelsi, A. and Hillmer, S. C. (1990). Bench-marking time series with reliable bench-marks. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 39(3):367–379.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 38(3):290–295.

United Nations General Assembly (2015). Transforming our world: the 2030 agenda for sustainable development. *United Nations: New York, NY, USA*.

USAID (2014). Demographic and health surveys. United States Agency for International Development. `http://www.dhsprogram.com`.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: an improved r for assessing convergence of mcmc (with discussion). *Bayesian analysis*, 16(2):667–718.

Verhulst, A., Prieto, J. R., Alam, N., Eilerts-Spinelli, H., Erchick, D. J., Gerland, P., Katz, J., Lankoande, B., Liu, L., Pison, G., et al. (2022). Divergent age patterns of under-5 mortality in south asia and sub-saharan africa: a modelling study. *The Lancet global health*, 10(11):e1566–e1574.

Wakefield, J. (2013). *Bayesian and frequentist regression methods*, volume 23. Springer.

Wakefield, J., Fuglstad, G.-A., Riebler, A., Godwin, J., Wilson, K., and Clark, S. J. (2019). Estimating under-five mortality in space and time in a developing world context. *Statistical Methods in Medical Research*, 28(9):2614–2634.

Wakefield, J., Okonek, T., and Pedersen, J. (2020). Small area estimation for disease prevalence mapping. *International Statistical Review*, 88(2):398–418.

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

Wang, J., Fuller, W. A., and Qu, Y. (2008). Small area estimation under a restriction. *Survey methodology*, 34(1):29. Available at: `https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2008001/article/10619-eng.pdf?st=TuZsIBto` (accessed January 2023).

Williams, M. and Berg, E. (2013). Incorporating user input into optimal constraining procedures for survey estimates. *Journal of Official Statistics*, 29(3):375.

Wilmoth, J., Zureick, S., Canudas-Romo, V., Inoue, M., and Sawyer, C. (2012). A flexible two-dimensional mortality model for use in indirect estimation. *Population studies*, 66(1):1–28.

Wu, Y., Li, Z. R., Mayala, B. K., Wang, H., Gao, P., Paige, J., Fuglstad, G.-A., Moe, C., Godwin, J., Donohue, R. E., Croft, T. N., and Wakefield, J. (2021). Spatial Modeling for Subnational Administrative Level 2 Small-Area Estimation. *DHS Spatial Analysis Reports*, 21.

You, Y. and Rao, J. (2002). A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. *Canadian Journal of Statistics*, 30(3):431–439.

You, Y., Rao, J., and Dick, J. (2002). Benchmarking hierarchical bayes small area estimators with application in census undercoverage estimation. In *Proceedings of the Survey Methods Section*, pages 86–90. Citeseer. Available at: `https://ssc.ca/sites/default/files/survey/documents/SSC2002_Y_You.pdf` (accessed January 2023).

Zhang, J. L. and Bryant, J. (2020). Fully Bayesian Benchmarking of Small Area Estimation Models. *Journal of Official Statistics*, 36(1):197–223.

Appendix A

APPENDIX FOR CHAPTER 2

A.1 HIV Application

A.1.1 Unit-level figures and tables

The subnational breakdown of unbenchmarked and benchmarked, unit-level models can be seen in Figure A.1 and Table A.1.

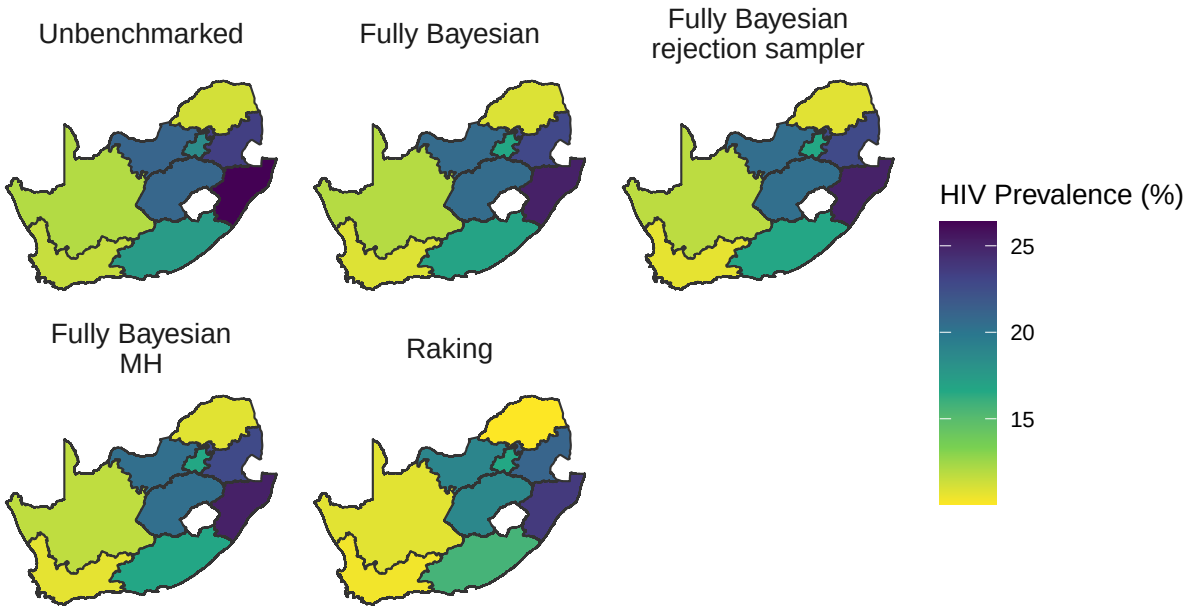


Figure A.1: Comparison of HIV prevalence estimates from benchmarked and unbenchmarked unit-level models.

A table represetation of the aggregated national estimates presented in Figure A.2 can be found in Table A.2.

Table A.1: Admin1 HIV prevalence estimates from unbenchmarked and benchmarked unit-level models. Point estimates provided are medians, with 95% credible intervals. Provinces are arranged in the order of lowest unbenchmarked median to highest.

Province	**Unbenched (%)**	**FB (%)**	**FB: Rejection sampler (%)**	**FB: MH (%)**	**Raking (%)**
Limpopo	11.2 (9.0, 13.9)	11.0 (8.8, 13.6)	10.8 (8.8, 13.4)	10.8 (8.7, 13.4)	10.1 (8.1, 12.5)
Western Cape	11.5 (8.6, 15.3)	10.9 (8.2, 14.4)	10.7 (8.0, 14.0)	10.7 (8.2, 14.0)	10.4 (7.7, 13.8)
Northern Cape	12.0 (9.2, 15.4)	12.0 (9.2, 15.6)	11.8 (9.0, 15.3)	11.7 (8.9, 15.4)	10.8 (8.3, 13.8)
Eastern Cape	17.5 (15.6, 19.4)	16.9 (15.3, 18.7)	16.8 (15.1, 18.5)	16.7 (15.1, 18.5)	15.7 (14.1, 17.4)
Gauteng	18.6 (15.3, 22.5)	16.7 (14.1, 19.6)	16.7 (14.1, 19.5)	16.7 (14.1, 19.4)	16.7 (13.7, 20.2)
Free State	21.0 (17.9, 24.2)	20.6 (17.7, 24.1)	20.5 (17.5, 23.7)	20.5 (17.7, 23.7)	18.8 (16.1, 21.8)
North West	21.1 (18.0, 24.5)	20.7 (17.8, 24.0)	20.5 (17.6, 23.8)	20.5 (17.6, 23.7)	18.9 (16.2, 22.0)
Mpumalanga	23.4 (20.2, 27.0)	22.9 (19.8, 26.3)	22.8 (19.7, 26.2)	22.7 (19.8, 26.1)	21.0 (18.2, 24.3)
KwaZulu-Natal	26.4 (23.1, 30.0)	25.1 (22.1, 28.3)	25.1 (22.2, 28.2)	25.1 (22.2, 28.1)	23.7 (20.7, 26.9)

Table A.2: Aggregated national level HIV prevalence estimates from Thembisa, unbenchmarked, and benchmarked results from the unit-level model. 95% credible intervals are given next to posterior medians.

Model	**Median (%)**	**SD (%)**
Thembisa	17.1 (15.6, 18.3)	0.61
Unbenched	19.1 (17.6, 20.6)	0.76
FB	18.0 (17.1, 19.0)	0.48
FB: Rejection sampler	17.9 (17.0, 18.9)	0.46
FB: MH	17.9 (17.0, 18.8)	0.47
Raking	17.1 (15.8, 18.5)	0.68

A.1.2 Area-level figures and tables

In Figure A.2, we display national level estimates from Thembisa, the unbenchmarked, and benchmarked results from the area-level model. Results from this model are very similar to results from the unit-level model, found in the Chapter 2.

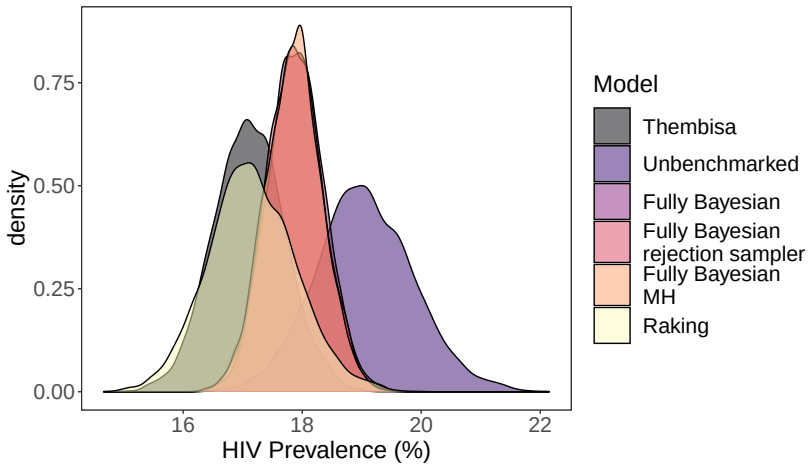


Figure A.2: Aggregated national level HIV prevalence estimates from Thembisa, unbenchmarked, and benchmarked results from the area-level model. All densities are based on 5000 samples

Table A.3: Aggregated national level HIV prevalence estimates from Thembisa, unbenchmarked, and benchmarked results from the area-level model. 95% credible intervals are given next to posterior medians.

Model	**Median (%)**	**SD (%)**
Thembisa	17.1 (15.6, 18.3)	0.61
Unbenched	19.0 (17.5, 20.7)	0.81
FB	17.9 (17.0, 18.8)	0.47
FB: Rejection sampler	17.9 (17.0, 18.8)	0.47
FB: MH	17.9 (16.9, 18.8)	0.47
Raking	17.1 (15.8, 18.7)	0.73

Table A.4: Admin1 HIV prevalence estimates from unbenchmarked and benchmarked area-level models. Point estimates provided are medians, with 95% credible intervals. Provinces are arranged in the order of lowest unbenchmarked median to highest.

Province	**Unbenched (%)**	**FB (%)**	**FB: Rejection sampler (%)**	**FB: MH (%)**	**Raking (%)**
Limpopo	10.4 (8.3, 13.1)	10.1 (8.0, 12.8)	10.2 (8.0, 12.8)	10.1 (8.0, 12.8)	9.4 (7.5, 11.8)
Northern Cape	12.5 (8.9, 17.4)	12.2 (8.6, 17.0)	12.3 (8.7, 17.1)	12.4 (8.8, 17.3)	11.3 (8.0, 15.7)
Western Cape	12.6 (8.2, 18.8)	11.1 (7.5, 16.4)	11.0 (7.4, 16.2)	11.1 (7.5, 16.0)	11.3 (7.4, 17.0)
Eastern Cape	16.8 (14.6, 19.4)	16.3 (14.3, 18.5)	16.3 (14.1, 18.7)	16.2 (14.2, 18.6)	15.1 (13.1, 17.4)
Gauteng	19.3 (15.9, 23.6)	17.5 (14.7, 20.6)	17.4 (14.7, 20.5)	17.3 (14.5, 20.3)	17.4 (14.3, 21.2)
North West	20.2 (17.5, 23.3)	19.9 (17.2, 22.8)	19.9 (17.2, 22.9)	20.0 (17.3, 22.9)	18.2 (15.7, 21.0)
Free State	21.0 (17.9, 24.4)	20.6 (17.7, 24.1)	20.7 (17.6, 24.0)	20.7 (17.6, 24.0)	18.9 (16.1, 21.9)
Mpumalanga	23.4 (20.2, 27)	22.9 (19.7, 26.4)	22.9 (19.6, 26.4)	22.8 (19.6, 26.3)	21.1 (18.2, 24.3)
KwaZulu-Natal	25.5 (21.7, 29.4)	24.0 (20.7, 27.6)	24.0 (20.7, 27.4)	24.0 (20.6, 27.5)	22.9 (19.5, 26.5)

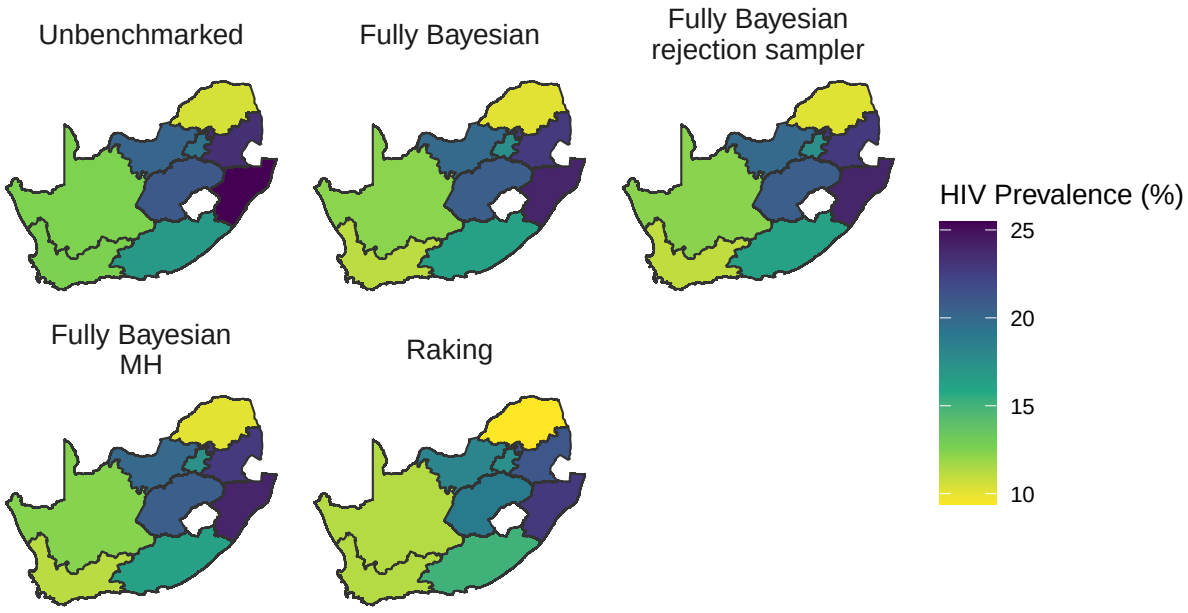


Figure A.3: Comparison of HIV prevalence estimates from benchmarked and unbenchmarked area-level models.

A.1.3 Area-level and unit-level comparison plots

In Figure A.4 we display the densities for the national aggregated HIV prevalence estimates from unbenchmarked and benchmarked models, comparing across unit-level and area-level models. As noted in the main text, the differences between unit-level and area-level estimates are very small, although there are relatively large differences between the unbenchmarked and benchmarked models themselves.

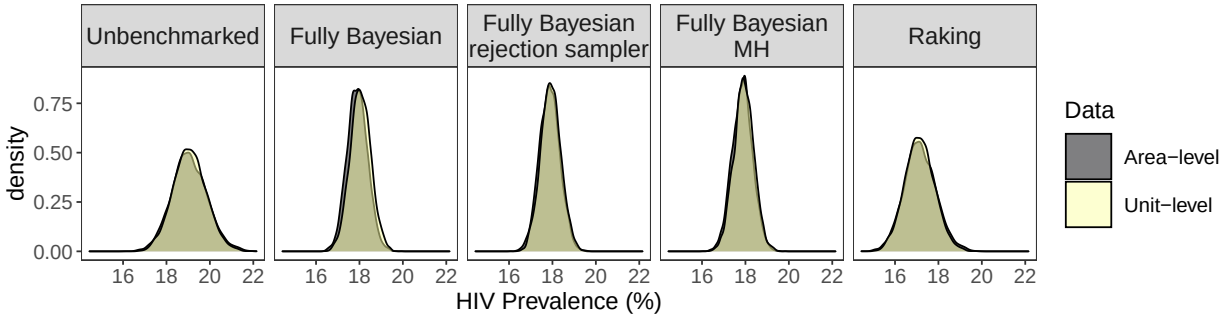


Figure A.4: Aggregated national level HIV prevalence estimates from unbenchmarked and benchmarked models. All densities are based on 5000 samples.

In Figures A.5 through A.13, we show the differences between the unit-level and area-level, benchmarked and unbenchmarked sub-national estimates for each of the nine provinces. For most provinces, the differences between area-level and unit-level models are small, though we note that estimates from the unit-level models may have higher precision.

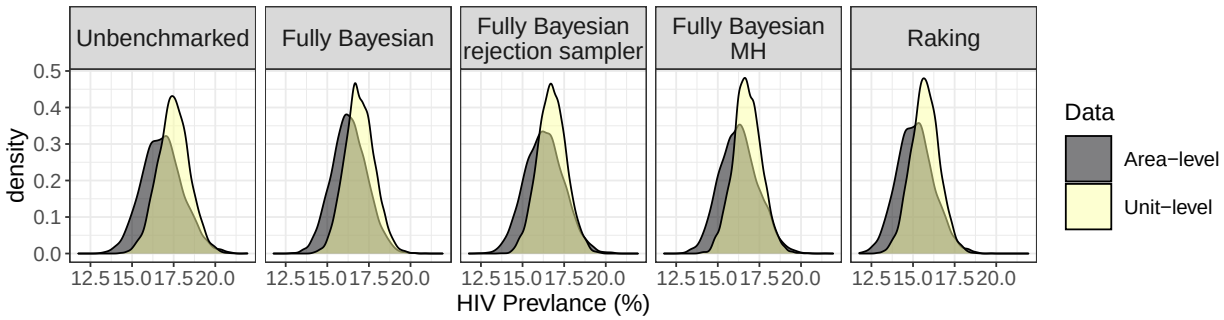


Figure A.5: Subnational HIV prevalence estimates for the Eastern Cape province from unbenchmarked and benchmarked models. All densities are based on 5000 samples.

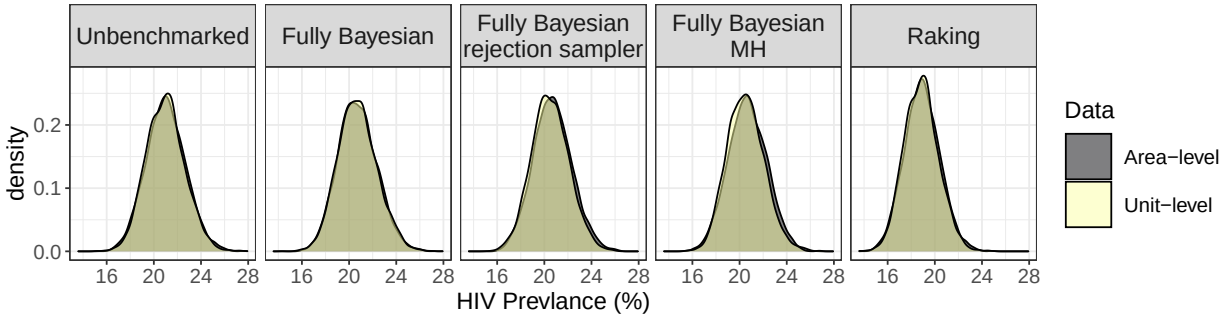


Figure A.6: Subnational HIV prevalence estimates for the Free State province from unbenchmarked and benchmarked models. All densities are based on 5000 samples.

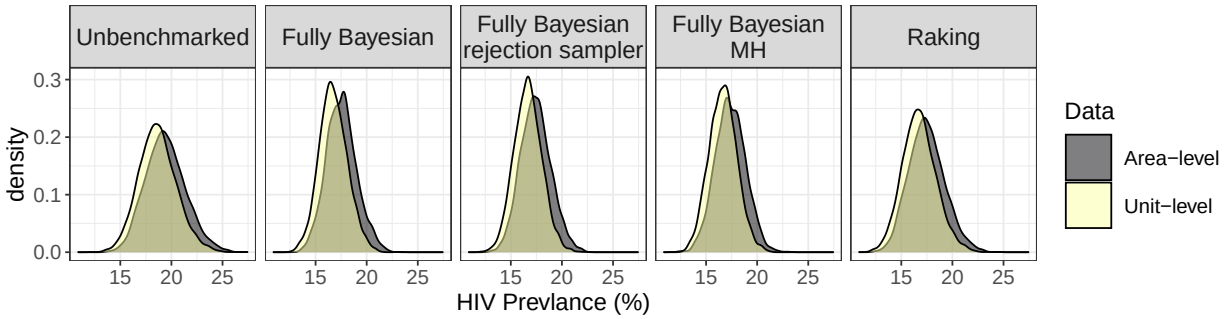


Figure A.7: Subnational HIV prevalence estimates for the Gauteng province from unbenchmarked and benchmarked models. All densities are based on 5000 samples.

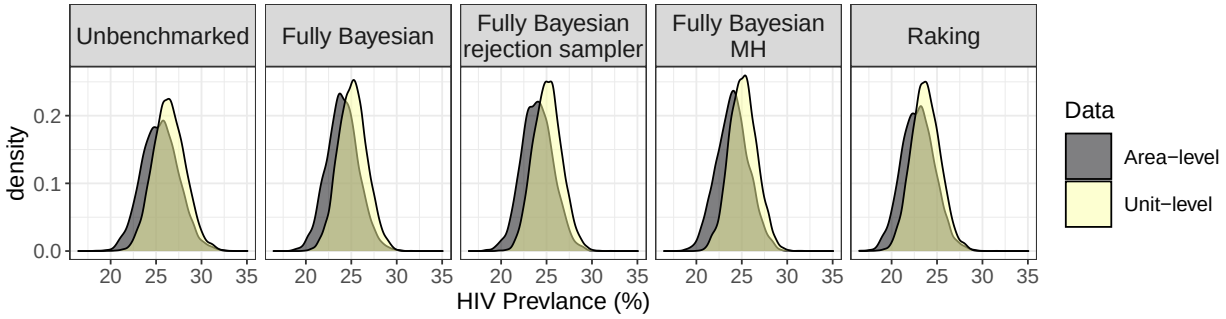


Figure A.8: Subnational HIV prevalence estimates for the KwaZulu-Natal province from unbenchmarked and benchmarked models. All densities are based on 5000 samples.

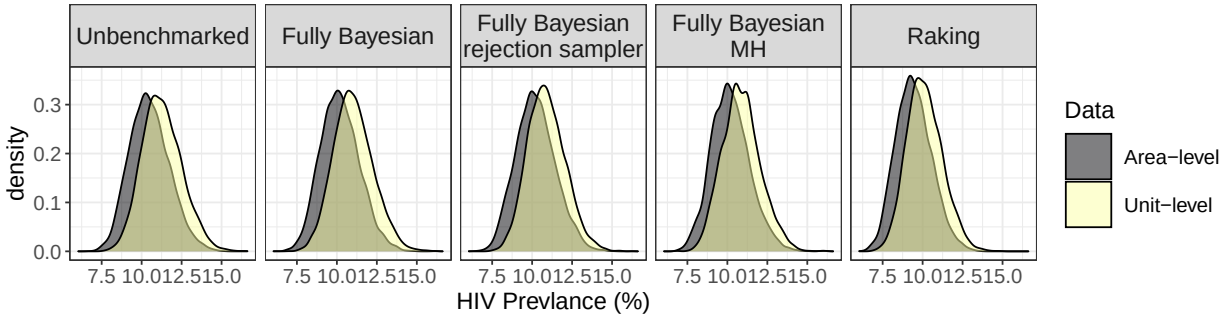


Figure A.9: Subnational HIV prevalence estimates for the Limpopo province from unbenchmarked and benchmarked models. All densities are based on 5000 samples.

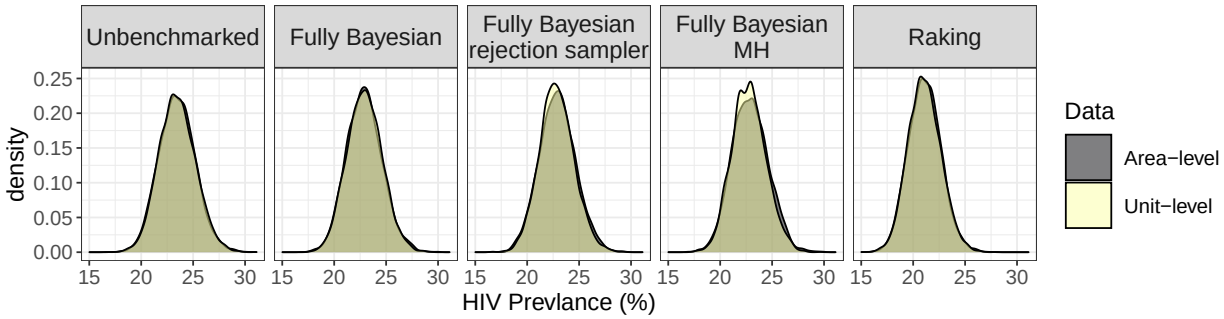


Figure A.10: Subnational HIV prevalence estimates for the Mpumalanga province from unbenchmarked and benchmarked models. All densities are based on 5000 samples.

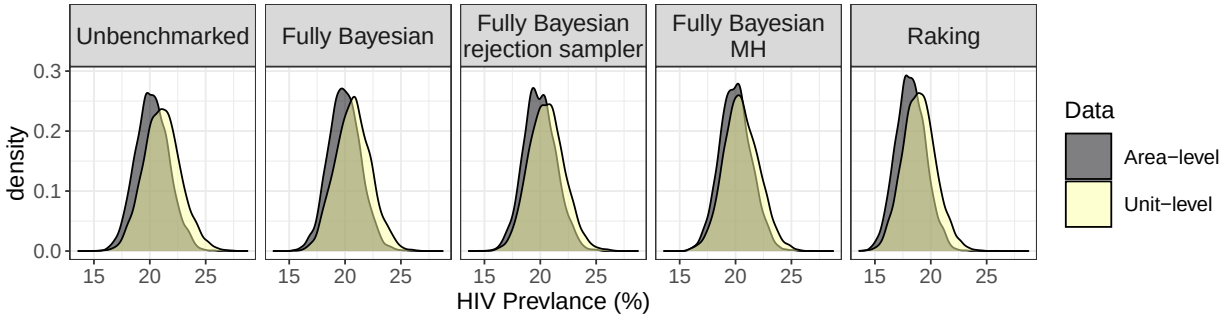


Figure A.11: Subnational HIV prevalence estimates for the North West province from unbenchmarked and benchmarked models. All densities are based on 5000 samples.

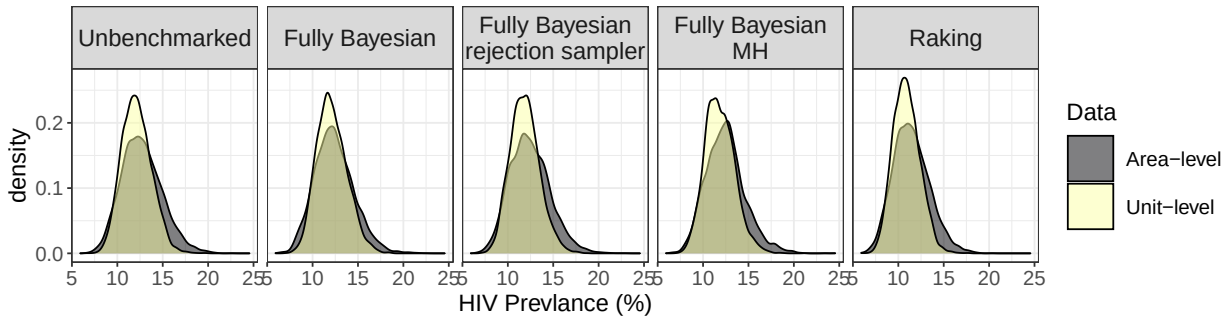


Figure A.12: Subnational HIV prevalence estimates for the Northern Cape province from unbenchmarked and benchmarked models. All densities are based on 5000 samples.

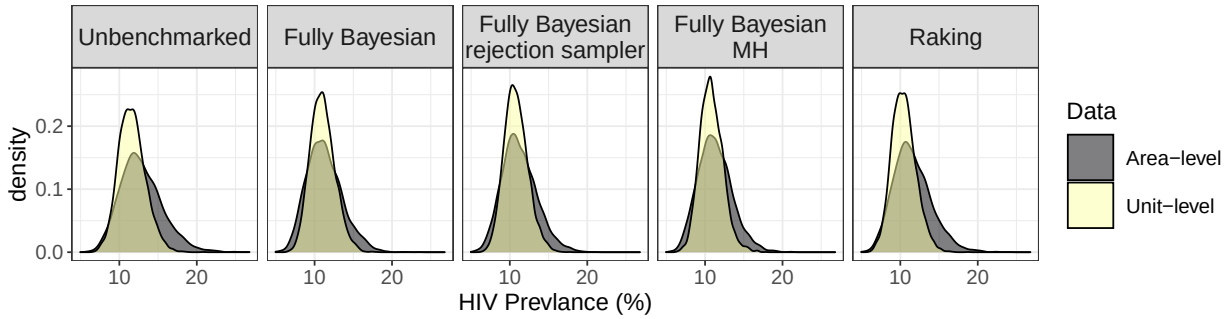


Figure A.13: Subnational HIV prevalence estimates for the Western Cape province from unbenchmarked and benchmarked models. All densities are based on 5000 samples.

In Figures A.14 through A.16, we compare benchmarked and unbenchmarked estimates and their uncertainty for area-level and unit-level models.

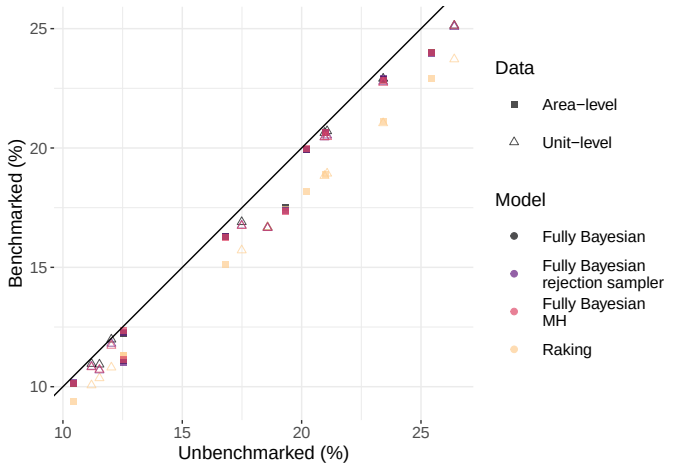


Figure A.14: Comparison of unbenchmarked and benchmarked posterior medians by model and data type at a subnational level.

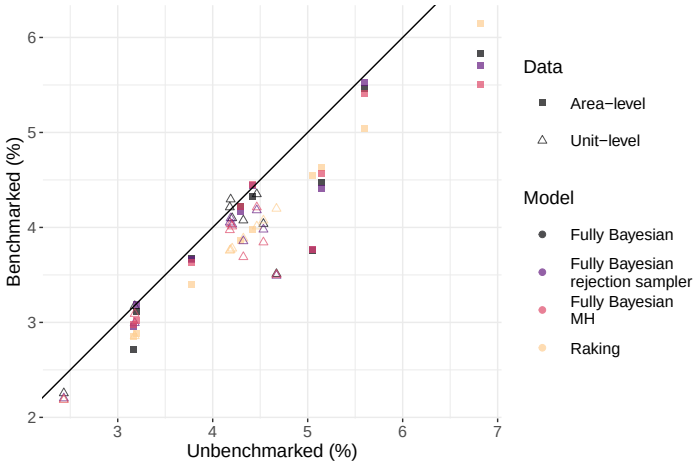


Figure A.15: Comparison of unbenchmarked and benchmarked posterior 95% CI widths by model and data type at a subnational level.

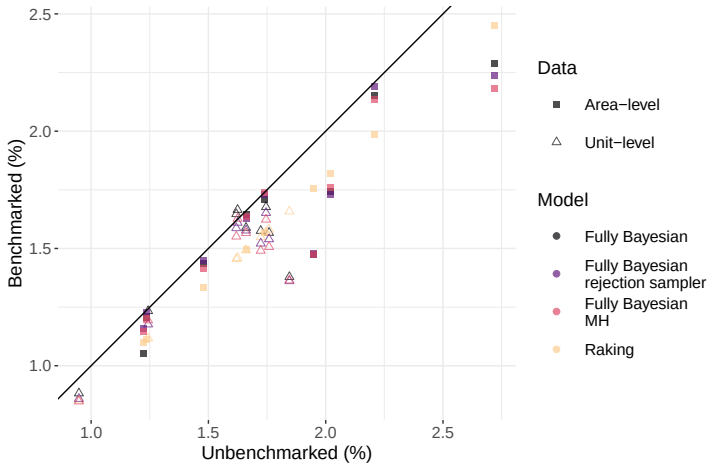


Figure A.16: Comparison of unbenchmarked and benchmarked posterior standard deviations by model and data type at a subnational level.

In Figures A.17 through A.19, we display benchmarked and unbenchmarked estimates and their uncertainty for area-level and unit-level models.

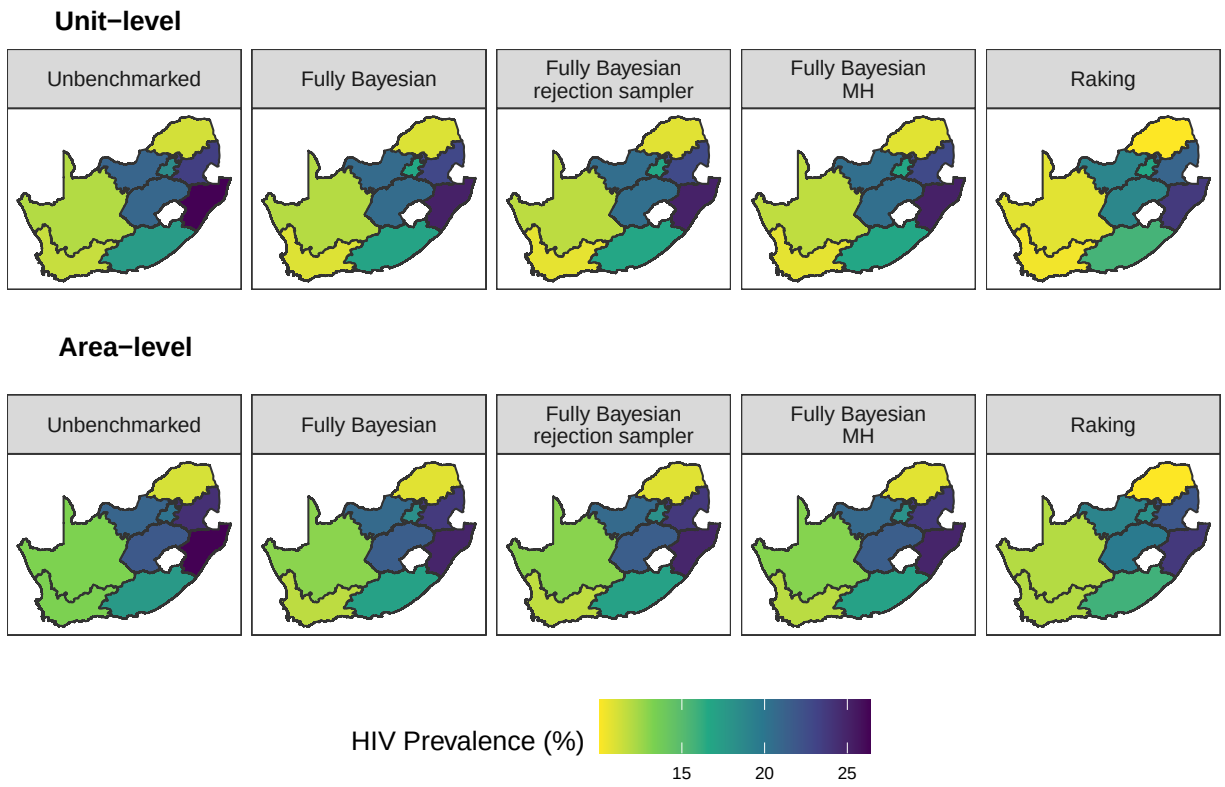


Figure A.17: Unbenchmarked and benchmarked posterior medians by model and data type at a subnational level.

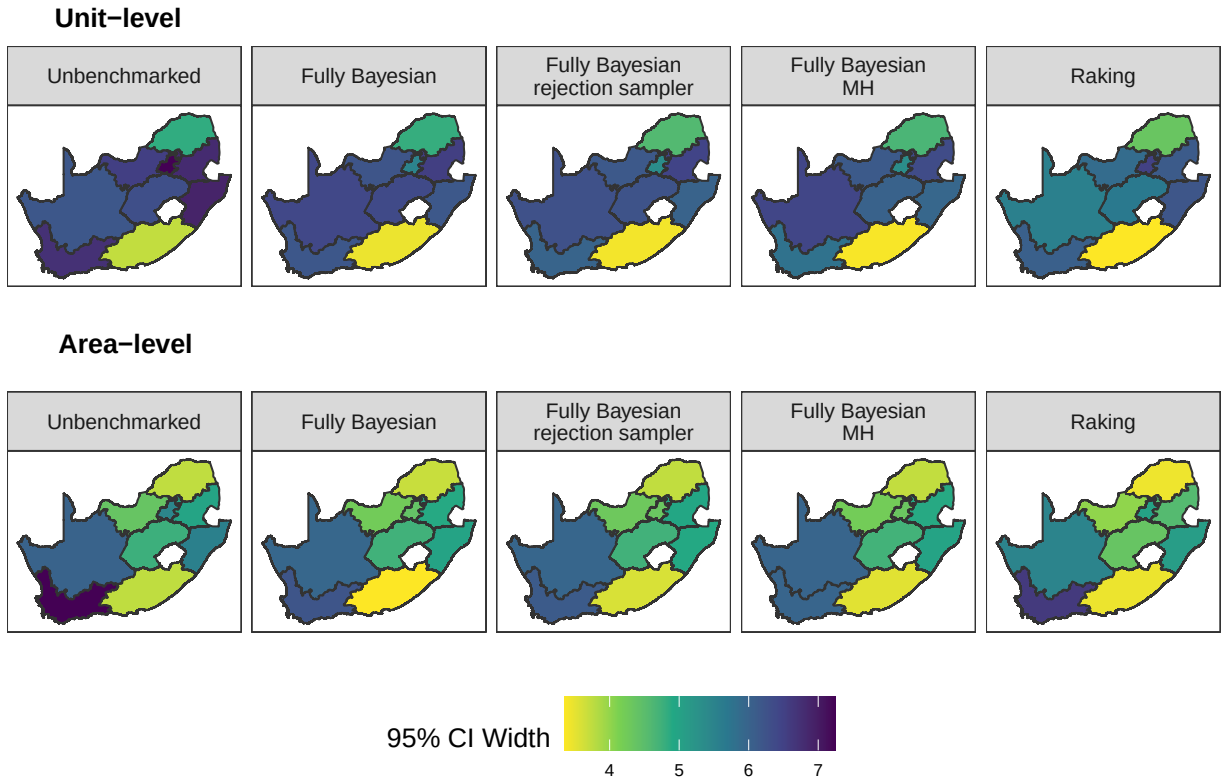


Figure A.18: Unbenchmarked and benchmarked posterior 95% CI widths by model and data type at a subnational level.

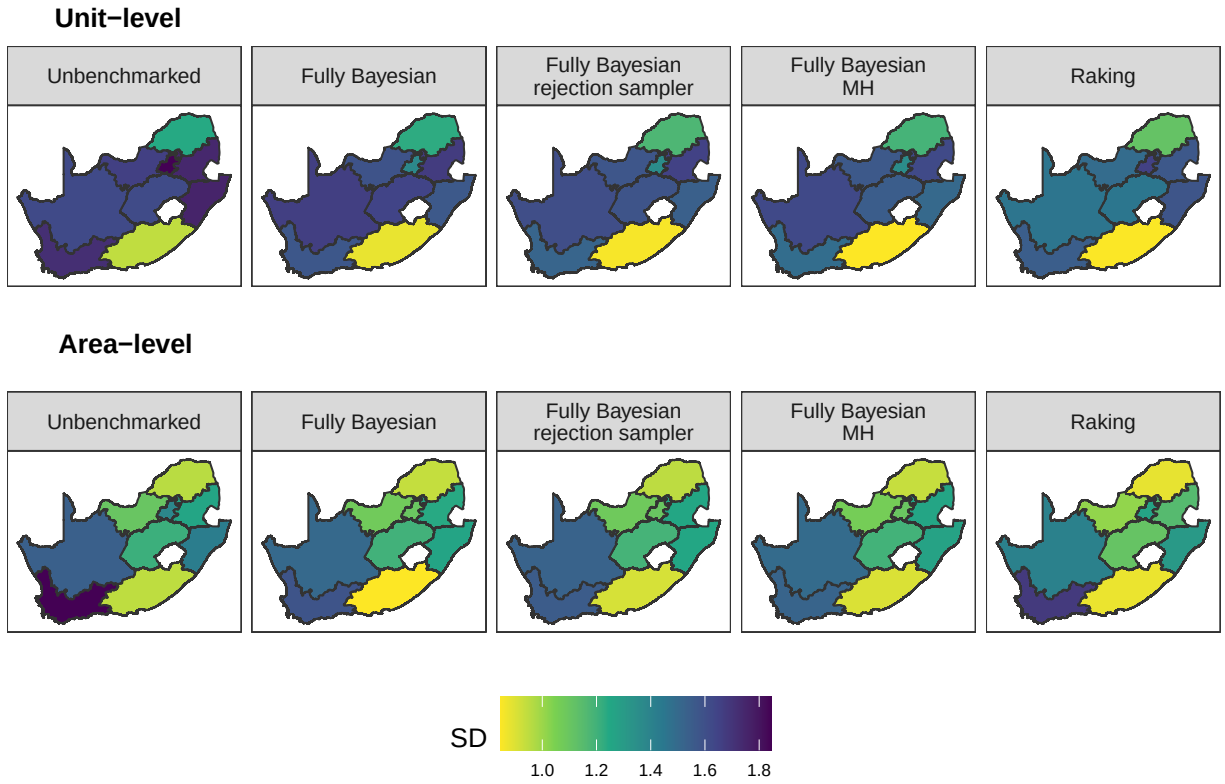


Figure A.19: Unbenchmarked and benchmarked posterior standard deviations by model and data type at a subnational level.

A.1.4 Model Validation

Below we display the model validation results for the HIV application for both unit- and area-level models. We compare posterior medians of the predictive distribution in each area (having left that area out of model fitting) to the direct estimate in each area, respectively. The direct estimates in each area are simply the Horvitz-Thompons estimators in each are i, with design variance on the logit scale calculated as $\hat{V}_i^{DES} = \hat{\sigma}_i^{2HT}/(\hat{\theta}_i^{2HT}(1-\hat{\theta}_i^{2HT}))^2$. We obtain the posterior predictive distribution in each area i by fitting the unbenchmarked model having left out the data for area i, and then adding the design variance to those samples on the logit scale. The resulting posterior predictive samples, on the logit scale, are thus a sum of a posterior draw from $\text{logit}(\hat{\theta}_i)$ and $N(0, \hat{V}_i^{DES})$. Credible intervals are

calculated from the draws of the posterior predictive distributions.

We note that both the 50% and 80% intervals suggest that both unit- and area-level models are reasonably well-suited to our data, with the 80% credible intervals capturing 8/9 direct estimates in each model, and the 50% credible intervals capturing 3/9 and 4/9 in the unit- and area-level models, respectively. This suggests that the area-level model may be more suited to the data, which is unsurprising given that the area-level model accounts for the survey design directly. For the unit-level model (as noted in the discussion in the main paper) that there are likely variables in the survey design that are not being accounted for in the unit-level model that may lead to undercoverage.

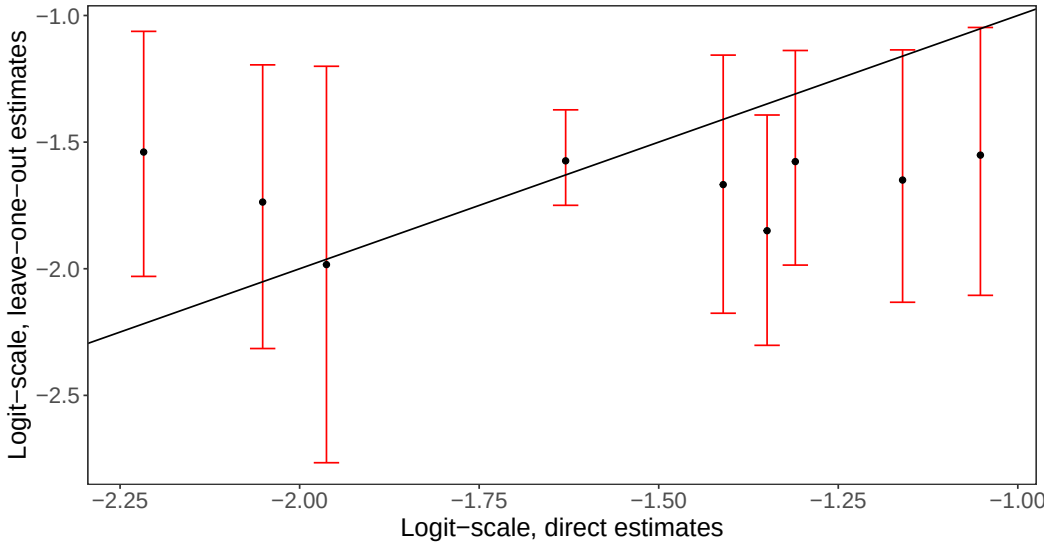


Figure A.20: Scatterplot comparing leave-one-out posterior predictive estimates and direct estimates on the logit scale for the unit-level HIV model, with 80% confidence intervals.

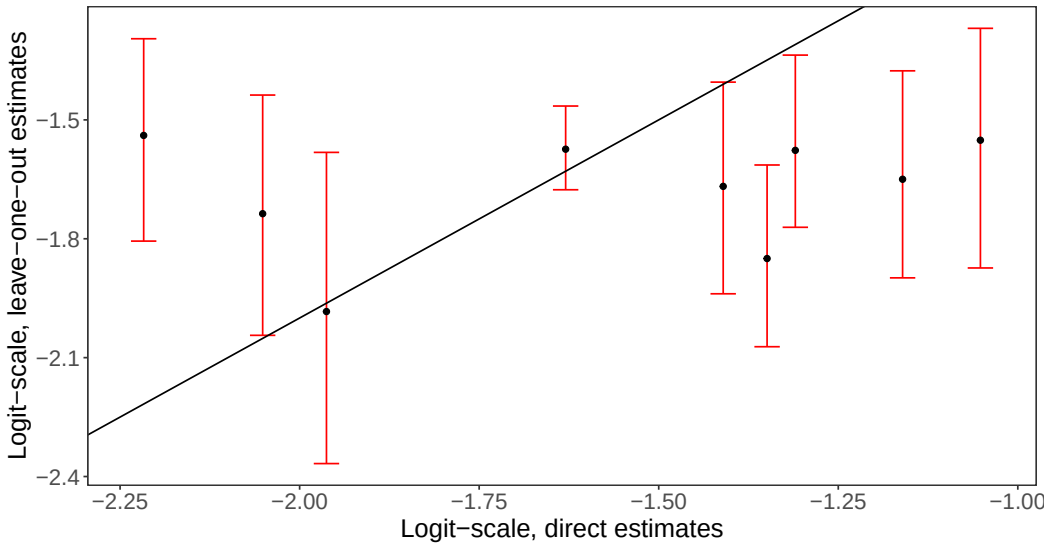


Figure A.21: Scatterplot comparing leave-one-out posterior predictive estimates and direct estimates on the logit scale for the unit-level HIV model, with 50% confidence intervals.

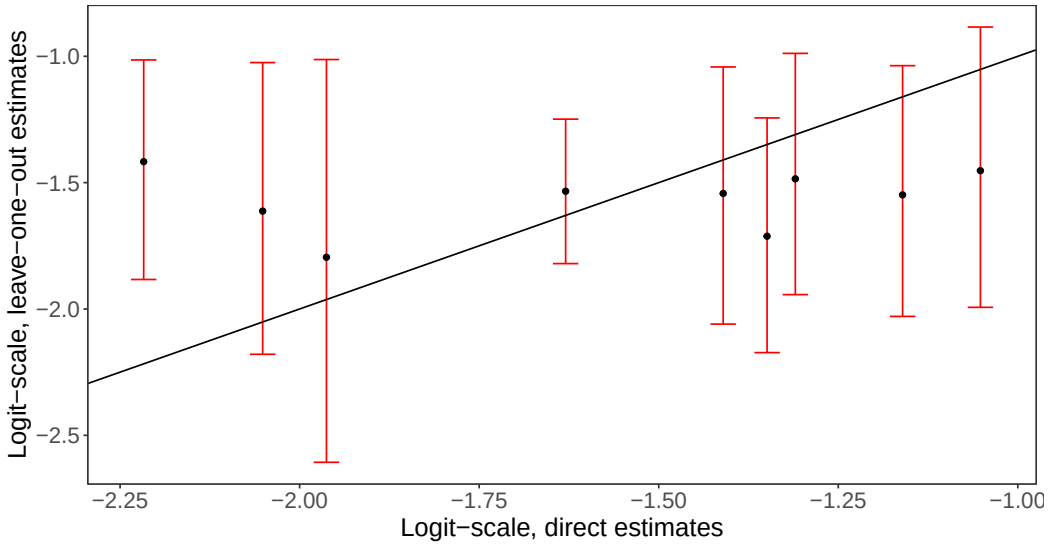


Figure A.22: Scatterplot comparing leave-one-out posterior predictive estimates and direct estimates on the logit scale for the area-level HIV model, with 80% confidence intervals.

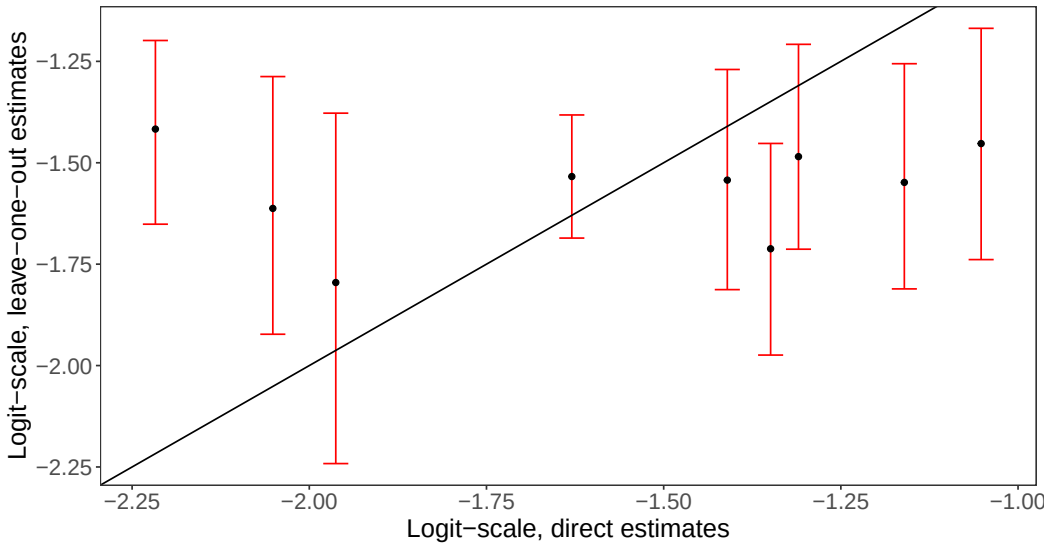


Figure A.23: Scatterplot comparing leave-one-out posterior predictive estimates and direct estimates on the logit scale for the area-level HIV model, with 50% confidence intervals.

A.2 U5MR Application

Below we describe the data, models, and software used in the U5MR application.

A.2.1 Data

Spatial boundary files for Namibia are obtained from GADM, the Database of Global Administrative Areas (Database of Global Administrative Areas (GADM), 2019). We use the 2013 administrative boundary (the most recent DHS survey year) for all years in our U5MR model for Namibia, and note that this implies 13 Administrative 1 (admin1) regions. Of note, the Kavango region split in 2013 to form Kavango East and Kavango West. We obtain estimates for the Kavango region as a whole, as this is consistent with the 2013 GADM file, the most recent DHS survey we have available for Namibia, and current subnational UN IGME estimates for Namibia.

For our U5MR application, we use data from the 2000, 2006-2007, and 2013 DHS surveys for Namibia. These three surveys were chosen for their GPS data availability, and to align

with the analysis carried out to produce official subnational U5MR estimates by UN IGME (Inter-agency Group for Child Mortality Estimation, 2020), and therefore allow for comparison between the benchmarking methods used in that paper and those proposed here. The surveys followed a multi-stage, stratified design and were designed to provide estimates at the admin1 level, which consists of 13 regions. Again for consistency with official UN IGME estimates, we make predictions for years 2014-2019.

We obtain national level benchmarks for 2000-2019 in Namibia from the UN IGME national estimates from the B3 model (Alkema and New, 2014). The national level U5MR estimates and their 90% confidence intervals can be found in Table A.6. The likelihood for our benchmarks for the fully Bayesian benchmarking approach requires standard errors for the benchmarks, which we approximate via the assumption that the benchmark is asymptotically normally distributed. These standard errors are presented in Table A.6. The data sources used to produce the B3 estimates include surveys/censuses with full birth histories, summary birth histories, and household deaths. A full table of the data sources used is presented in Table A.5. Of note, the three full birth history DHS surveys we use for our model (2000, 2006-2007, 2013) are included in the B3 model. As many additional data sources went into the UN IGME estimates, we consider this setting to be external benchmarking and therefore treat our estimates as independent of the national level benchmarks despite this inclusion. Implications of the potential internal benchmarking scenario are outside the scope of this paper.

We use population count data from WorldPop for our U5MR application (Tatem, 2017; Stevens et al., 2015). Population counts are available at a 100 meter resolution, and using spatial boundary files from GADM, we aggregate population counts to the admin1 level. The population count data from WorldPop prior to 2020 is not provided with any measure of uncertainty, and so we treat the estimated population proportions as fixed. The spatial distribution of the under five population from 2000-2019 in Namibia can be seen in Figure A.24. In Figure A.25, the same data is displayed as a stacked bar chart over time.

Table A.5: Data sources included in the B3 model for Namibia. MM = maternal mortality.

Survey/Census	**Year**	**Collection Method**
Census	1991	Summary Birth History
DHS (MM adjusted)	1992	Full Birth History
DHS	1992	Summary Birth History
DHS	1992	Full Birth History
DHS (MM adjusted)	2000	Full Birth History
DHS	2000	Summary Birth History
DHS	2000	Full Birth History
Census	2001	Household Deaths
DHS (MM adjusted)	2006-2007	Full Birth History
DHS	2006-2007	Summary Birth History
DHS	2006-2007	Full Birth History
DHS	2006-2007	Household Deaths
Census	2011	Summary Birth History
Census	2011	Household Deaths
DHS (MM adjusted)	2013	Full Birth History
DHS	2013	Full Birth History
Inter-censal Demographic Survey	2016	Summary Birth History
Inter-censal Demographic Survey	2016	Household Deaths

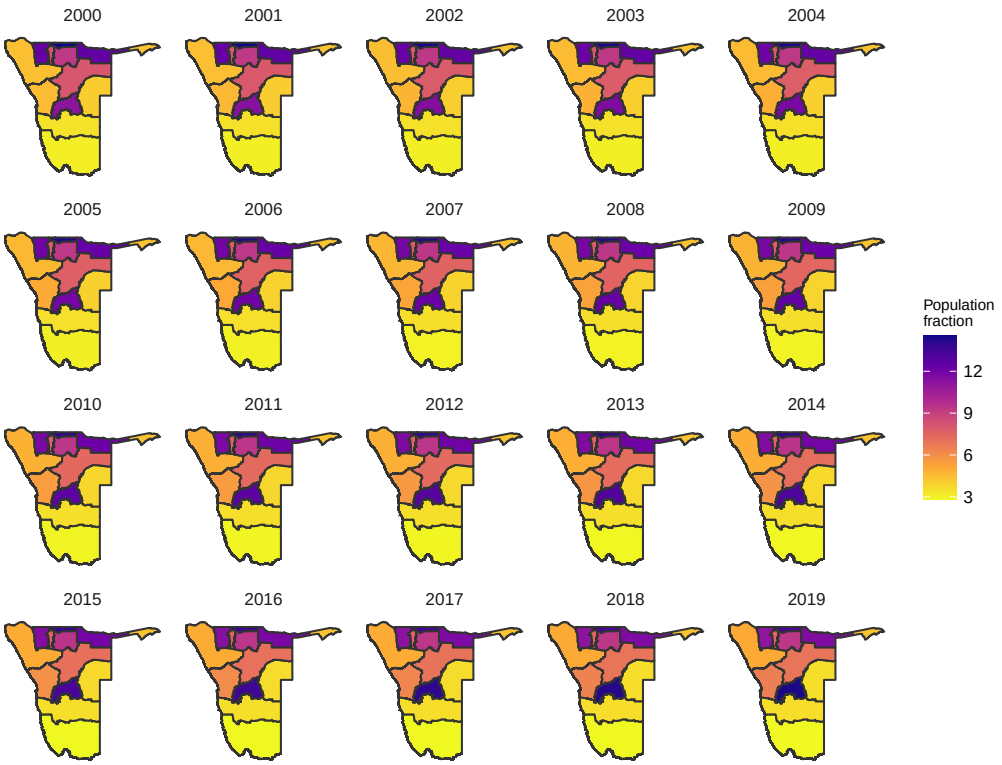
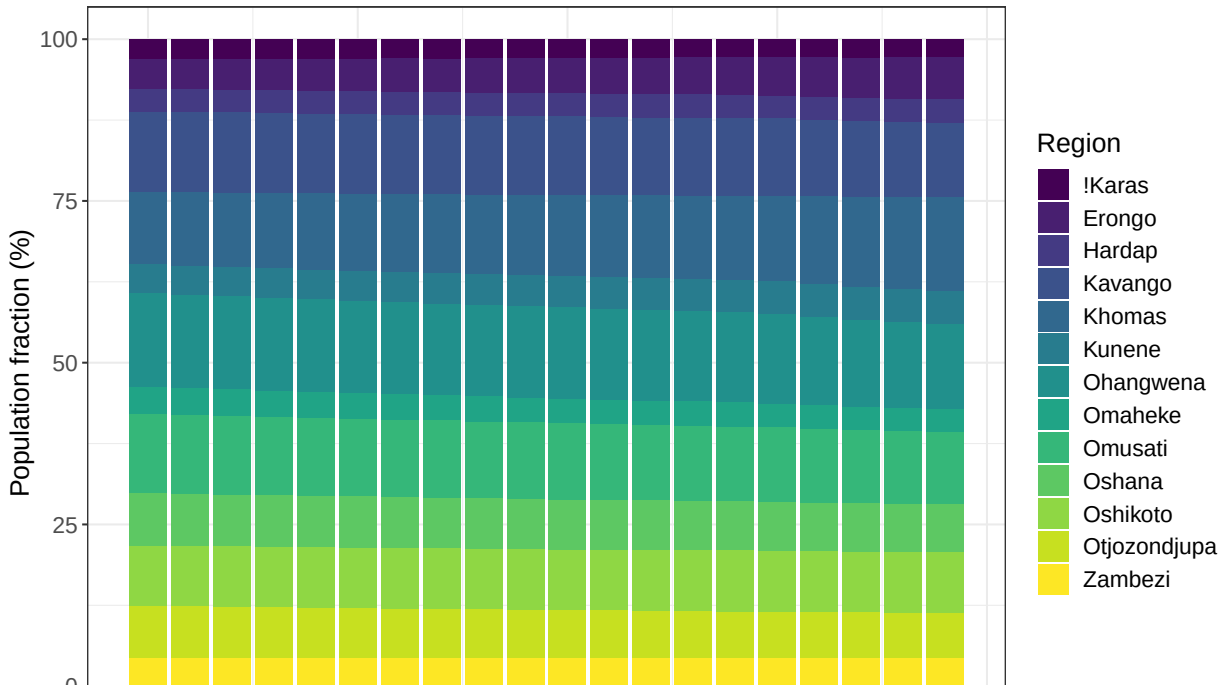


Figure A.24: Percentage of the under-5 population living in each region by year.

Table A.6: National benchmarks for U5MR from the UN IGME B3 model for Namibia. Standard errors are computed using the upper bound of the 90% confidence interval via the assumption that the benchmark is normally distributed. U5MR is reported as deaths per 1000 live births.

Year	U5MR	CI (90%)	SE (%)*
2000	75.4	(67.9, 84.4)	5.5
2001	74.9	(67.5, 83.8)	5.4
2002	74.3	(67.0, 83.2)	5.4
2003	74.0	(66.5, 82.8)	5.3
2004	73.7	(66.1, 82.3)	5.2
2005	70.2	(62.8, 78.6)	5.1
2006	65.6	(58.4, 73.8)	5.0
2007	60.7	(53.3, 69.4)	5.3
2008	55.8	(47.9, 65.0)	5.6
2009	51.8	(43.5, 62.3)	6.4
2010	49.7	(40.4, 62.0)	7.5
2011	50.4	(39.4, 66.0)	9.5
2012	52.2	(39.3, 71.4)	11.7
2013	50.1	(36.3, 71.8)	13.2
2014	48.1	(33.7, 72.3)	14.7
2015	47.8	(32.3, 74.7)	16.3
2016	46.3	(30.0, 75.6)	17.8
2017	44.5	(28.0, 75.1)	18.6
2018	43.3	(26.2, 76.0)	19.9
2019	42.4	(24.8, 77.0)	21.1



A.2.2 Unbenchmarked models

Area-level: Spatial Fay-Herriot

We obtain design-based, direct estimates ${}_{60}\hat{q}_{0it}^{HT}$ of U5MR and their variances for each area i and time t, and smooth the direct estimates via a spatio-temporal extension to the classic Fay-Herriot model described in Li et al. (2019) (Fay and Herriot, 1979; Mercer et al., 2015). Direct estimates are combined across surveys using a standard meta-analysis approach and an HIV bias correction is done, both of which are described in Li et al. (2019) in detail.

Unit-level: Binomial

From each DHS survey, we have records of birth and death dates for each child born to a mother sampled in the survey. These records are expanded into monthly binary outcomes that indicate whether death occurs for a given child in a given month, allowing us to perform a discrete time survival analysis with discrete hazards by age groups. In this way, children who survive longer contribute more observations to our data. The binary outcomes are collapsed to binomial observations in clusters c within area i in year t from survey s for age groups defined as 0-1, 1-11, 12-23, 24-35, 36-47, and 48-59 months. Let $m = 0, 1, \ldots, 59$ denote age in months, and let

$$a[m] = \begin{cases} 1, 0 \leq m < 1 \\ 2, 1 \leq m < 12 \\ 3, 12 \leq m < 24 \\ 4, 24 \leq m < 36 \\ 5, 36 \leq m < 48 \\ 6, 48 \leq m < 60 \end{cases} \qquad a^*[m] = \begin{cases} 1, 0 \leq m < 1 \\ 2, 1 \leq m < 12 \\ 3, 12 \leq m < 60 \end{cases}$$

We have number of deaths observed $y_{1m,i[c]ts}$ in $n_{m,i[c]ts}$ months of life in cluster c within area i, year t in agemonth m from survey s. We consider the unbenchmarked model

$$\begin{aligned} y_{1m,i[c]ts} \mid n_{m,i[c]ts}, {}_1q_{m,i[c]ts} &\sim \text{Binomial}(n_{m,i[c]ts}, {}_1q_{m,i[c]ts}) \\ \eta_{m,i[c]ts} = \text{logit}({}_1q_{m,i[c]ts}) &= \mu_{a[m]} + \alpha_t + \gamma_{a^*[m]t} + b_i \\ &\quad + \beta_i \times t + \delta_{it} + \nu_s + e_{i[c]} + \log(\text{HIV}_{ts}), \end{aligned} \tag{A.1}$$

and aggregate to a space/time/age level via

$${}_1q_{m,it} = \text{expit}\left(\frac{\mu_{a[m]} + \alpha_t + \gamma_{a^*[m]t} + b_i + \beta_i \times t + \delta_{it}}{\sqrt{1 + h^2\sigma_e^2}}\right),$$

where ${}_1q_{m,it}$ is the monthly hazard of death for age m. The linear predictor contains fixed intercepts for age bands $\mu_{a[m]}$, an iid temporal random effect $\alpha_t \sim N(0, \tau_\alpha)$, and age group-specific temporal random effects following a random walk 2 (RW2) $\gamma_{a^*[m]t}$ (Lindgren and Rue, 2008). The term b_i is a spatial random effect following a BYM2 prior (Riebler et al., 2016; Besag et al., 1991). A sum-to-zero constraint is placed on the structured component of the BYM2 random effect to ensure identifiability. Two separate terms allow for space-time interactions: β_i are random, area-specific slopes, and δ_{it} is a Type IV Knorr-Held interaction term (Knorr-Held, 2000). Finally we have a survey fixed effect ν_s which we constrain to sum to zero, and a log offset for HIV, varying by survey and time. These log offsets account for the differing survival probabilities of children born to HIV-positive mothers (Walker et al., 2012). The inclusion of a log offset term in our linear predictor corresponds approximately to a multiplicative bias correction for mothers who could not be sampled due to death from HIV (Wakefield et al., 2019).

We obtain predictions of U5MR at each time point ${}_{60}q_{0t}$ by calculating posterior draws

$k = 1, \ldots, K$ of all terms from Equation A.1 and the formula

$$ {}_{60}q_{0it}^{(k)} = 1 - \prod_{m=1}^{59} \left(1 - \text{expit} \left(\frac{\mu_{a[m]}^{(k)} + \alpha_t^{(k)} + \gamma_{a^*[m]t}^{(k)} + b_i^{(k)} + \beta_i^{(k)} \times t + \delta_{it}^{(k)}}{\sqrt{1 + h^2 \sigma_e^{2(k)}}} \right) \right). $$

Note that we do not include the survey fixed effect, cluster random effect, or log offset in our predictions.

For the variance parameters in all random effects, we use PC priors with parameters $U = 1$, $\alpha = 0.01$. For the age-specific intercepts, we use the default INLA prior, which is normal, centered at 0 with precision 0.001. For the ϕ parameter in the BYM2 spatial random effect, we use a PC prior with parameters $U = 0.5$, $\alpha = 2/3$. These prior and hyperprior choices are used in all benchmarked models below, when applicable.

A.2.3 Benchmarked models

Approaches

Raking:

For the unit-level model, we obtain posterior medians ${}_{60}\hat{q}_{0it}^{M}$ from the unbenchmarked model for each area i and time point t, and compute

$$ R_t^{R1} = \frac{\sum_{i=1}^{m} w_{it} \times {}_{60}\hat{q}_{0it}^{M}}{y_{2t}} $$

using weights w_{it} and national benchmarks are each time point y_{2t}. We then refit the unbenchmarked model including R_t^{R1} as a log offset term in the linear predictor to obtain benchmarked posterior draws ${}_{60}\hat{q}_{0it}^{R1(k)}$.

For the area-level model, we adjust the unbenchmarked direct estimates with the ratio R_t^{R1} in the same manner we adjust for HIV, as described in Li et al. (2019), and then proceed fitting the smoothed direct model with these HIV- and benchmark-adjusted direct estimates.

Fully Bayesian: Rejection sampler:

We obtain $k = 1, \ldots, K$ posterior draws ${}_{60}\hat{q}_{0it}^{(k)}$ from the unbenchmarked model for each area i and year t, and apply the algorithm described in the proposed approach section of Chapter 2 to obtain benchmarked posterior draws ${}_{60}\hat{q}_{0it}^{FB2}$.

Fully Bayesian: Metropolis-Hastings algorithm:

We obtain $k = 1, \ldots, K$ posterior draws $\hat{\theta}_i^{(k)}$ from an *adjusted* unbenchmarked model for each area i, and apply the algorithm described in the proposed approach section of Chapter 2 using weights w_i, and the benchmark y_2 with associated standard error $\sigma_{y_2}^2$, to obtain benchmarked posterior draws $\hat{\theta}_i^{FB3(k)}$. The adjusted unbenchmarked model uses the prior $\pi^+(\beta_0) \sim \text{Normal}(\text{logit}(y_2), 0.001^{-1})$.

A.2.4 Software

We implement the unbenchmarked and raking benchmarked U5MR models in INLA via the R package `SUMMER` (Rue et al., 2009; Li et al., 2020), and the proposed benchmarking approach and Bayes estimate approach using a combination of INLA and R. INLA is a popular tool for conducting Bayesian, space-time analyses, and provides great computational gains compared to traditional MCMC methods. While nonlinear predictors can be incorporated via the R package `inlabru`, INLA itself cannot make use of nonlinear predictors by design (Bachl et al., 2019). As such, the fully Bayesian benchmarking approach described in Zhang and Bryant (2020) cannot be fit in INLA. Since the unbenchmarked U5MR model is straightforward to fit in INLA using the `SUMMER` package, this application serves as an example of how the proposed rejection sampling approach to fully Bayesian benchmarking may be particularly useful for modelers who have already produced unbenchmarked estimates in an a programme that doesn't allow them to incorporate additional likelihood(s) for the benchmarks.

A.2.5 U5MR Results

The proposed rejection sampling approach to fully Bayesian benchmarking performs similarly in the U5MR example to the HIV prevalence example in that the posterior density of the aggregated national level estimate is a compromise between the national IGME density and the unbenchmarked density for both the unit-level and area-level models. Figure A.27 displays the ratio of unbenchmarked medians from the unit-level and area-level models to IGME national medians across time, with estimates and their confidence intervals shown in A.26. During "prediction" years where there is no data informing estimates (2014 onward), the ratio for the unit-level model is particularly large, and we expect to see the greatest effect in benchmarking during these years. The ratio is particularly large likely due to the flexibility of the RW2 in time in this model, and the ratio is less large in prediction years for the area-level model. The ratio for the area-level model overall is greater for almost all years than the ratio for the unit-level model.

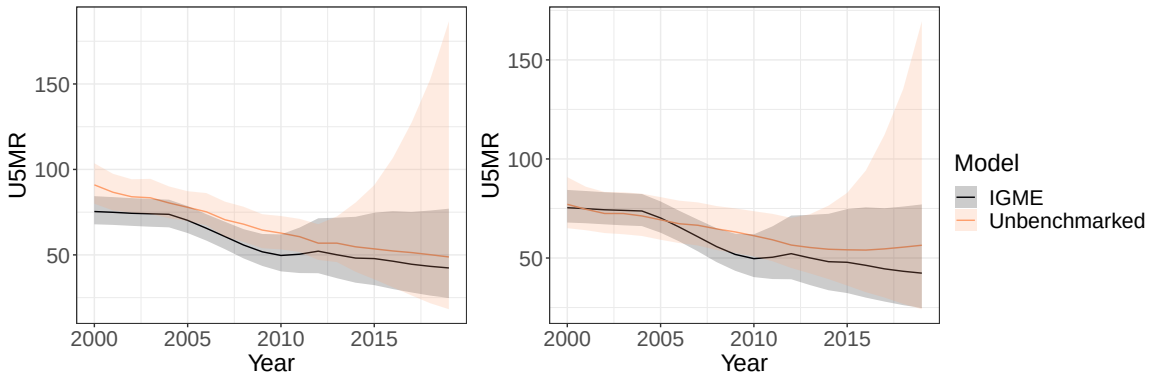


Figure A.26: Aggregated national unbenchmarked medians from the area-level model (left) and unit-level model (right) compared to IGME national medians across time.

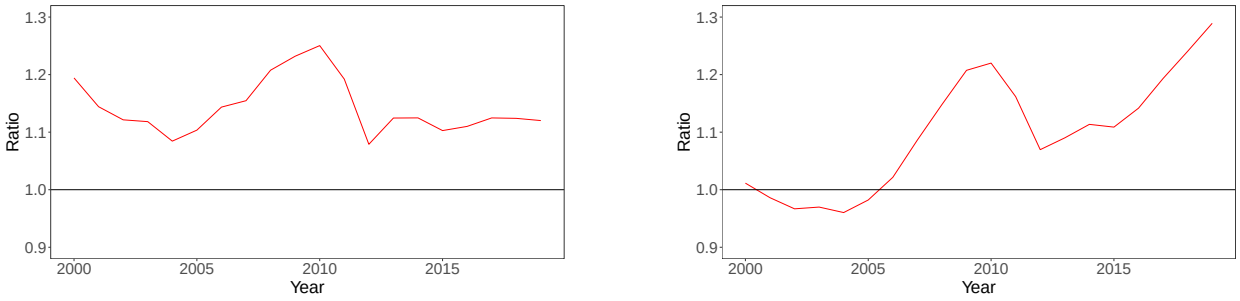


Figure A.27: Ratio of aggregated national unbenchmarked medians from the area-level model (left) and unit-level model (right) to IGME national medians across time. A value greater than 1 corresponds to the unbenchmarked estimate being higher than the IGME estimate.

Figures A.28 through A.47 display the aggregated national level estimates produced from each benchmarked and unbenchmarked method for both unit- and area-level models. The posterior densities from the fully Bayesian rejection sampling approach are less uncertain than the posterior densities from the unbenchmarked model due to the additional information contained in the benchmarking constraint being incorporated in the benchmarked approach. The overlap between the IGME density and the Bayes estimate approach is expected as the Bayes estimate approach performs exact benchmarking. A table of these national estimates for all years and models can be found in Table A.7.

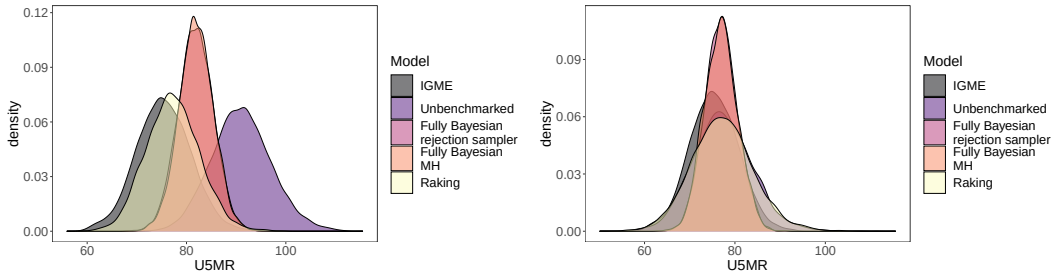


Figure A.28: Aggregated national level U5MR estimates from IGME, unbenchmarked, and benchmarked models for area-level (left) and unit-level (right) models, for 2000. All densities are based on 5000 samples. U5MR is reported as deaths per 1000 live births.

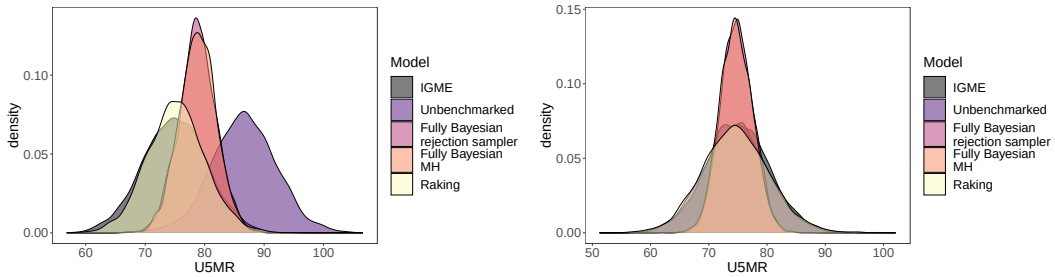


Figure A.29: Aggregated national level U5MR estimates from IGME, unbenchmarked, and benchmarked models for area-level (left) and unit-level (right) models, for 2001. All densities are based on 5000 samples. U5MR is reported as deaths per 1000 live births.

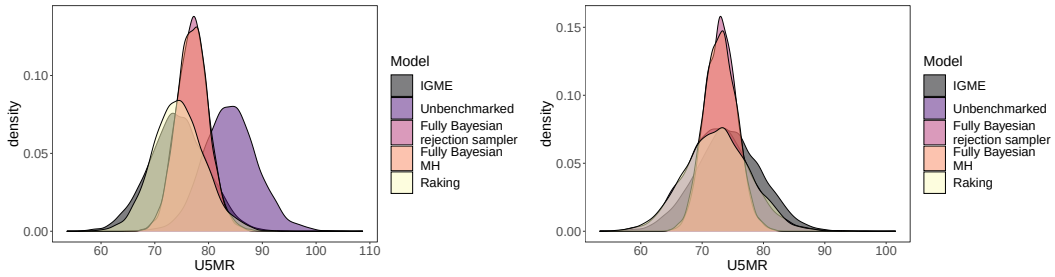


Figure A.30: Aggregated national level U5MR estimates from IGME, unbenchmarked, and benchmarked models for area-level (left) and unit-level (right) models, for 2002. All densities are based on 5000 samples. U5MR is reported as deaths per 1000 live births.

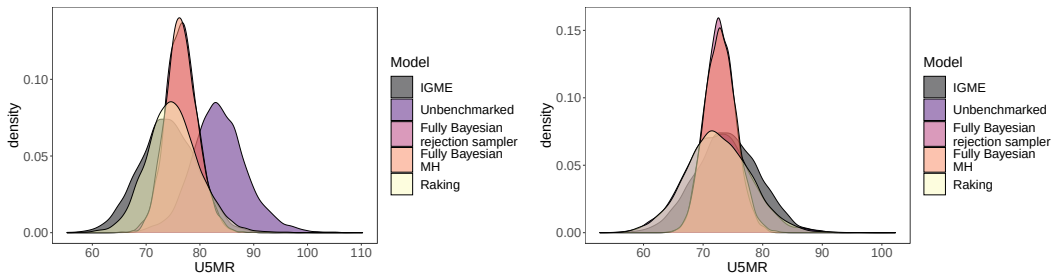


Figure A.31: Aggregated national level U5MR estimates from IGME, unbenchmarked, and benchmarked models for area-level (left) and unit-level (right) models, for 2003. All densities are based on 5000 samples. U5MR is reported as deaths per 1000 live births.

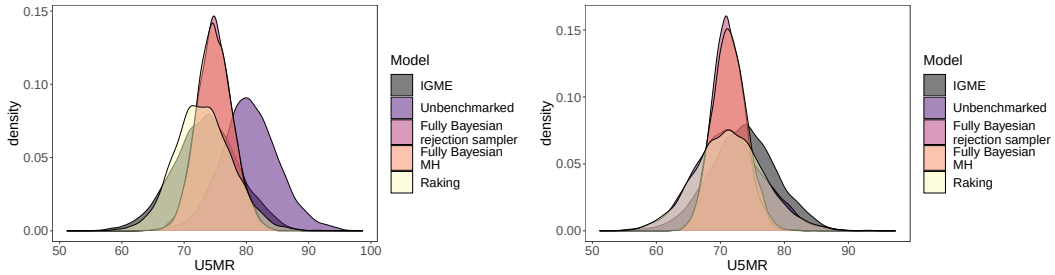


Figure A.32: Aggregated national level U5MR estimates from IGME, unbenchmarked, and benchmarked models for area-level (left) and unit-level (right) models, for 2004. All densities are based on 5000 samples. U5MR is reported as deaths per 1000 live births.

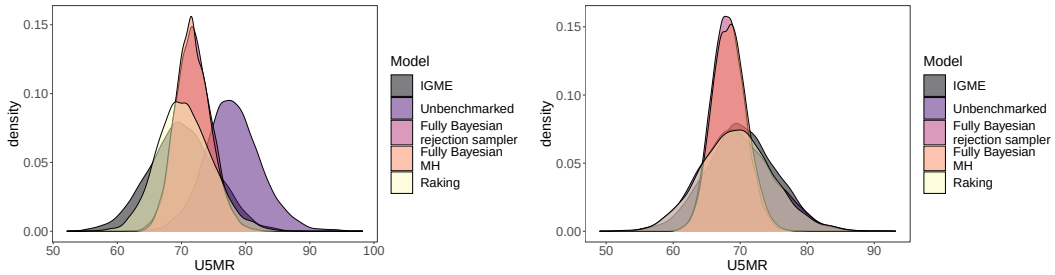


Figure A.33: Aggregated national level U5MR estimates from IGME, unbenchmarked, and benchmarked models for area-level (left) and unit-level (right) models, for 2005. All densities are based on 5000 samples. U5MR is reported as deaths per 1000 live births.

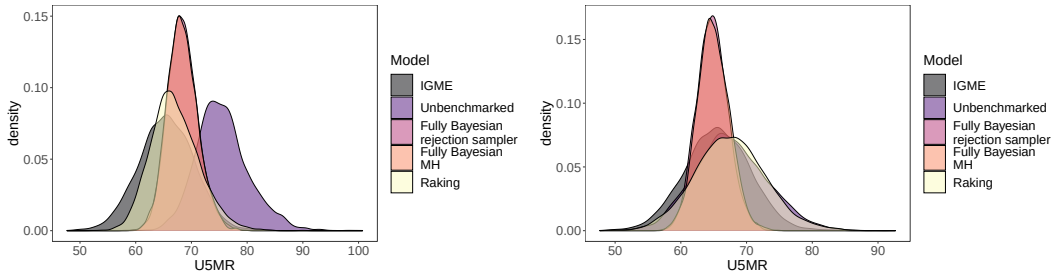


Figure A.34: Aggregated national level U5MR estimates from IGME, unbenchmarked, and benchmarked models for area-level (left) and unit-level (right) models, for 2006. All densities are based on 5000 samples. U5MR is reported as deaths per 1000 live births.

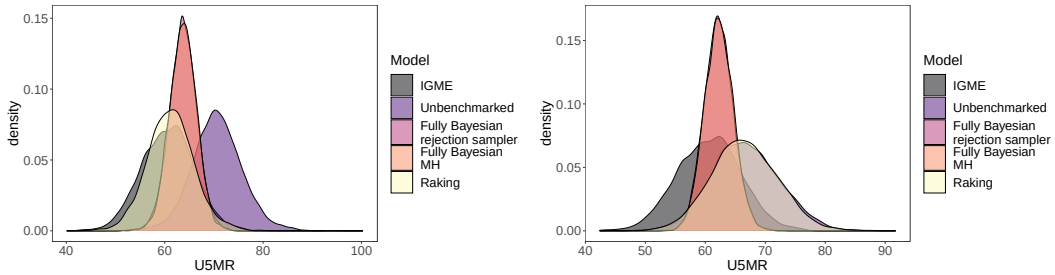


Figure A.35: Aggregated national level U5MR estimates from IGME, unbenchmarked, and benchmarked models for area-level (left) and unit-level (right) models, for 2007. All densities are based on 5000 samples. U5MR is reported as deaths per 1000 live births.

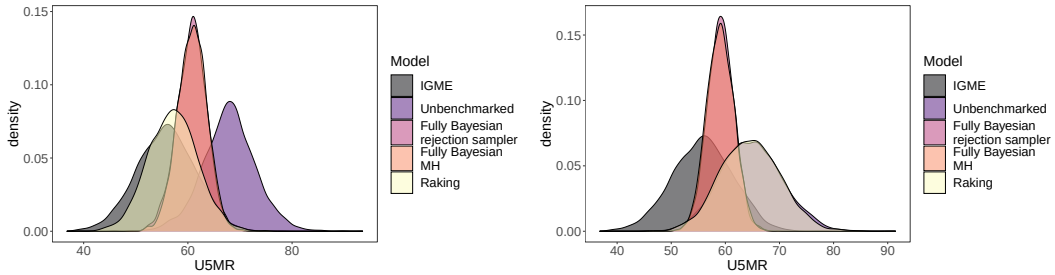


Figure A.36: Aggregated national level U5MR estimates from IGME, unbenchmarked, and benchmarked models for area-level (left) and unit-level (right) models, for 2008. All densities are based on 5000 samples. U5MR is reported as deaths per 1000 live births.

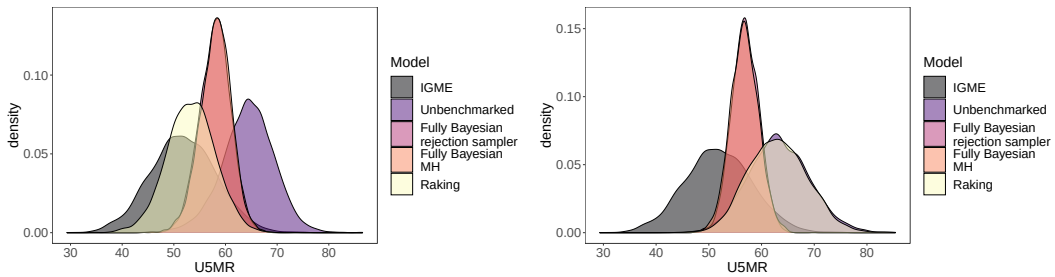


Figure A.37: Aggregated national level U5MR estimates from IGME, unbenchmarked, and benchmarked models for area-level (left) and unit-level (right) models, for 2009. All densities are based on 5000 samples. U5MR is reported as deaths per 1000 live births.

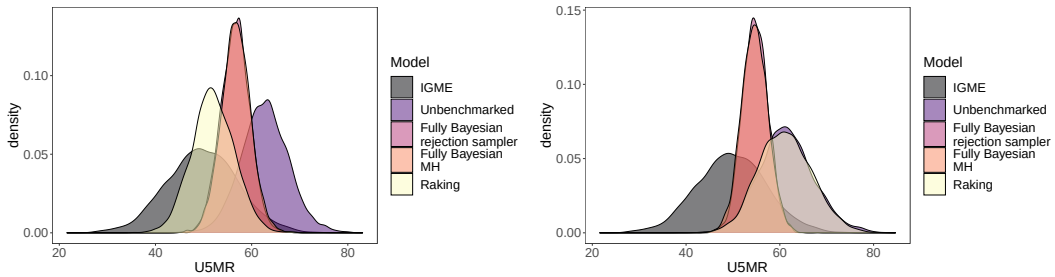


Figure A.38: Aggregated national level U5MR estimates from IGME, unbenchmarked, and benchmarked models for area-level (left) and unit-level (right) models, for 2010. All densities are based on 5000 samples. U5MR is reported as deaths per 1000 live births.

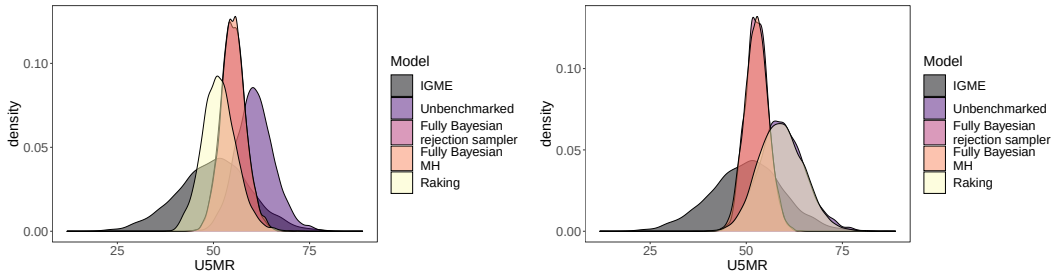


Figure A.39: Aggregated national level U5MR estimates from IGME, unbenchmarked, and benchmarked models for area-level (left) and unit-level (right) models, for 2011. All densities are based on 5000 samples. U5MR is reported as deaths per 1000 live births.

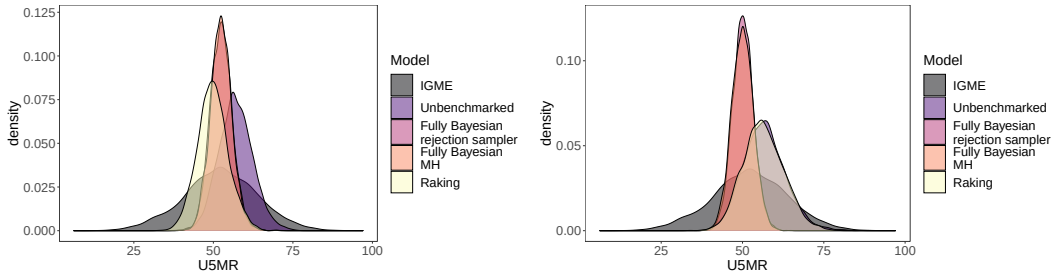


Figure A.40: Aggregated national level U5MR estimates from IGME, unbenchmarked, and benchmarked models for area-level (left) and unit-level (right) models, for 2012. All densities are based on 5000 samples. U5MR is reported as deaths per 1000 live births.

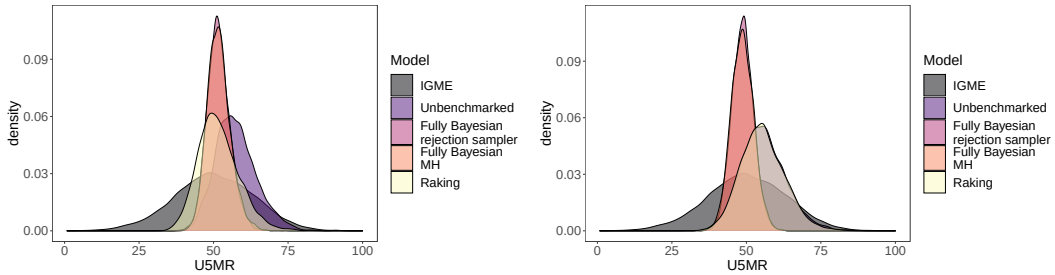


Figure A.41: Aggregated national level U5MR estimates from IGME, unbenchmarked, and benchmarked models for area-level (left) and unit-level (right) models, for 2013. All densities are based on 5000 samples. U5MR is reported as deaths per 1000 live births.

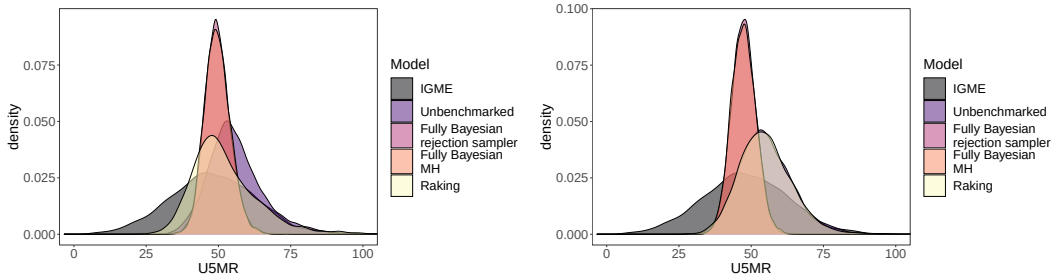


Figure A.42: Aggregated national level U5MR estimates from IGME, unbenchmarked, and benchmarked models for area-level (left) and unit-level (right) models, for 2014. All densities are based on 5000 samples. U5MR is reported as deaths per 1000 live births.

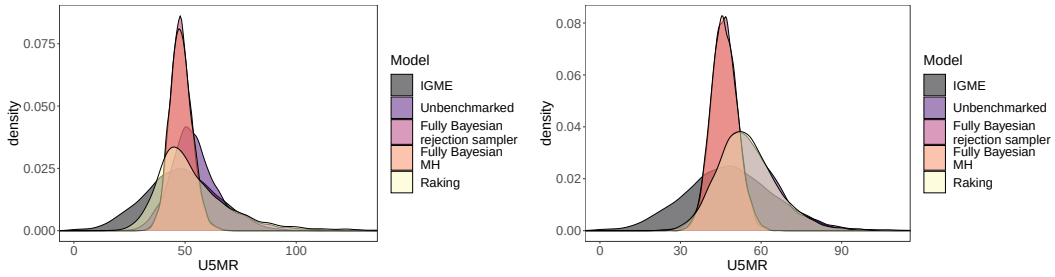


Figure A.43: Aggregated national level U5MR estimates from IGME, unbenchmarked, and benchmarked models for area-level (left) and unit-level (right) models, for 2015. All densities are based on 5000 samples. U5MR is reported as deaths per 1000 live births.

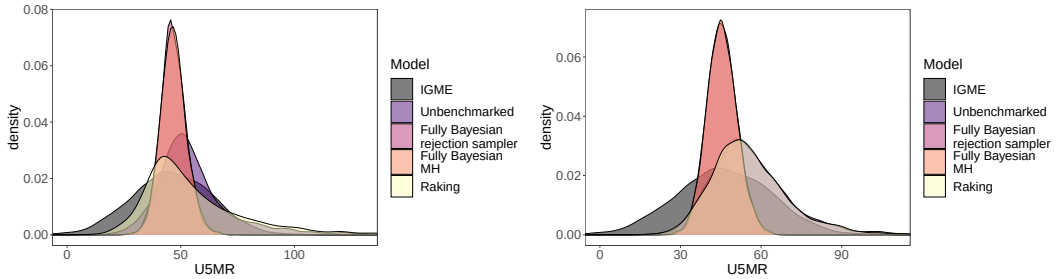


Figure A.44: Aggregated national level U5MR estimates from IGME, unbenchmarked, and benchmarked models for area-level (left) and unit-level (right) models, for 2016. All densities are based on 5000 samples. U5MR is reported as deaths per 1000 live births.

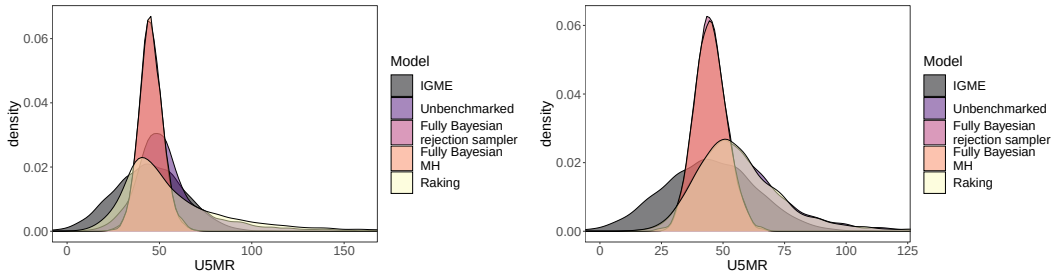


Figure A.45: Aggregated national level U5MR estimates from IGME, unbenchmarked, and benchmarked models for area-level (left) and unit-level (right) models, for 2017. All densities are based on 5000 samples. U5MR is reported as deaths per 1000 live births.

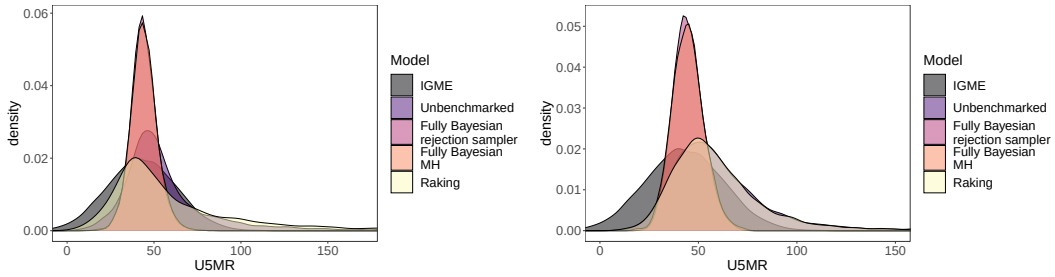


Figure A.46: Aggregated national level U5MR estimates from IGME, unbenchmarked, and benchmarked models for area-level (left) and unit-level (right) models, for 2018. All densities are based on 5000 samples. U5MR is reported as deaths per 1000 live births.

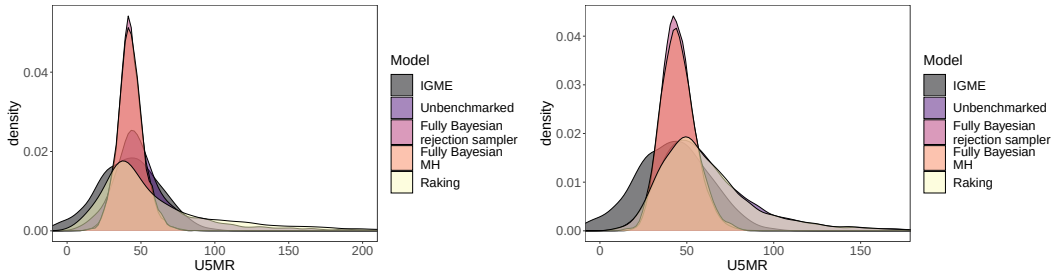


Figure A.47: Aggregated national level U5MR estimates from IGME, unbenchmarked, and benchmarked models for area-level (left) and unit-level (right) models, for 2019. All densities are based on 5000 samples. U5MR is reported as deaths per 1000 live births.

The proportion of samples accepted in the fully Bayesian rejection sampler approach was 0.5% for the unit-level model, and 0.001% for the area-level model. The proportion of samples accepted in the fully Bayesian MH approach was 7% for the unit-level model, and 0.3% for the area-level model.

Table A.7: Aggregated national level U5MR estimates from IGME, unbenchmarked, and benchmarked models for 2010. 95% credible intervals are given next to posterior medians. U5MR is reported as deaths per 1000 live births. AL = Area-level, UL = Unit-level.

Year	Model	AL: Median	UL: Median	AL: SD	UL: SD
2000	IGME	75.4 (67.9, 84.4)	75.4 (67.9, 84.4)	5.5	5.5
	Unbenchmarked	91 (79.9, 103.6)	77.1 (65, 90.8)	6.0	6.6
	FB: Rejection Sampler	81.9 (75, 88.7)	76.9 (70, 84.1)	3.5	3.6
	FB: MH	81.9 (75.1, 88.7)	76.9 (70, 84)	3.5	3.7
	Raking	77.2 (67.2, 87.9)	77.1 (64.2, 91.1)	5.4	6.8
2001	IGME	74.9 (67.5, 83.8)	74.9 (67.5, 83.8)	5.4	5.4
	Unbenchmarked	86.6 (76, 97.4)	74.6 (64, 86)	5.4	5.5
	FB: Rejection Sampler	78.8 (72.7, 85.1)	74.8 (69.6, 80.2)	3.2	2.8
	FB: MH	78.9 (72.6, 85)	74.6 (69.3, 80.3)	3.1	2.8
	Raking	75.2 (65.6, 85.1)	74.6 (64, 86.2)	4.9	5.7
2002	IGME	74.3 (67, 83.2)	74.3 (67, 83.2)	5.4	5.4
	Unbenchmarked	84 (74.7, 94.2)	72.5 (62.6, 83.3)	5.0	5.3
	FB: Rejection Sampler	77.1 (71.4, 83.1)	73 (68.2, 78.1)	3.0	2.6
	FB: MH	77 (71.3, 83.4)	73 (67.9, 78.2)	3.0	2.7
	Raking	74.3 (65.5, 84.4)	72.6 (62.3, 83)	4.8	5.3
2003	IGME	74 (66.5, 82.8)	74 (66.5, 82.8)	5.3	5.3
	Unbenchmarked	83.4 (74, 94.5)	72.5 (62, 83.2)	5.1	5.4
	FB: Rejection Sampler	76.6 (71, 82.8)	72.8 (68.1, 78)	3.0	2.5
	FB: MH	76.5 (71.1, 82.8)	72.9 (68, 78.4)	3.0	2.7
	Raking	74.8 (66, 85.3)	72.4 (62.2, 83.4)	4.9	5.4
2004	IGME	73.7 (66.1, 82.3)	73.7 (66.1, 82.3)	5.2	5.2
	Unbenchmarked	80.4 (71.5, 90.1)	71.2 (61.1, 82.4)	4.6	5.5
	FB: Rejection Sampler	74.9 (69.5, 80.8)	71.1 (66.6, 76.5)	2.9	2.5
	FB: MH	74.8 (69.5, 80.9)	71.4 (66.5, 76.8)	2.9	2.7
	Raking	73 (64.5, 82.5)	71.4 (61.1, 82.5)	4.6	5.4
2005	IGME	70.2 (62.8, 78.6)	70.2 (62.8, 78.6)	5.1	5.1
	Unbenchmarked	77.8 (69.8, 87.3)	69.3 (59.2, 80.7)	4.4	5.5
	FB: Rejection Sampler	71.8 (66.7, 77.6)	68.1 (63.7, 73.3)	2.8	2.5
	FB: MH	71.7 (66.8, 77.7)	68.4 (63.7, 73.7)	2.8	2.6
	Raking	70.4 (62.7, 79.6)	69.4 (59.2, 80.4)	4.3	5.4
2006	IGME	65.6 (58.4, 73.8)	65.6 (58.4, 73.8)	5.0	5.0
	Unbenchmarked	75.2 (67.3, 86.2)	67.3 (57.5, 78.9)	4.7	5.5
	FB: Rejection Sampler	68.3 (63.5, 74.3)	64.8 (60.3, 69.8)	2.7	2.4
	FB: MH	68.2 (63.4, 74.2)	64.9 (60.2, 69.8)	2.7	2.4
	Raking	66.7 (59.4, 76.4)	67.5 (57.2, 78.6)	4.3	5.4
2007	IGME	60.7 (53.3, 69.4)	60.7 (53.3, 69.4)	5.3	5.3
	Unbenchmarked	70.6 (60.8, 81.2)	66.5 (56.3, 78.1)	5.1	5.6
	FB: Rejection Sampler	63.7 (58.1, 69.3)	62.2 (57.7, 67)	2.8	2.4
	FB: MH	63.8 (58.4, 69)	62.3 (57.6, 67.1)	2.7	2.4
	Raking	61.2 (52.1, 71)	66.4 (56.1, 77.5)	4.7	5.5

Year	Model	AL: Median	UL: Median	AL: SD	UL: SD
2008	IGME	55.8 (47.9, 65)	55.8 (47.9, 65)	5.6	5.6
	Unbenchmarked	68 (57.6, 78.2)	64.7 (54, 76.2)	5.0	5.7
	FB: Rejection Sampler	60.8 (54.8, 66.2)	59.1 (54.3, 63.8)	2.9	2.4
	FB: MH	60.9 (55.1, 66.2)	59.1 (54.2, 64)	2.8	2.5
	Raking	57.1 (48, 66.5)	64.6 (54, 75.5)	4.8	5.5
2009	IGME	51.8 (43.5, 62.3)	51.8 (43.5, 62.3)	6.4	6.4
	Unbenchmarked	64.5 (53.9, 73.8)	63.2 (52.5, 75.1)	5.0	5.7
	FB: Rejection Sampler	58.1 (52, 63.8)	56.9 (51.9, 61.9)	3.0	2.6
	FB: MH	58.2 (51.8, 63.7)	56.7 (51.6, 62.1)	3.0	2.7
	Raking	53.3 (44, 62.3)	63 (52.8, 74.2)	4.7	5.6
2010	IGME	49.7 (40.4, 62)	49.7 (40.4, 62)	7.5	7.5
	Unbenchmarked	62.8 (53.2, 72.7)	61.2 (50.8, 73.6)	4.9	5.8
	FB: Rejection Sampler	56.6 (50.8, 62.4)	54.7 (49.5, 60.1)	3.0	2.7
	FB: MH	56.7 (50.7, 62.7)	54.6 (49, 60.3)	3.0	2.8
	Raking	51.9 (43.5, 60.7)	61 (50.6, 72.7)	4.4	5.6
2011	IGME	50.4 (39.4, 66)	50.4 (39.4, 66)	9.5	9.5
	Unbenchmarked	60.7 (51.7, 71.1)	59.2 (48.4, 72.3)	4.9	6.0
	FB: Rejection Sampler	55.1 (49.2, 61.6)	52.7 (47.3, 58.6)	3.2	2.9
	FB: MH	55.2 (49.1, 61.9)	52.7 (46.8, 58.8)	3.2	3.1
	Raking	51.4 (43.3, 60.8)	58.9 (48.3, 70.9)	4.4	5.8
2012	IGME	52.2 (39.3, 71.4)	52.2 (39.3, 71.4)	11.7	11.7
	Unbenchmarked	56.9 (47, 68)	56.5 (45, 70.3)	5.3	6.5
	FB: Rejection Sampler	52.4 (45.9, 59.3)	50.1 (44, 56.6)	3.4	3.2
	FB: MH	52.4 (46, 59.5)	50 (43.6, 56.5)	3.4	3.3
	Raking	49.7 (40.7, 60)	56.2 (45, 69.2)	4.8	6.3
2013	IGME	50.1 (36.3, 71.8)	50.1 (36.3, 71.8)	13.2	13.2
	Unbenchmarked	56.9 (45.8, 72.5)	55.3 (42.3, 72.1)	6.9	7.5
	FB: Rejection Sampler	51.5 (44.9, 60)	48.7 (41.9, 56)	3.8	3.6
	FB: MH	51.6 (44.8, 60.3)	48.6 (41.4, 56.1)	3.9	3.7
	Raking	51 (40, 66.7)	55.2 (42.3, 70.9)	6.9	7.3
2014	IGME	48.1 (33.7, 72.3)	48.1 (33.7, 72.3)	14.7	14.7
	Unbenchmarked	54.7 (40.1, 80.6)	54.4 (39.4, 76.4)	10.2	9.4
	FB: Rejection Sampler	49.5 (41.7, 59.1)	47.4 (39.5, 55.7)	4.4	4.1
	FB: MH	49.5 (41.5, 59.2)	47.3 (39.3, 55.8)	4.5	4.3
	Raking	50.2 (35.3, 80.8)	54.3 (39, 74.7)	11.7	9.2
2015	IGME	47.8 (32.3, 74.7)	47.8 (32.3, 74.7)	16.3	16.3
	Unbenchmarked	53.5 (35.7, 90.8)	54.1 (36.1, 82.9)	14.9	12.2
	FB: Rejection Sampler	48 (38.9, 58.4)	46.2 (37.2, 56)	4.9	4.8
	FB: MH	47.9 (39, 58.7)	46.1 (36.9, 56.4)	5.0	4.9
	Raking	50 (31, 101.7)	53.9 (35.6, 81.1)	18.9	12.1
2016	IGME	46.3 (30, 75.6)	46.3 (30, 75.6)	17.8	17.8
	Unbenchmarked	52.3 (31, 106.8)	54 (32.8, 94.2)	21.5	16.5
	FB: Rejection Sampler	46.4 (36, 58)	45.3 (35.2, 56.7)	5.6	5.5
	FB: MH	46.6 (36.3, 58.9)	45.3 (34.6, 57.1)	5.7	5.6
	Raking	49.7 (25.9, 132.1)	53.8 (32.3, 92.1)	30.7	16.7

Year	Model	AL: Median	UL: Median	AL: SD	UL: SD
2017	IGME	44.5 (28, 75.1)	44.5 (28, 75.1)	18.6	18.6
	Unbenchmarked	51.3 (26.3, 127.4)	54.6 (30.1, 112.2)	30.1	23.0
	FB: Rejection Sampler	45.1 (33.1, 58.7)	44.7 (32.9, 58.4)	6.4	6.4
	FB: MH	45.1 (33.4, 59.3)	44.7 (32.2, 58.7)	6.5	6.7
	Raking	49.4 (21.6, 169.1)	54.3 (29.4, 109.8)	46.4	23.6
2018	IGME	43.3 (26.2, 76)	43.3 (26.2, 76)	19.9	19.9
	Unbenchmarked	50.1 (21.7, 152.7)	55.5 (26.9, 135.3)	41.2	31.7
	FB: Rejection Sampler	43.7 (29.7, 59.9)	44.2 (30.7, 60.9)	7.4	7.7
	FB: MH	43.9 (29.8, 60.1)	44.5 (29.6, 61.9)	7.6	8.1
	Raking	49.1 (17, 220.1)	55 (26.7, 133.2)	64.8	33.1
2019	IGME	42.4 (24.8, 77)	42.4 (24.8, 77)	21.1	21.1
	Unbenchmarked	48.8 (18.2, 186.6)	56.4 (24.2, 169.5)	53.6	43.0
	FB: Rejection Sampler	42.5 (26.3, 62.2)	44.2 (28.4, 65.6)	8.9	9.4
	FB: MH	42.6 (26.4, 62.9)	44.4 (27.1, 67.1)	8.9	10.2
	Raking	48.4 (13.5, 284.7)	55.9 (23.8, 164.4)	83.9	45.7

Comparisons between unit-level and area-level models for each admin1 region across time can be found in Figures A.48 through A.51.

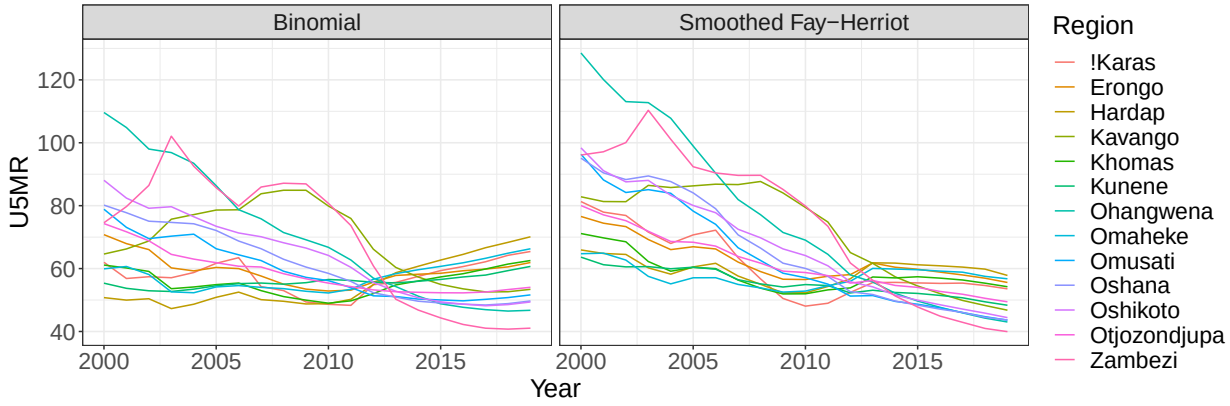


Figure A.48: Comparison of median, unbenchmarked U5MR estimates from unit- and area-level models across time. U5MR is reported as deaths per 1000 live births.

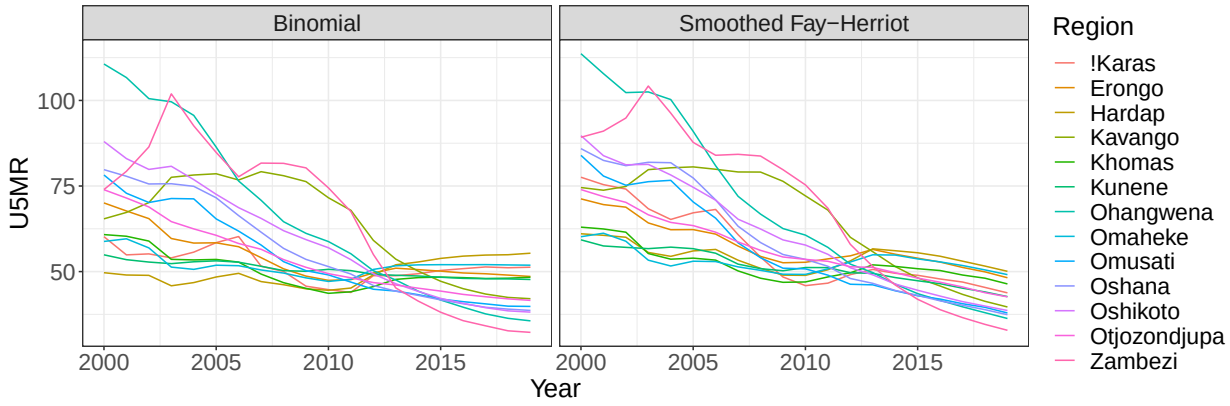


Figure A.49: Comparison of median, benchmarked U5MR estimates from unit- and area-level models across time for the Fully Bayesian Metropolis-Hastings approach. U5MR is reported as deaths per 1000 live births.

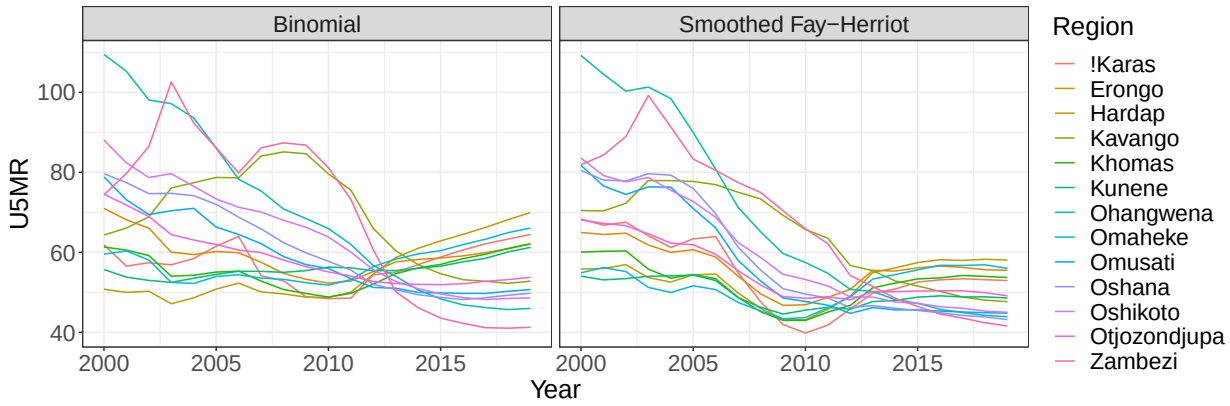


Figure A.50: Comparison of median, benchmarked U5MR estimates from unit- and area-level models across time for the raking approach. U5MR is reported as deaths per 1000 live births.

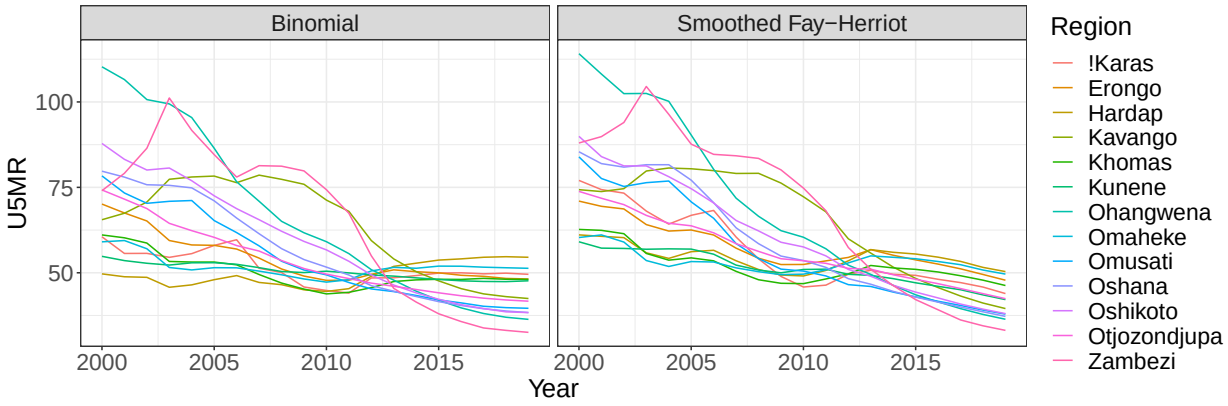


Figure A.51: Comparison of median, benchmarked U5MR estimates from unit- and area-level models across time for the rejection sampler approach. U5MR is reported as deaths per 1000 live births.

Subnational maps of unbenchmarked and benchmarked estimates from 2000 to 2019 can be found in Figures A.52 through A.71.

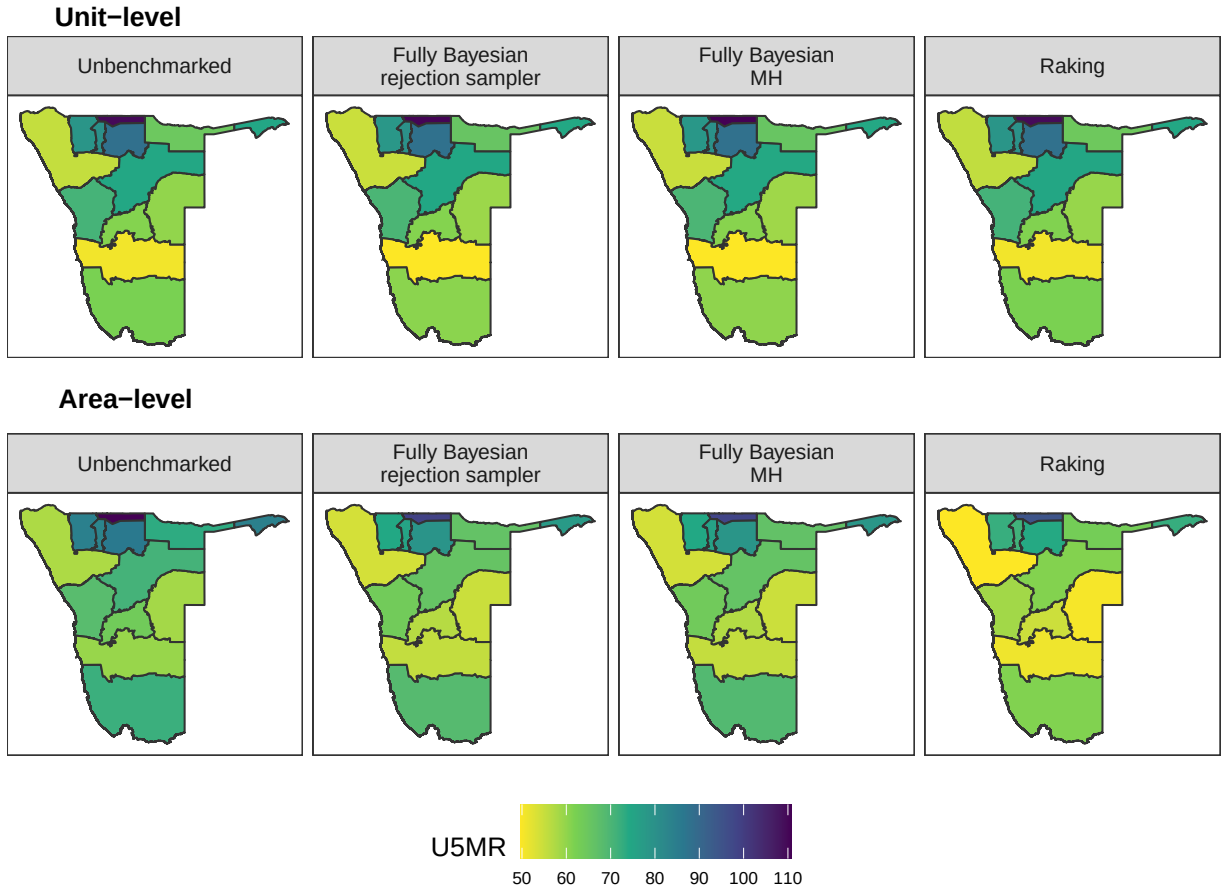


Figure A.52: Comparison of median U5MR estimates from benchmarked and unbenchmarked unit- and area-level models for 2000. U5MR is reported as deaths per 1000 live births.

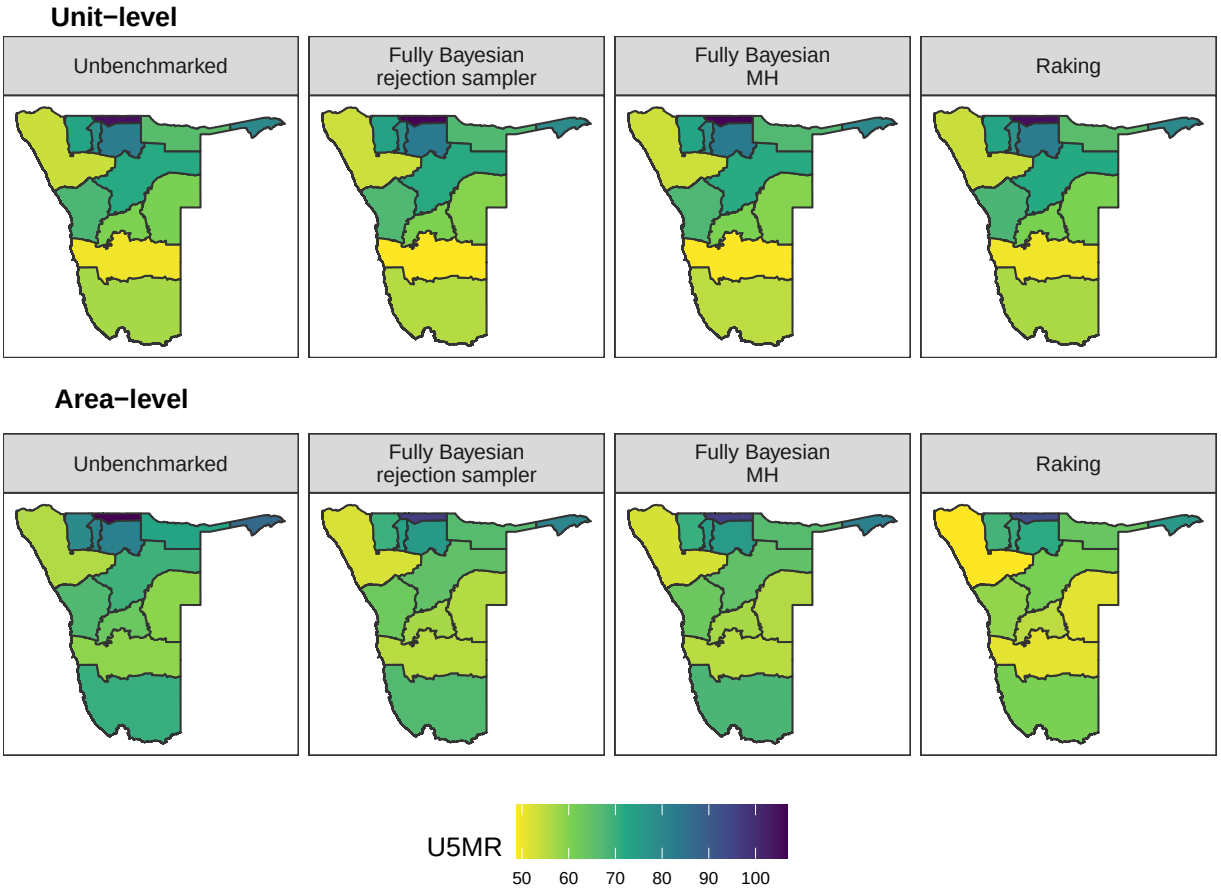


Figure A.53: Comparison of median U5MR estimates from benchmarked and unbenchmarked unit- and area-level models for 2001. U5MR is reported as deaths per 1000 live births.

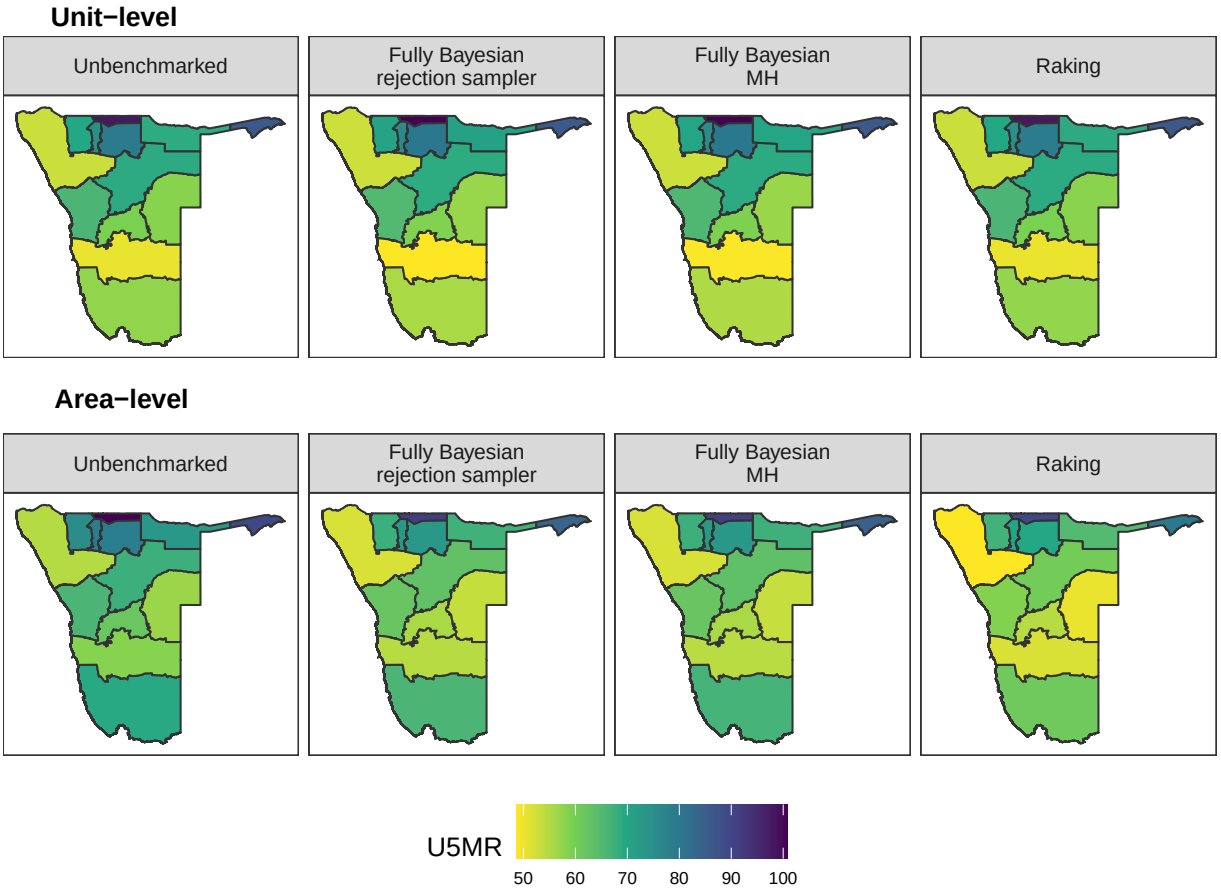


Figure A.54: Comparison of median U5MR estimates from benchmarked and unbenchmarked unit- and area-level models for 2002. U5MR is reported as deaths per 1000 live births.

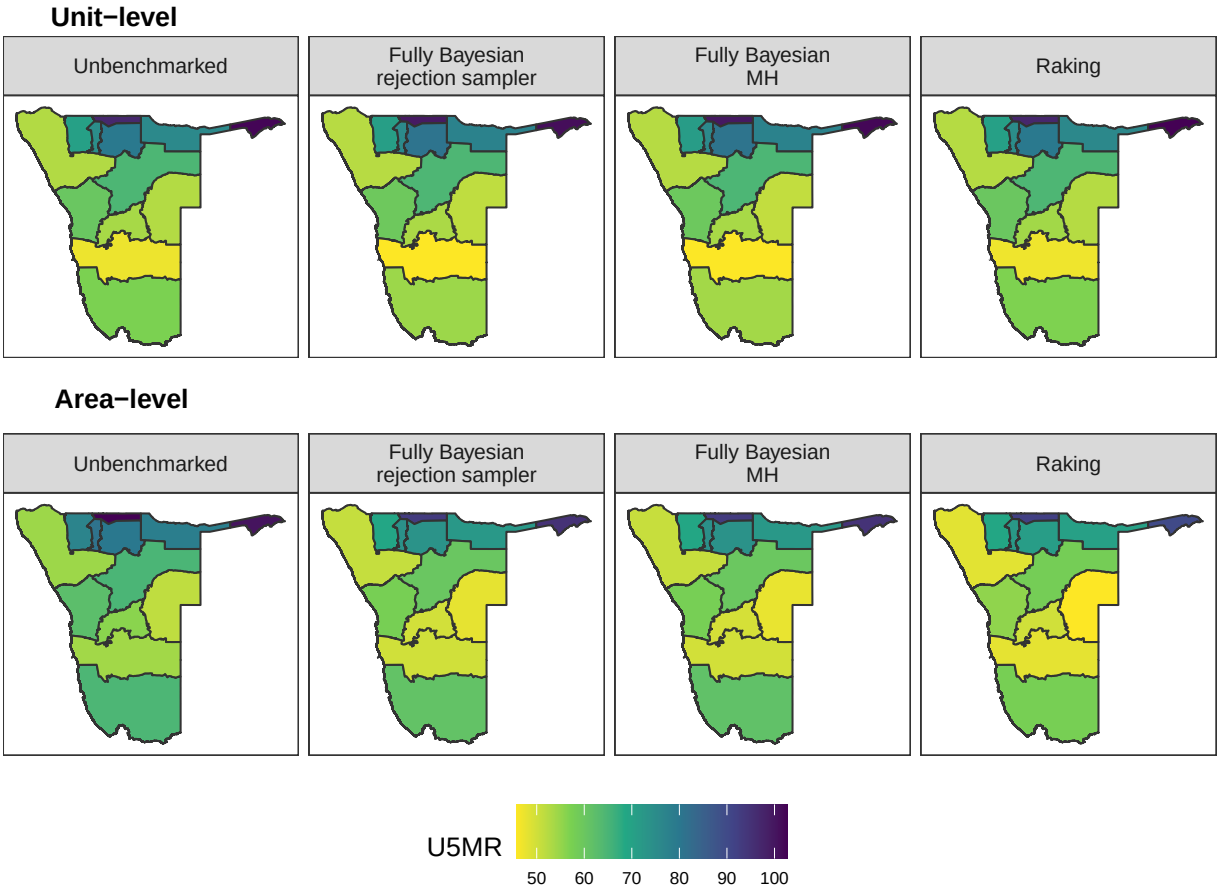


Figure A.55: Comparison of median U5MR estimates from benchmarked and unbenchmarked unit- and area-level models for 2003. U5MR is reported as deaths per 1000 live births.

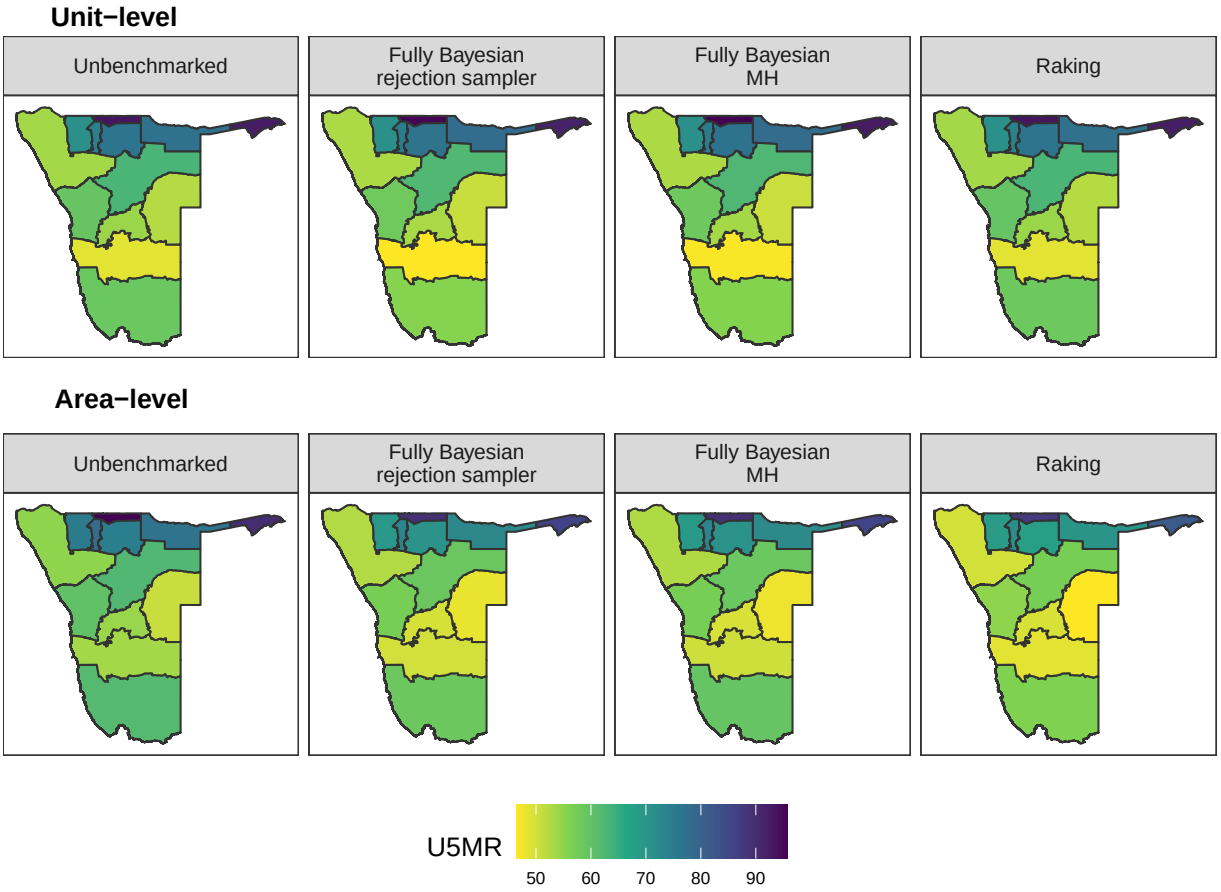


Figure A.56: Comparison of median U5MR estimates from benchmarked and unbenchmarked unit- and area-level models for 2004. U5MR is reported as deaths per 1000 live births.

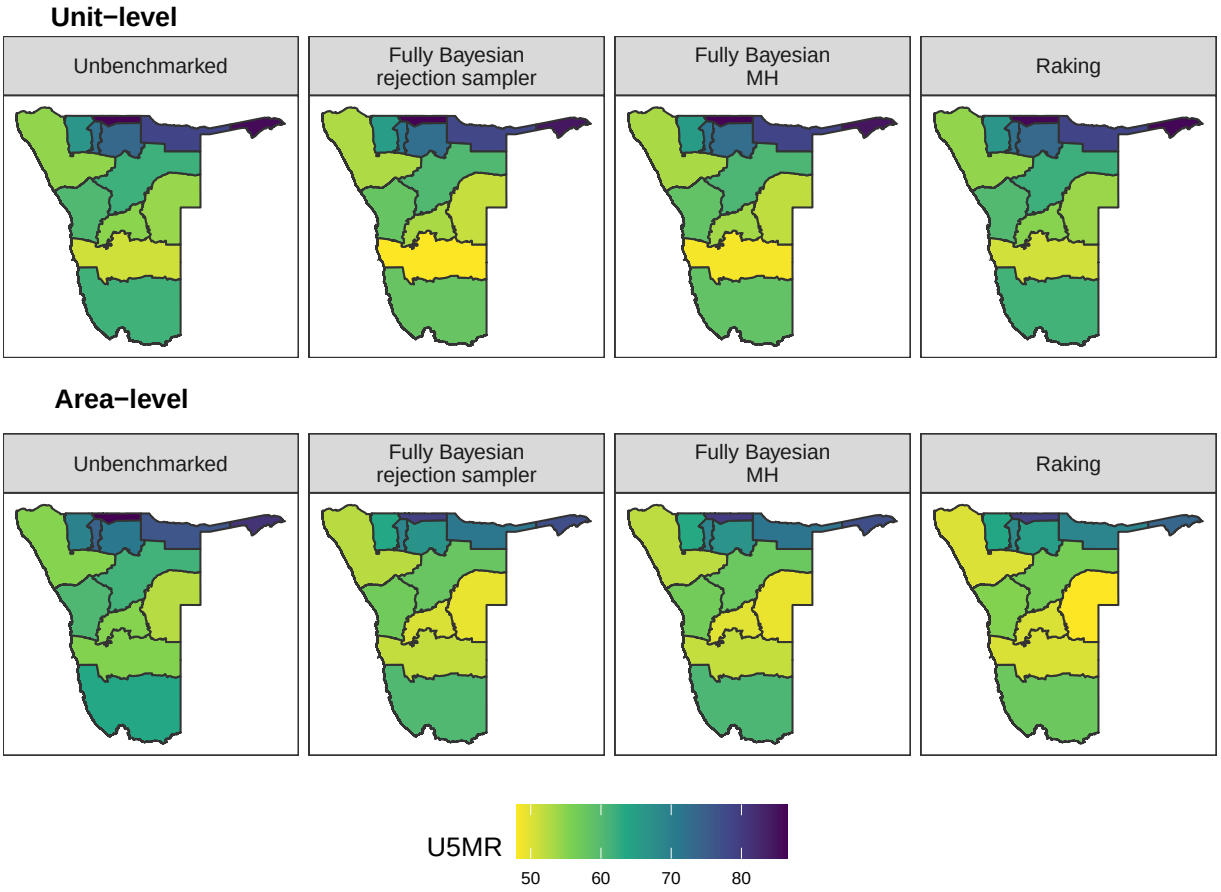


Figure A.57: Comparison of median U5MR estimates from benchmarked and unbenchmarked unit- and area-level models for 2005. U5MR is reported as deaths per 1000 live births.

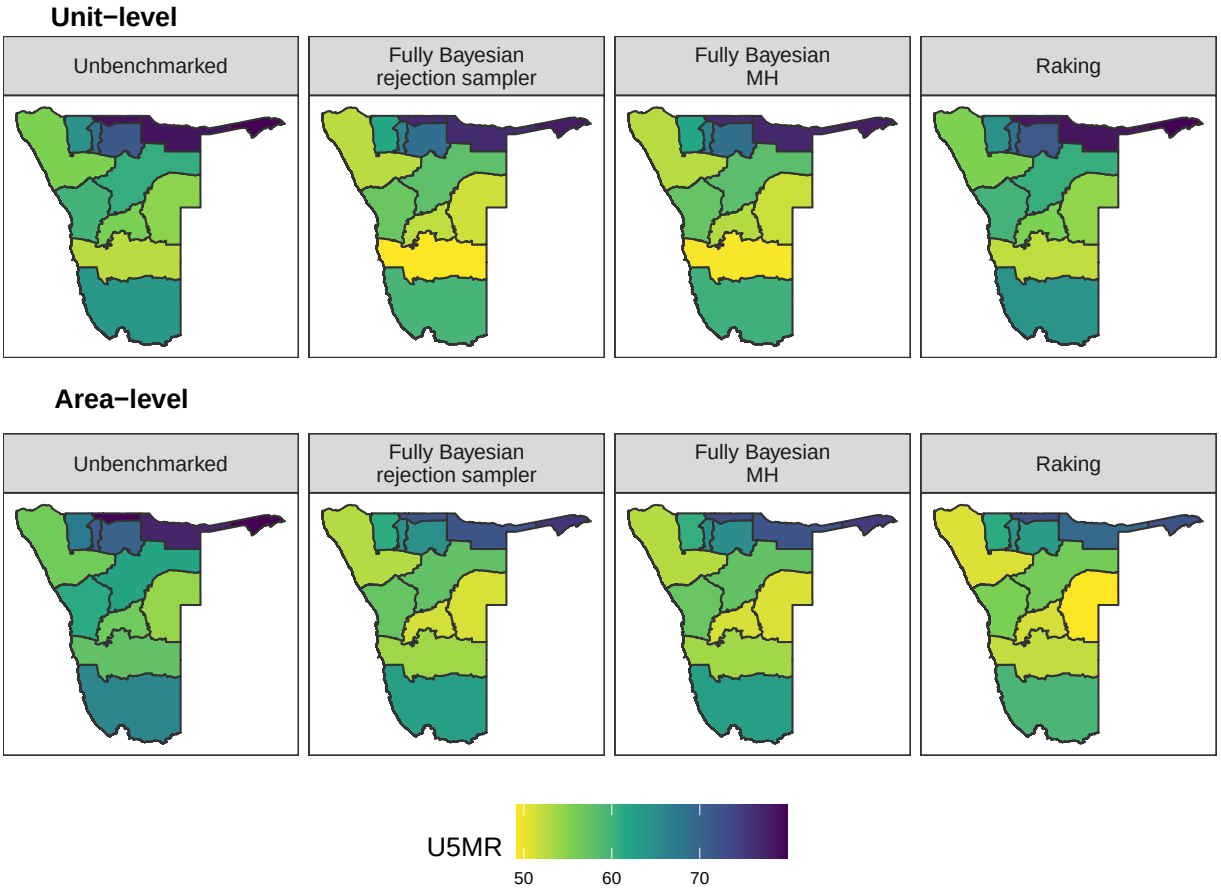


Figure A.58: Comparison of median U5MR estimates from benchmarked and unbenchmarked unit- and area-level models for 2006. U5MR is reported as deaths per 1000 live births.

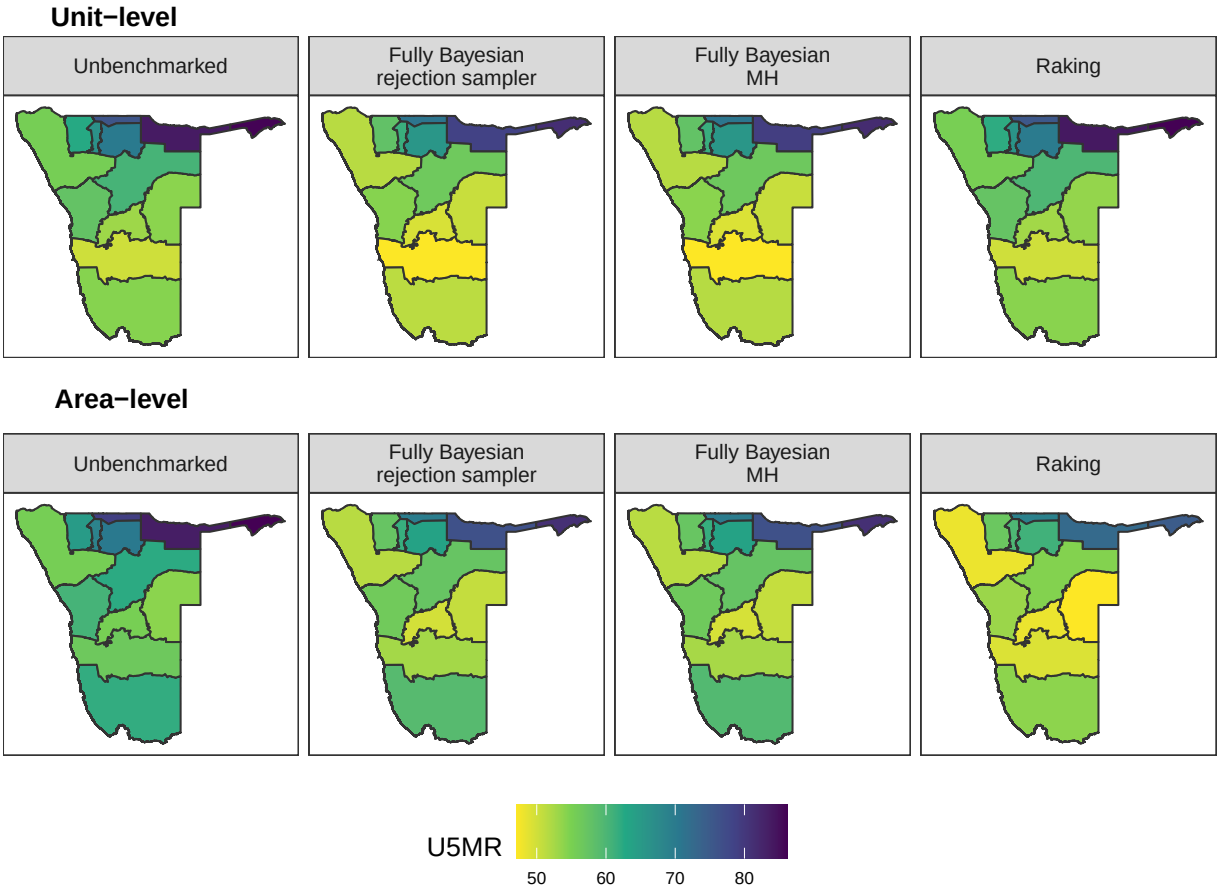


Figure A.59: Comparison of median U5MR estimates from benchmarked and unbenchmarked unit- and area-level models for 2007. U5MR is reported as deaths per 1000 live births.

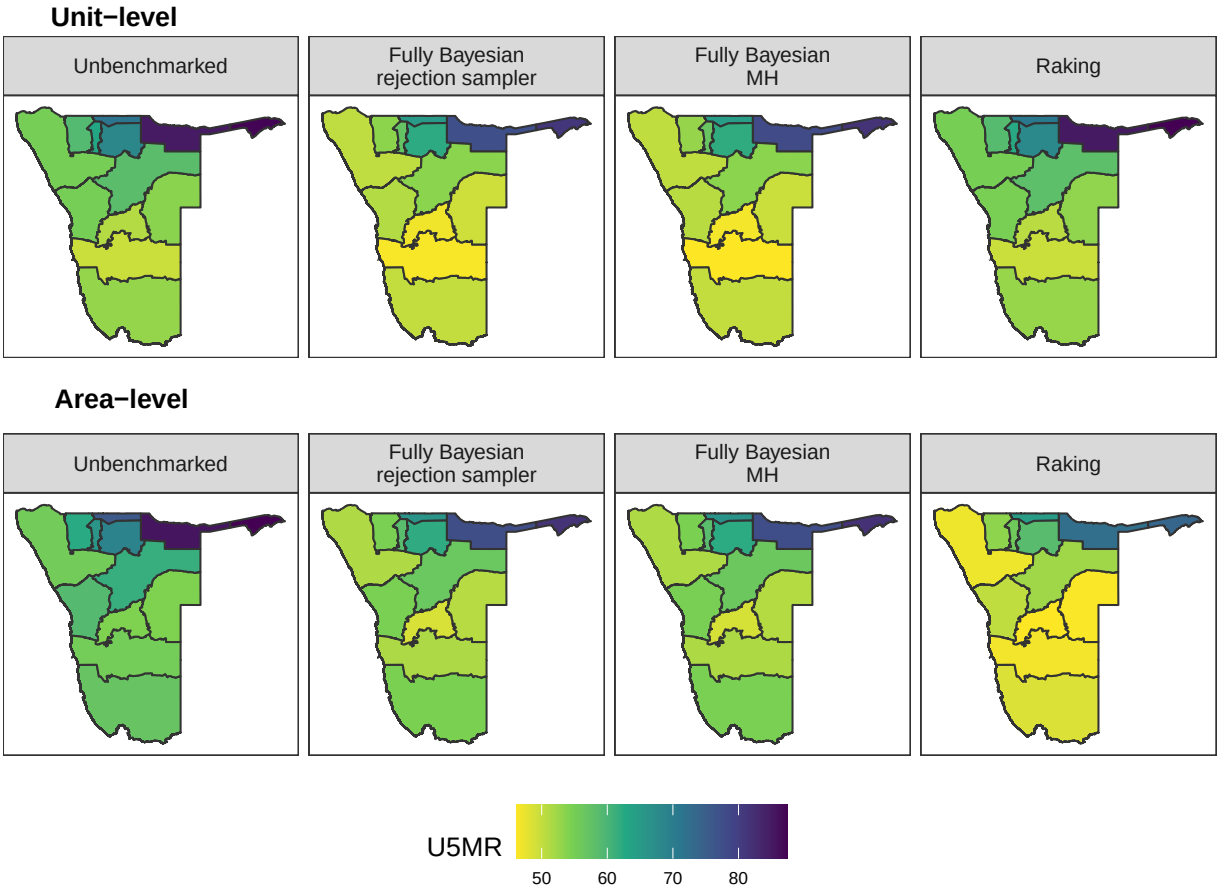


Figure A.60: Comparison of median U5MR estimates from benchmarked and unbenchmarked unit- and area-level models for 2008. U5MR is reported as deaths per 1000 live births.

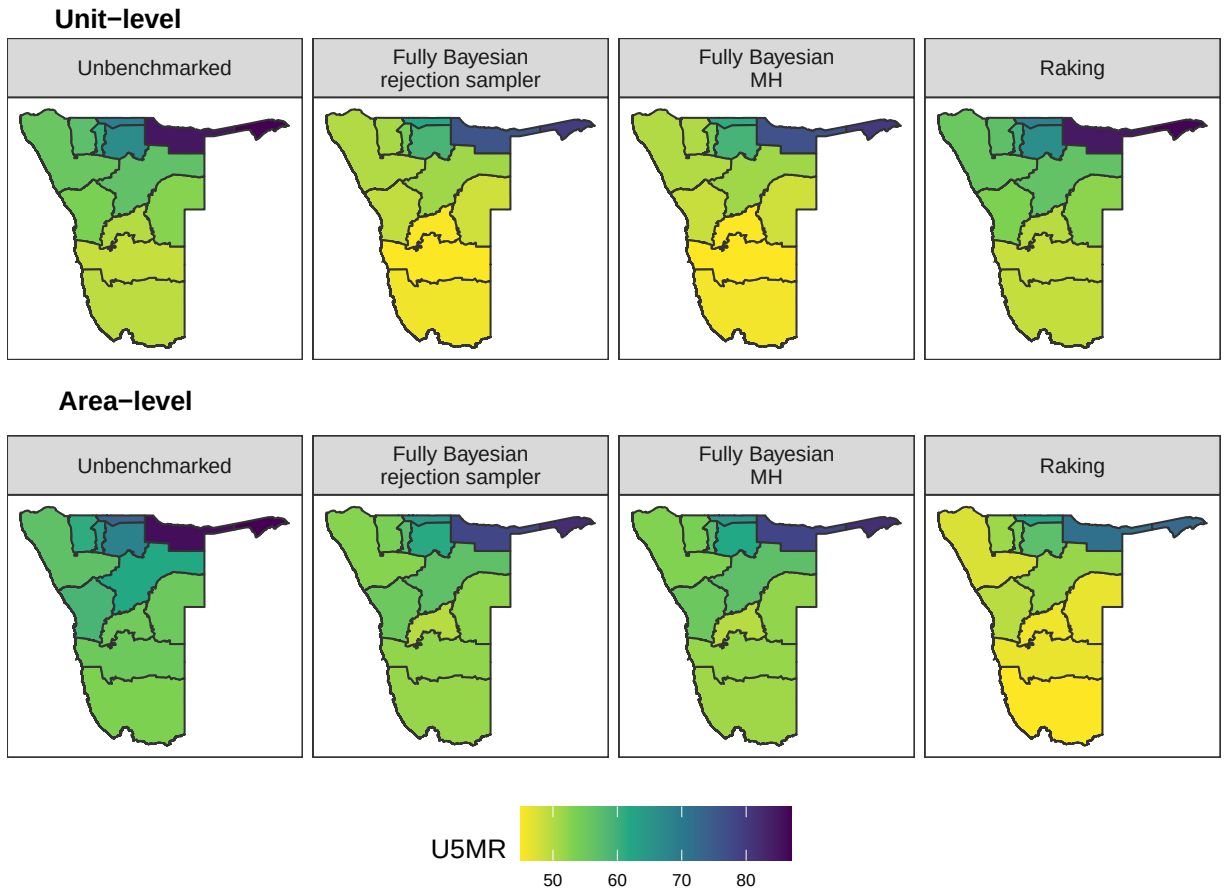


Figure A.61: Comparison of median U5MR estimates from benchmarked and unbenchmarked unit- and area-level models for 2009. U5MR is reported as deaths per 1000 live births.

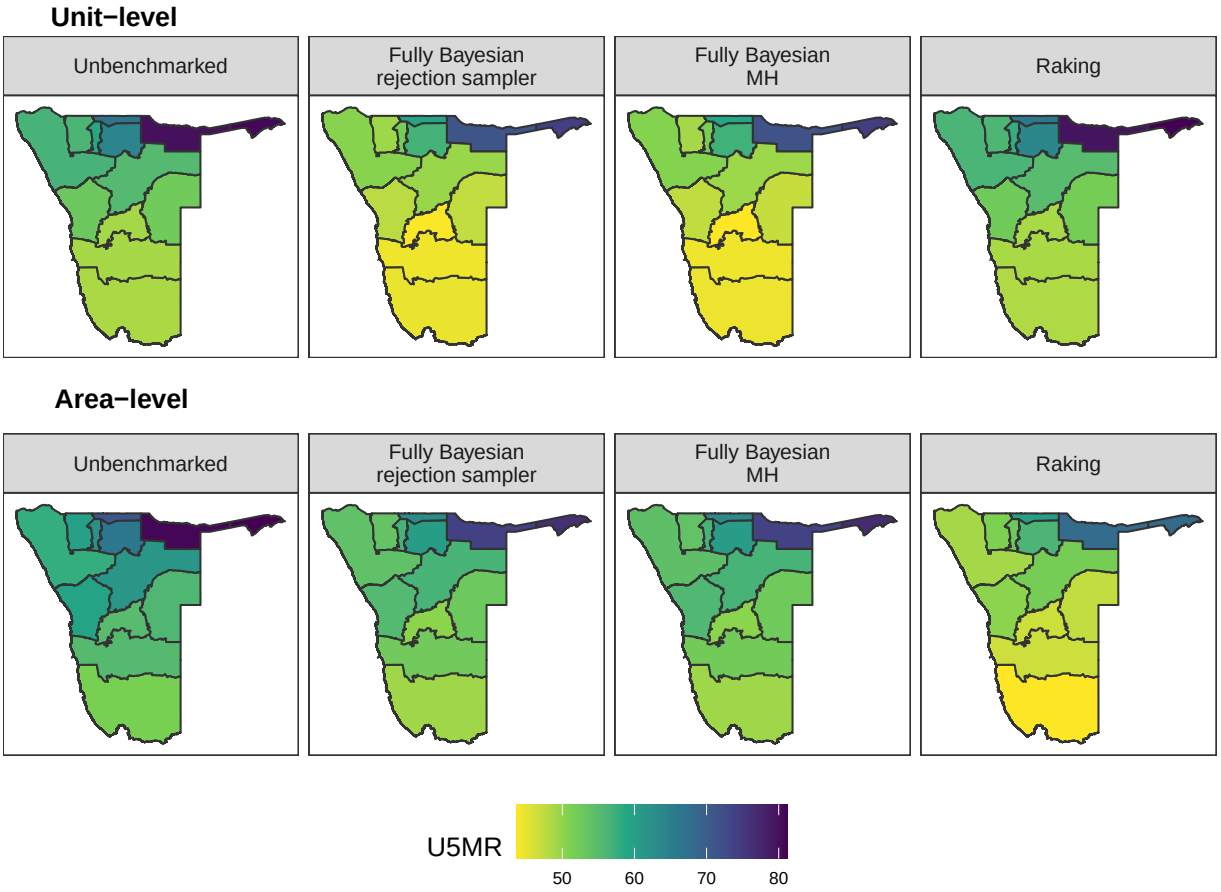


Figure A.62: Comparison of median U5MR estimates from benchmarked and unbenchmarked unit- and area-level models for 2010. U5MR is reported as deaths per 1000 live births.

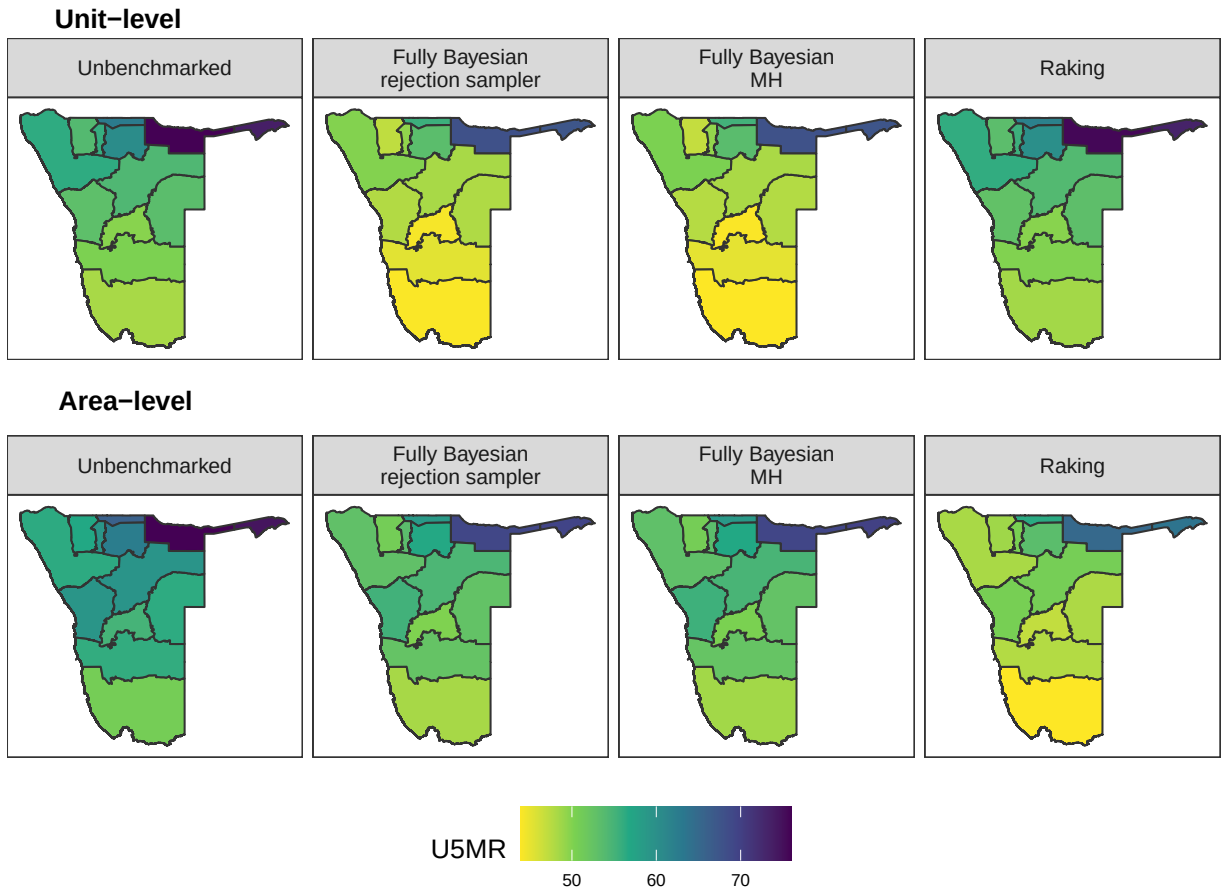


Figure A.63: Comparison of median U5MR estimates from benchmarked and unbenchmarked unit- and area-level models for 2011. U5MR is reported as deaths per 1000 live births.

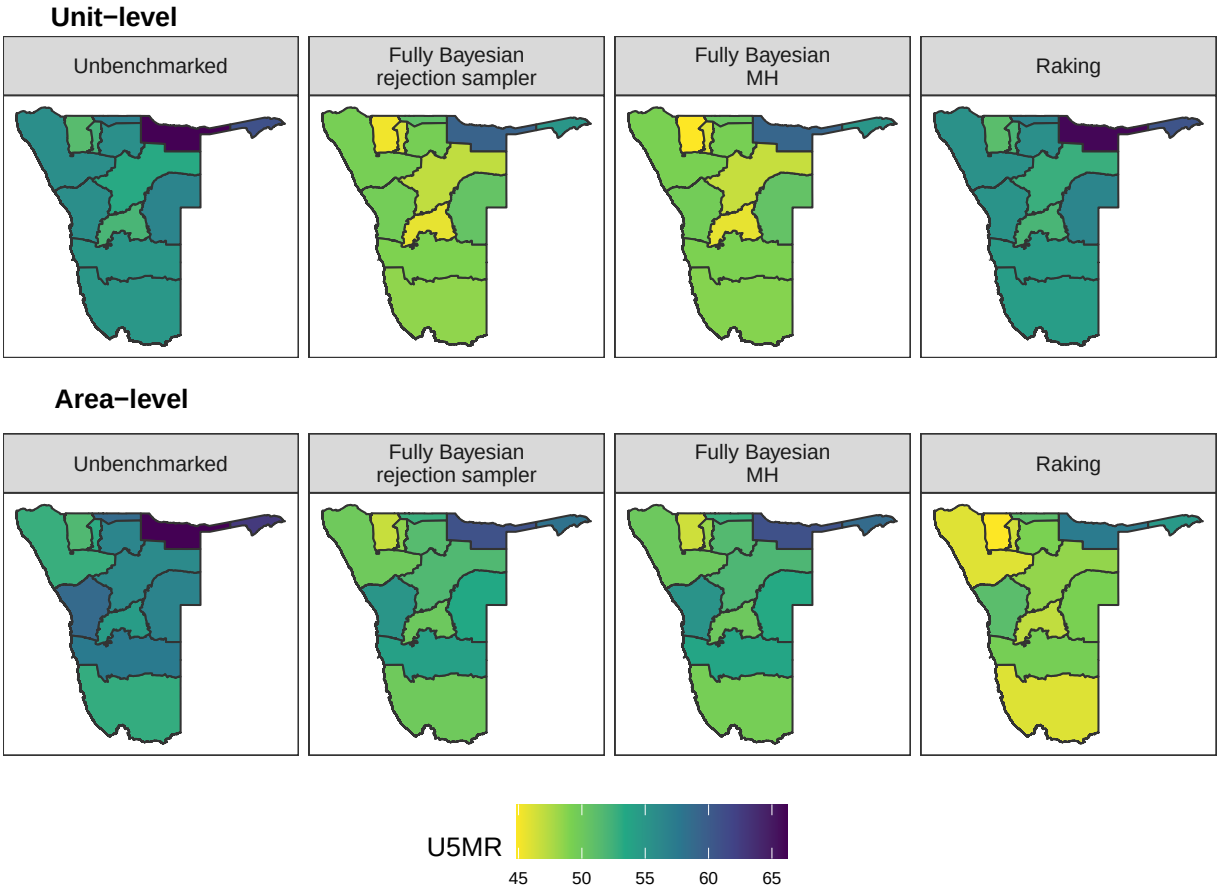


Figure A.64: Comparison of median U5MR estimates from benchmarked and unbenchmarked unit- and area-level models for 2012. U5MR is reported as deaths per 1000 live births.

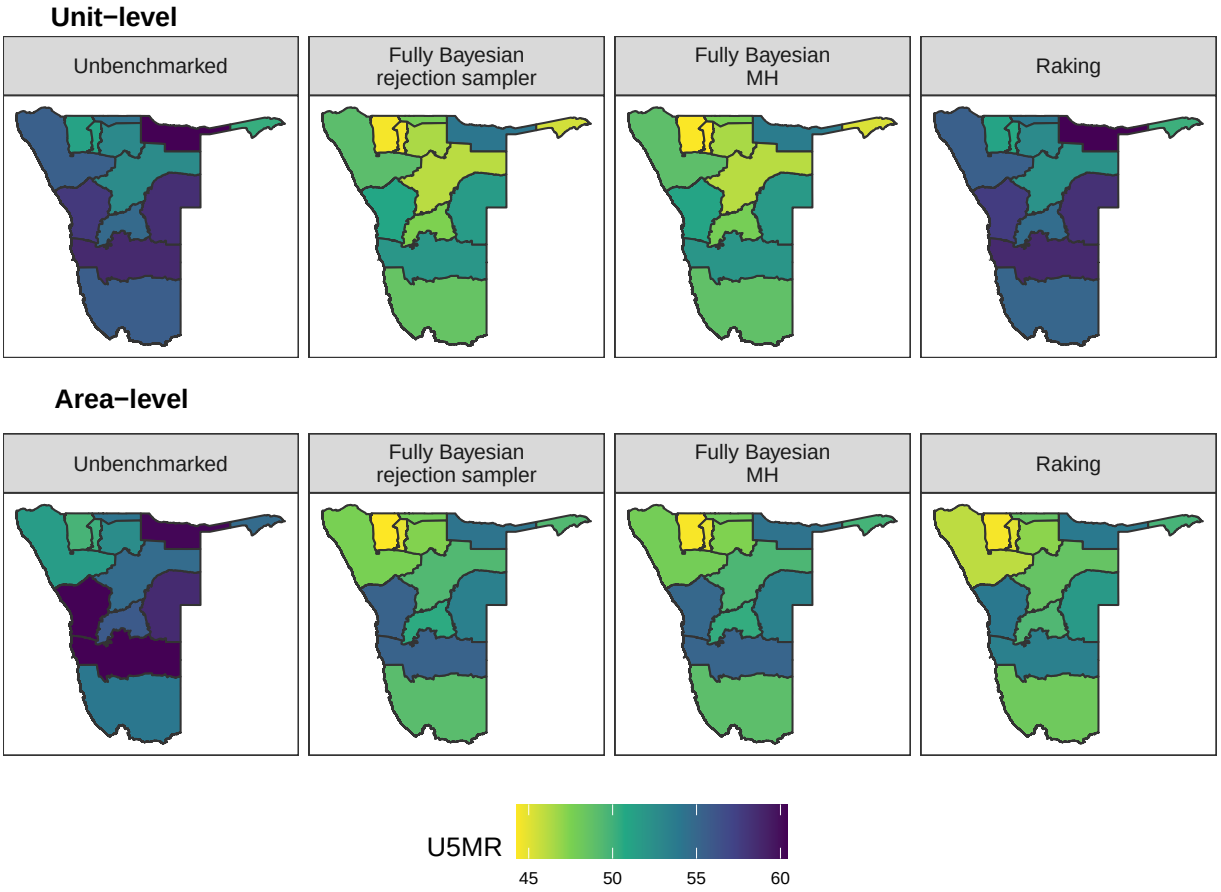


Figure A.65: Comparison of median U5MR estimates from benchmarked and unbenchmarked unit- and area-level models for 2013. U5MR is reported as deaths per 1000 live births.

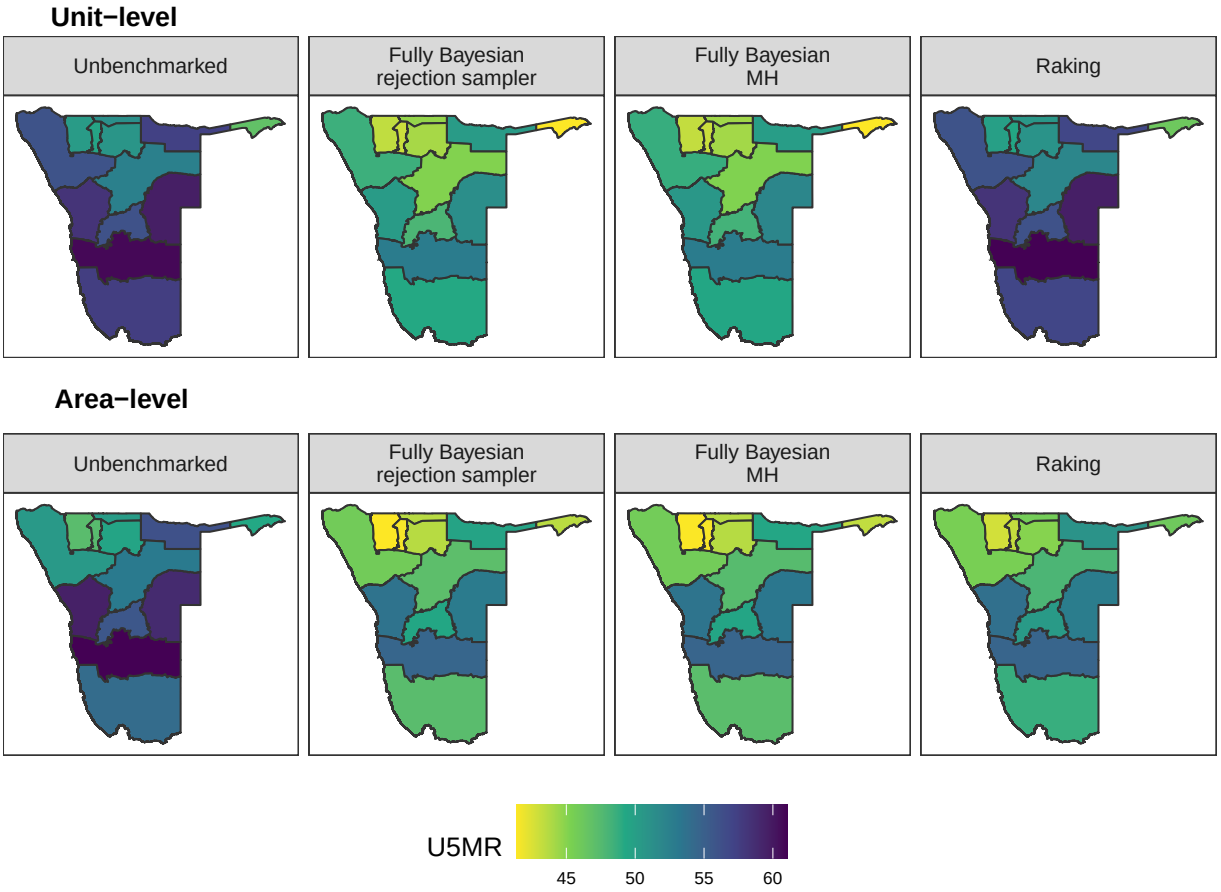


Figure A.66: Comparison of median U5MR estimates from benchmarked and unbenchmarked unit- and area-level models for 2014. U5MR is reported as deaths per 1000 live births.

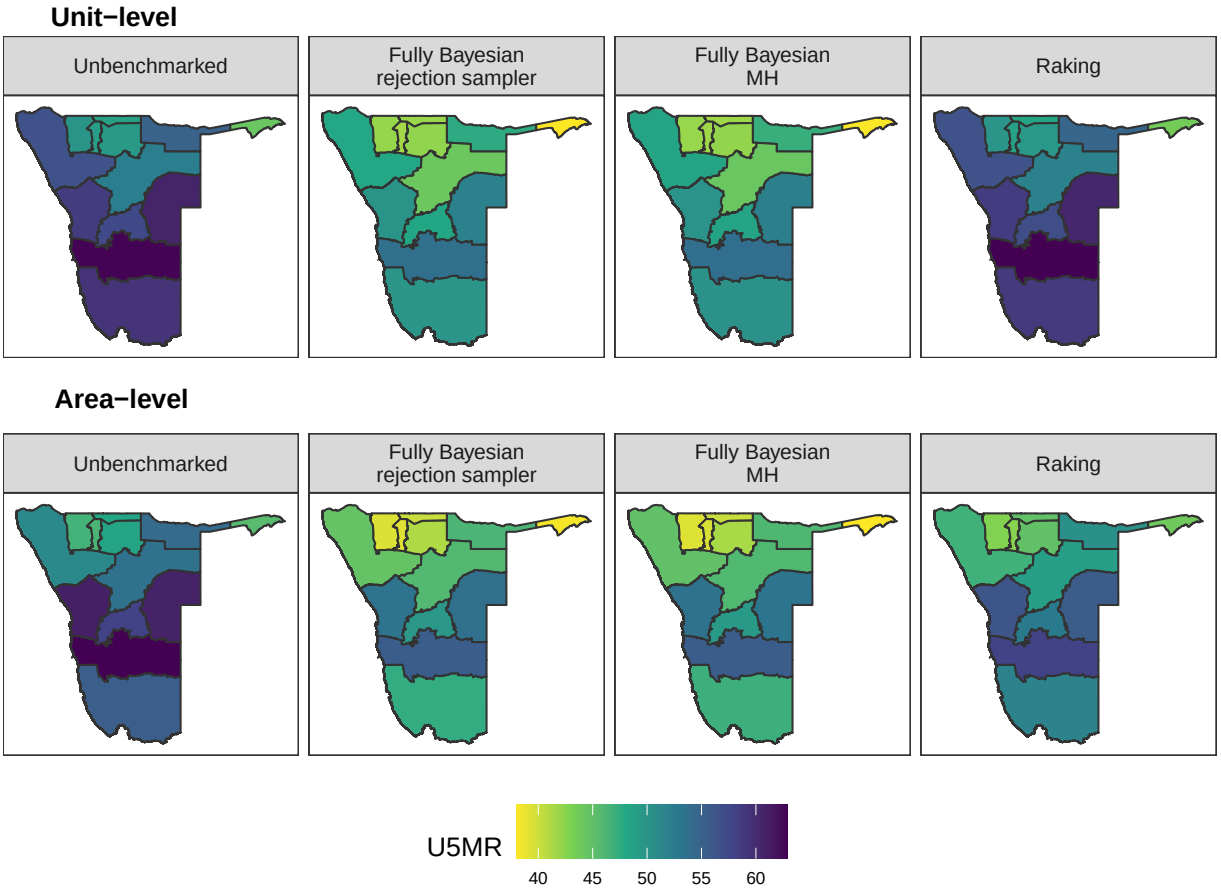


Figure A.67: Comparison of median U5MR estimates from benchmarked and unbenchmarked unit- and area-level models for 2015. U5MR is reported as deaths per 1000 live births.

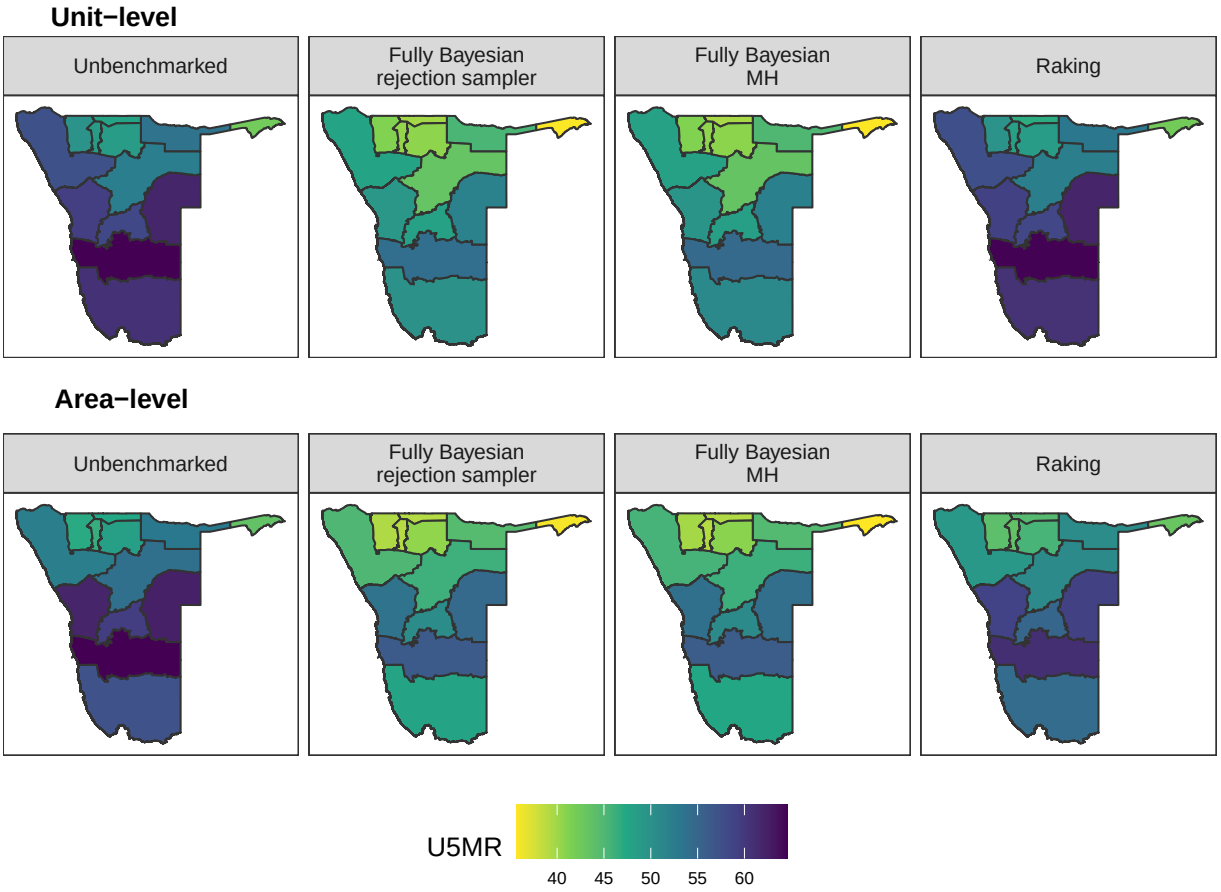


Figure A.68: Comparison of median U5MR estimates from benchmarked and unbenchmarked unit- and area-level models for 2016. U5MR is reported as deaths per 1000 live births.

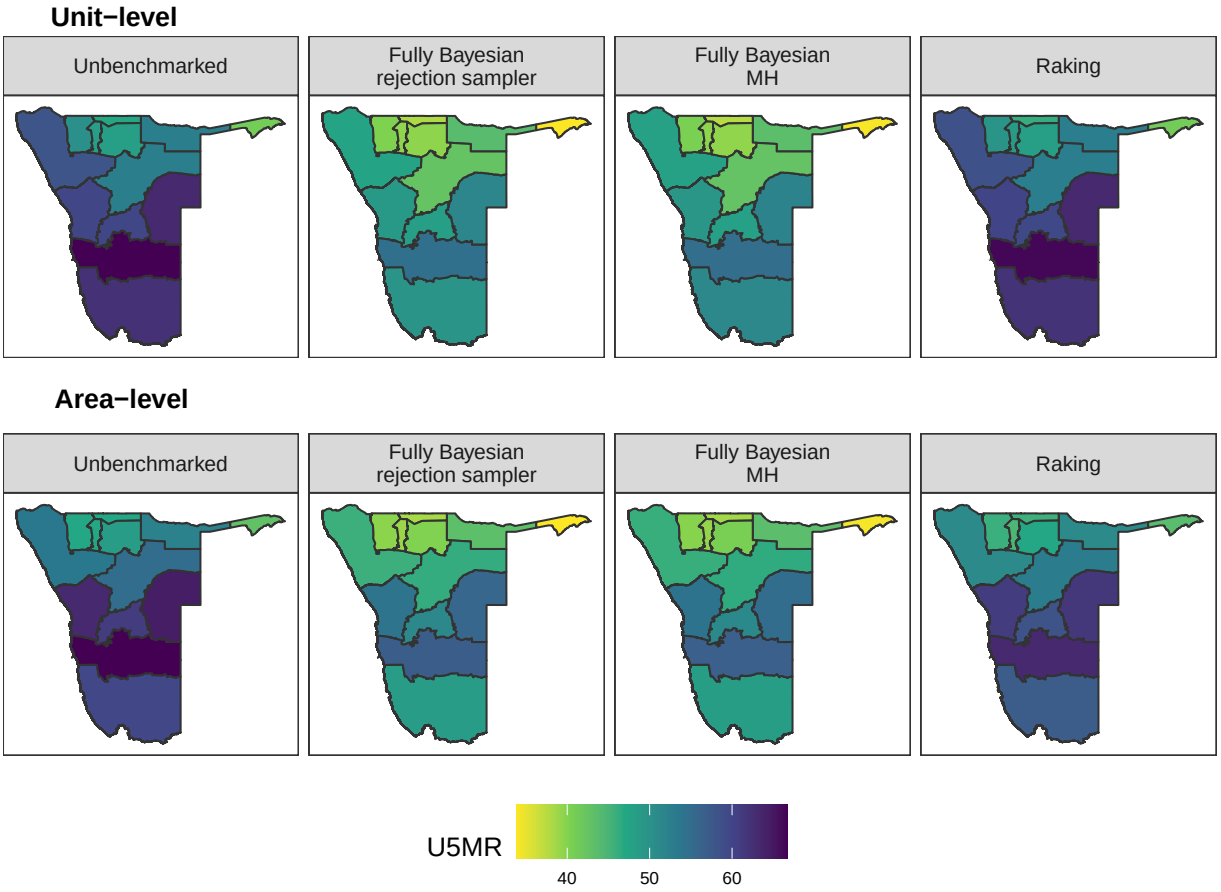


Figure A.69: Comparison of median U5MR estimates from benchmarked and unbenchmarked unit- and area-level models for 2017. U5MR is reported as deaths per 1000 live births.

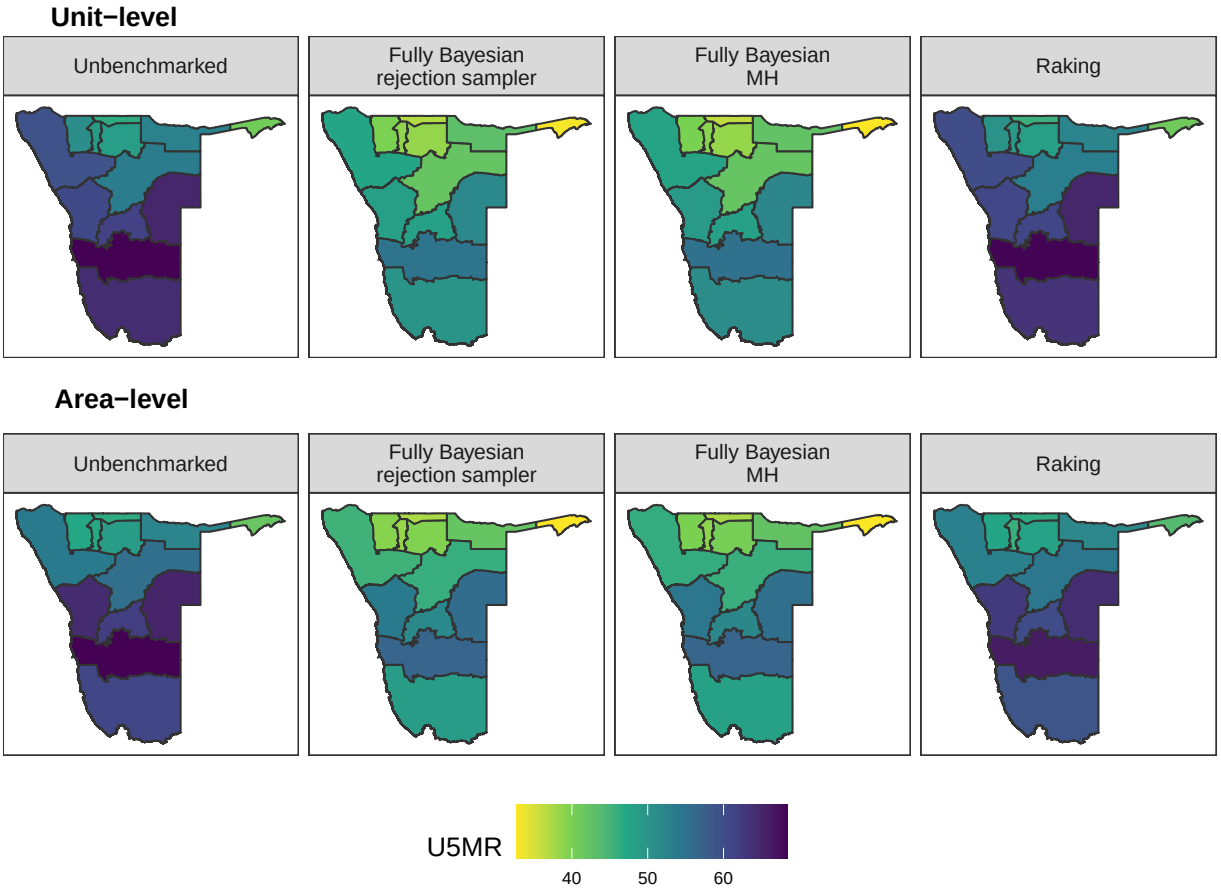


Figure A.70: Comparison of median U5MR estimates from benchmarked and unbenchmarked unit- and area-level models for 2018. U5MR is reported as deaths per 1000 live births.

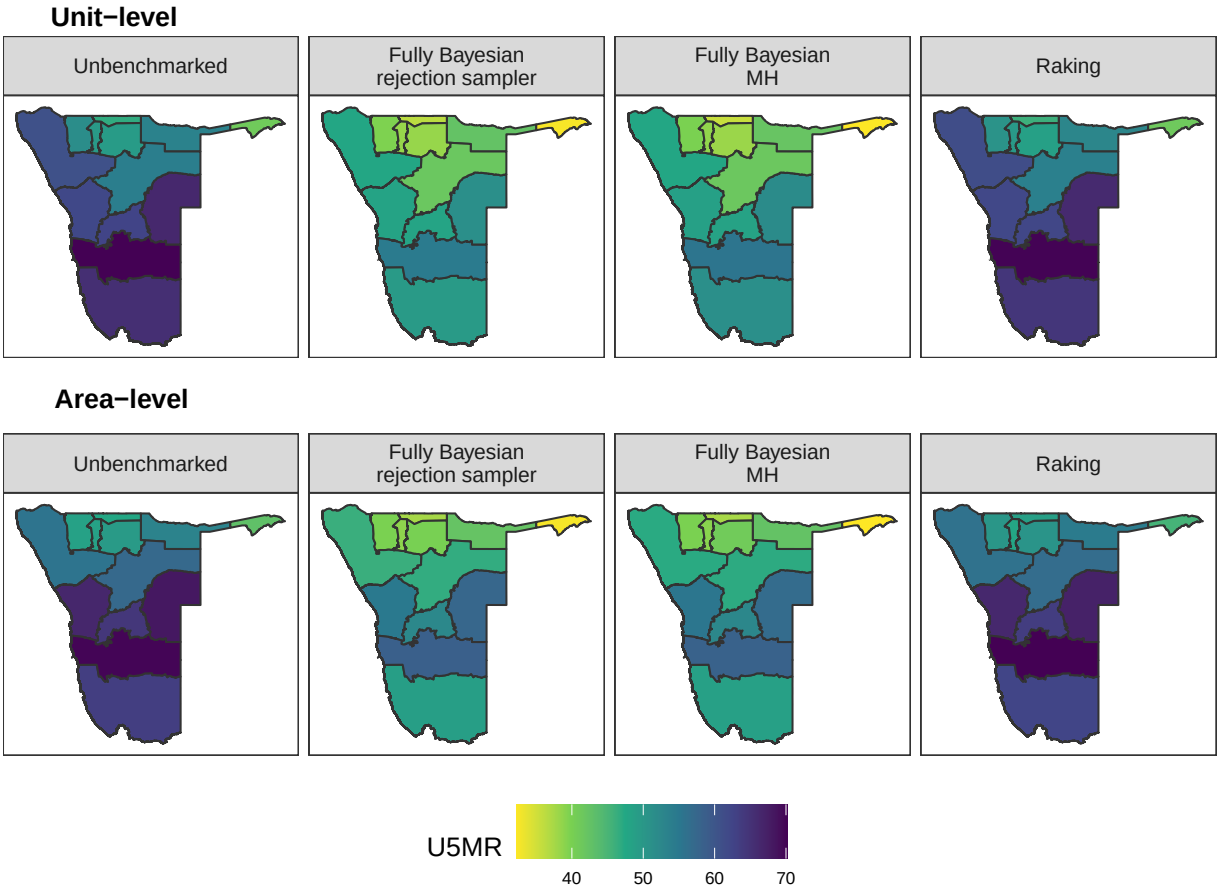


Figure A.71: Comparison of median U5MR estimates from benchmarked and unbenchmarked unit- and area-level models for 2019. U5MR is reported as deaths per 1000 live births.

A.2.6 Model Validation

Below we display the model validation results for the HIV application for both unit- and area-level models. We compare posterior medians of the predictive distribution in each area (having left that area out of model fitting) to the direct estimate in each area, respectively.

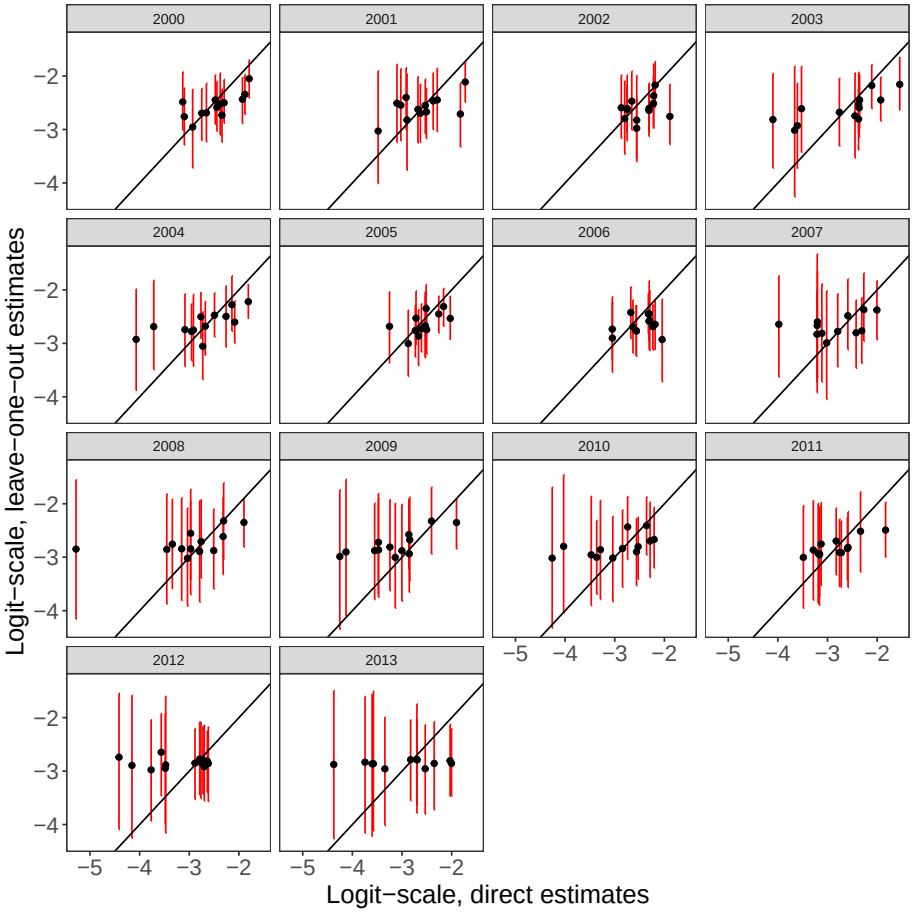


Figure A.72: Scatterplot comparing leave-one-out posterior predictive estimates and direct estimates on the logit scale for the unit-level U5MR model, with 80% confidence intervals.

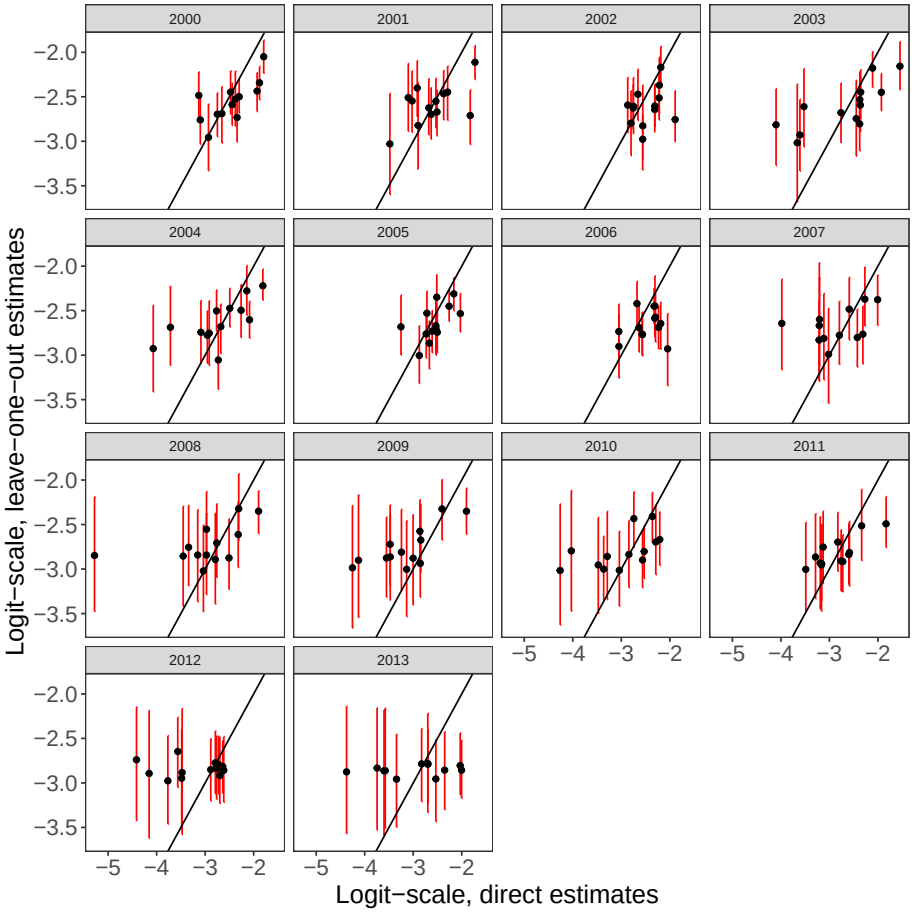


Figure A.73: Scatterplot comparing leave-one-out posterior predictive estimates and direct estimates on the logit scale for the unit-level U5MR model, with 50% confidence intervals.

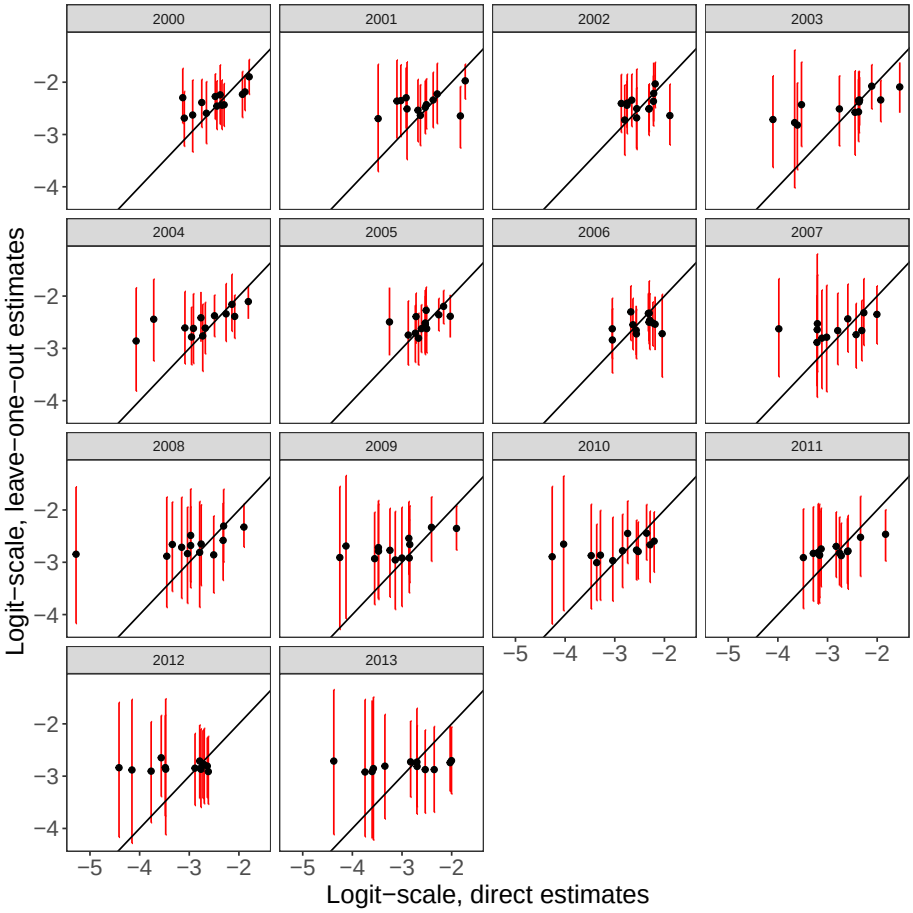


Figure A.74: Scatterplot comparing leave-one-out posterior predictive estimates and direct estimates on the logit scale for the area-level U5MR model, with 80% confidence intervals.

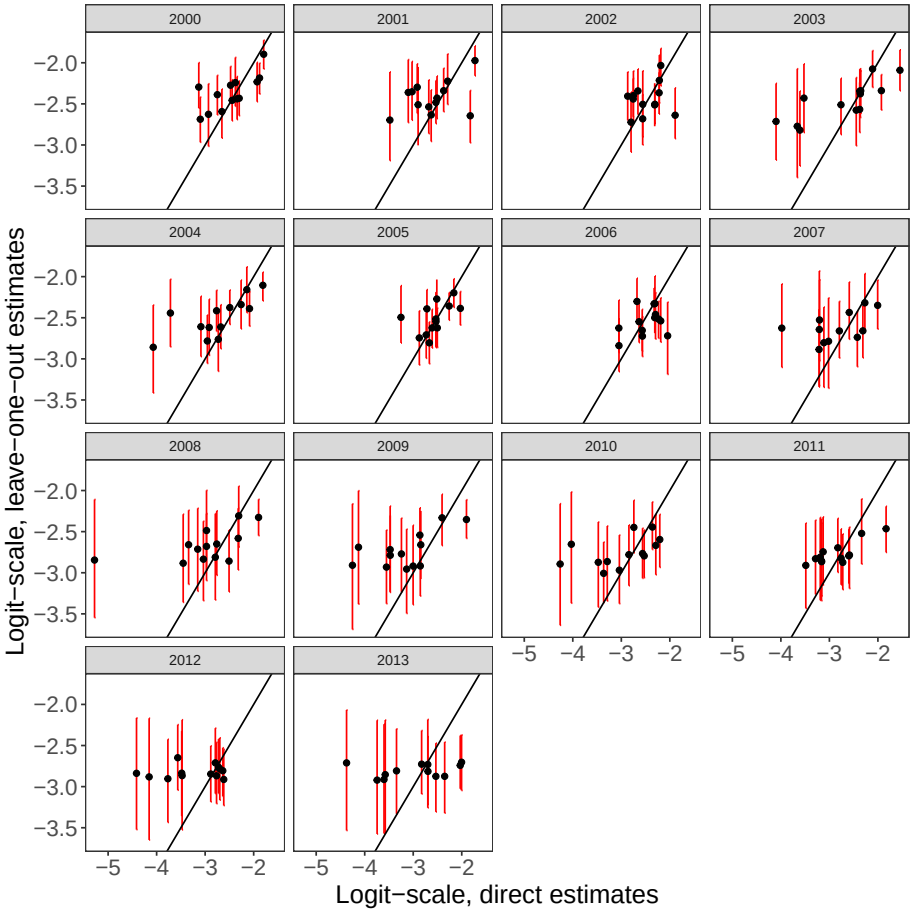


Figure A.75: Scatterplot comparing leave-one-out posterior predictive estimates and direct estimates on the logit scale for the area-level U5MR model, with 50% confidence intervals.

A.3 Simulation

To demonstrate the improved computational speed and flexibility of our approach compared to Zhang and Bryant (2020), we compare run-times for our approach used with INLA to that of Zhang and Bryant (2020) implemented in STAN; a probabilistic programming language that uses a variant of Hamiltonian Monte Carlo to do full Bayesian inference (Carpenter et al., 2017). We chose not to compare the computational speed of our approach to that of Datta et al. (2011) or the raking approach, as both of those benchmarking approaches target different benchmarked estimates than our proposed method. Note however that both the rak-

ing and benchmarked Bayes estimate approach will be the fastest approach to benchmarking in general, as they involve a very quick step to adjusting unbenchmarked draws/estimates, and do not rely on acceptance rates (as our methods do to achieve a reasonable number of effective samples) or MCMC methods.

We simulate unit-level (cluster) binomial observations, using the nine provinces of South Africa as our spatial structure. For each simulation setting, binomial probabilities were given by $p_i = 0.28, 0.29, \ldots, 0.35, 0.36$, with 100 binomial trials in each cluster. Equal probability weights were given to each province. We varied: the number of samples taken in each area, $n = \{5, 10, 100, 1000\}$, the national-level benchmark, $y_2 = \{0.29, 0.3\}$, and the variance of the national-level benchmark, $\sigma^2_{y_2} = \{0.01, 0.0001\}$. For each simulation setting, we generated 10 unique datasets using the given parameters.

The unbenchmarked model we fit to the generated data is the unit-level model used later in our application to South Africa, where we have binomial observations for clusters c within area i. Let $y_{1i[c]}$ be the number of cases observed in $n_{i[c]}$ total observations in cluster c within area i. We consider the unbenchmarked, unit-level model

$$\begin{aligned} y_{1i[c]} \mid n_{i[c]}, \theta_{i[c]} &\sim \text{Binomial}(n_{i[c]}, \theta_{i[c]}), \\ \eta_{i[c]} = \text{logit}(\theta_{i[c]}) &= \beta_0 + b_i + e_{i[c]} \end{aligned}$$

where $\theta_{i[c]}$ is case prevalence in area i and cluster c, β_0 is an intercept term, b_i follows a BYM2 model (Riebler et al., 2016), denoted $b_i \sim \text{BYM2}(\tau_b, \phi)$, and $e_i[c] \overset{iid}{\sim} \text{N}(0, \sigma^2_e)$ is an iid cluster-level random effect. We used hyperpriors $\phi \sim \text{Beta}(0.5, 0.5)$, $\tau_b = \text{PC}(U = 1, \alpha = 0.01)$, $1/\sigma^2_e \sim \text{loggamma}(0.1, 0.1)$ in model fitting. Note that although the model we fit was not used to generate the data, as we are interested only in comparing run times from each method, and ensuring that the benchmarked distributions are the same from each method, this does not affect the validity of our simulation results.

The modified unbenchmarked model, used for the Metropolis-Hastings algorithm is the same as above with the exception of the prior for the intercept being $\pi^+(\beta_0) \overset{d}{=}$

$\text{N}(\text{logit}(y_2), \sqrt{0.1})$, where y_2 is the national-level benchmark. The benchmarked model we fit using the Zhang and Bryant (2020) approach includes the additional likelihood

$$y_2 \mid \boldsymbol{\theta}, \sigma^2_{y_2} \sim \text{N}\left(\bar{\boldsymbol{\theta}}, \sigma^2_{y_2}\right),$$

and the proposed rejection sampler and Metropolis-Hastings algorithm approaches are carried out as described in subsections 3.4.1 and 3.4.2. Code for reproducing the simulation can be found at github.com/taylorokonek/benchmarking-paper-sim.

To fairly compare run-times, we compare the time it takes to:

1. Fit the fully Benchmarked model per Zhang and Bryant (2020) in STAN, and obtain a bulk-ESS of 1000.

2. Fit the unbenchmarked model in INLA, draw posterior samples, and obtain 1000 accepted samples using the rejection sampler.

3. Fit the modified unbenchmarked model in INLA, draw posterior samples, and obtain a bulk-ESS of 1000.

For both Stan and the Metropolis-Hastings algorithm, we run four chains with 1000 burn-in samples each, and the appropriate number of samples after to obtain the desired bulk-ESS.

As expected, our proposed approaches outperform that of Zhang and Bryant (2020) in most simulation settings, as noted in Figures A.76 and A.77. In particular, the amount of time needed to obtain 1000 samples from the benchmarked posterior distribution is much lower for both of our proposed approaches than that of Zhang and Bryant (2020) when the number of samples in each area is large. When the national variance is smaller, as in Figure A.77, the Zhang and Bryant (2020) tends to outperform both the rejection sampler and MH algorithm at low sample sizes (5 and 10 per area), but not at larger sample sizes. Finally, note that, in settings with smaller national variance, the Metropolis-Hastings algorithm outperforms the rejection sampler approach in terms of computational speed, while the reverse

is true when national variance is large. The speed of the MH algorithm could potentially be optimized by modifying prior for the intercept, $\pi^{+}(\beta_0)$, and we suggest that when possible, multiple priors should be tested if the acceptance rate in the MH algorithm is lower than desired.

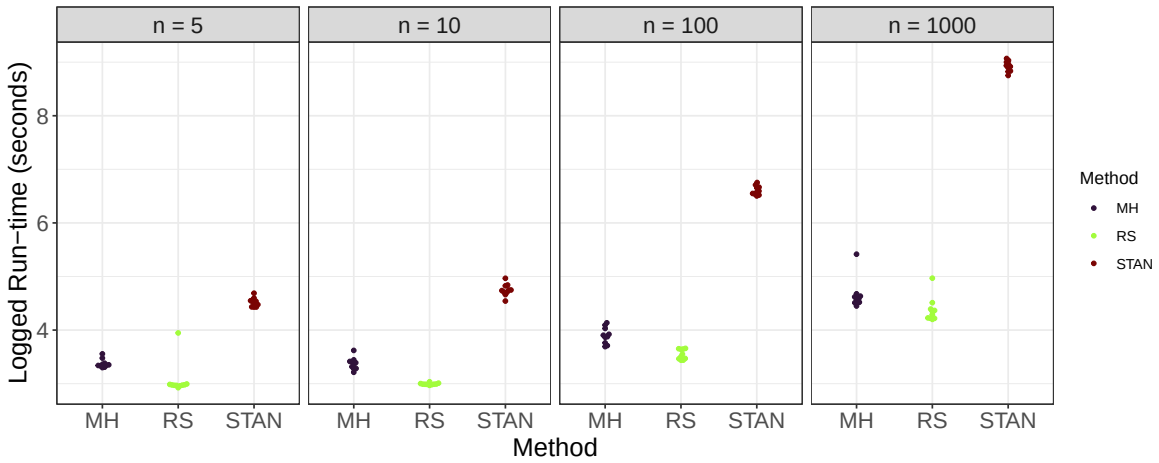


Figure A.76: Comparative, total run-time (in seconds) needed to obtain 1000 samples from the rejection sampling (RS) approach, or 1000 bulk-ESS from the Metropolis-Hastings (MH) approach or the approach of Zhang and Bryant (2020) (STAN). Each beeswarm plot contains 10 observations, with data generated under the given simulation setting for 10 different seeds. Setting: $y_2 = 0.29$, $\sigma^2_{y_2} = 0.01$

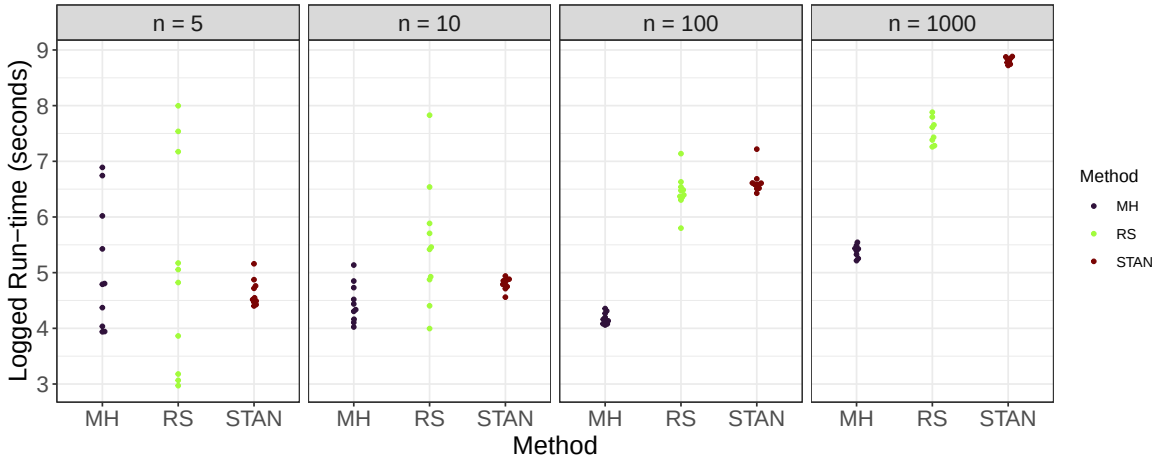


Figure A.77: Comparative, total run-time (in seconds) needed to obtain 1000 samples from the rejection sampling (RS) approach, or 1000 bulk-ESS from the Metropolis-Hastings (MH) approach or the approach of Zhang and Bryant (2020) (STAN). Each beeswarm plot contains 10 observations, with data generated under the given simulation setting for 10 different seeds. Setting: $y_2 = 0.29$, $\sigma^2_{y_2} = 0.0001$

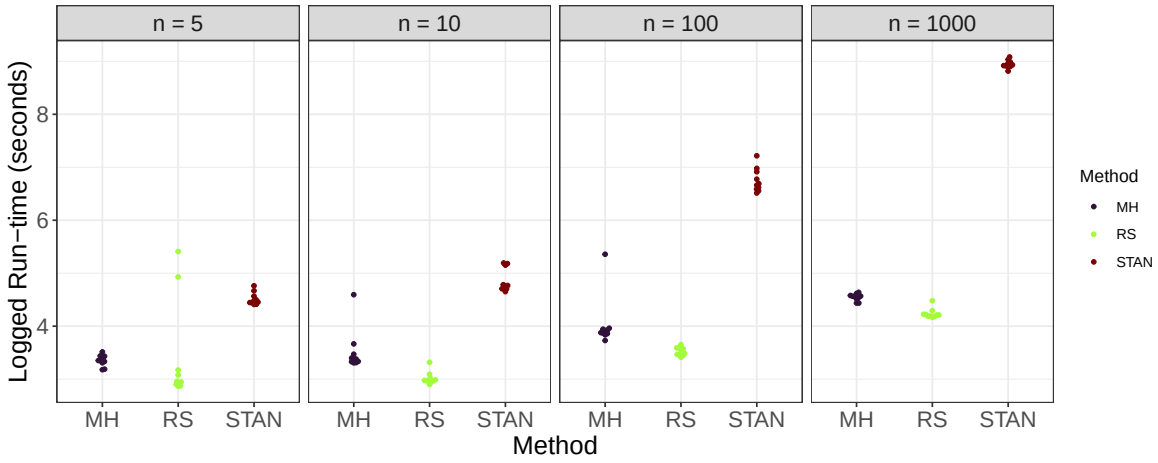


Figure A.78: Comparative, total run-time (in seconds) needed to obtain 1000 samples from the rejection sampling (RS) approach, or 1000 bulk-ESS from the Metropolis-Hastings (MH) approach or the approach of Zhang and Bryant (2020) (STAN). Each beeswarm plot contains 10 observations, with data generated under the given simulation setting for 10 different seeds. Setting: $y_2 = 0.30$, $\sigma^2_{y_2} = 0.01$

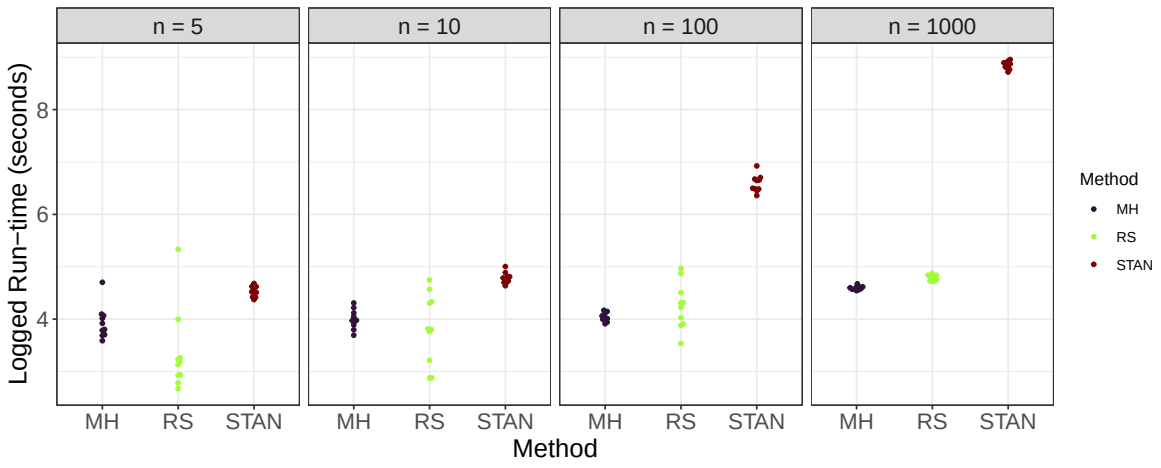


Figure A.79: Comparative, total run-time (in seconds) needed to obtain 1000 samples from the rejection sampling (RS) approach, or 1000 bulk-ESS from the Metropolis-Hastings (MH) approach or the approach of Zhang and Bryant (2020) (STAN). Each beeswarm plot contains 10 observations, with data generated under the given simulation setting for 10 different seeds. Setting: $y_2 = 0.30$, $\sigma^2_{y_2} = 0.0001$

Appendix B

APPENDIX FOR CHAPTER 3

B.1 Age-heaping in DHS surveys

For the four DHS surveys we consider in our application (Malawi 2015-2016, Burkina Faso 2010, Senegal 2010, Namibia 2013), we display the total death counts at each age recorded in the entire survey in Figure B.1. Note that we expect peaks at 24, 36, and 48 months because they capture a full year of deaths as opposed to only single months, but the peaks observed at 12 months reflect age heaping as they cover the same age span as the age groups surrounding it. The small number of counts observed at unexpected age months (25 months, for example), are the few exceptions to the typical interval-censoring scheme used in DHS surveys.

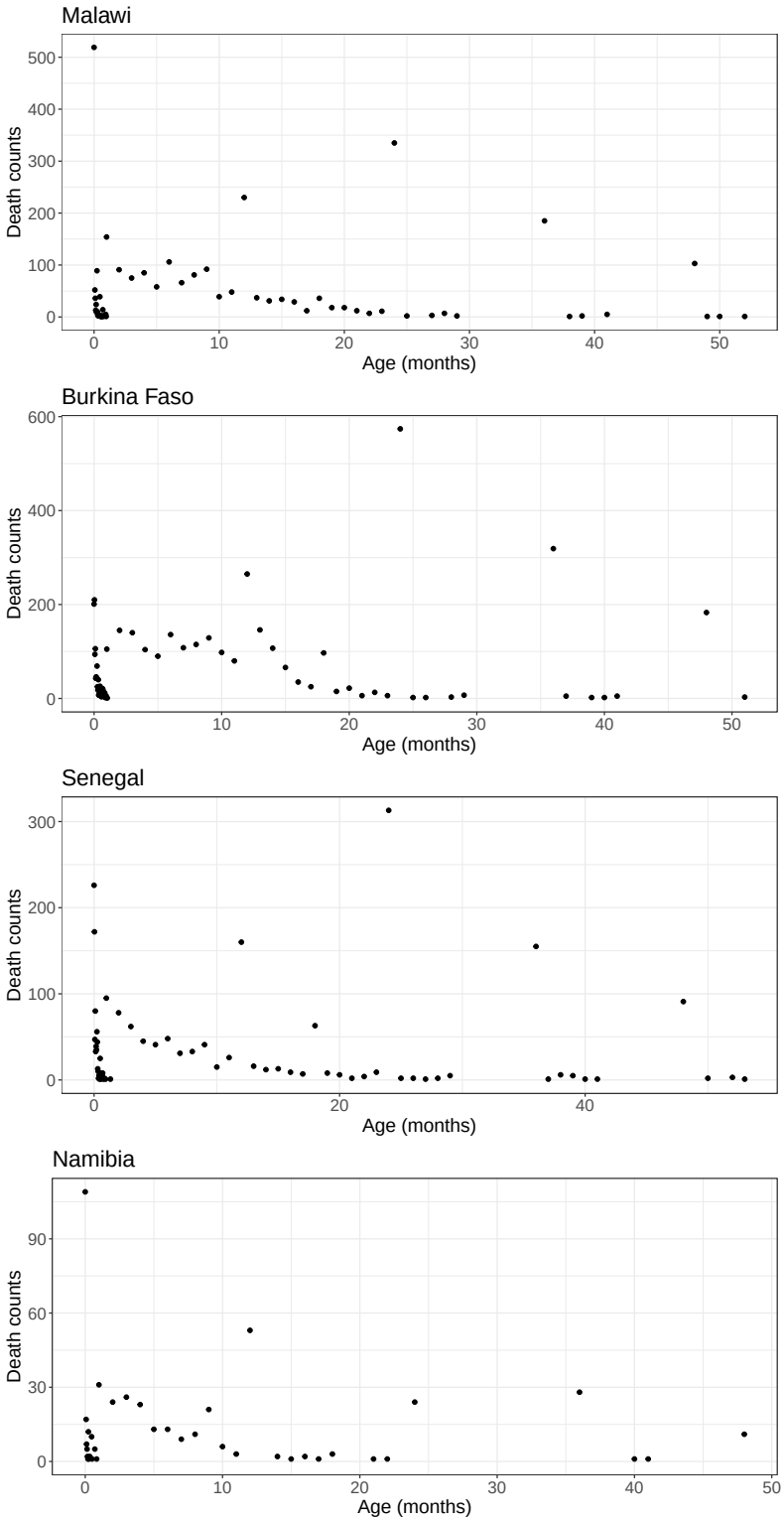


Figure B.1: Total death counts at each age group within DHS surveys.

B.2 Influence Functions

Recall that for a MLE $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_J)$ from a parametric model with $j = 1, \ldots, J$ parameters, we can write $\hat{\theta}_j$ as asymptotically linear

$$\hat{\theta}_j - \theta_j = \frac{1}{n} \sum_{i=1}^{n} w_i \Delta_i + o_p(n^{-1/2})$$

with influence functions Δ_i given by

$$\Delta_i = \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log(\mathrm{L})(\hat{\boldsymbol{\theta}}) \right] \left[\mathbf{H}_{\log(\mathrm{L})} \right],$$

where $\frac{\partial}{\partial \boldsymbol{\theta}} \log(\mathrm{L})(\hat{\boldsymbol{\theta}})$ is an $n \times J$ dimensional matrix of score functions for each individual $i = 1, \ldots n$ with respect to parameters $j = 1, \ldots J$, and $\mathbf{H}_{\log(\mathrm{L})}$ denotes the Hessian of the weighted log likelihood.

Since our log likelihood is given by

$$\log(\mathrm{L}) = \sum_{i=1}^{n} \left[(1 - I_i)(-H_i(t_i)) + I_i \log(\exp(-H_i(t_{0i})) - \exp(-H_i(t_{1i}))) \right]$$

we can then obtain the score for an individual i for each parameter θ_j as

$$\begin{aligned} \frac{\partial}{\partial \theta_j} \log(\mathrm{L})(\hat{\boldsymbol{\theta}}) = (1 - I_i)(-\frac{\partial}{\partial \theta_j} H_i(t_i)) \\ + I_i \left(\frac{-\exp(-H_i(t_{0i})) \frac{\partial}{\partial \theta_j} H_i(t_{0i}) + \exp(-H_i(t_{1i})) \frac{\partial}{\partial \theta_j} H_i(t_{1i})}{\exp(-H_i(t_{0i})) - \exp(-H_i(t_{1i}))} \right). \end{aligned}$$

In practice, it is more computationally efficient to calculate the gradient analytically, though we may calculate the gradient numerically as well.

z	F				M			
i	1	2	3	4	1	2	3	4
c_{zi}	0.006	-0.285	0.346	-0.934	0.009	-0.194	0.161	-0.911
β_{z1i}	0.016	0.501	-0.797	1.935	0.010	0.320	-0.425	1.88
β_{z2i}	-0.005	-0.157	0.202	-0.537	-0.003	-0.109	0.116	-0.527
β_{z3i}	-0.001	-0.029	-0.023	-0.106	-0.001	-0.021	0.026	-0.100
β_{z4i}	-0.0001	-0.002	-0.002	-0.007	-0.00005	-0.002	0.002	-0.006
β_{z5i}	-0.003	-0.002	-0.079	-0.049	-0.002	-0.008	0.110	-0.057
β_{z6i}	0.0004	0.013	-0.024	-0.014	0.0001	0.002	-0.002	-0.004
β_{z7i}	-0.00002	0.002	0.003	0.002	-0.00001	0.001	-0.001	-0.00004
β_{z8i}	-0.0004	-0.006	0.043	-0.004	-0.00003	-0.001	-0.004	-0.002

Table B.1: Estimated values for $\hat{w}_{zi}$, from Tables D1 and D2 in the online appendix of Clark (2019).

B.3 SVD model: details

B.3.1 Variance for proposed SVD approach

We compute the variance for the proposed SVD approach similarly to the variance calculation for the direct estimate approach in Mercer et al. (2015), but more generally where we allow the number of covariates in our model to not directly correspond to age groups. This slightly complicates the variance calculation. Assume the same set-up as in the Supplement of Mercer et al. (2015) where our estimate of interest is

$$\begin{aligned}\eta = \text{logit}({}_5q_0) &= \log\left(\frac{{}_5q_0}{1-{}_5q_0}\right) \\ &= \log\left(\frac{1-\prod_{i=1}^{n}(e^{\mathbf{XB}}+1)^{-k}}{\prod_{i=1}^{n}(e^{\mathbf{XB}}+1)^{-k}}\right) \\ &= \log\left(\prod_{i=1}^{n}(e^{\mathbf{XB}}+1)^{k}-1\right)\end{aligned}$$

where $\mathbf{X}$ is a $n \times m$ matrix of m LSVs used as covariates in our logistic regression model for n yearly age groups, and B is a $m \times 1$ matrix of the finite population estimates for the coefficients of the LSVs. We assume that each age group $i = 1, \ldots, n$ consists of k months.

We obtain the asymptotic distribution of the MLE $\hat{\mathbf{B}} \sim N(\mathbf{B}, \Sigma)$ from the `svyglm()` function in the `survey` package in `R`, and use the delta method to obtain the asymptotic distribution for $\hat{\eta}$ as

$$\hat{\eta} \sim \mathrm{N}(\mathrm{logit}({}_5q_0), \hat{V}_{\mathrm{DES}})$$

where

$$\hat{V}_{\mathrm{DES}} = \left(\frac{\partial \eta}{\partial \mathbf{B}}\right)^{\top} \hat{\Sigma} \left(\frac{\partial \eta}{\partial \mathbf{B}}\right)$$

where $\hat{\Sigma}$ is $Var(\hat{\mathbf{B}})$ obtained from `svyglm()`.
Again similarly to Mercer et al. (2015), we define

$$\begin{aligned} \gamma &= \prod_{i=1}^{n} (e^{\mathbf{XB}} + 1)^k, \\ &= \prod_{i=1}^{n} (e^{\sum_{j=1}^{m} B_j X_{ij}} + 1)^k. \end{aligned}$$

Noting that $\eta = \log(\gamma - 1)$, algebra will show that

$$\frac{\partial \eta}{\partial B_j} = \frac{\gamma}{\gamma - 1} \left[k \sum_{i=1}^{n} \left(X_{ij} \mathrm{expit} \left(\sum_{l=1}^{m} B_l X_{il} \right) \right) \right].$$

B.4 Additional Results

Below we display additional results from Weibull, generalized Gamma, piecewise exponential, lognormal, Gompertz, and exponentially-truncated shifted power (ETSP) models for Burkina Faso, Malawi, Senegal, and Namibia.

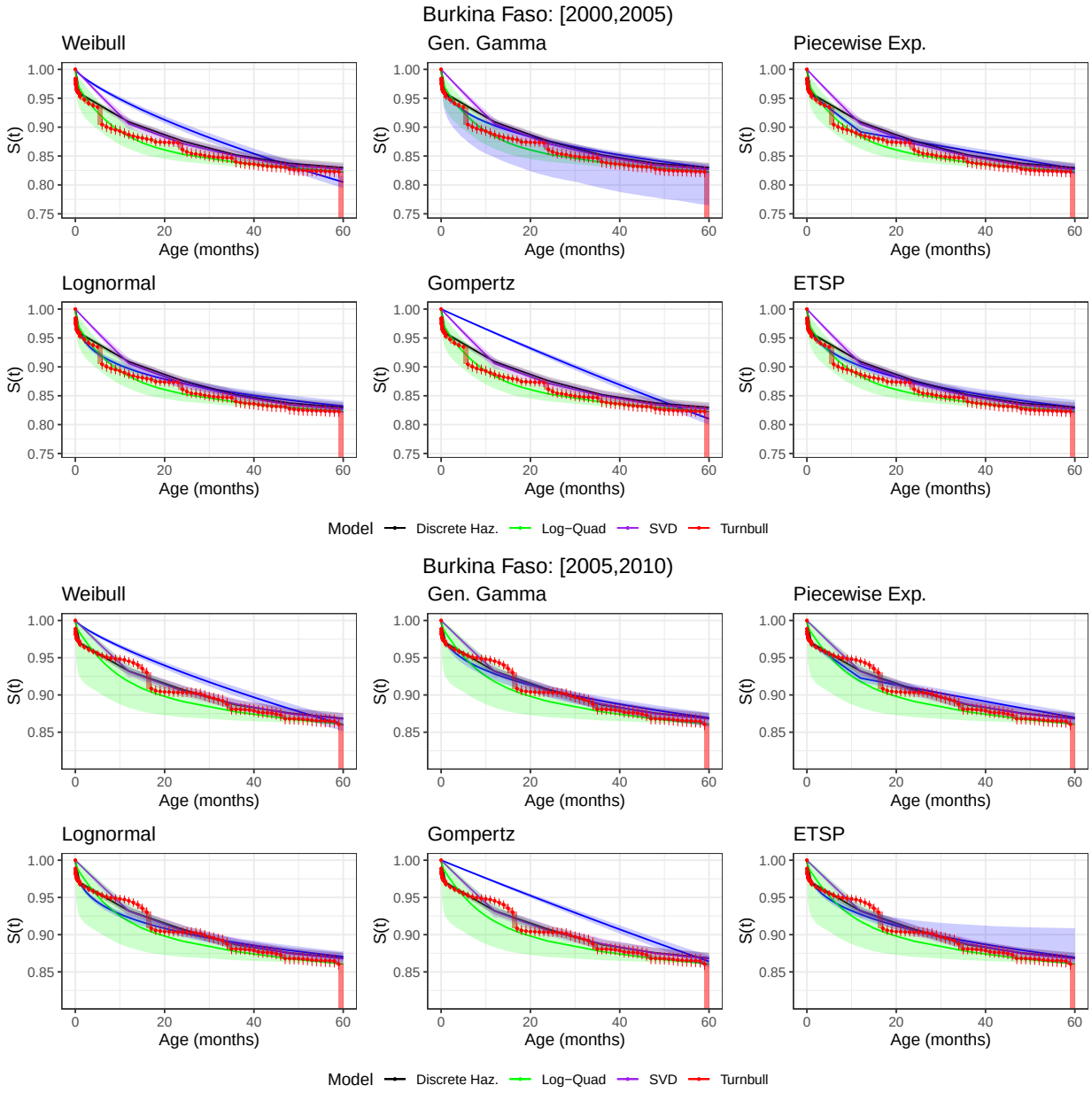


Figure B.2: Estimated survival curves for Burkina Faso in $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom) from ages 0 to 60 months. Parametric, pseudo-likelihood estimates are in blue. All confidence bands are 95% confidence intervals based on finite population variances, with the exception of the log-quad model where uncertainty is calculated as in Guillot et al. (2022).

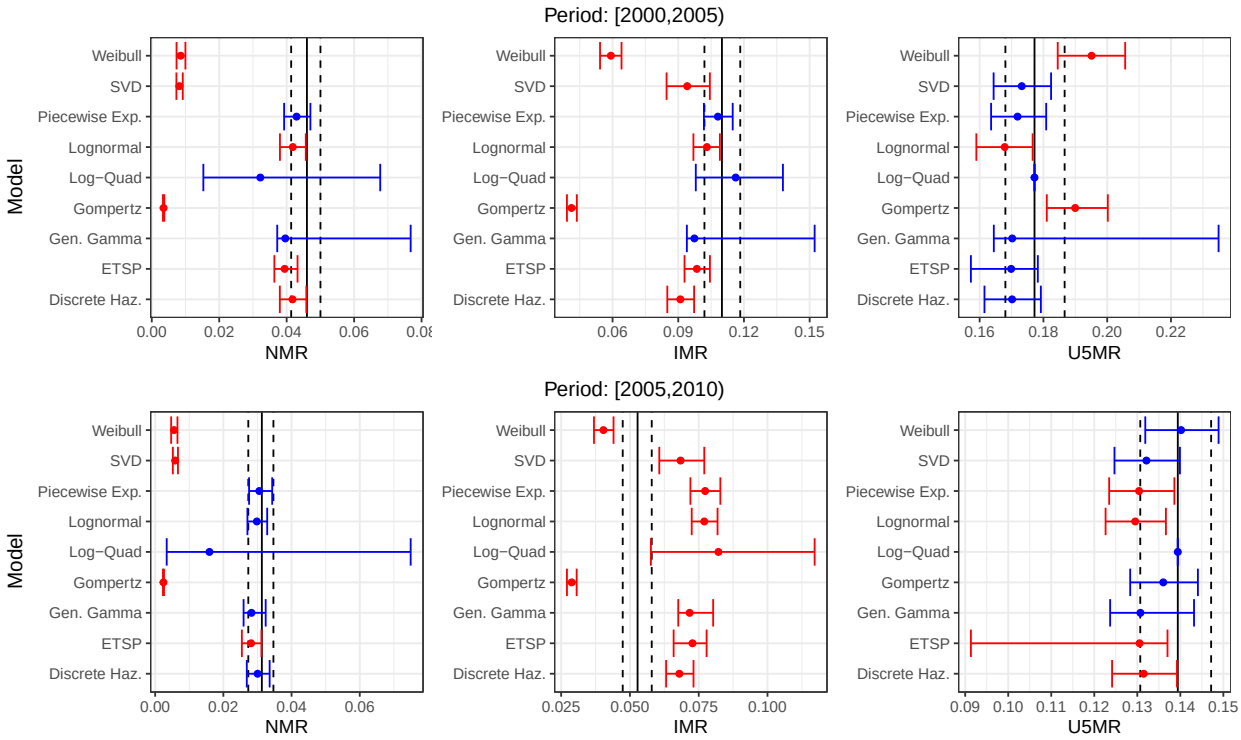


Figure B.3: Estimates of NMR, IMR, and U5MR for Burkina Faso in periods $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom). Turnbull point estimates are denoted by vertical black lines. All 95% confidence intervals are based on finite population variances, with the exception of the log-quad model where uncertainty is calculated as in Guillot et al. (2022).

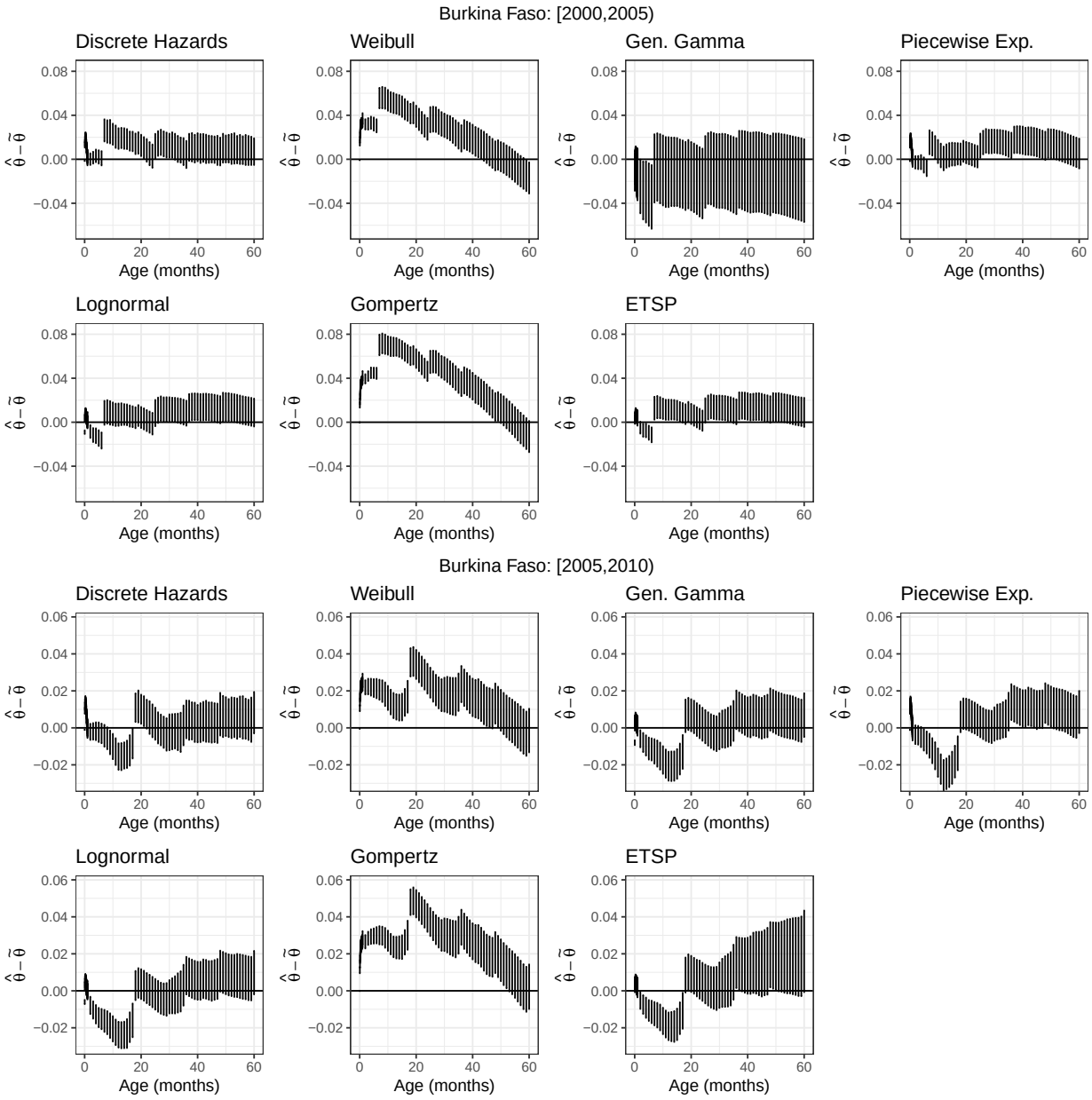


Figure B.4: Empirical distributions of differences in survival curves for Burkina Faso in $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom) from ages 0 to 60 months between parametric estimates $\hat{\theta}$ and the Turnbull estimate $\tilde{\theta}$.

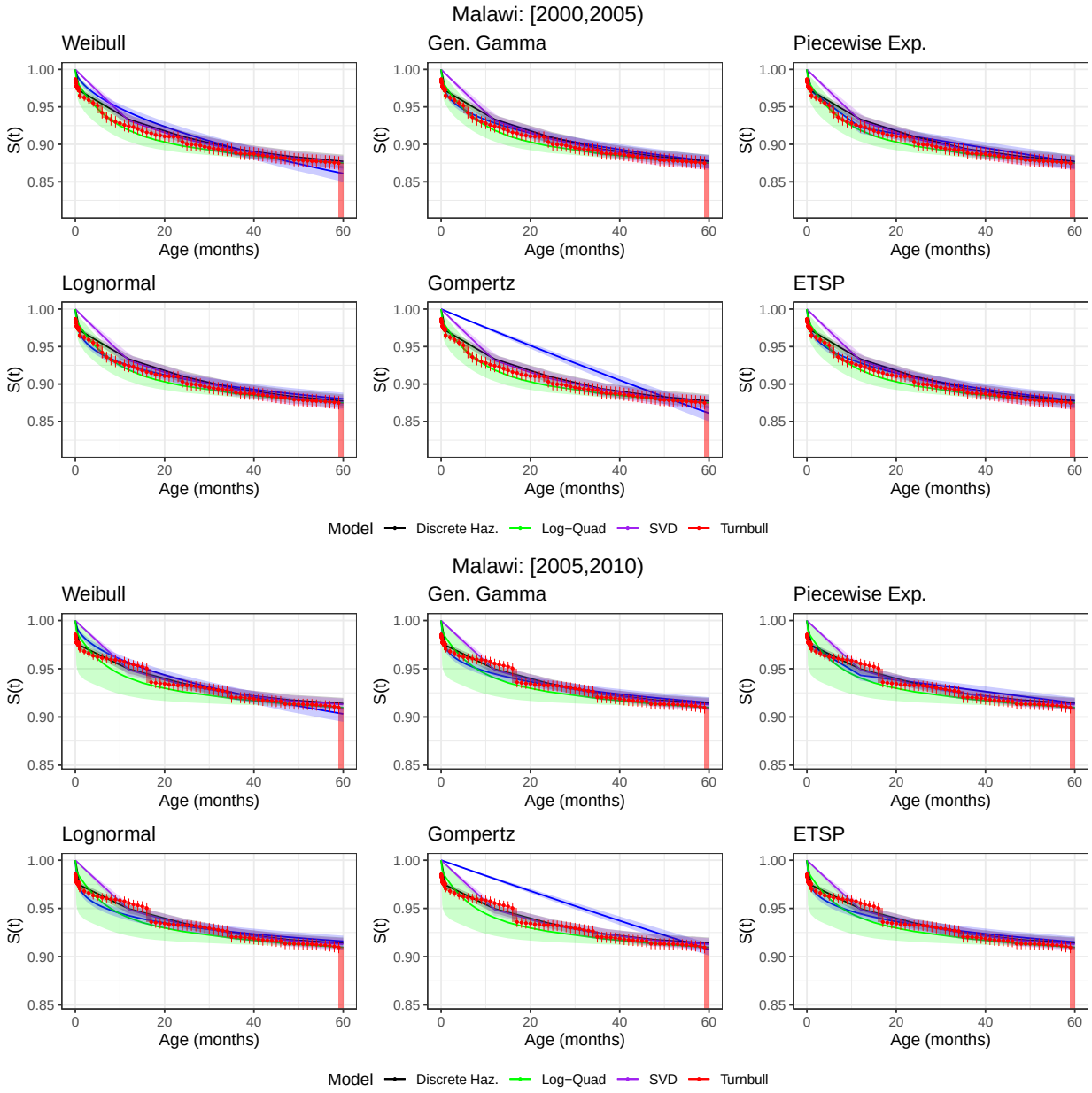


Figure B.5: Estimated survival curves for Malawi in $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom) from ages 0 to 60 months. Parametric, pseudo-likelihood estimates are in blue. All confidence bands are 95% confidence intervals based on finite population variances, with the exception of the log-quad model where uncertainty is calculated as in Guillot et al. (2022).

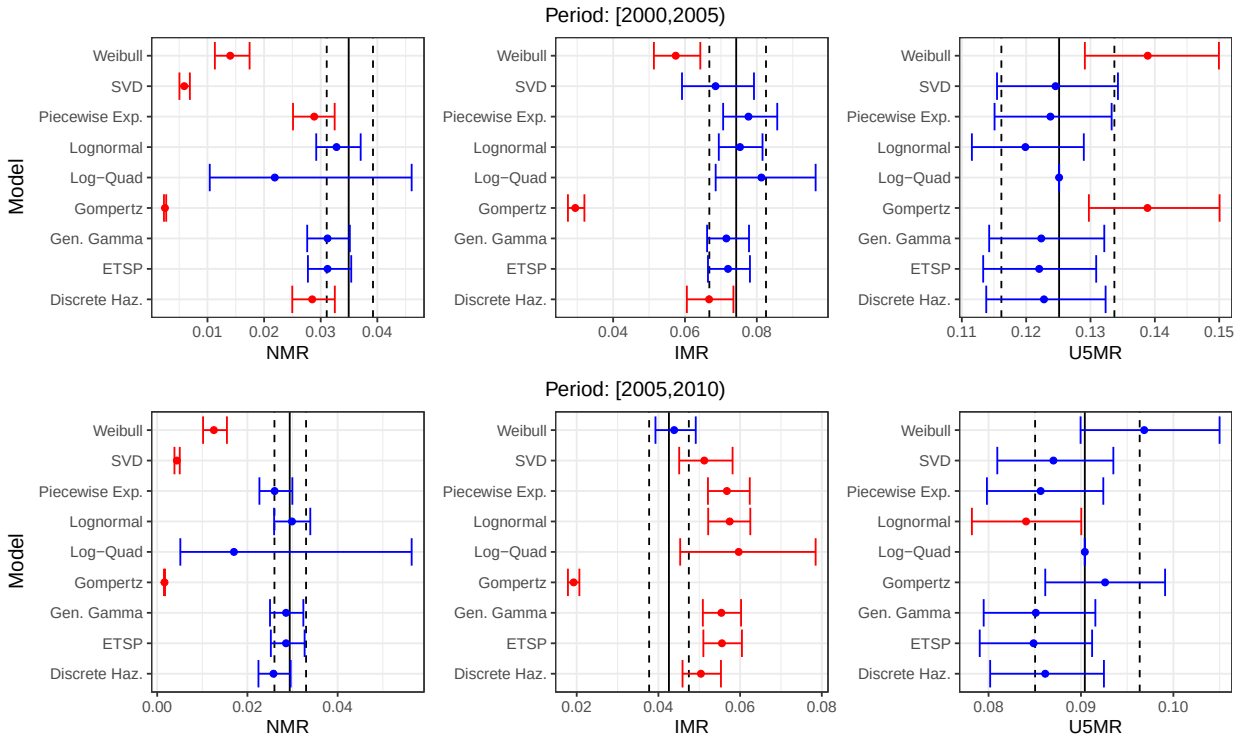


Figure B.6: Estimates of NMR, IMR, and U5MR for Malawi in periods $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom). Turnbull point estimates are denoted by vertical black lines. All 95% confidence intervals are based on finite population variances, with the exception of the log-quad model where uncertainty is calculated as in Guillot et al. (2022).

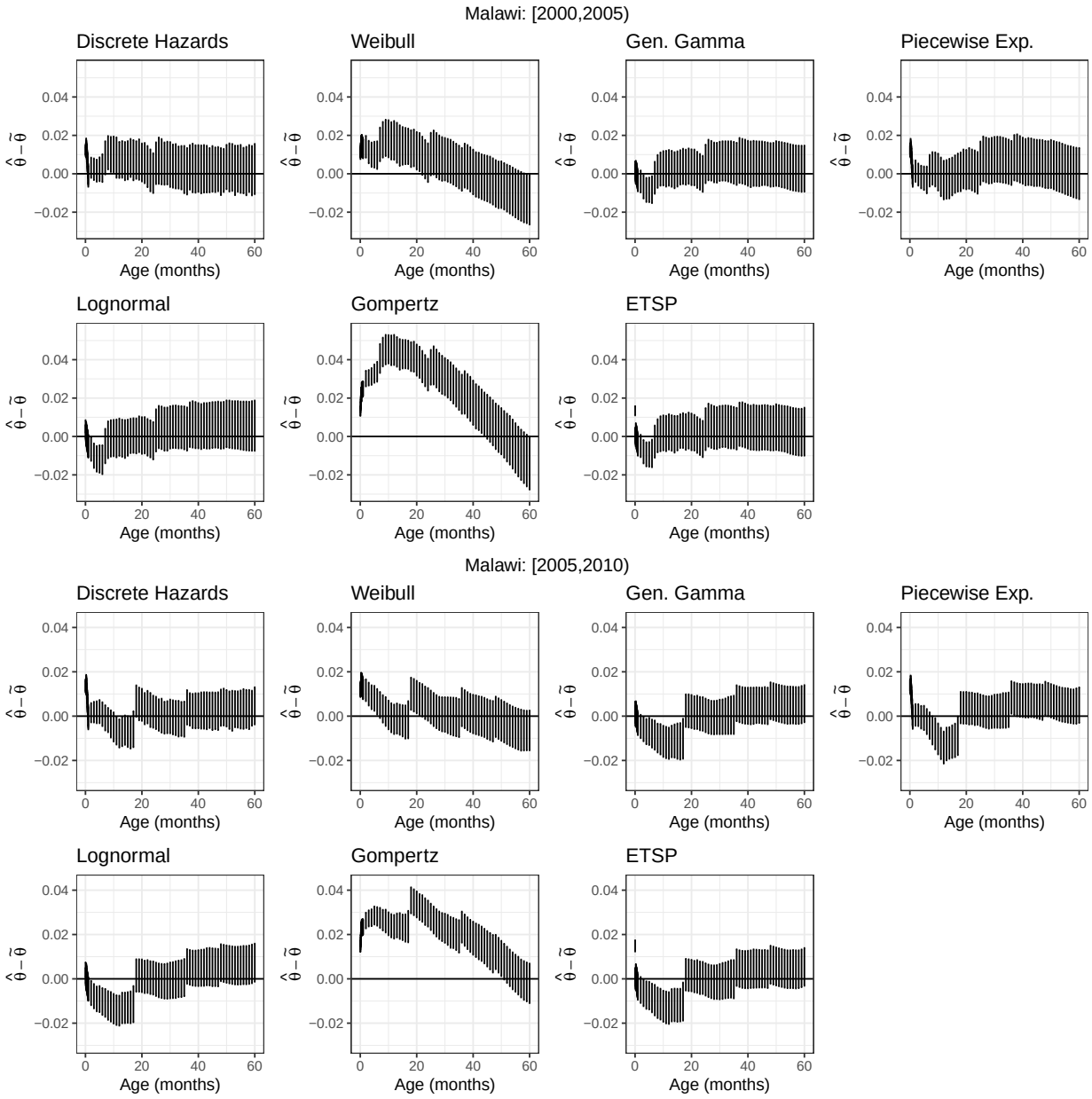


Figure B.7: Empirical distributions of differences in survival curves for Malawi in $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom) from ages 0 to 60 months between parametric estimates $\hat{\theta}$ and the Turnbull estimate $\tilde{\theta}$.

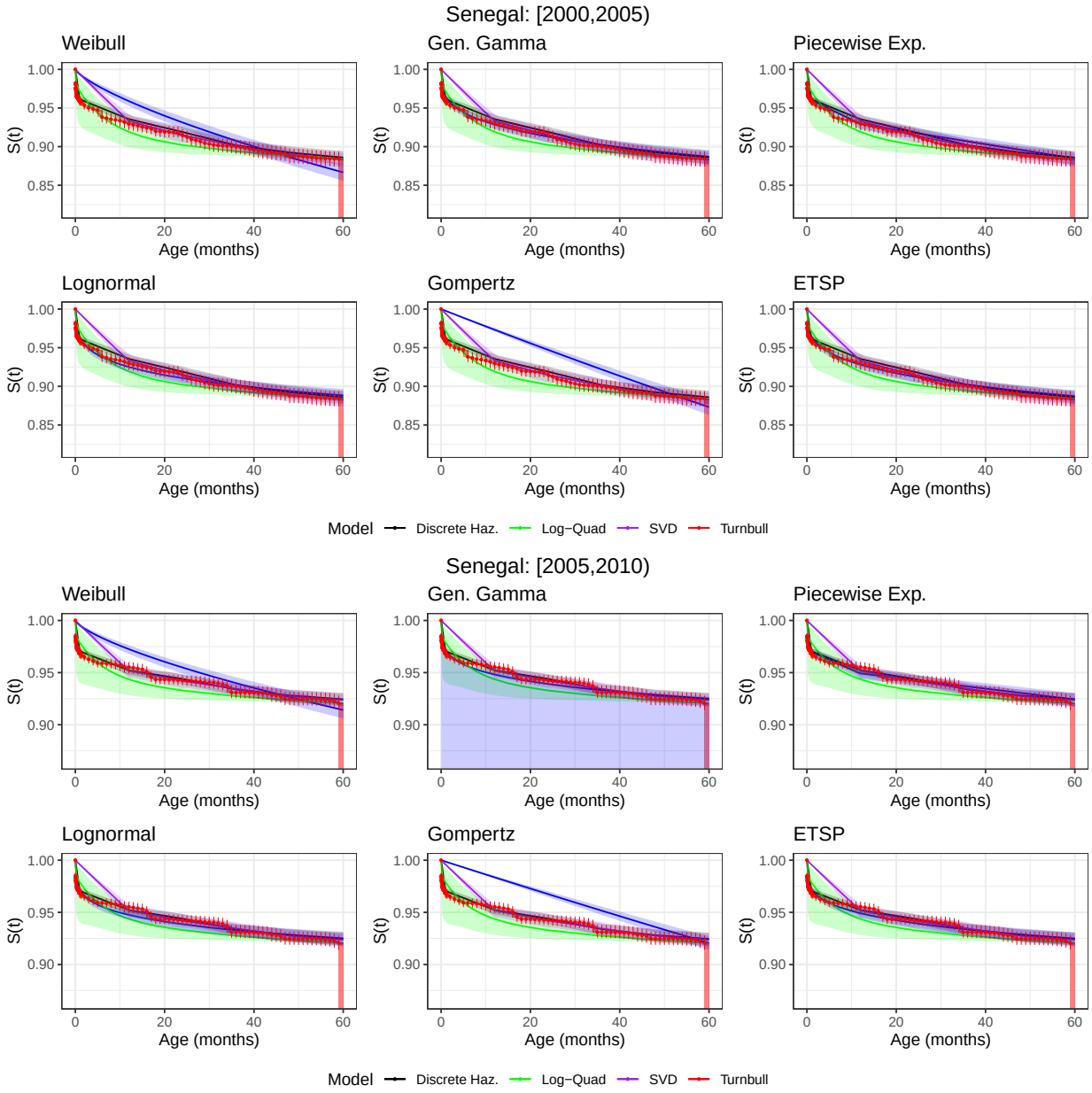


Figure B.8: Estimated survival curves for Senegal in $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom) from ages 0 to 60 months. Parametric, pseudo-likelihood estimates are in blue. All confidence bands are 95% confidence intervals based on finite population variances, with the exception of the log-quad model where uncertainty is calculated as in Guillot et al. (2022).

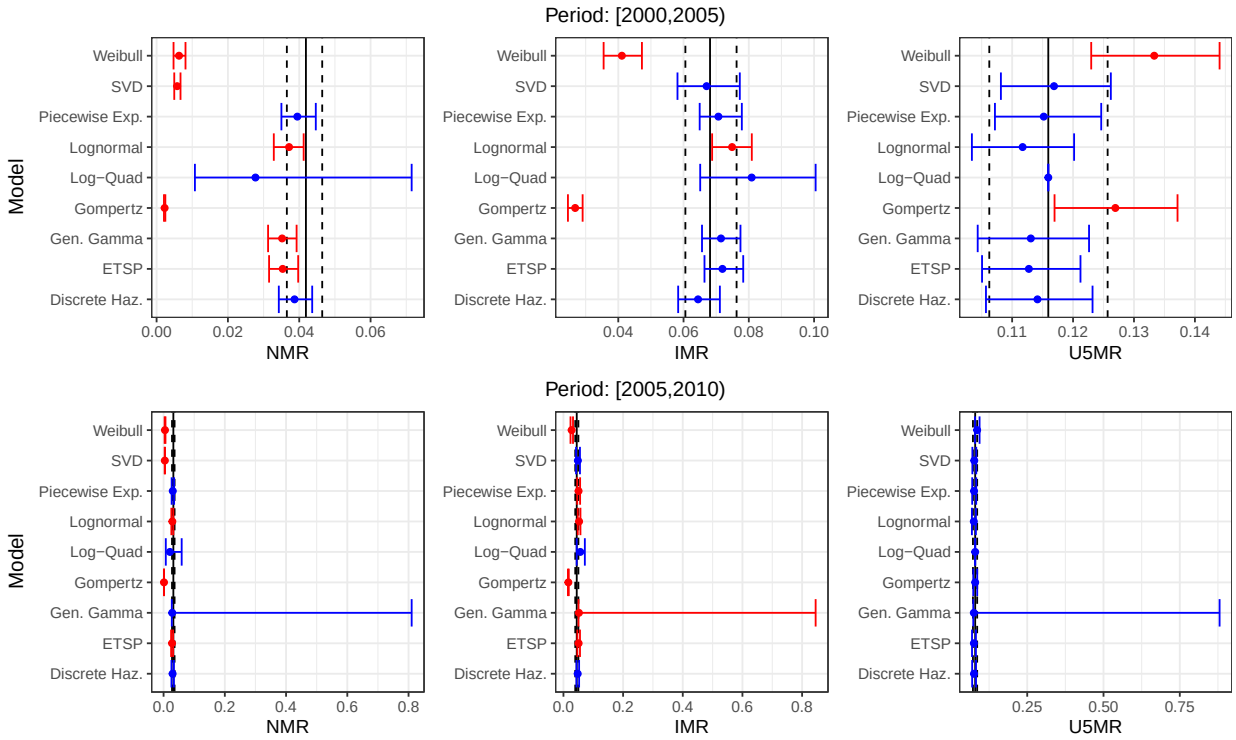


Figure B.9: Estimates of NMR, IMR, and U5MR for Senegal in periods $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom). Turnbull point estimates are denoted by vertical black lines. All 95% confidence intervals are based on finite population variances, with the exception of the log-quad model where uncertainty is calculated as in Guillot et al. (2022).

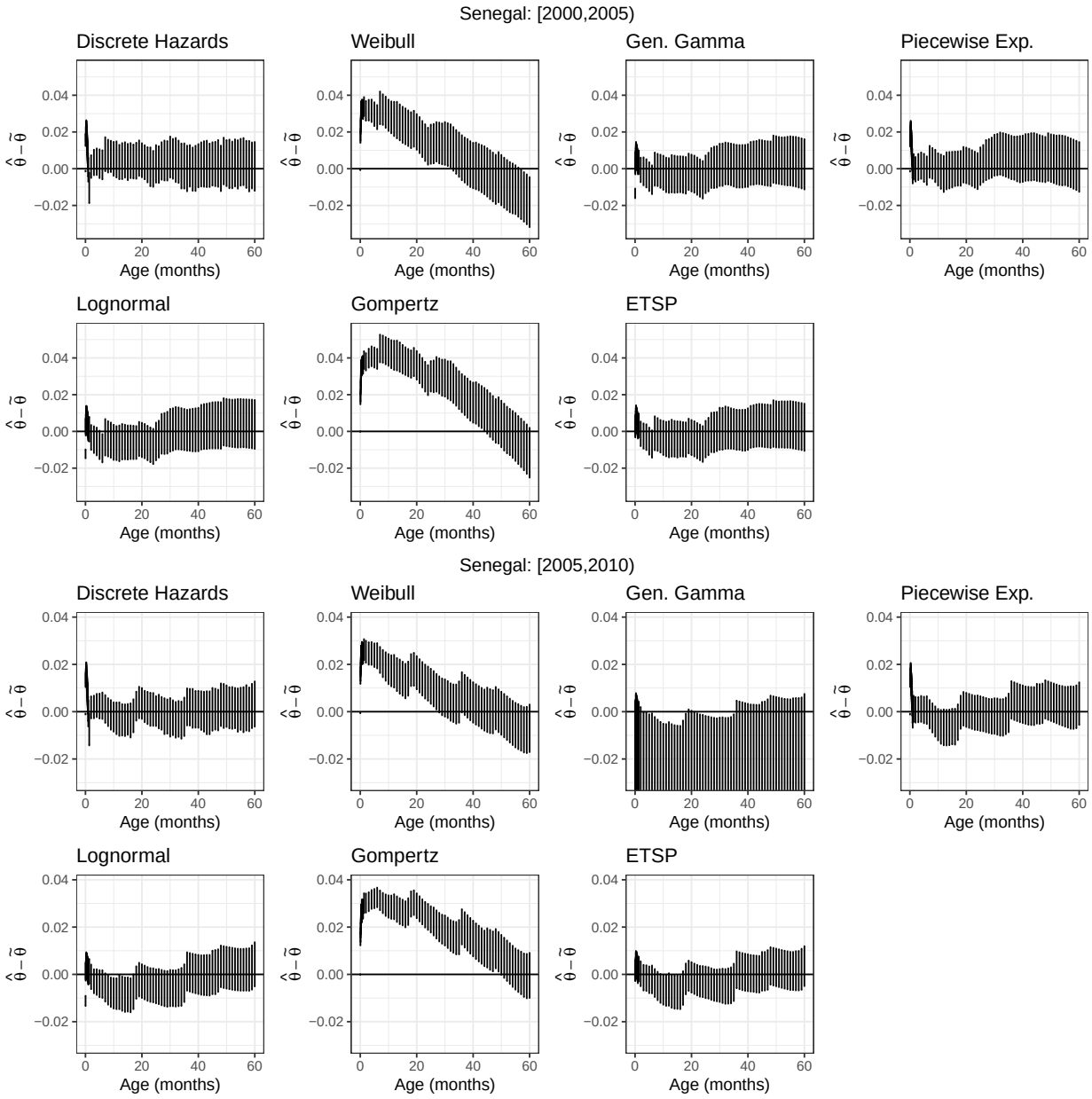


Figure B.10: Empirical distributions of differences in survival curves for Senegal in $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom) from ages 0 to 60 months between parametric estimates $\hat{\theta}$ and the Turnbull estimate $\tilde{\theta}$. Note that for $[2005, 2010)$ the differences have been cut off at -0.03 for clarity, though the differences extend much further negative for the generalized gamma model.

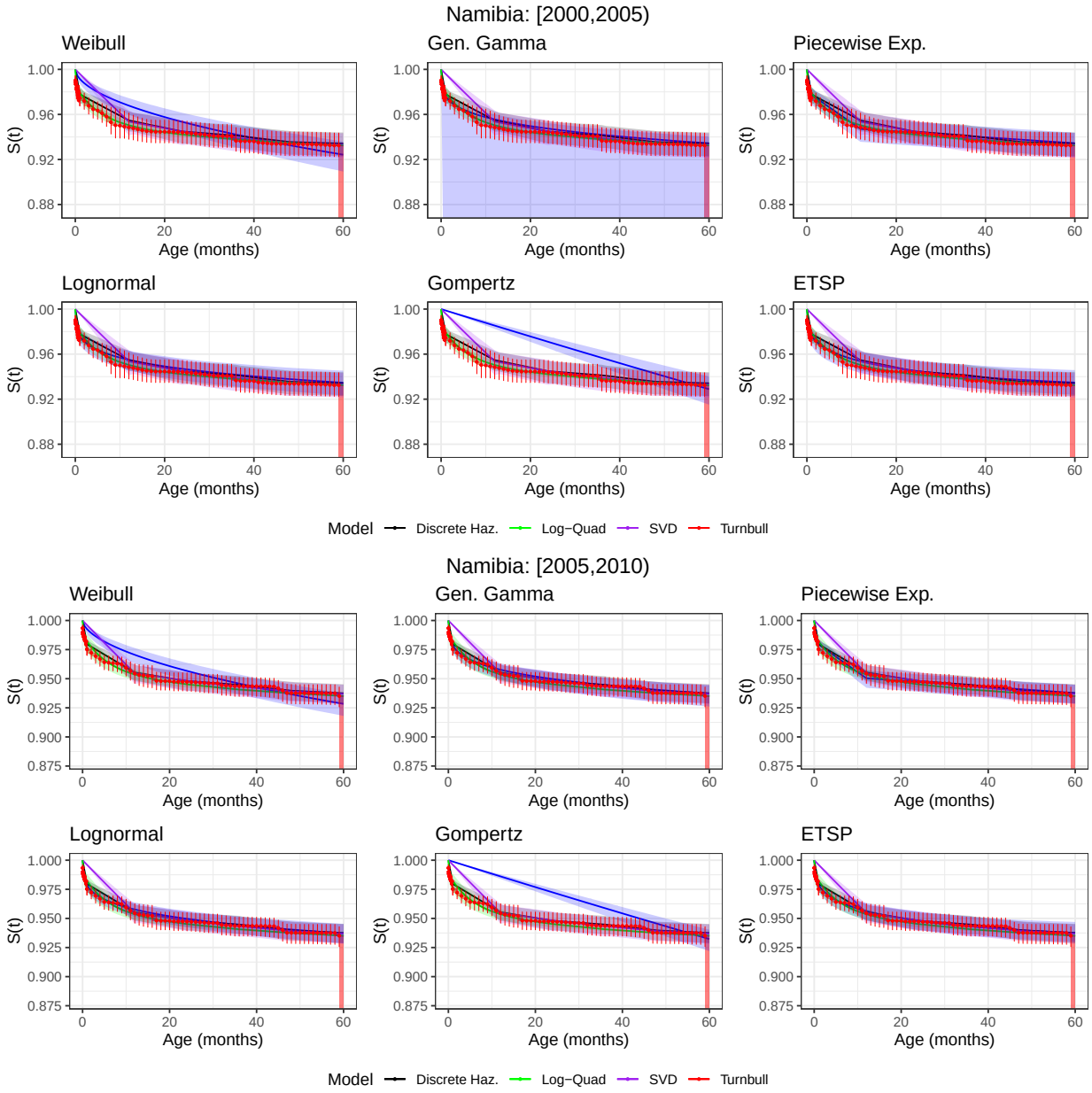


Figure B.11: Estimated survival curves for Namibia in $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom) from ages 0 to 60 months. Parametric, pseudo-likelihood estimates are in blue. All confidence bands are 95% confidence intervals based on finite population variances, with the exception of the log-quad model where uncertainty is calculated as in Guillot et al. (2022).

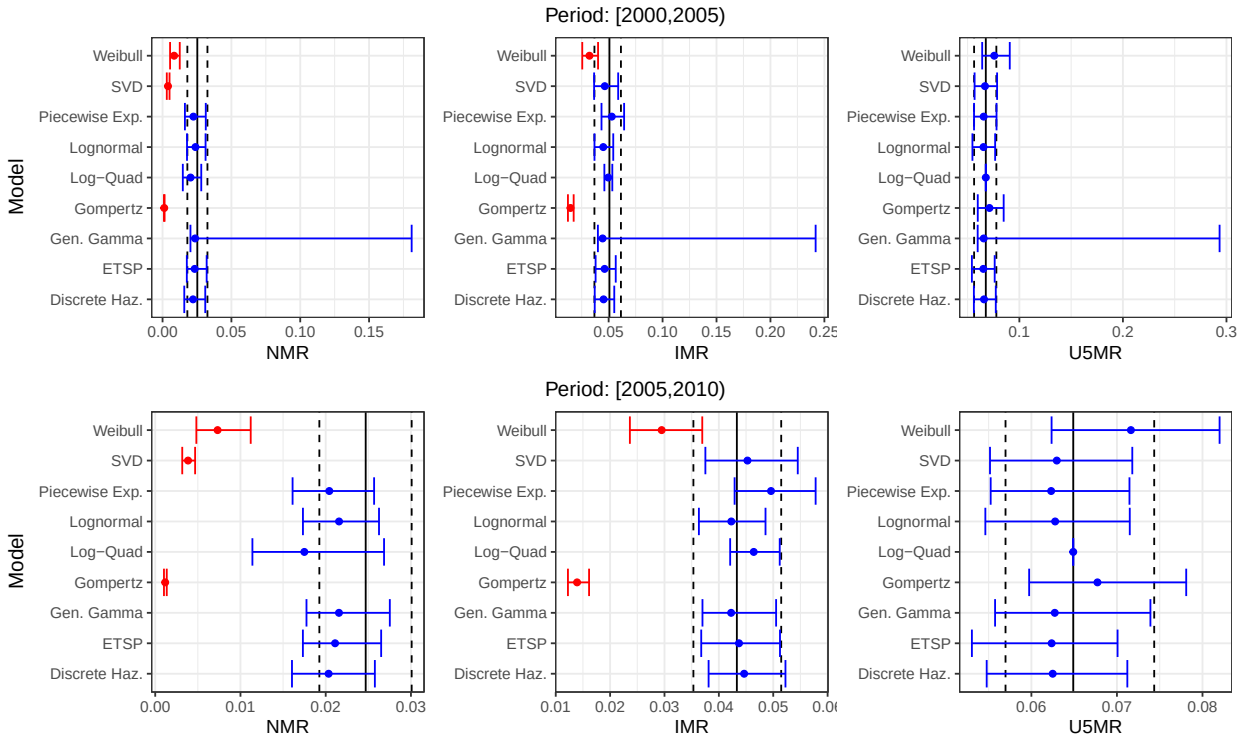


Figure B.12: Estimates of NMR, IMR, and U5MR for Namibia in periods $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom). Turnbull point estimates are denoted by vertical black lines. All 95% confidence intervals are based on finite population variances, with the exception of the log-quad model where uncertainty is calculated as in Guillot et al. (2022).

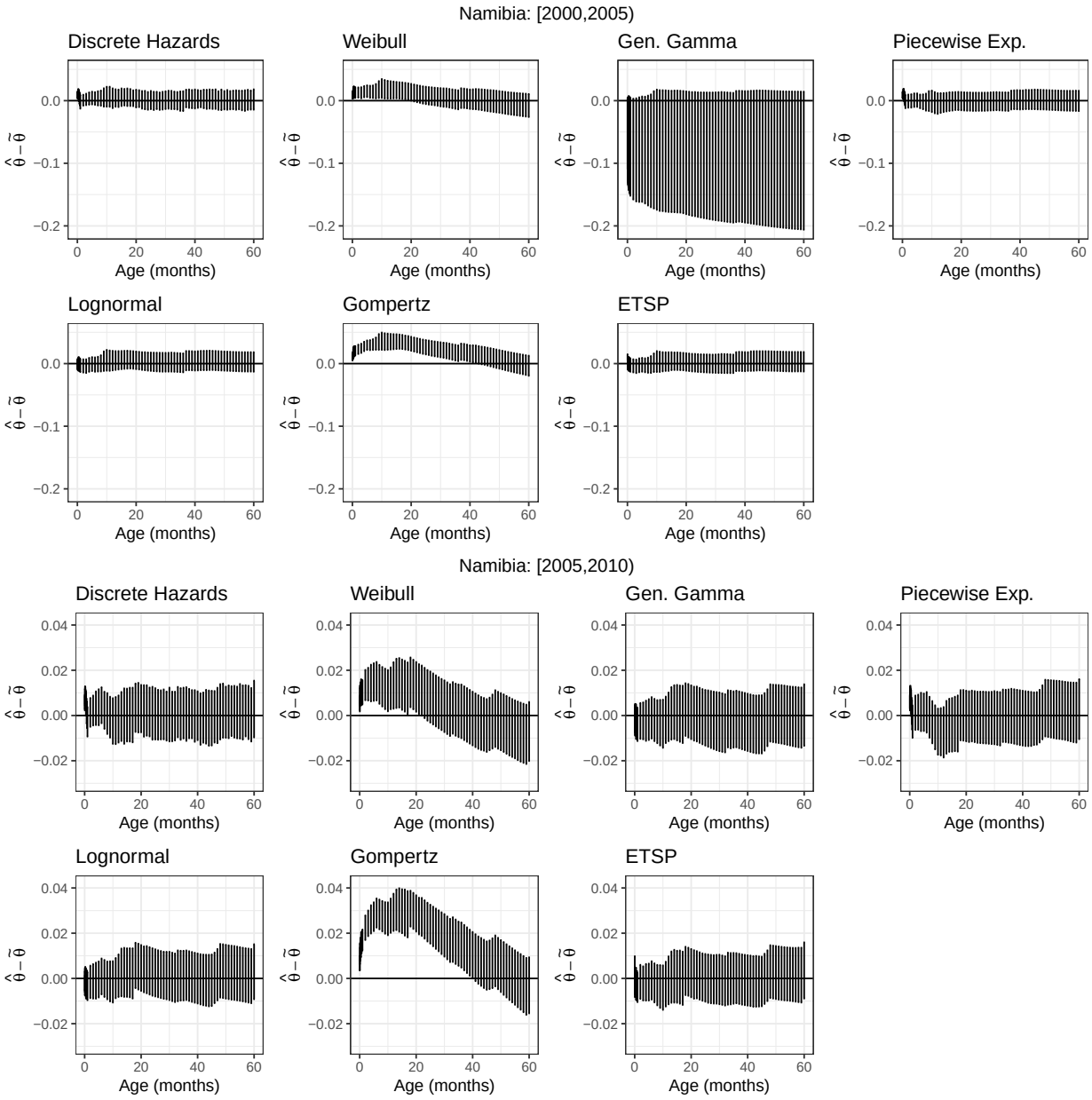


Figure B.13: Empirical distributions of differences in survival curves for Namibia in $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom) from ages 0 to 60 months between parametric estimates $\hat{\theta}$ and the Turnbull estimate $\tilde{\theta}$.

B.5 Comparison to models unadjusted for age heaping

We repeat our application, fitting the same models without addressing age-heaping at 12 months (by interval-censoring observations recorded as having died between 6 and 18 months for that entire 12 month period).

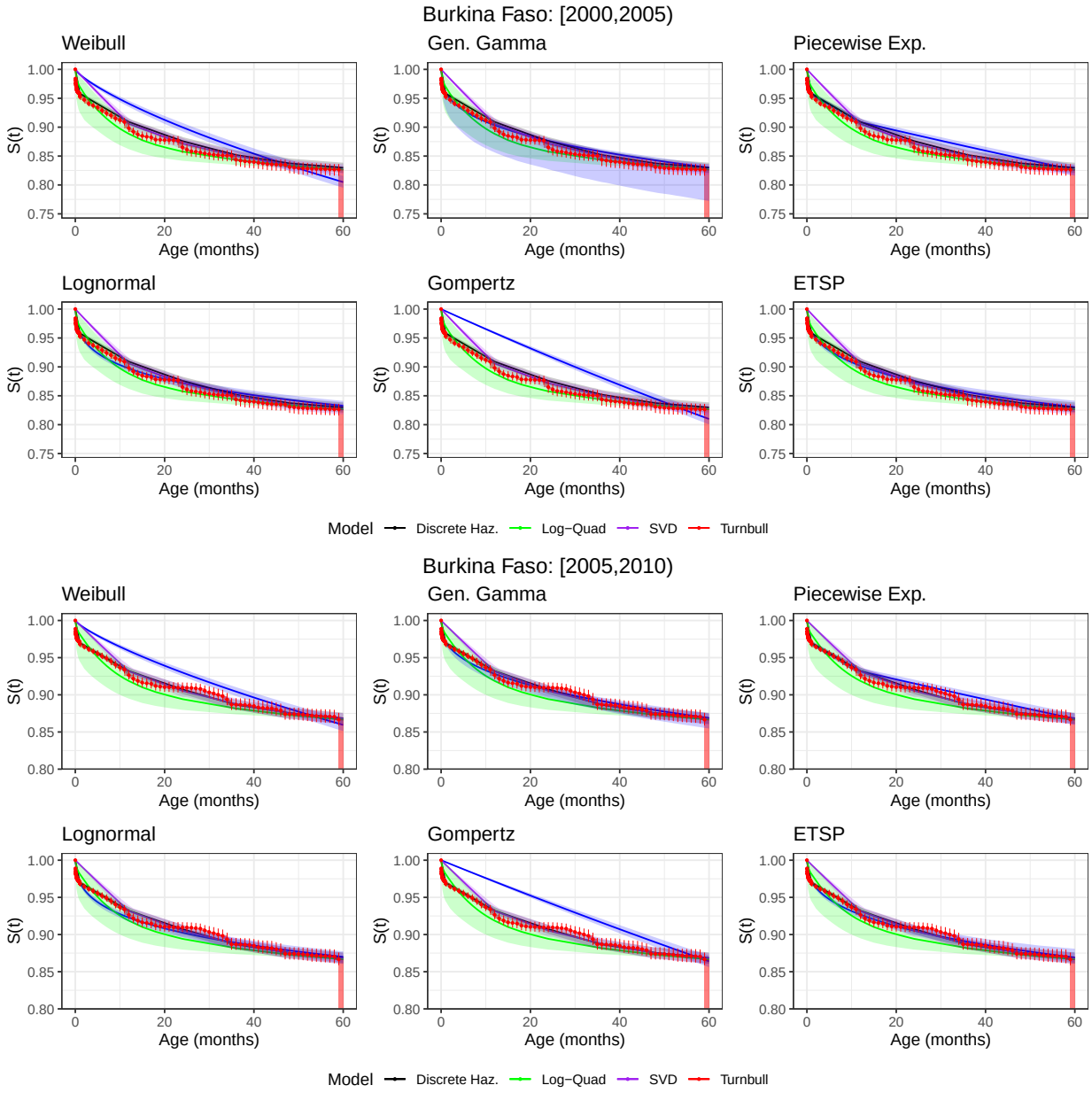


Figure B.14: Estimated survival curves for Burkina Faso in $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom) from ages 0 to 60 months, not adjusted for age heaping at 12 months. Parametric, pseudo-likelihood estimates are in blue. All confidence bands are 95% confidence intervals based on finite population variances, with the exception of the log-quad model where uncertainty is calculated as in Guillot et al. (2022).

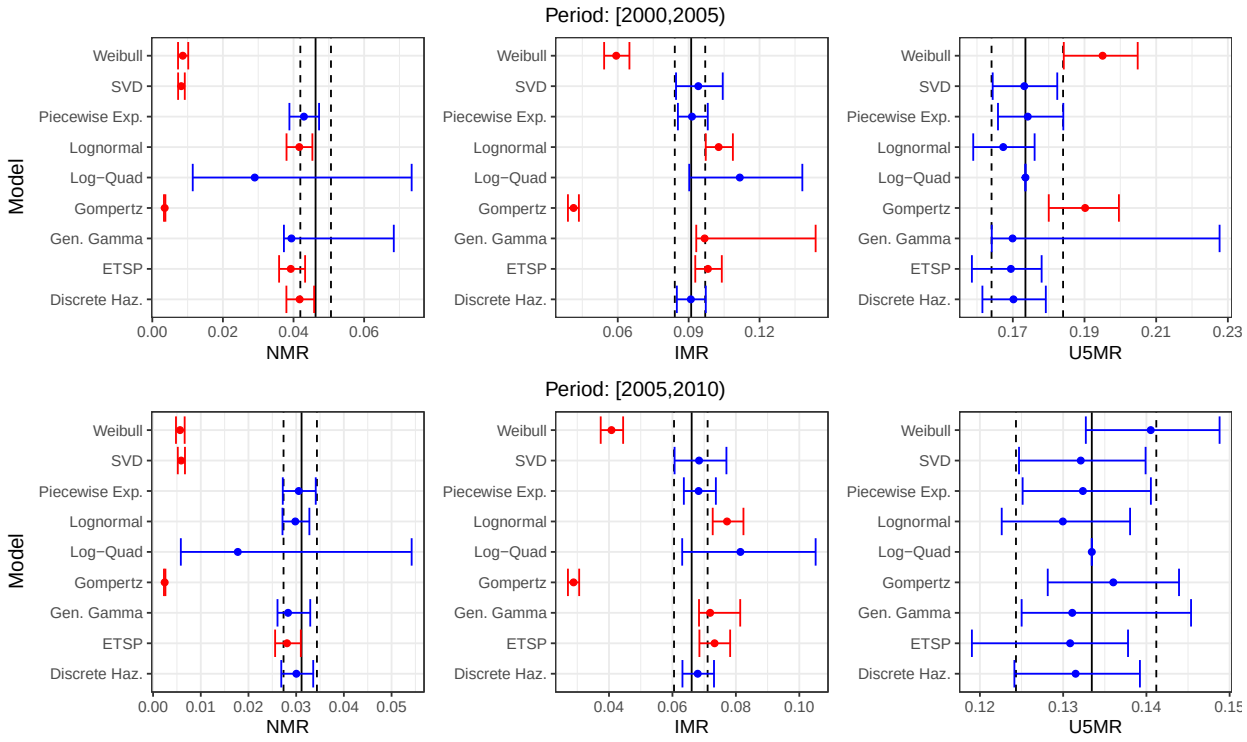


Figure B.15: Estimates of NMR, IMR, and U5MR for Burkina Faso in periods $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom), not adjusted for age heaping at 12 months. Turnbull point estimates are denoted by vertical black lines. All 95% confidence intervals are based on finite population variances, with the exception of the log-quad model where uncertainty is calculated as in Guillot et al. (2022).

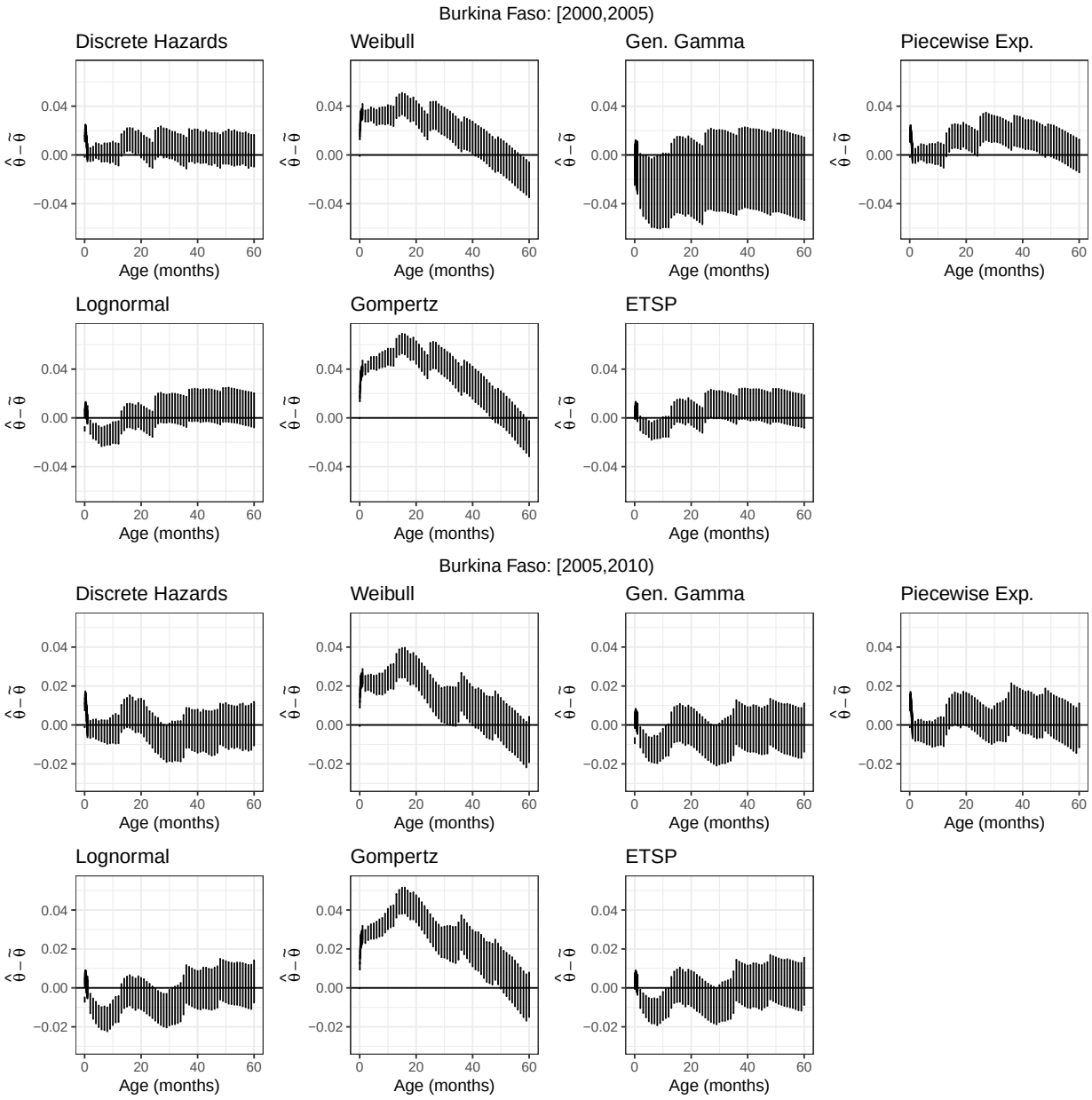


Figure B.16: Empirical distributions of differences in survival curves for Burkina Faso in $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom) from ages 0 to 60 months between parametric estimates (not adjusted for age heaping) $\hat{\theta}$ and the Turnbull estimate $\tilde{\theta}$.

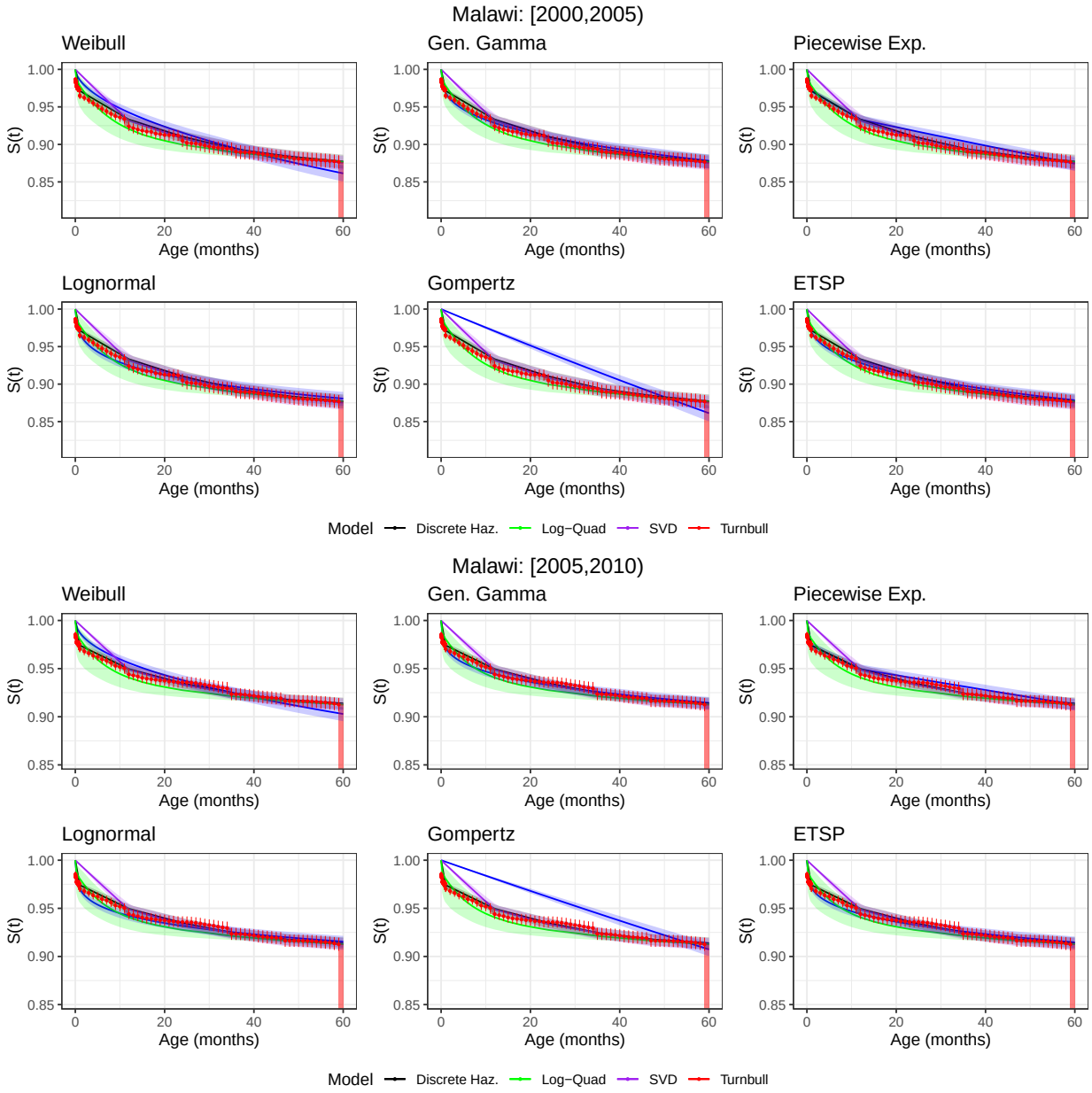


Figure B.17: Estimated survival curves for Malawi in $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom) from ages 0 to 60 months, not adjusted for age heaping at 12 months. Parametric, pseudo-likelihood estimates are in blue. All confidence bands are 95% confidence intervals based on finite population variances, with the exception of the log-quad model where uncertainty is calculated as in Guillot et al. (2022).

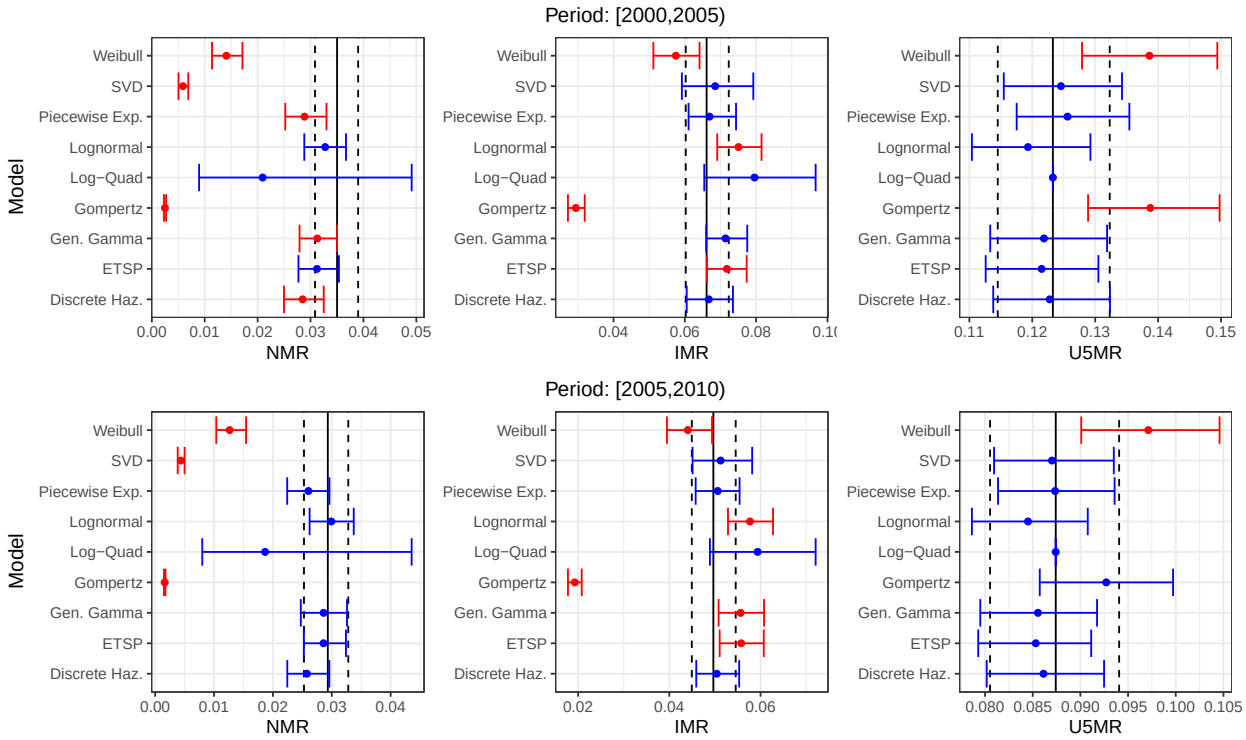


Figure B.18: Estimates of NMR, IMR, and U5MR for Malawi in periods $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom), not adjusted for age heaping at 12 months. Turnbull point estimates are denoted by vertical black lines. All 95% confidence intervals are based on finite population variances, with the exception of the log-quad model where uncertainty is calculated as in Guillot et al. (2022).

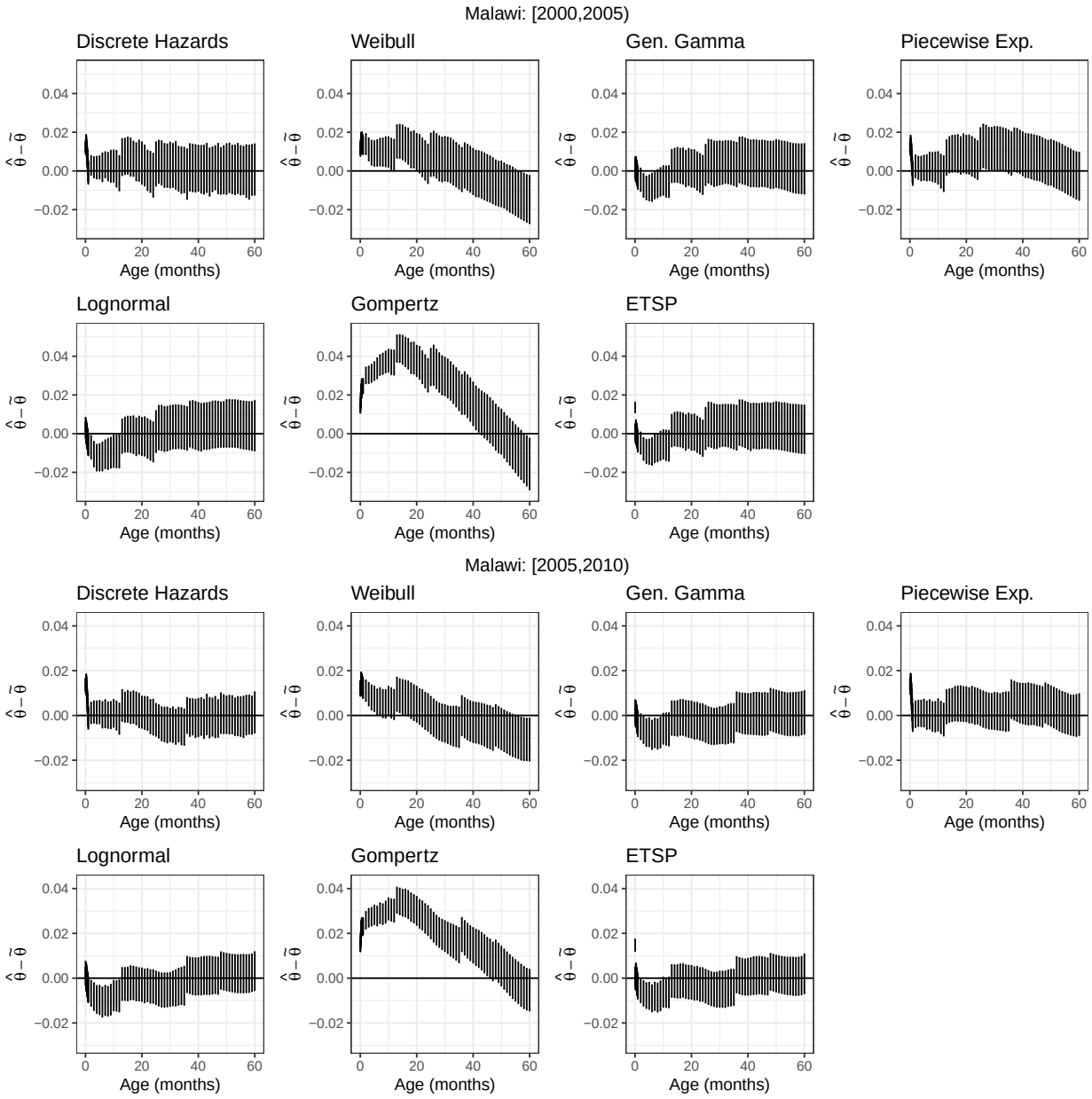


Figure B.19: Empirical distributions of differences in survival curves for Malawi in $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom) from ages 0 to 60 months between parametric estimates (not adjusted for age heaping) $\hat{\theta}$ and the Turnbull estimate $\tilde{\theta}$.

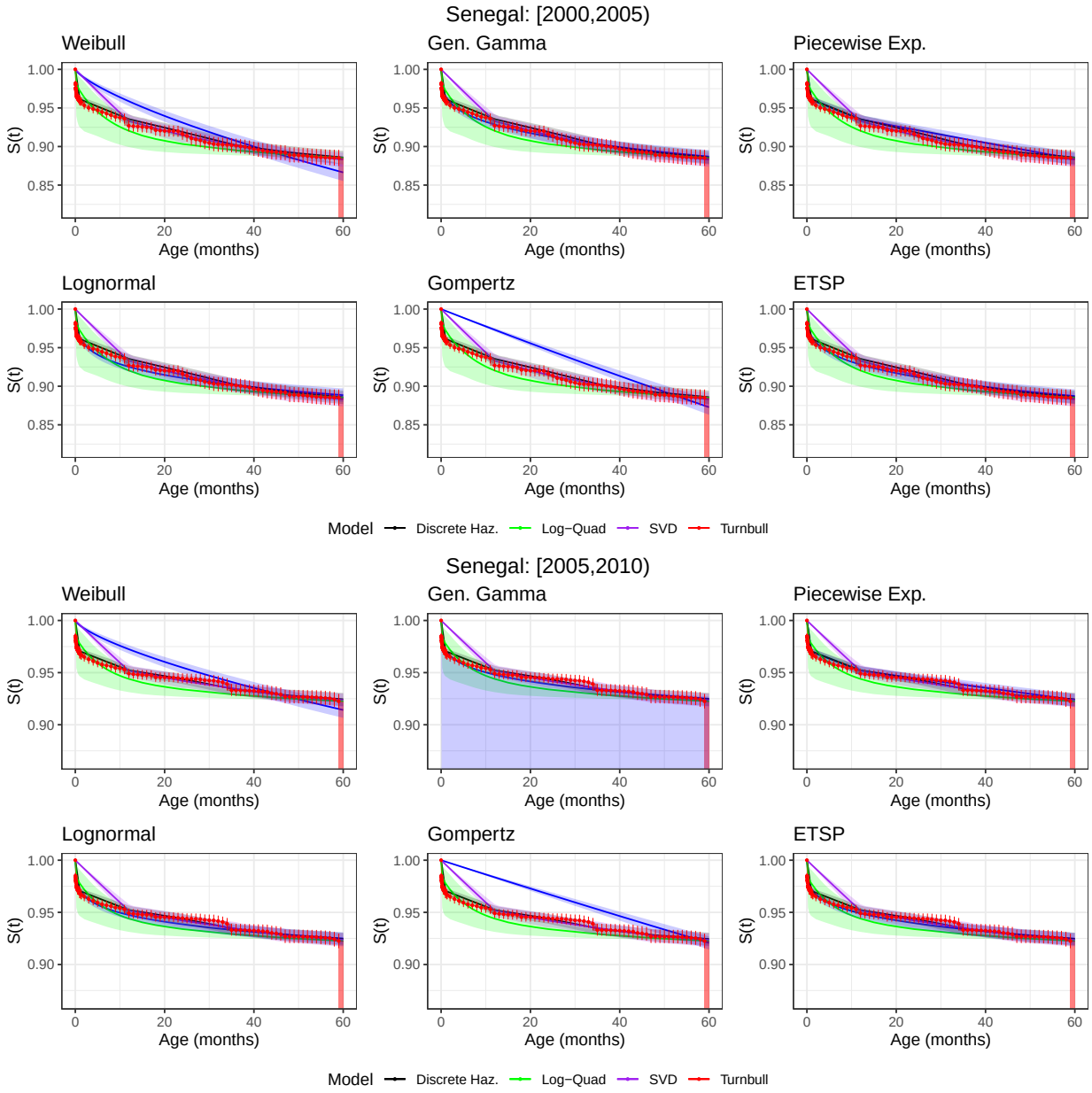


Figure B.20: Estimated survival curves for Senegal in $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom) from ages 0 to 60 months, not adjusted for age heaping at 12 months. Parametric, pseudo-likelihood estimates are in blue. All confidence bands are 95% confidence intervals based on finite population variances, with the exception of the log-quad model where uncertainty is calculated as in Guillot et al. (2022).

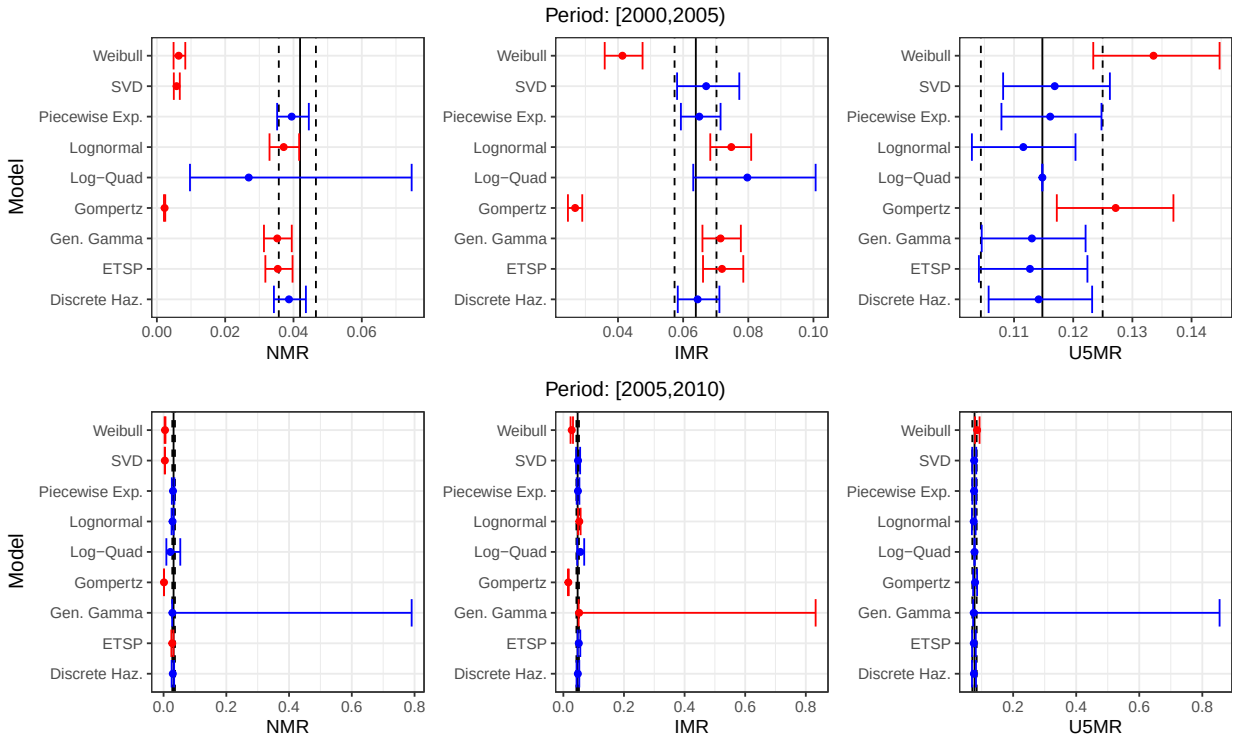


Figure B.21: Estimates of NMR, IMR, and U5MR for Senegal in periods $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom), not adjusted for age heaping at 12 months. Turnbull point estimates are denoted by vertical black lines. All 95% confidence intervals are based on finite population variances, with the exception of the log-quad model where uncertainty is calculated as in Guillot et al. (2022).

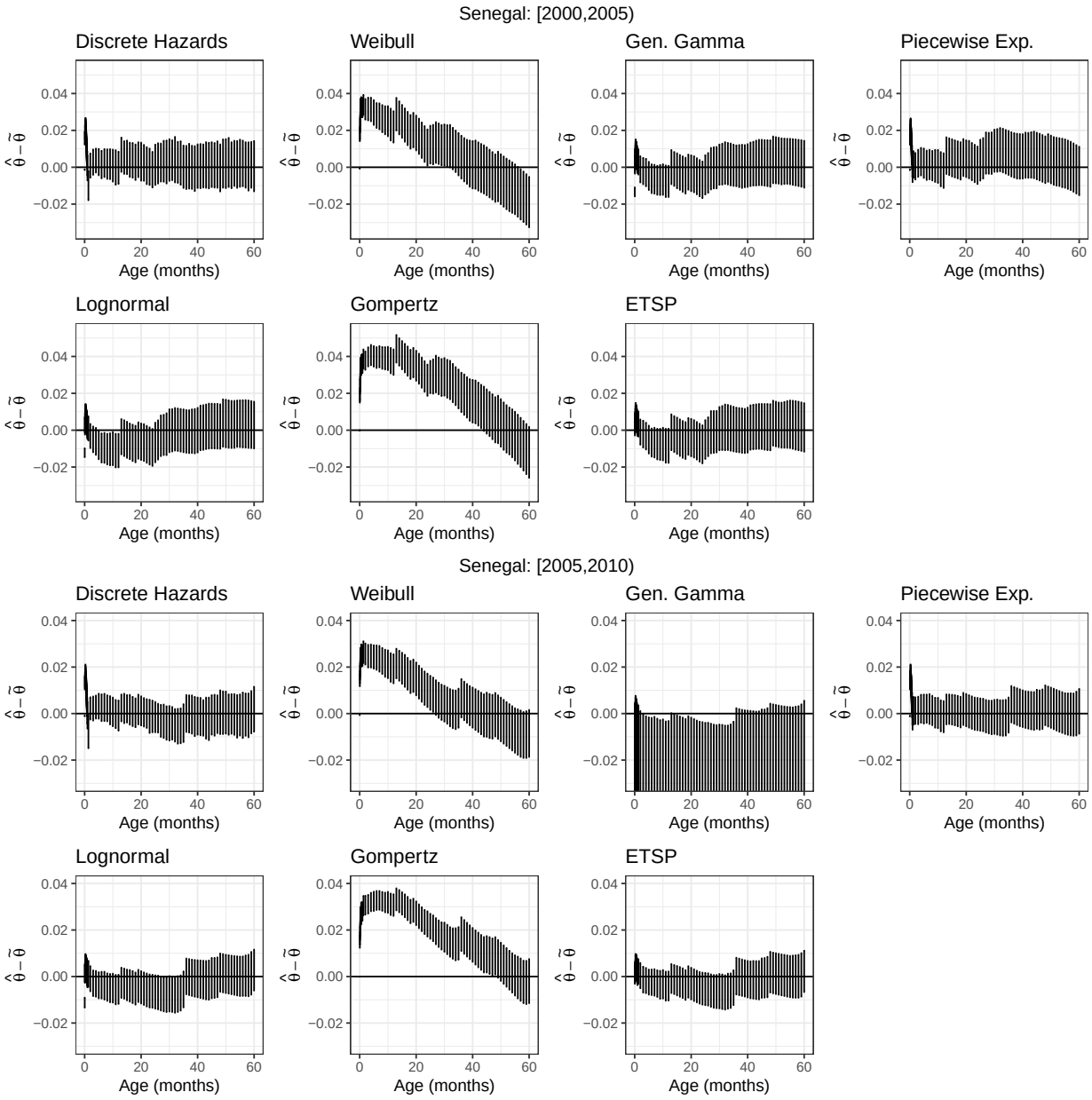


Figure B.22: Empirical distributions of differences in survival curves for Senegal in $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom) from ages 0 to 60 months between parametric estimates (not adjusted for age heaping) $\hat{\theta}$ and the Turnbull estimate $\tilde{\theta}$. Note that for $[2005, 2010)$ the differences have been cut off at -0.03 for clarity, though the differences extend much further negative for the generalized gamma model.

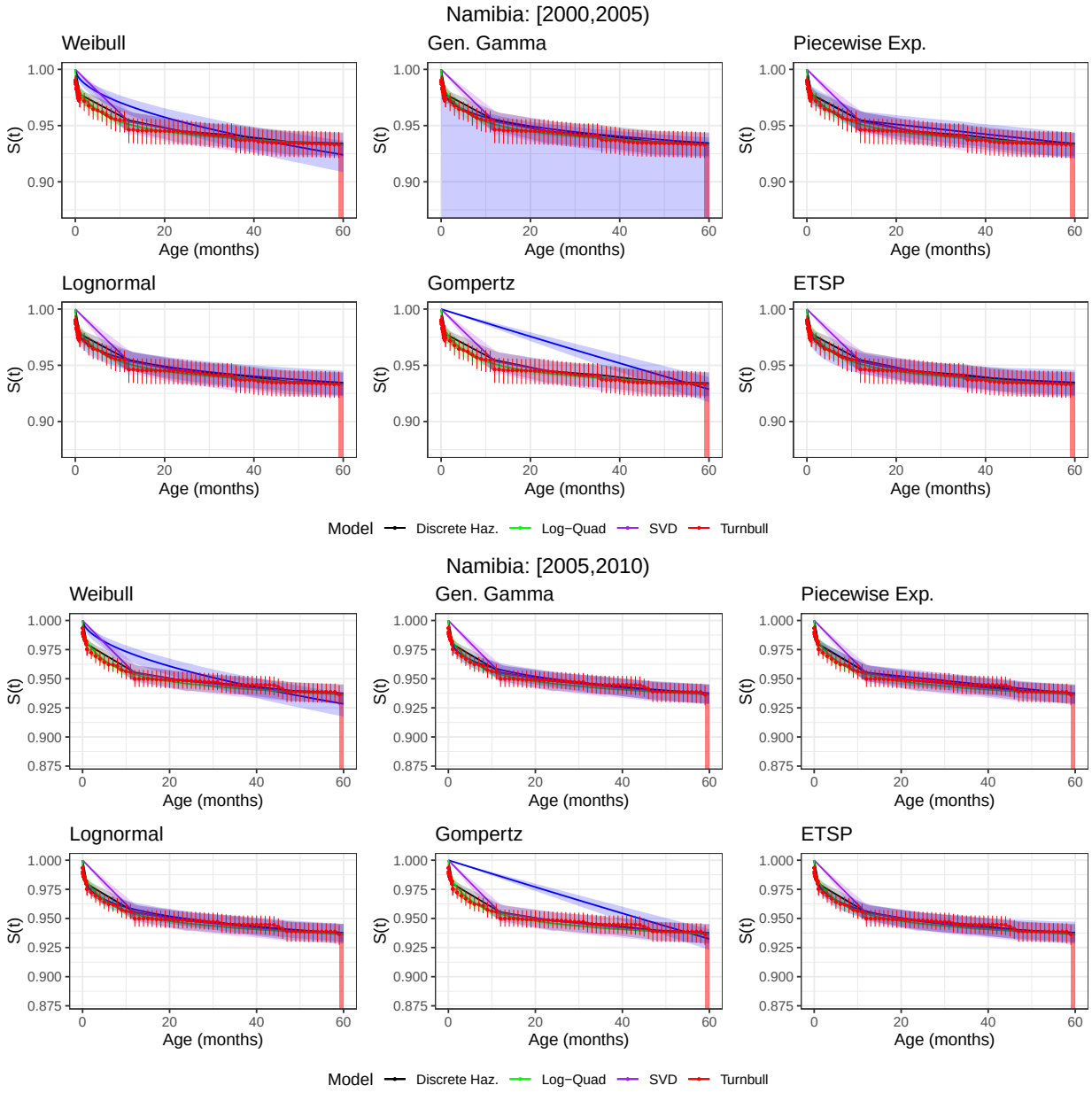


Figure B.23: Estimated survival curves for Namibia in $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom) from ages 0 to 60 months, not adjusted for age heaping at 12 months. Parametric, pseudo-likelihood estimates are in blue. All confidence bands are 95% confidence intervals based on finite population variances, with the exception of the log-quad model where uncertainty is calculated as in Guillot et al. (2022).

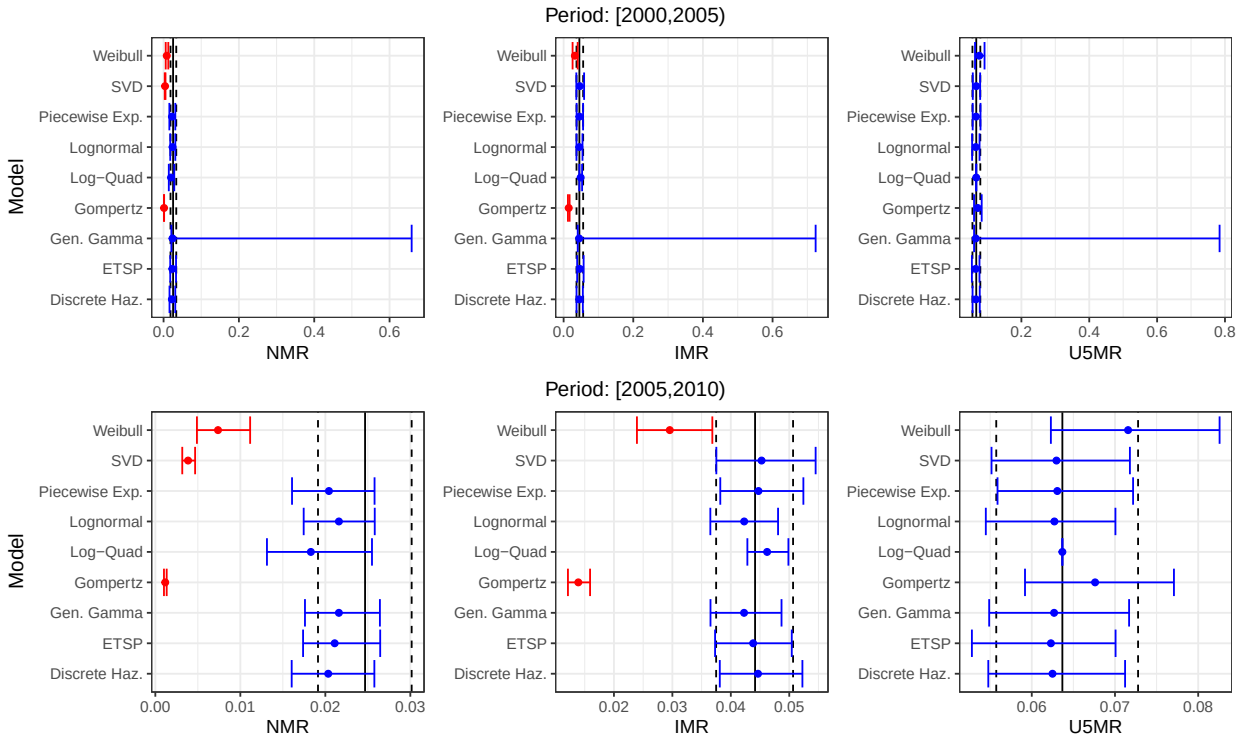


Figure B.24: Estimates of NMR, IMR, and U5MR for Namibia in periods $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom), not adjusted for age heaping at 12 months. Turnbull point estimates are denoted by vertical black lines. All 95% confidence intervals are based on finite population variances, with the exception of the log-quad model where uncertainty is calculated as in Guillot et al. (2022).

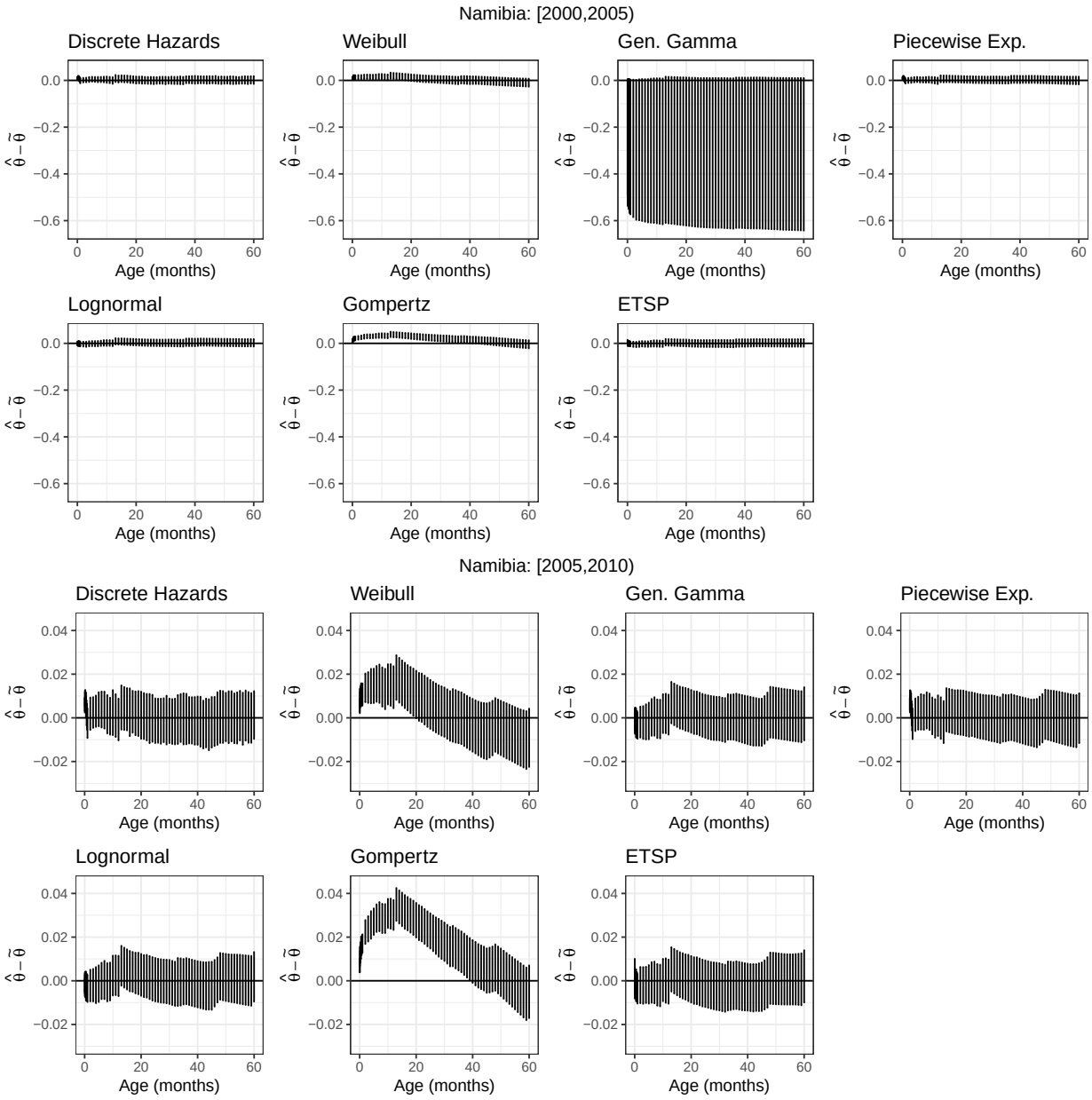


Figure B.25: Empirical distributions of differences in survival curves for Namibia in $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom) from ages 0 to 60 months between parametric estimates (not adjusted for age heaping) $\hat{\theta}$ and the Turnbull estimate $\tilde{\theta}$.

Country	Period	Weibull	Piecewise Exponential	Generalized Gamma	Lognormal	Gompertz	ETSP	Discrete Hazards
Burkina Faso	[2000, 2005)	18	33	**90**	67	14	67	64
	[2005, 2010)	23	65	**80**	64	11	70	**74**
Malawi	[2000, 2005)	44	63	**94**	**88**	19	**94**	**76**
	[2005, 2010)	55	**76**	**92**	**86**	16	**91**	**76**
Senegal	[2000, 2005)	33	**73**	**84**	**83**	20	**85**	**72**
	[2005, 2010)	37	**73**	63	**86**	15	**95**	**72**
Namibia	[2000, 2005)	59	**86**	**100**	**100**	29	**99**	**86**
	[2005, 2010)	53	**85**	**100**	**100**	27	**99**	**85**

Table B.2: Model validation results. Percentage of samples (out of 500) from $\hat{\theta} - \tilde{\theta}$ that contain 0 for all parametric models, countries, and periods, for models that do not adjust for age-heaping. Percentages greater than 70% are bolded.

In the following plots, we compare the parametric survival curves and Turnbull estimators when age heaping is adjusted for vs. unadjusted. Note that the point estimates for the survival curves are extremely similar for the Weibull, lognormal, Gompertz, generalized Gamma, and ETSP models, only differing in the third or fourth decimal place. This suggests that age-heaping occurring between 6 and 18 months does not greatly impact the overall shape of the survival curve. The age-heaping-adjusted piecewise exponential model differs quite significantly from the unadjusted piecewise exponential model at age 12 months, which is to be expected based on how we interval censored the data in the adjusted model. Also note that in all cases, the uncertainty surrounding the age-heaping-adjusted model is slightly larger than the uncertainty surrounding the unadjusted models, though perhaps not meaningfully larger.

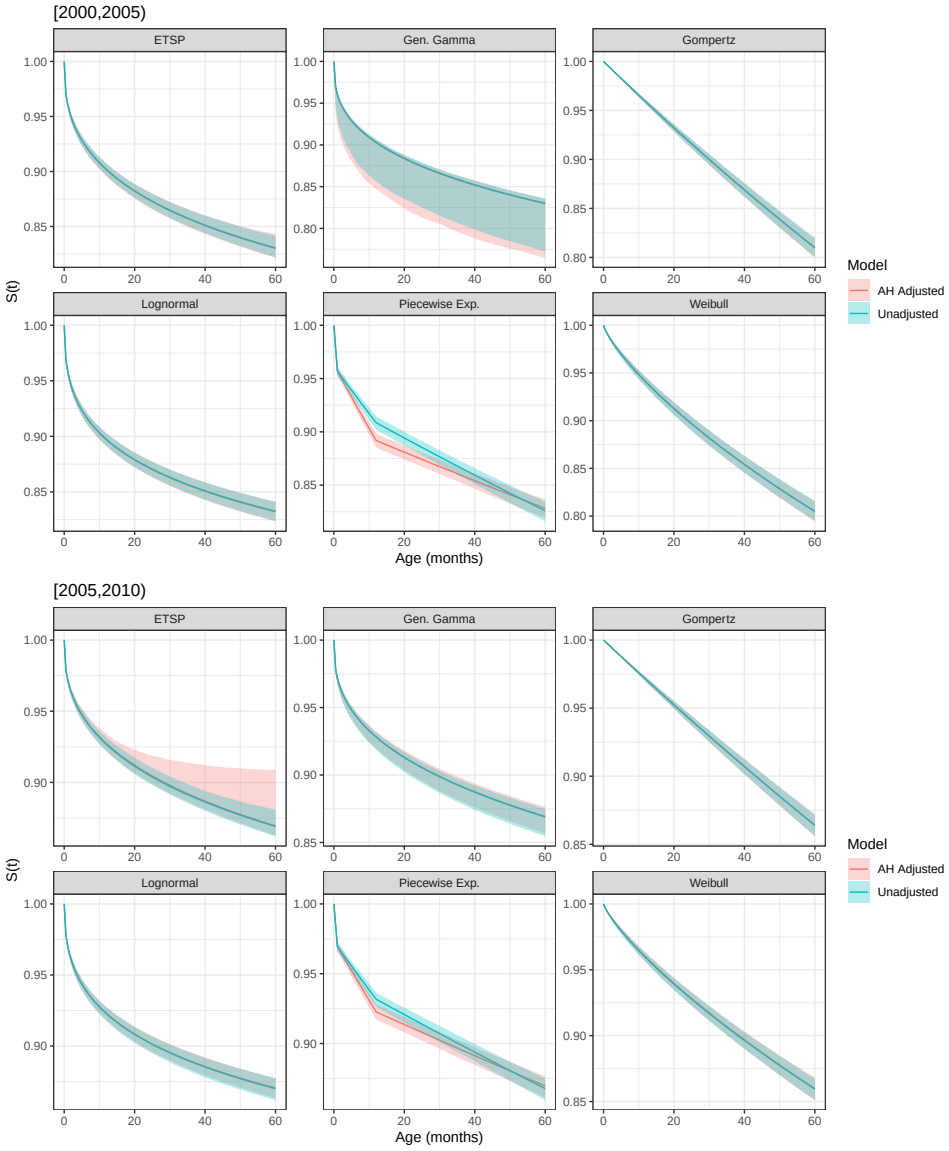


Figure B.26: Comparison of parametric models where data is adjusted for age-heaping at 12 months versus not for Burkina Faso in periods $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom).

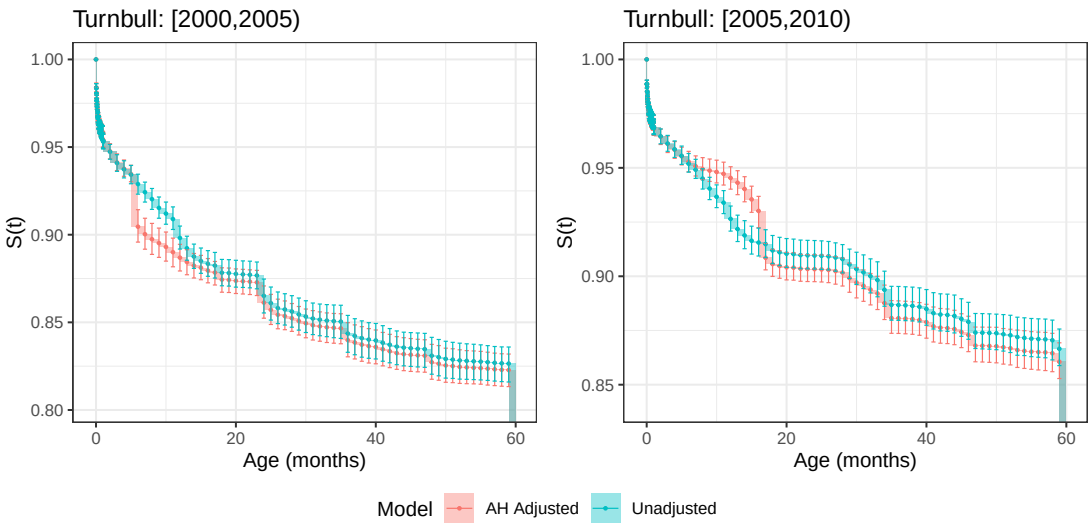


Figure B.27: Comparison of Turnbull estimator where data is adjusted for age-heaping at 12 months versus not for Burkina Faso in periods $[2000, 2005)$ (left) and $[2005, 2010)$ (right).

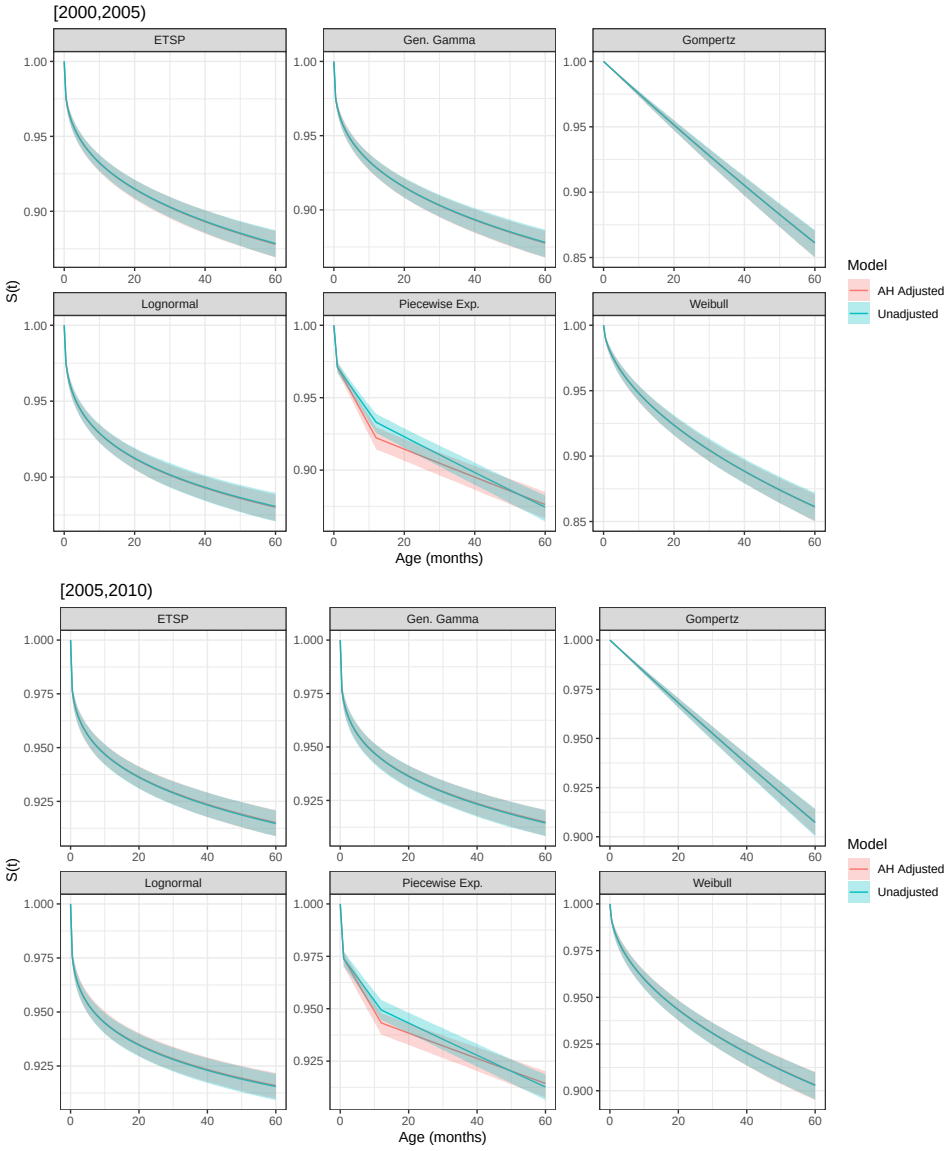


Figure B.28: Comparison of parametric models where data is adjusted for age-heaping at 12 months versus not for Malawi in periods $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom).

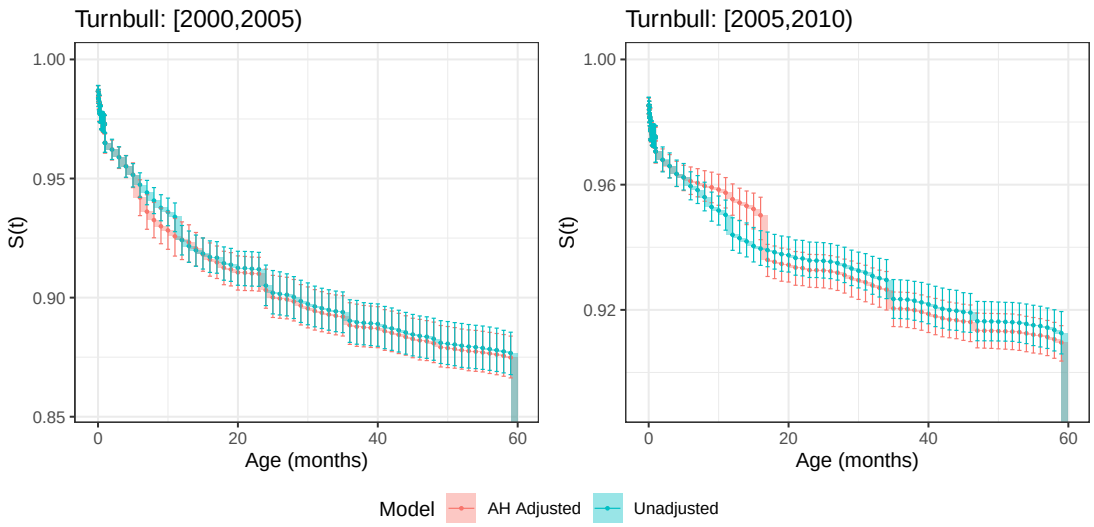


Figure B.29: Comparison of Turnbull estimator where data is adjusted for age-heaping at 12 months versus not for Malawi in periods $[2000, 2005)$ (left) and $[2005, 2010)$ (right).

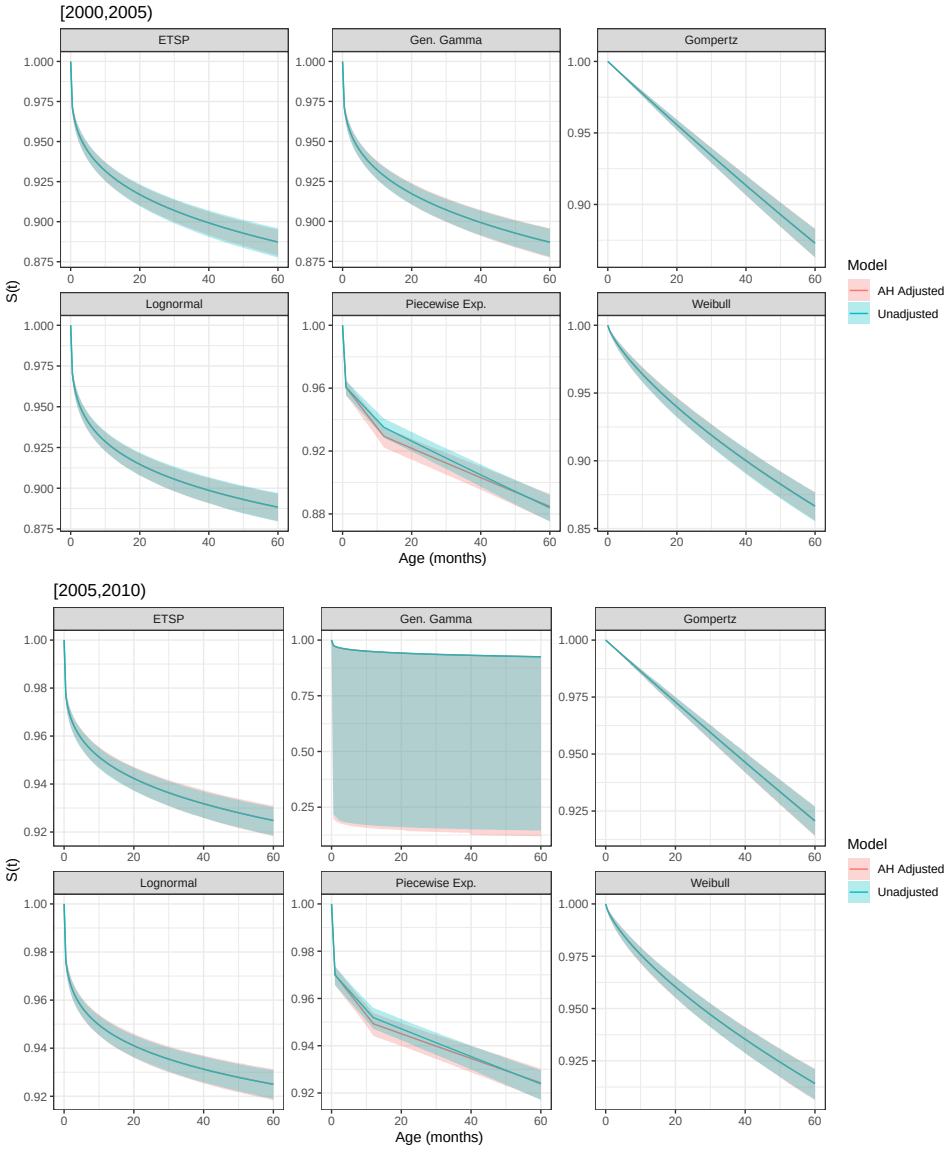


Figure B.30: Comparison of parametric models where data is adjusted for age-heaping at 12 months versus not for Senegal in periods $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom).

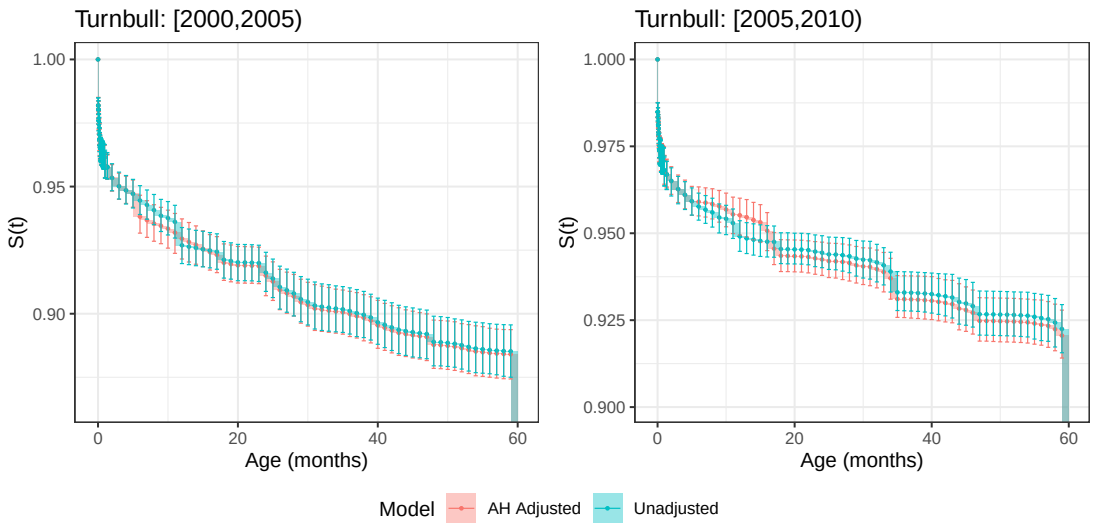


Figure B.31: Comparison of Turnbull estimator where data is adjusted for age-heaping at 12 months versus not for Senegal in periods $[2000, 2005)$ (left) and $[2005, 2010)$ (right).

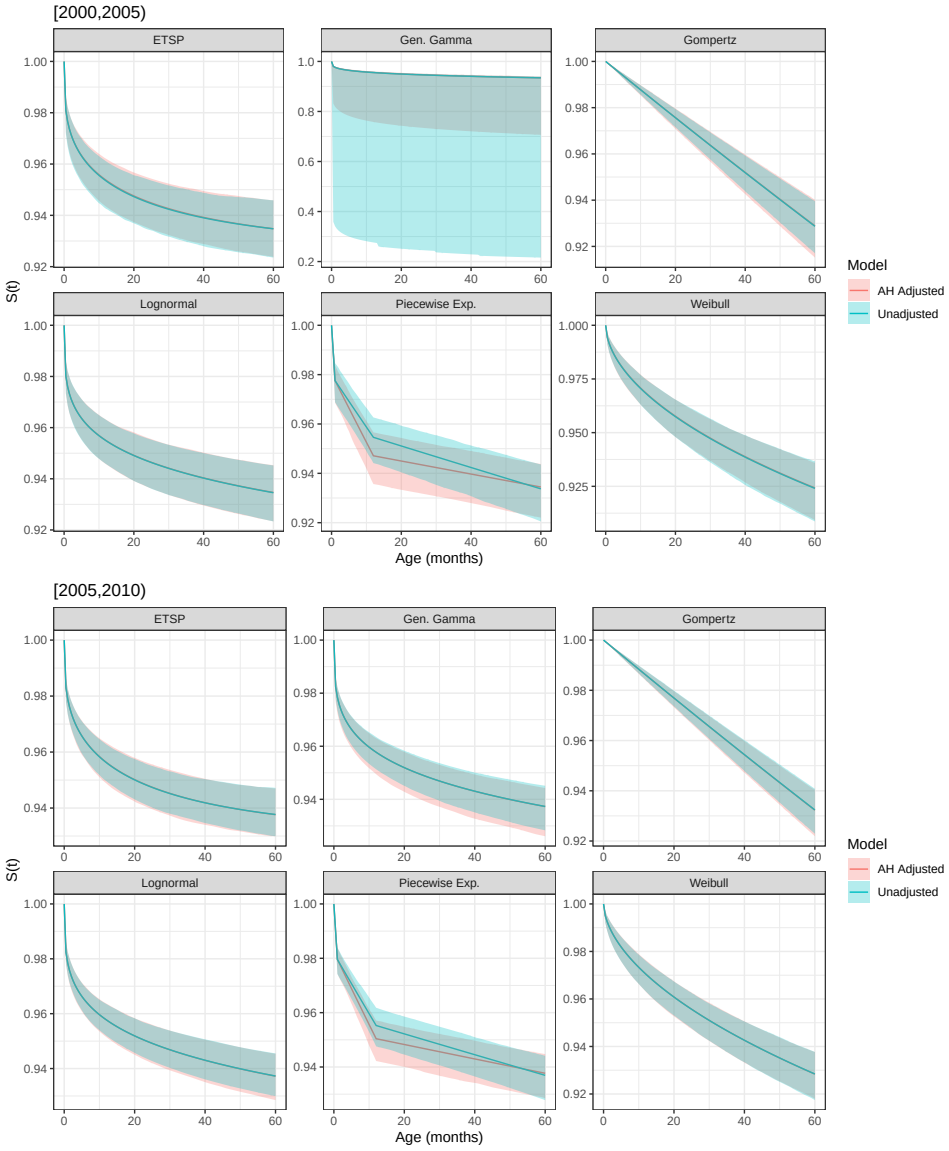


Figure B.32: Comparison of parametric models where data is adjusted for age-heaping at 12 months versus not for Namibia in periods $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom).

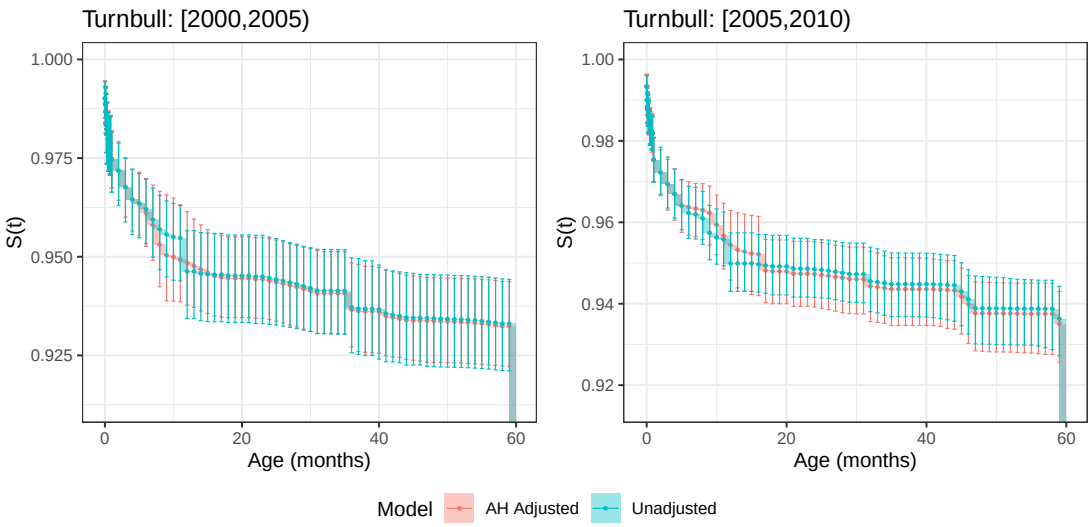


Figure B.33: Comparison of Turnbull estimator where data is adjusted for age-heaping at 12 months versus not for Namibia in periods $[2000, 2005)$ (left) and $[2005, 2010)$ (right).

Appendix C

APPENDIX FOR CHAPTER 4

C.1 Projection matrices for RW2 and ICAR priors

Note that for a RW1, the projection matrix $\mathbf{R}^\top$ is simply the lower triangular matrix from an LDL decomposition of the precision matrix $\mathbf{Q}$ for the RW1. We can obtain the projection matrices for a RW2 similarly. Note that for a RW2, the density is easily defined as a product of iid normally distributed random variables with precision τ in the space spanned by the $(t-2)$-dimensional vector

$$\begin{pmatrix} x_1 - 2x_2 + x_3 \\ x_2 - 2x_3 + x_4 \\ \vdots \\ x_{t-2} - 2x_{t-1} + x_t \end{pmatrix}$$

Then the projection matrix that takes us from $\mathbf{x}$ to the space spanned above is given by the $(t-2) \times t$ matrix

$$\mathbf{R}^\top = \begin{pmatrix} 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -2 & 1 \end{pmatrix},$$

where we note that $\mathbf{R}^\top$ is exactly the lower triangular matrix derived from an LDL decomposition of the known precision matrix for a RW2.

The projection matrix for an ICAR prior may be similarly obtained. Note that an

ICAR is defined on neighborhood differences, where the mean of an area a priori is equal to the average of its neighbors. Let n_i denote the number of neighbors belonging to area $i = 1, \ldots, N$, and let $i \sim j$ denote that areas $i \neq j$ are neighbors. Then we can write the density as a product of iid normally distributed random variables with precision τ in the space spanned by the N-dimensional vector

$$\begin{pmatrix} x_1 - \frac{1}{n_1} \sum_{1 \sim j} x_j \\ x_2 - \frac{1}{n_2} \sum_{2 \sim j} x_j \\ \vdots \\ x_N - \frac{1}{n_N} \sum_{N \sim j} x_j \end{pmatrix}$$

And then the projection matrix that takes us from $\mathbf{x}$ to the space spanned above is given by the $N \times N$ matrix defined by

$$\mathbf{R}_{ij} = \begin{cases} 1, & i = j \\ -1/n_i, & i \sim j, i < j \\ 0, & \text{else} \end{cases}$$

C.2 Correlated multivariate random walks in INLA

In listing C.1 we include code to construct the correlated multivariate random walk in INLA. After the model is defined with `inla.rgeneric.define()`, the use may then specify `f(idx, model = mvrw.cor.model, ...)` in the formula for the linear predictor in an INLA call to include the random effect. Note that to constrain the random effect, you must use the `extraconstr` argument, you cannot simply let `constr = TRUE`.

The user must specify four things: a matrix `W`, which contains 1's where the random effect is non-zero, and 0's where the random effect is zero; the projection matrix `R`, which will vary for first-order and second-order RWs (as well as ICARs); the number of discrete time points

`n`, and the order of the random walk. If the user plans to use this code for a correlated ICAR model, the order of the random walk should be set to 0, as the projection matrix does not project into a space of smaller dimension.

To adapt this code to provide a multivariate random walk on two parameters that have different precisions but no correlation, `rho` should be removed from the list in `interpret.theta`, the off-diagonals in `Sig` should be set to 0, the log prior for `rho` should be removed (line 51 of Listing C.1, and `initial` should return a vector of only two zeros.

The hyperpriors included in the model as written are PC priors (Simpson et al., 2017) with $U = 1, \alpha = 0.01$ on each precision parameter, and a Normal(0,1) prior on a transformed correlation parameter. These can be adapted as desired.

```
1   'inla.rgeneric.MVRW.cor.model' <- function(cmd = c("graph", "Q", "
      mu", "initial", "log.norm.const","log.prior", "quit"),
2     theta = NULL) {
3
4     #Internal function
5     interpret.theta <- function() {
6       return(
7       list(prec_x = exp(theta[1L]),
8            prec_w = exp(theta[2L]),
9            transformed_rho = theta[3L])
10      )
11    }
12
13    graph <- function() {
14      require(Matrix)
15      # W will just be input as the graph (an INLA sparse matrix)
16      return(W)
17    }
18
```

```
19      Q ← function() {
20          require(Matrix)
21
22          param ← interpret.theta()
23
24          rho_internal ← (2 / (1 + exp(-param$transformed_rho))) - 1
25
26          Sig ← matrix(c(1/param$prec_x, rho_internal * sqrt(1/param
                $prec_x) * sqrt(1/param$prec_w), rho_internal * sqrt(1/
                param$prec_x) * sqrt(1/param$prec_w), 1/param$prec_w),
                byrow = TRUE, nrow = 2)
27
28          temp ← t(R) %*% (diag(n-rw_order) %x% solve(Sig)) %*% R
29          temp ← inla.as.sparse(temp)
30          return( temp )
31      }
32
33      mu ← function()  {
34         return(numeric(0))
35      }
36
37      log.norm.const ← function() {
38         return(numeric(0))
39      }
40
41      log.prior ← function() {
42         param = interpret.theta()
43
44         U ← 1
```

```
45          alpha ← 0.01
46          lambda ← -log(alpha / U)
47
48          # PC priors with U = 1, alpha = 0.01  on precisions
49          res ← log(lambda) - log(2) - (3/2) * log(param$prec_x) -
               lambda * param$prec_x^(-1/2) +
50          log(param$prec_x) +
51          log(lambda) - log(2) - (3/2) * log(param$prec_w) - lambda *
                param$prec_w^(-1/2) +
52           log(param$prec_w) +
53           dnorm(param$transformed_rho, mean = 0, sd = 1, log = TRUE)
54
55           return(res)
56        }
57
58        initial ← function() {
59          return(c(0, 0, 0))
60        }
61
62        quit ← function() {
63          return(invisible())
64        }
65
66        res ← do.call(match.arg(cmd), args = list())
67          return(res)
68    }
69
70    mvrw.cor.model ← inla.rgeneric.define(
         inla.rgeneric.MVRW.cor.model,
```

```
71        W = inla.as.sparse(W), # Structure matrix of only 0's and 1's
72        R = R, # Projection matrix
73        n = n, # number of discrete time points
74        rw_order = 1)
75        # 1 if RW1, 2 if RW2 (make sure A is specified accordingly)
```

Listing C.1: INLA rgeneric code for correlated multivariate random walk.

C.3 Simulation study

To assess the performance of the proposed (correlated and uncorrelated) MVRW models, we conduct a simulation study. We consider the following combinations of parameter values, resulting in 16 simulation settings:

- ρ: 0, 0.1, 0.5, 0.9
- σ_x: 0.5, 5
- σ_w: 0.5, 5

In each simulation setting, we generate 100 observations $x_1, \ldots, x_{100}$ and $w_1, \ldots, w_{100}$ from the model given by

$$(\mathbf{x}, \mathbf{w}) \mid \boldsymbol{\eta} \sim \text{MultivariateNormal}(\boldsymbol{\eta}, \mathbf{I}) \tag{C.1}$$

$$\boldsymbol{\eta} = \psi(\sigma_x, \sigma_w, \rho) \tag{C.2}$$

where $\psi(\sigma_x, \sigma_w, \rho) \sim \text{MVRW2}(\sigma_x, \sigma_w, \rho)$. Samples from the MVRW2 are obtained via Algorithm 3.1 of Rue and Held (2005). We then fit the model given in Equation (C.1) (which we denote the "Correlated model"), and an "Uncorrelated model" (where ρ is set to zero in the MVRW2) to the data, and compute posterior means for the hyperparameters. Rather than a $N(0, 1)$ hyperprior on a transformed ρ, we instead use a Uniform$(-1, 1)$ hyperprior. We repeat the data generation and estimation process 50 times per simulation setting.

The posterior means across all simulations are given in Figure C.1. Note that the Correlated model correctly estimates σ_x, σ_w, and ρ on average, across all simulation settings, while the Uncorrelated model fails to estimate $\rho \neq 0$ (by design), and seems to underestimate σ_x and overestimate σ_w on average for the settings where $\sigma_x = 5$ and $\sigma_w = 5$, respectively.

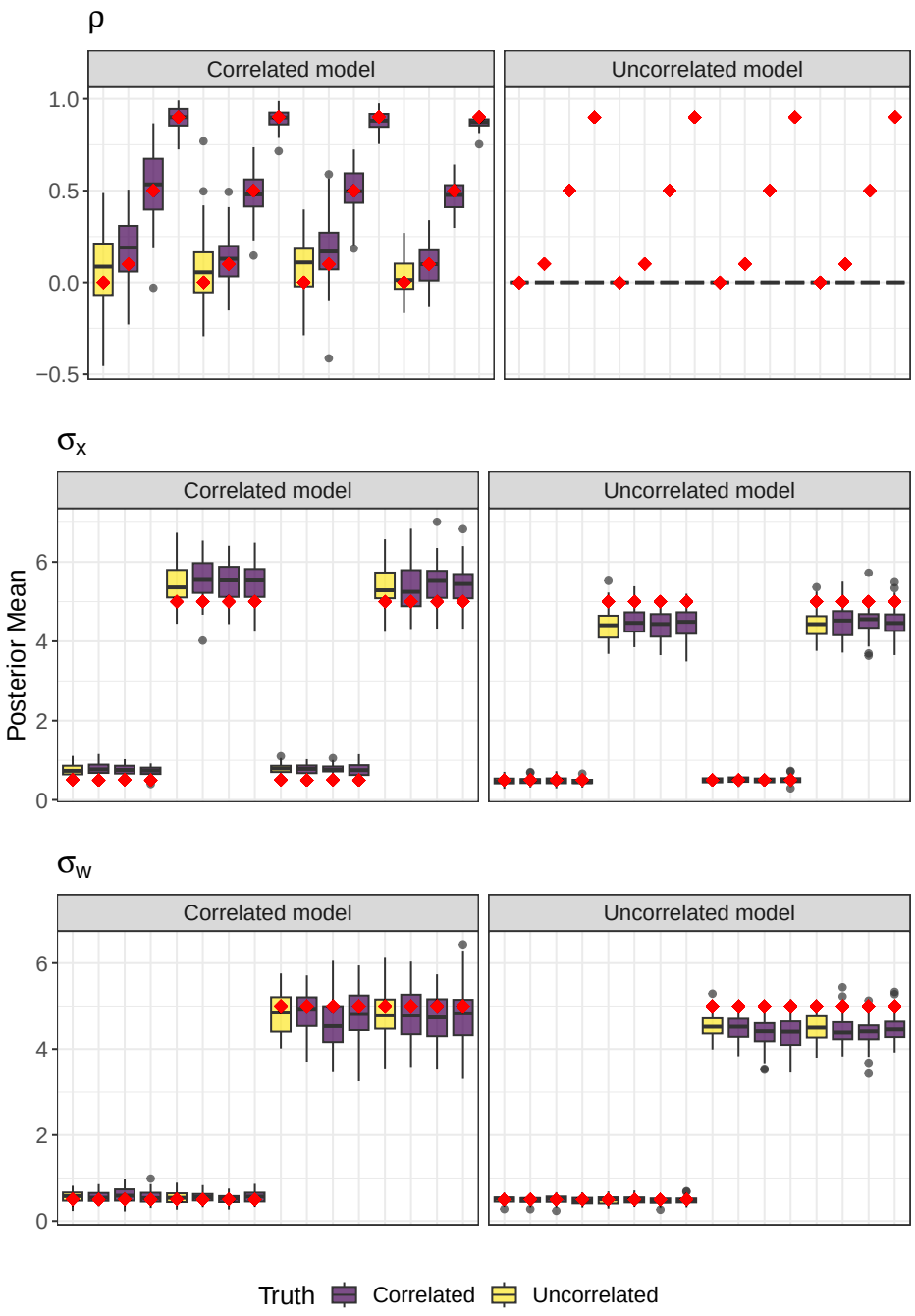


Figure C.1: Posterior mean estimates of hyperparameters ρ, σ_x, and σ_w across all simulation settings considered. True underlying parameter values are denoted with red diamonds. Simulation settings where the observations are truly uncorrelated are given by yellow boxplots, and settings that are truly correlated are given by purple boxplots. Each boxplot contains observations from 50 different simulated datasets.

C.4 Additional results

C.4.1 By model

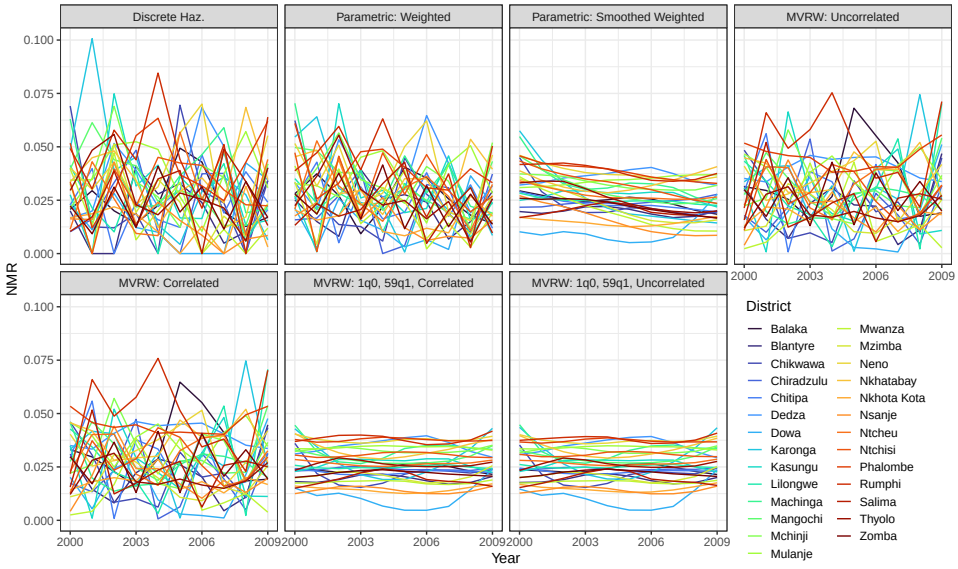


Figure C.2: Mean estimates of NMR from 2000-2009 for 27 districts of Malawi from different smoothing models considered.

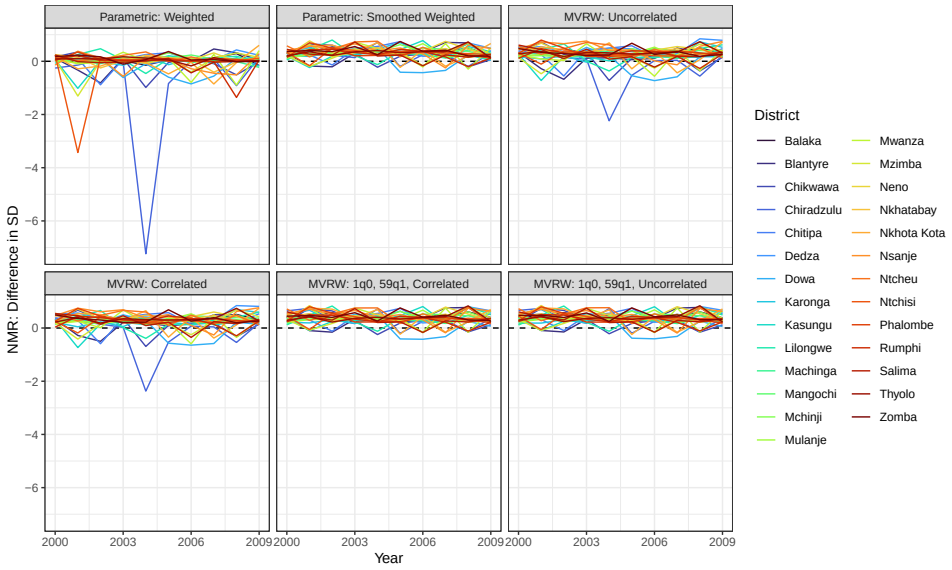


Figure C.3: Standard deviations of estimates of NMR from 2000-2009 for 27 districts of Malawi from different smoothing models considered.

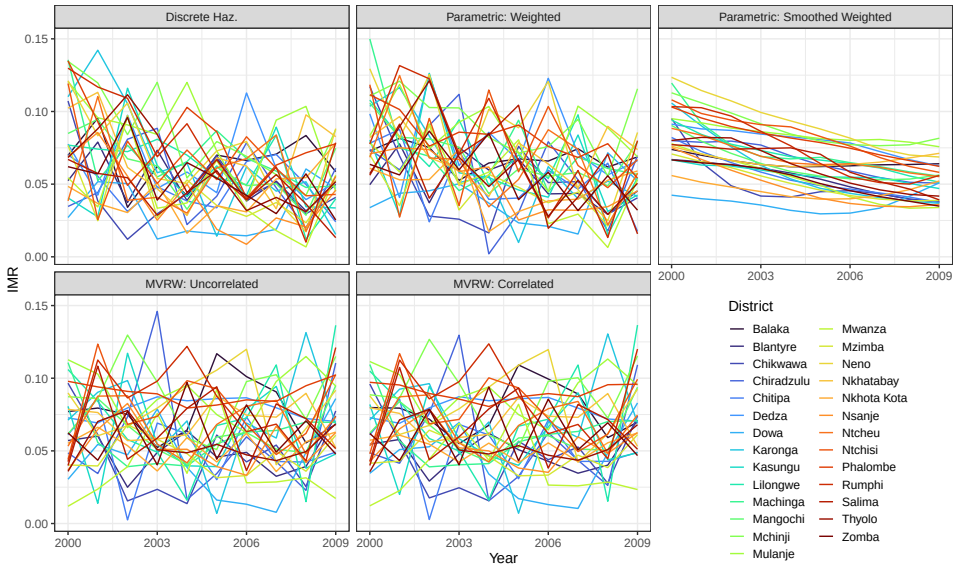


Figure C.4: Mean estimates of IMR from 2000-2009 for 27 districts of Malawi from different smoothing models considered.

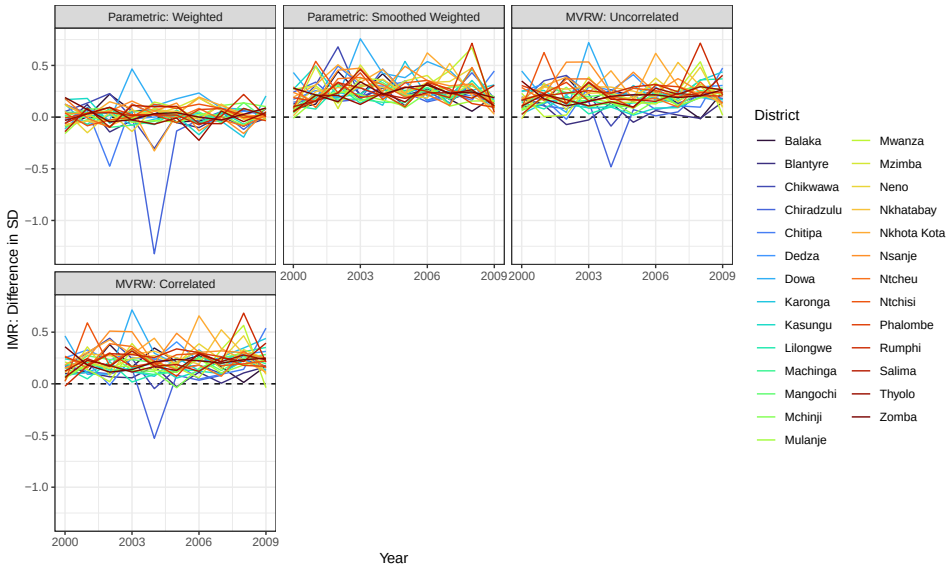


Figure C.5: Standard deviations of estimates of IMR from 2000-2009 for 27 districts of Malawi from different smoothing models considered.

C.4.2 By region

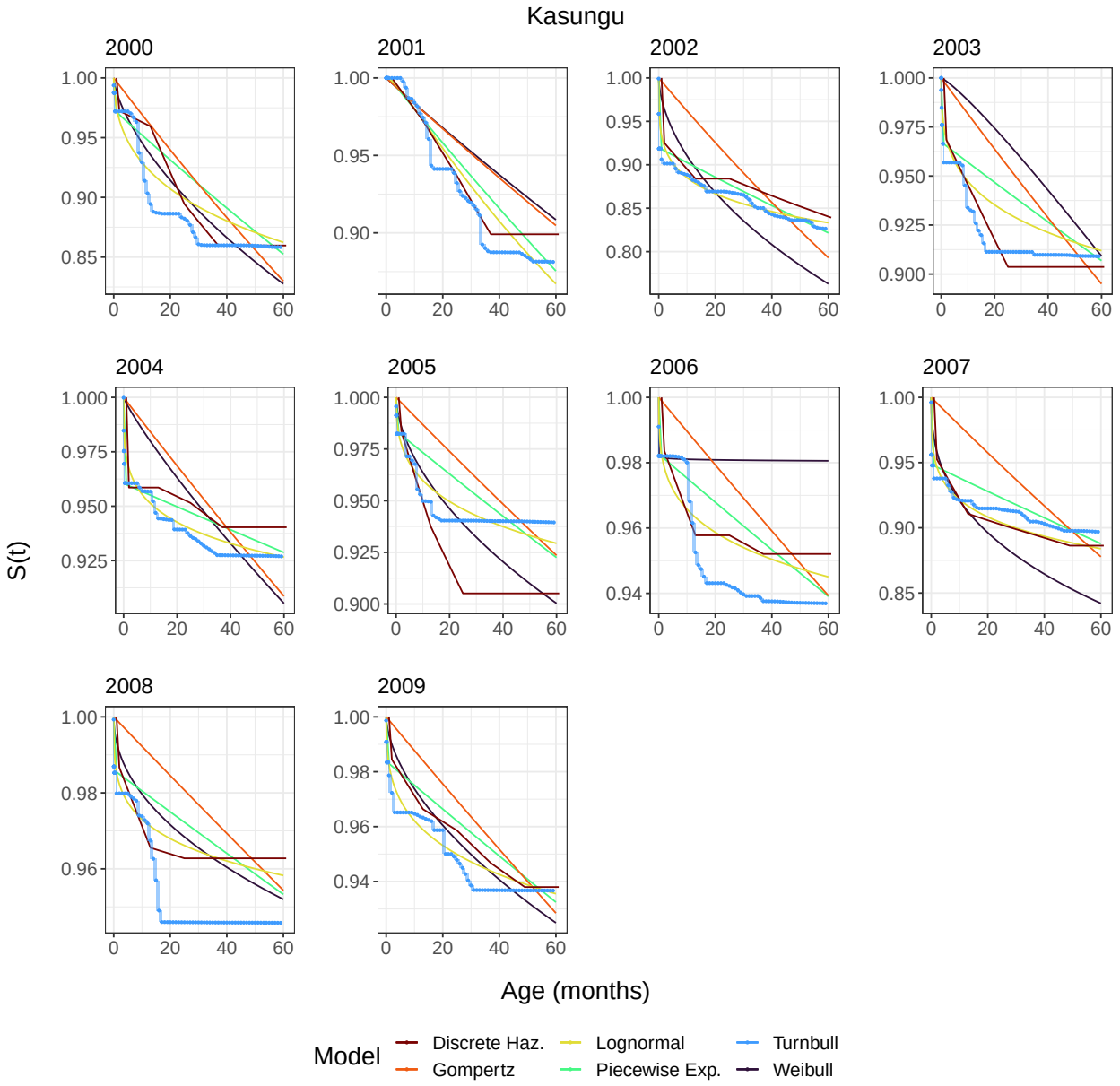


Figure C.6: Survey-weighted survival curves for the Kasungu district of Malawi from 2000-2009.

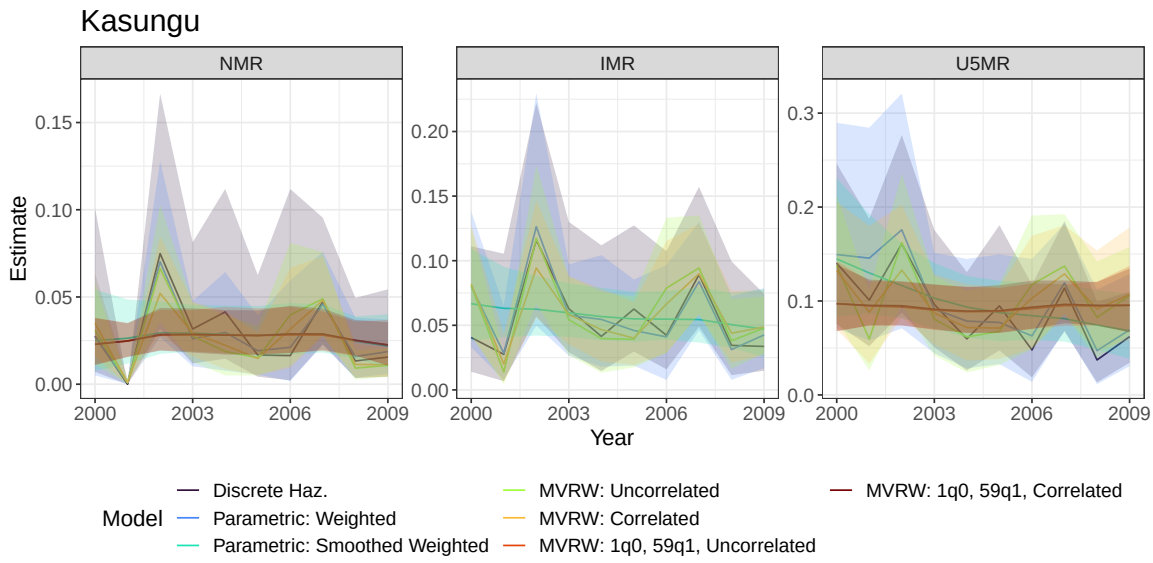


Figure C.7: Estimates of NMR, IMR, U5MR from direct estimates and smoothing models across time for the Kasungu district of Malawi from 2000-2009.

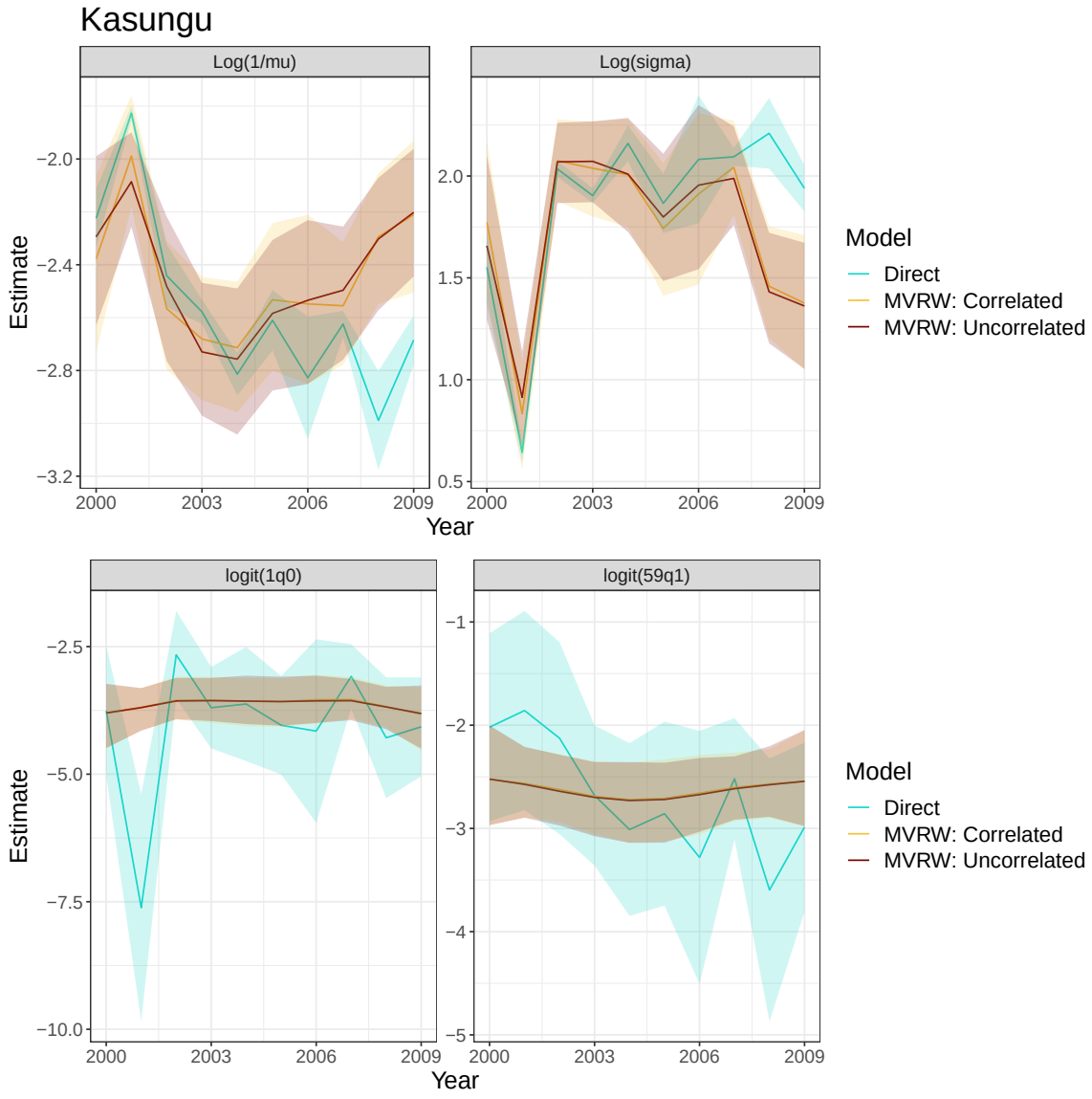


Figure C.8: Smoothed lognormal parameters (Top) and smoothed, transformed lognormal parameters (Bottom) for the Kasungu district of Malawi from 2000-2009.

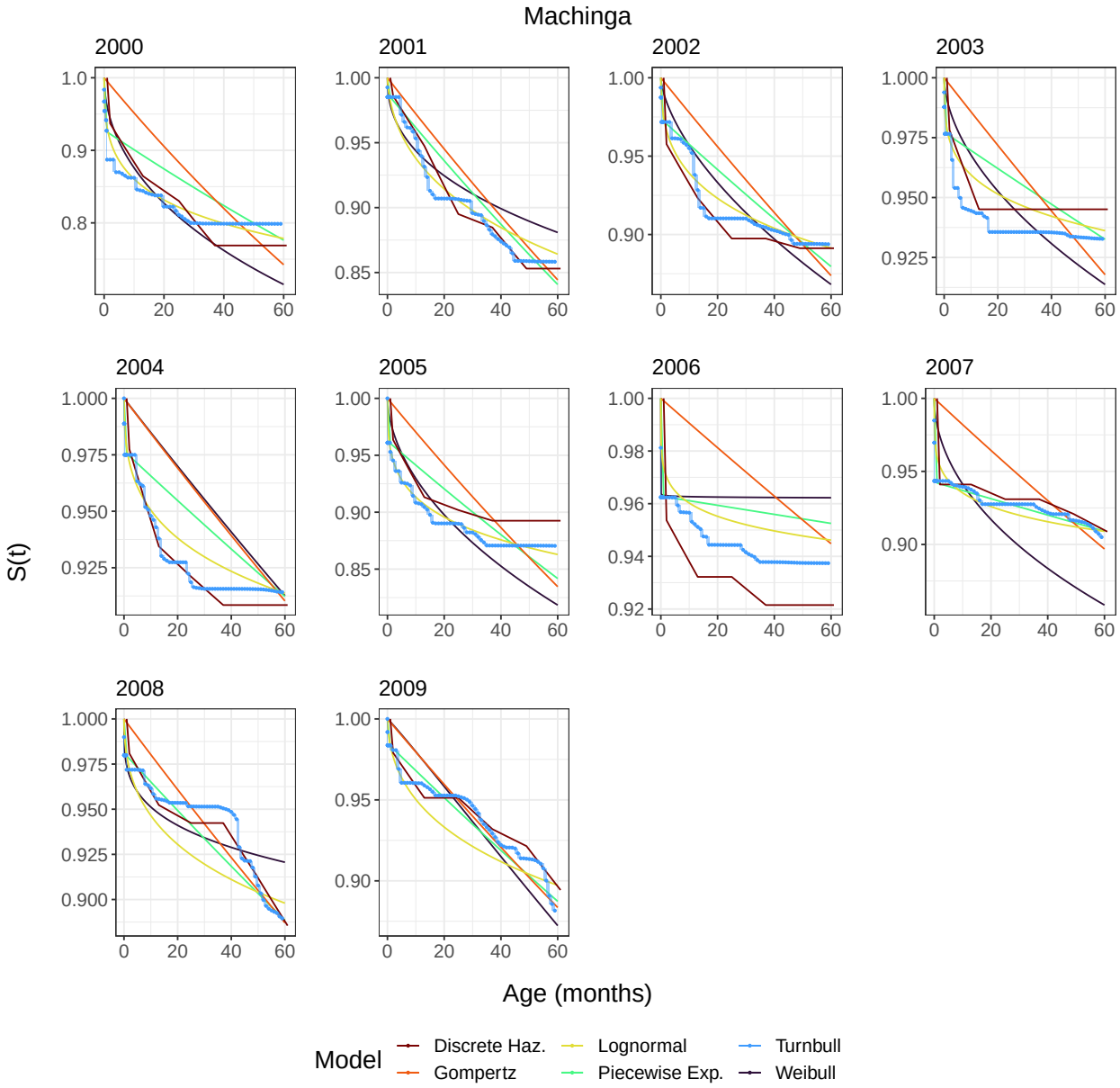


Figure C.9: Survey-weighted survival curves for the Machinga district of Malawi from 2000-2009.

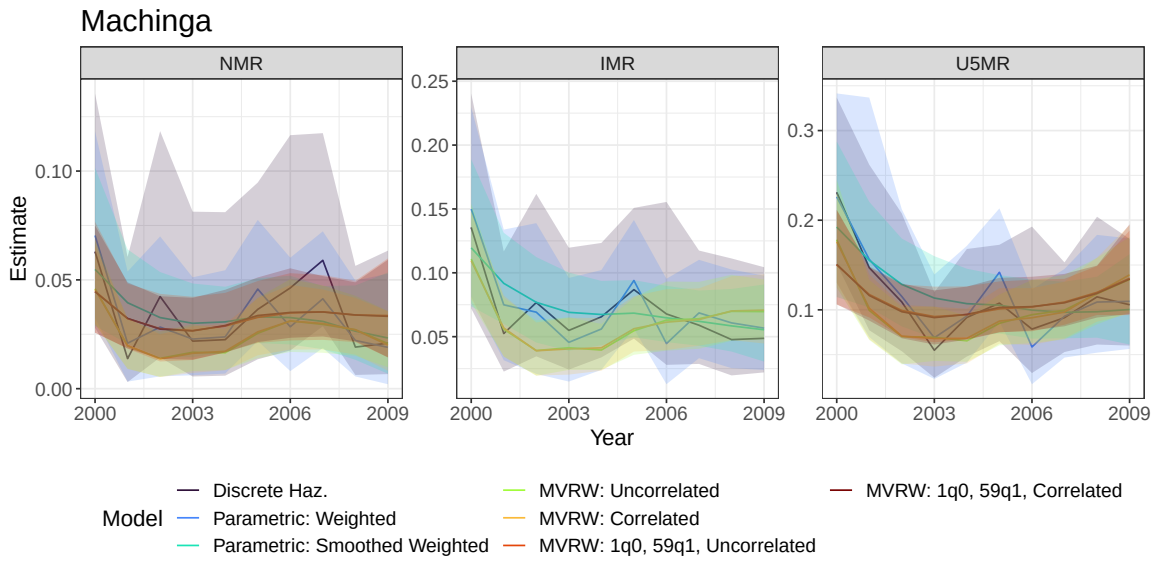


Figure C.10: Estimates of NMR, IMR, U5MR from direct estimates and smoothing models across time for the Machinga district of Malawi from 2000-2009.

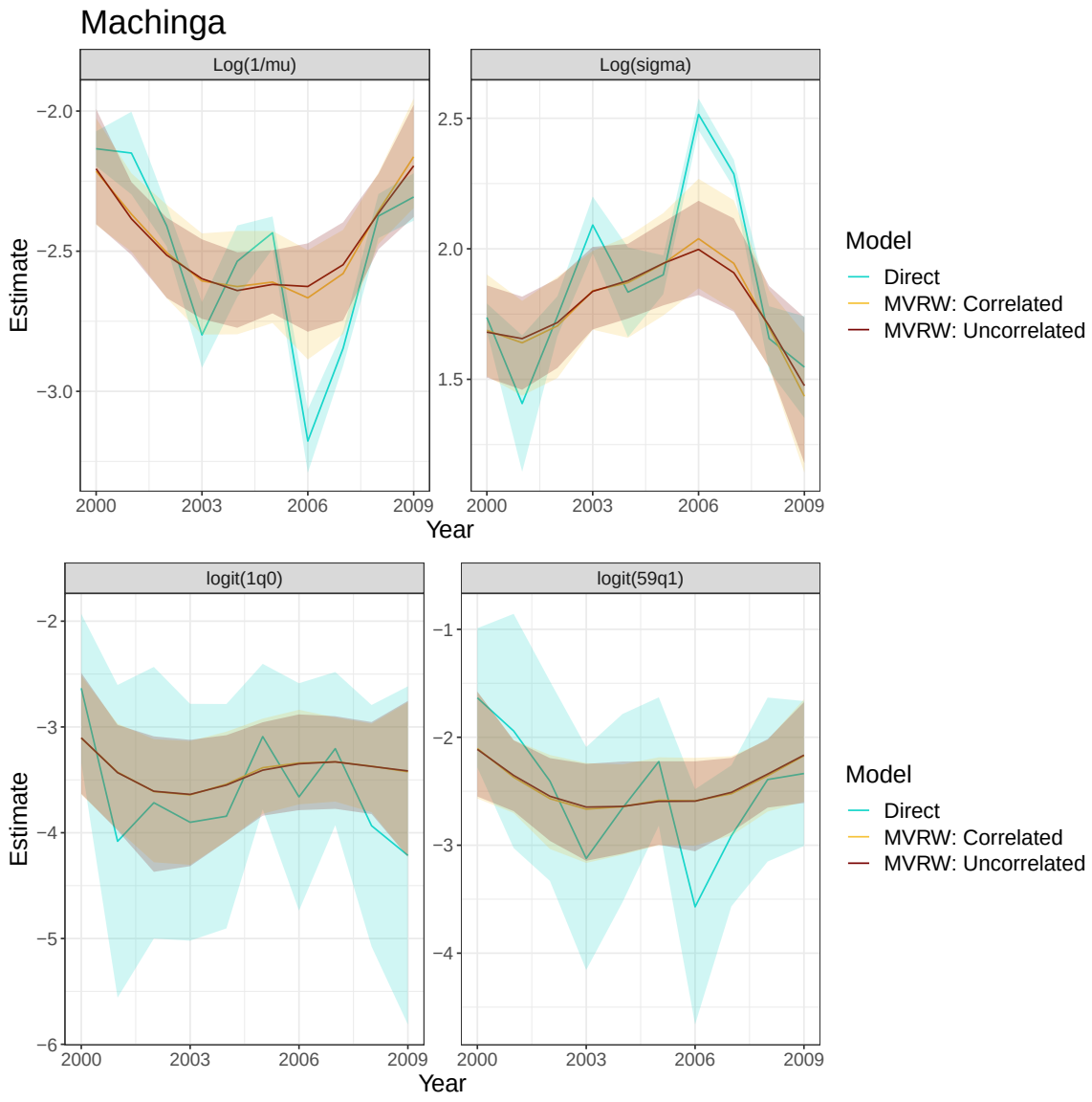


Figure C.11: Smoothed lognormal parameters (Top) and smoothed, transformed lognormal parameters (Bottom) for the Machinga district of Malawi from 2000-2009.

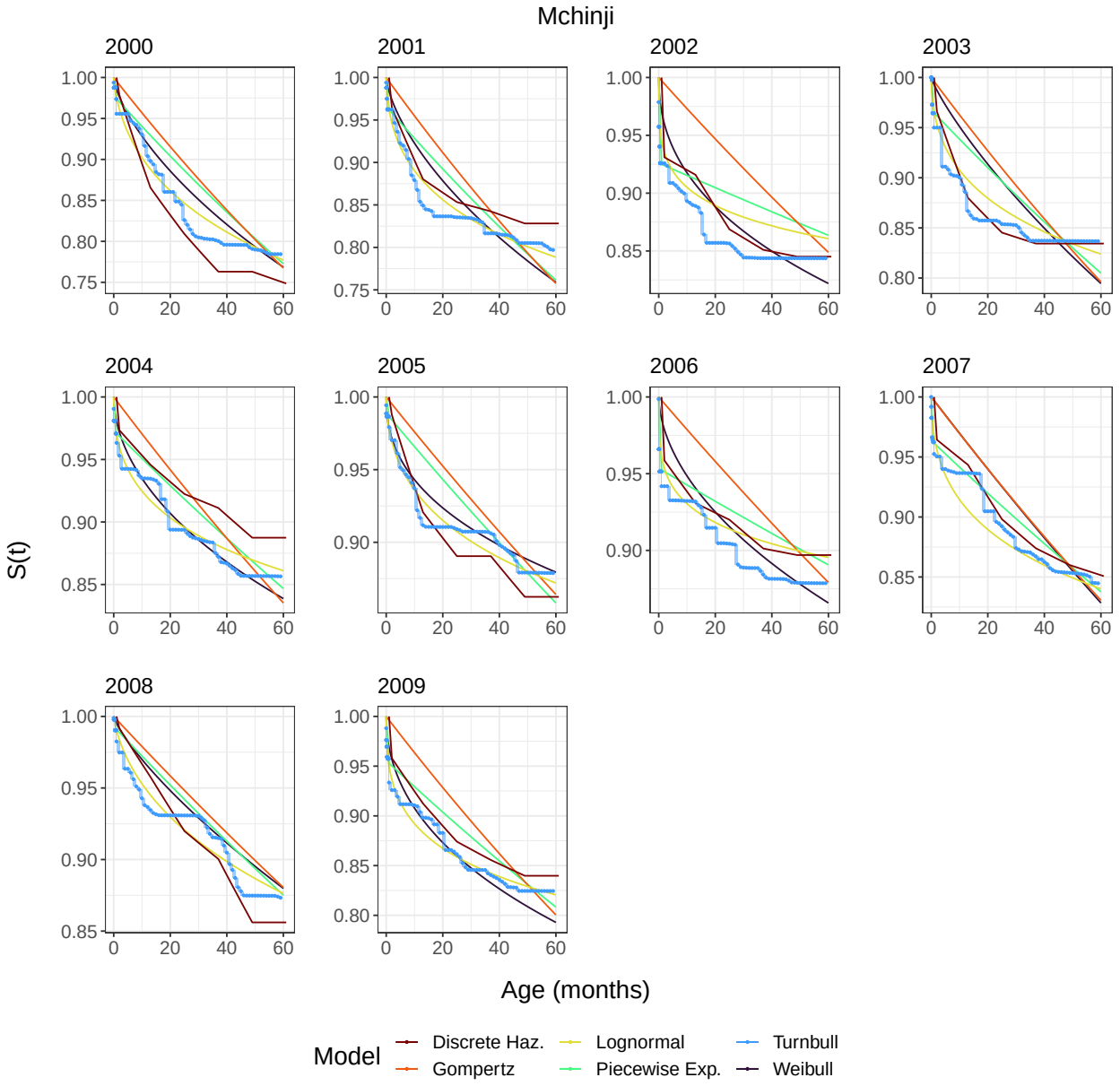


Figure C.12: Survey-weighted survival curves for the Mchinji district of Malawi from 2000-2009.

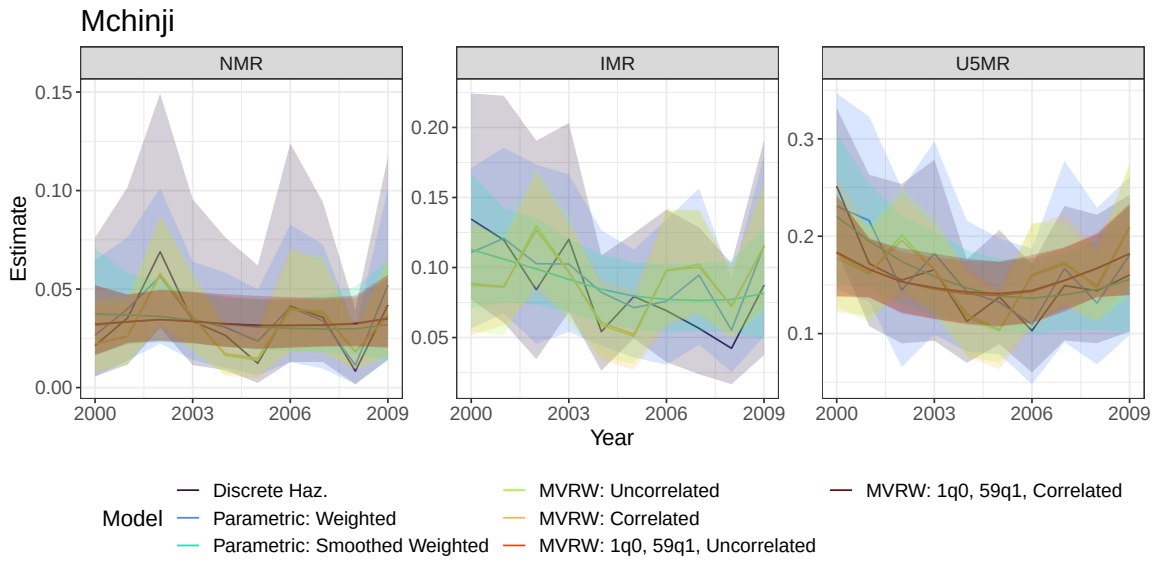


Figure C.13: Estimates of NMR, IMR, U5MR from direct estimates and smoothing models across time for the Mchinji district of Malawi from 2000-2009.

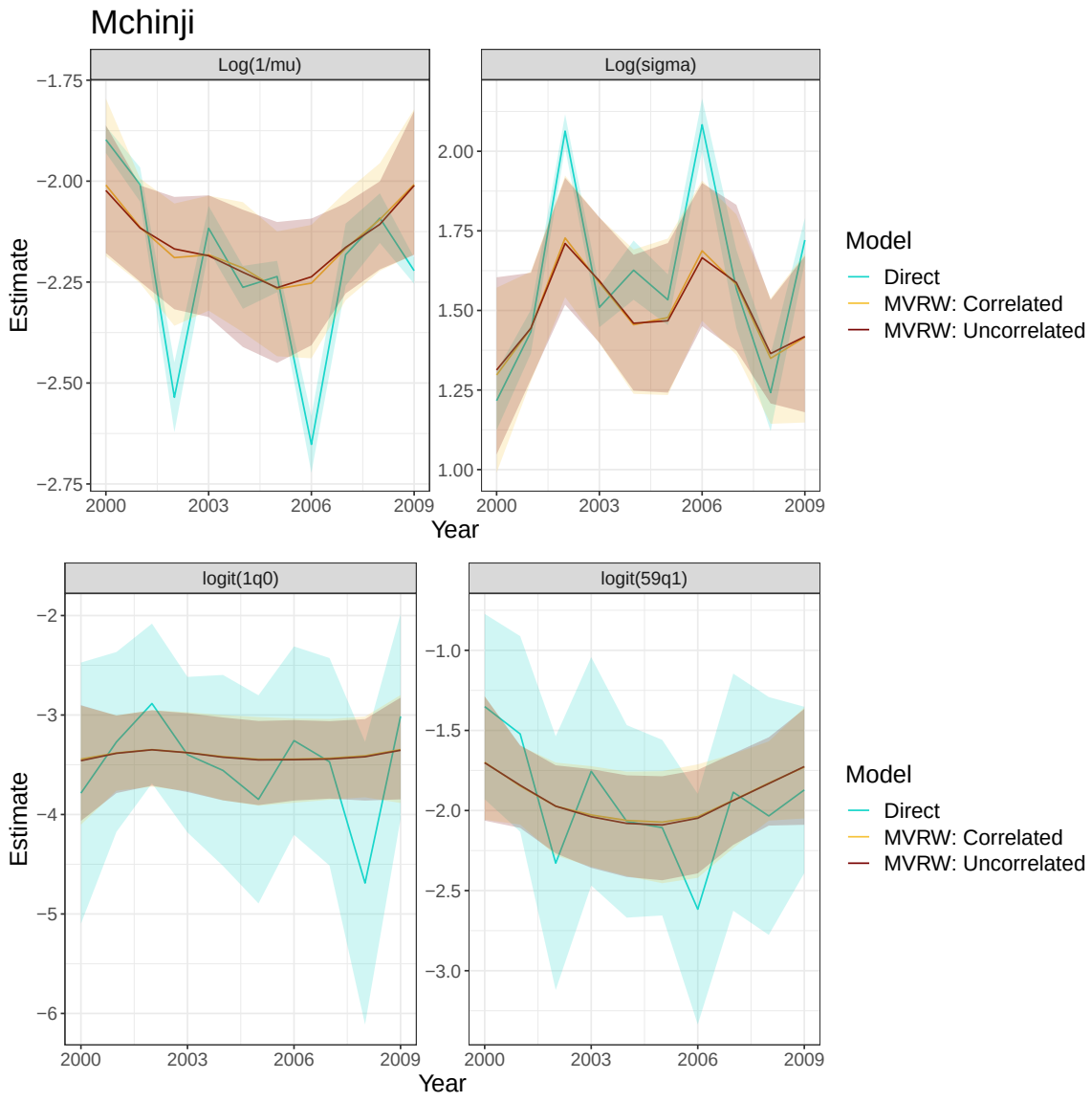


Figure C.14: Smoothed lognormal parameters (Top) and smoothed, transformed lognormal parameters (Bottom) for the Mchinji district of Malawi from 2000-2009.

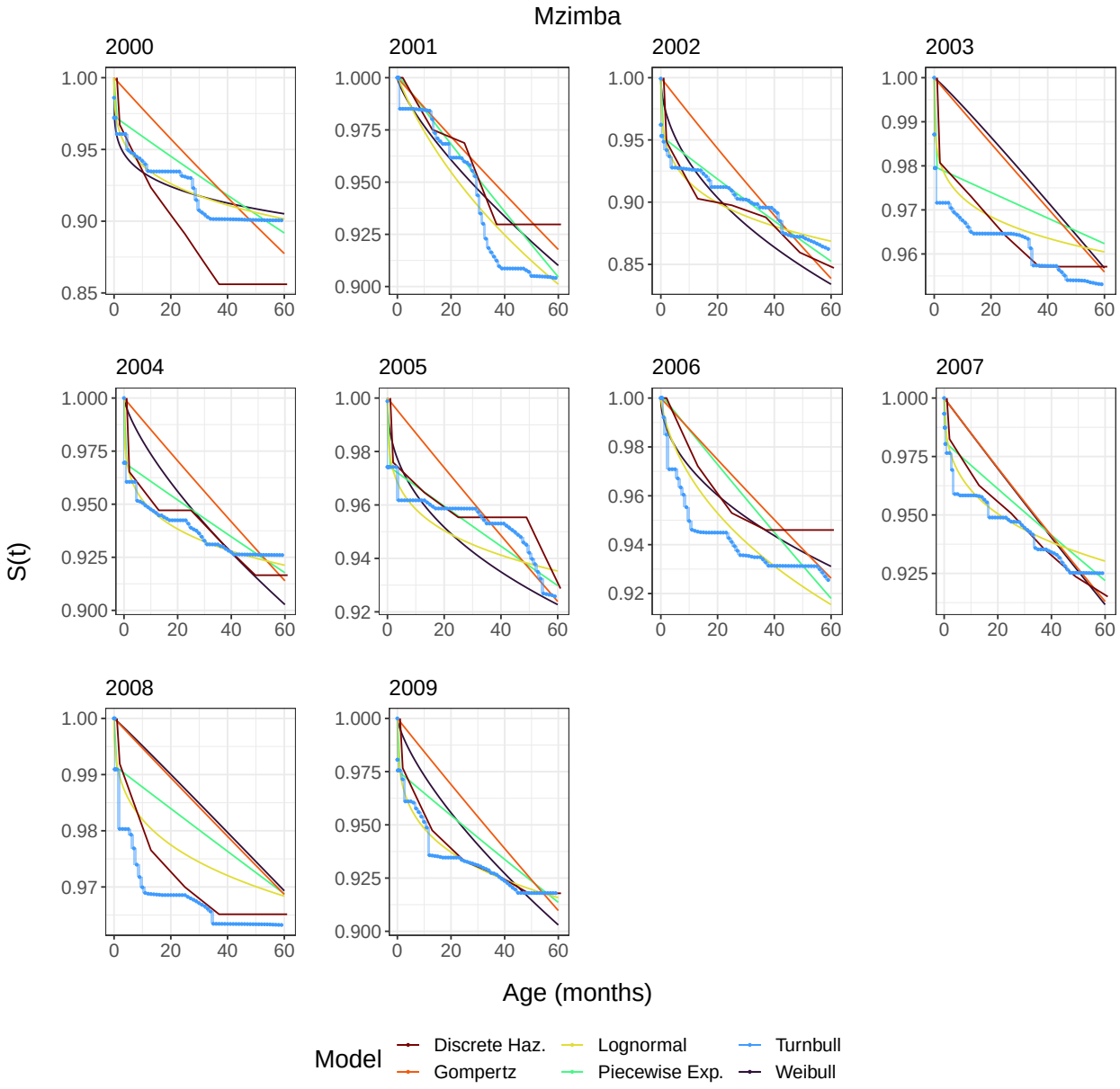


Figure C.15: Survey-weighted survival curves for the Mzimba district of Malawi from 2000-2009.

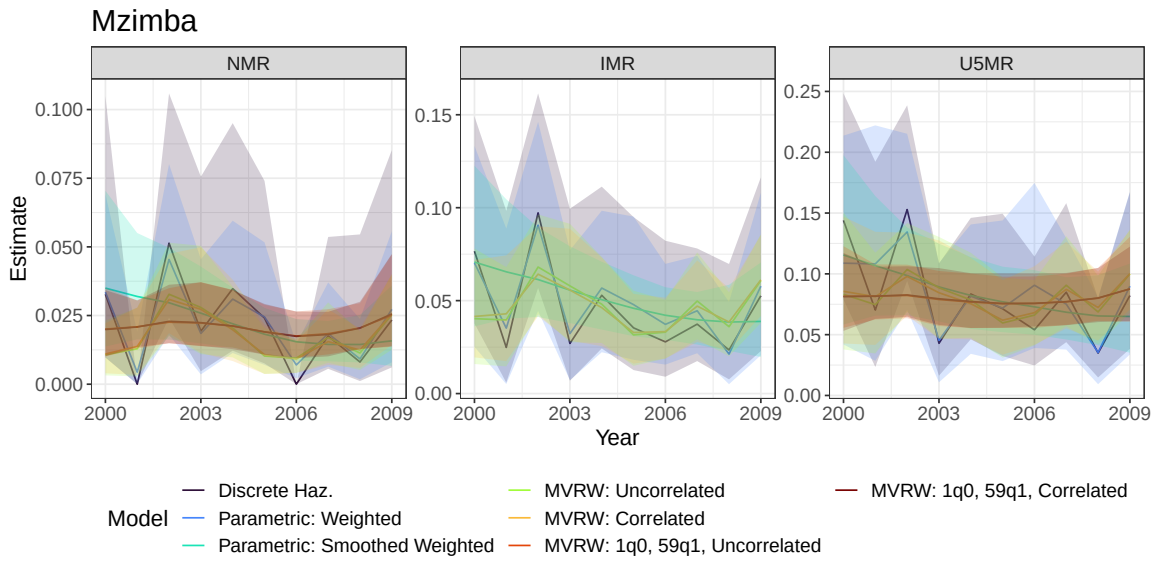


Figure C.16: Estimates of NMR, IMR, U5MR from direct estimates and smoothing models across time for the Mzimba district of Malawi from 2000-2009.

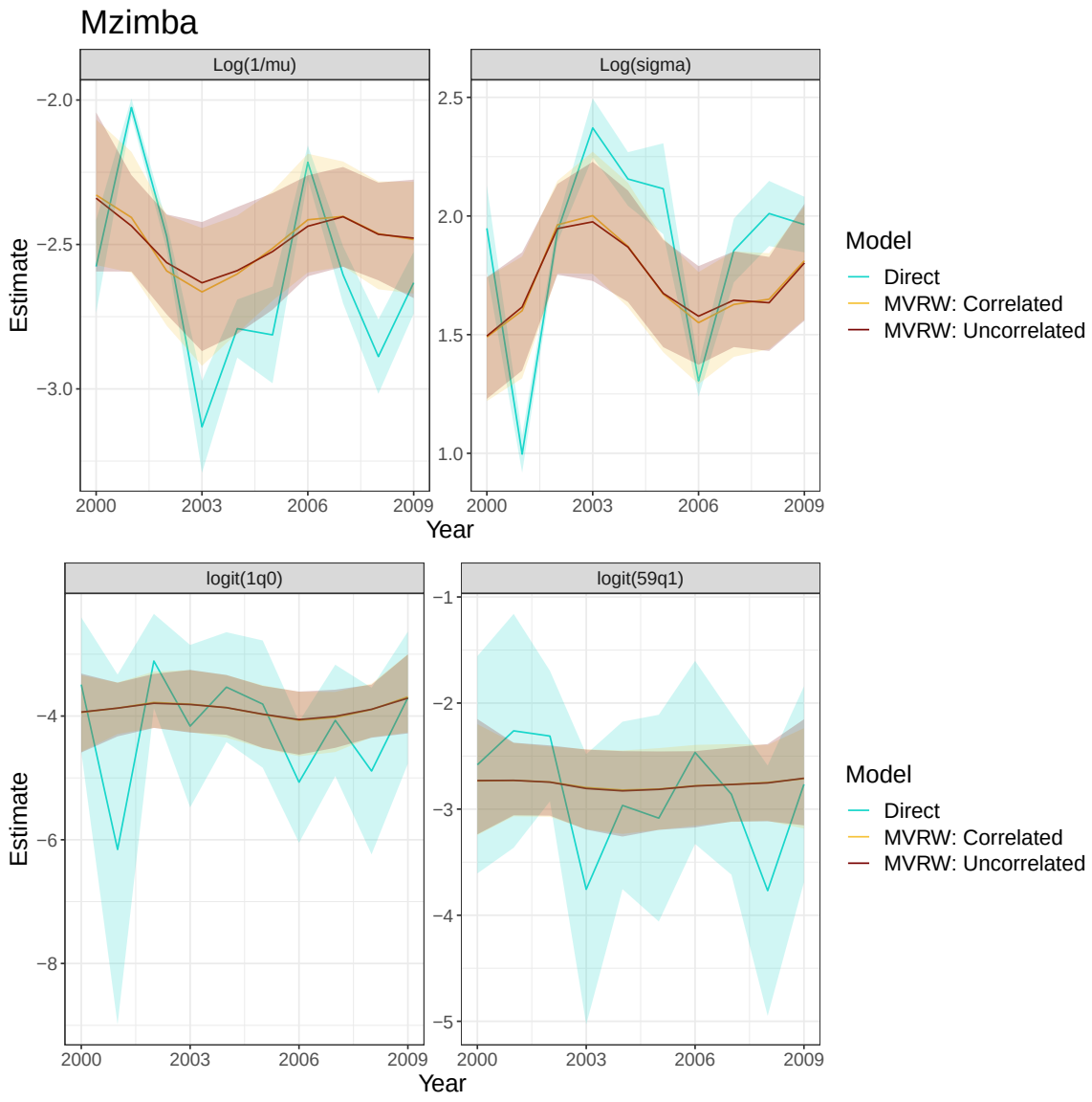


Figure C.17: Smoothed lognormal parameters (Top) and smoothed, transformed lognormal parameters (Bottom) for the Mzimba district of Malawi from 2000-2009.

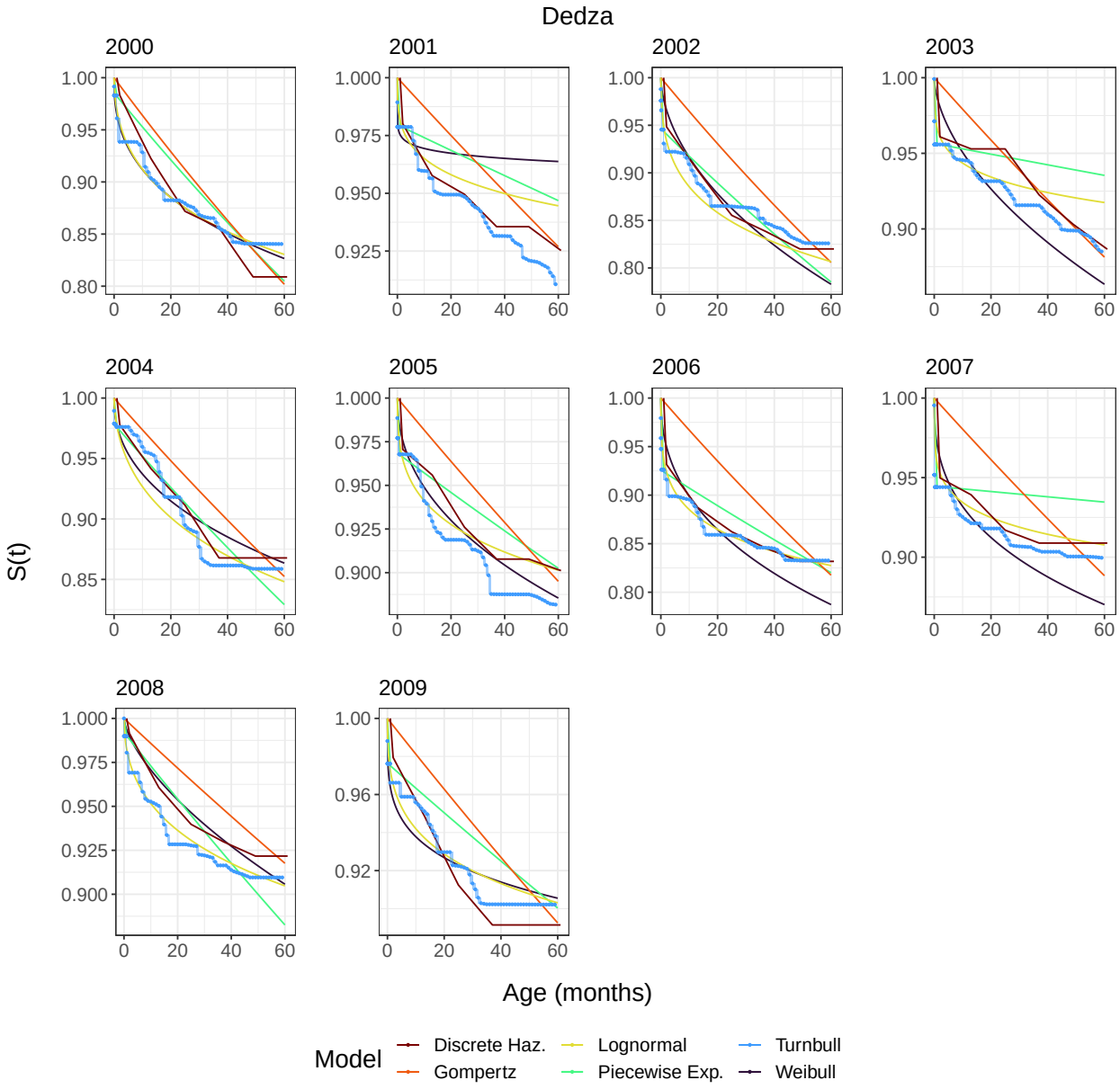


Figure C.18: Survey-weighted survival curves for the Dedza district of Malawi from 2000-2009.

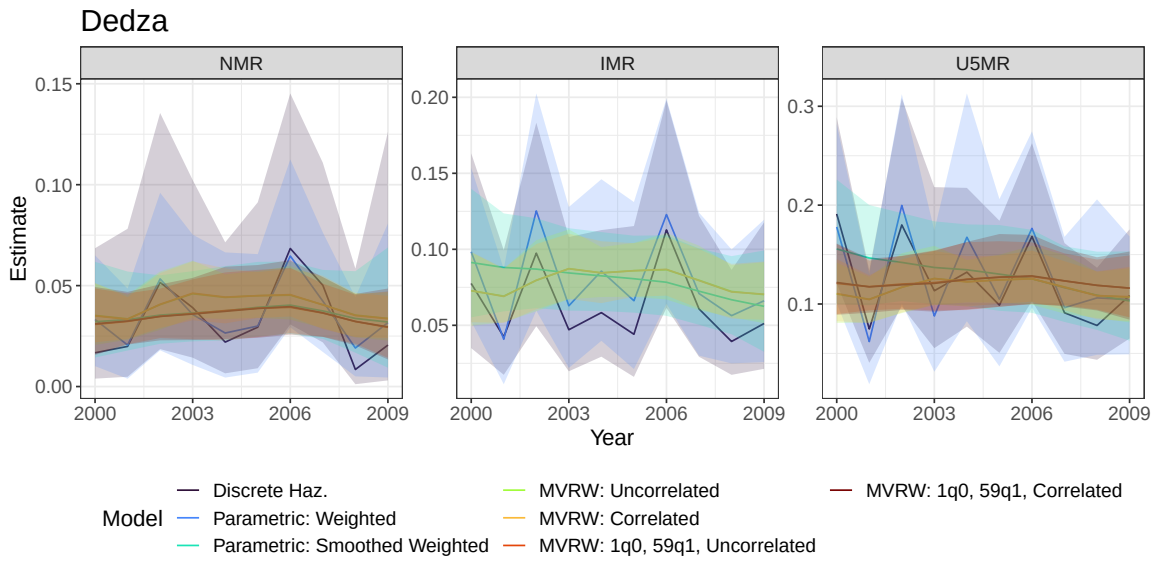


Figure C.19: Estimates of NMR, IMR, U5MR from direct estimates and smoothing models across time for the Dedza district of Malawi from 2000-2009.

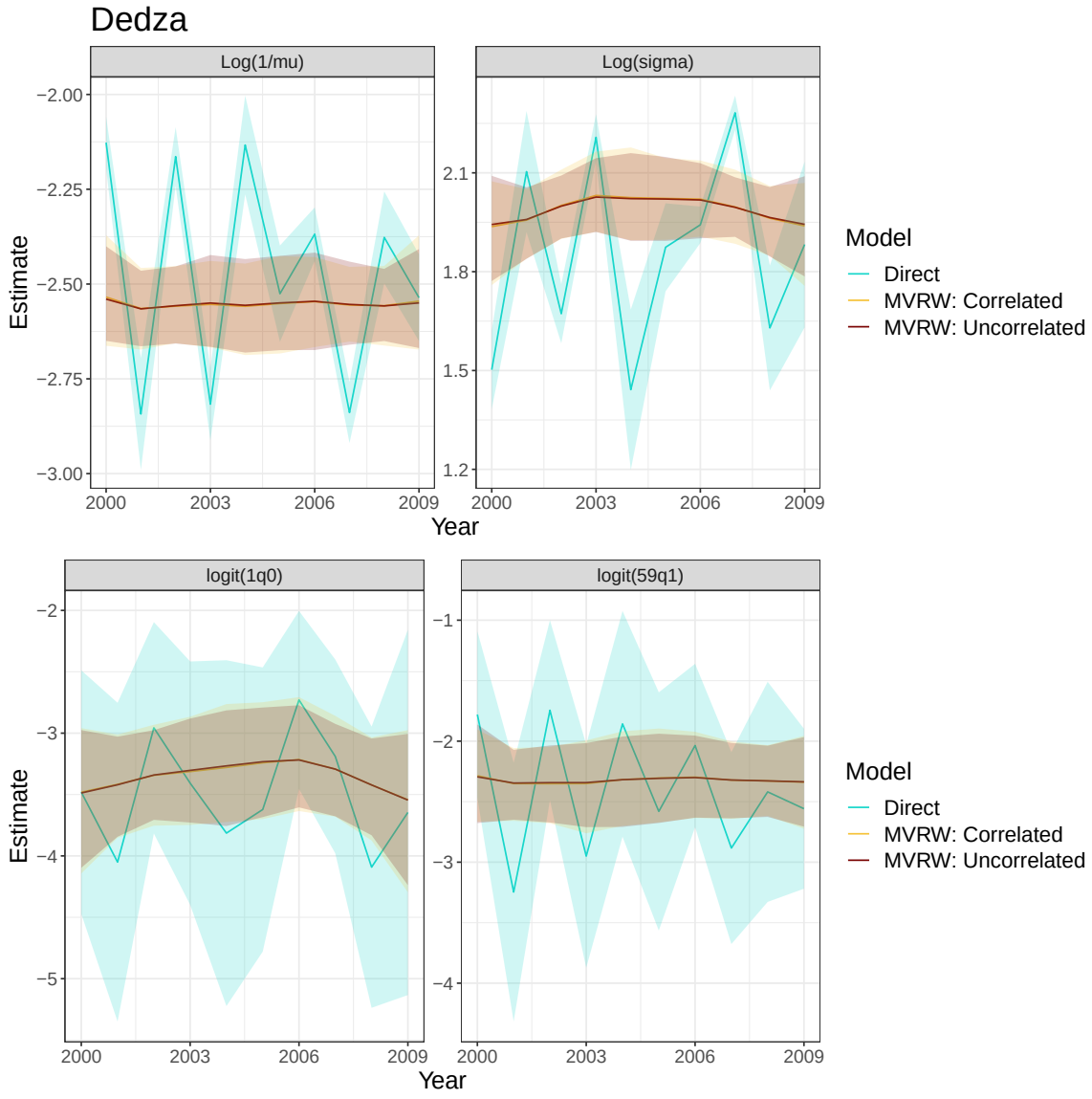


Figure C.20: Smoothed lognormal parameters (Top) and smoothed, transformed lognormal parameters (Bottom) for the Dedza district of Malawi from 2000-2009.

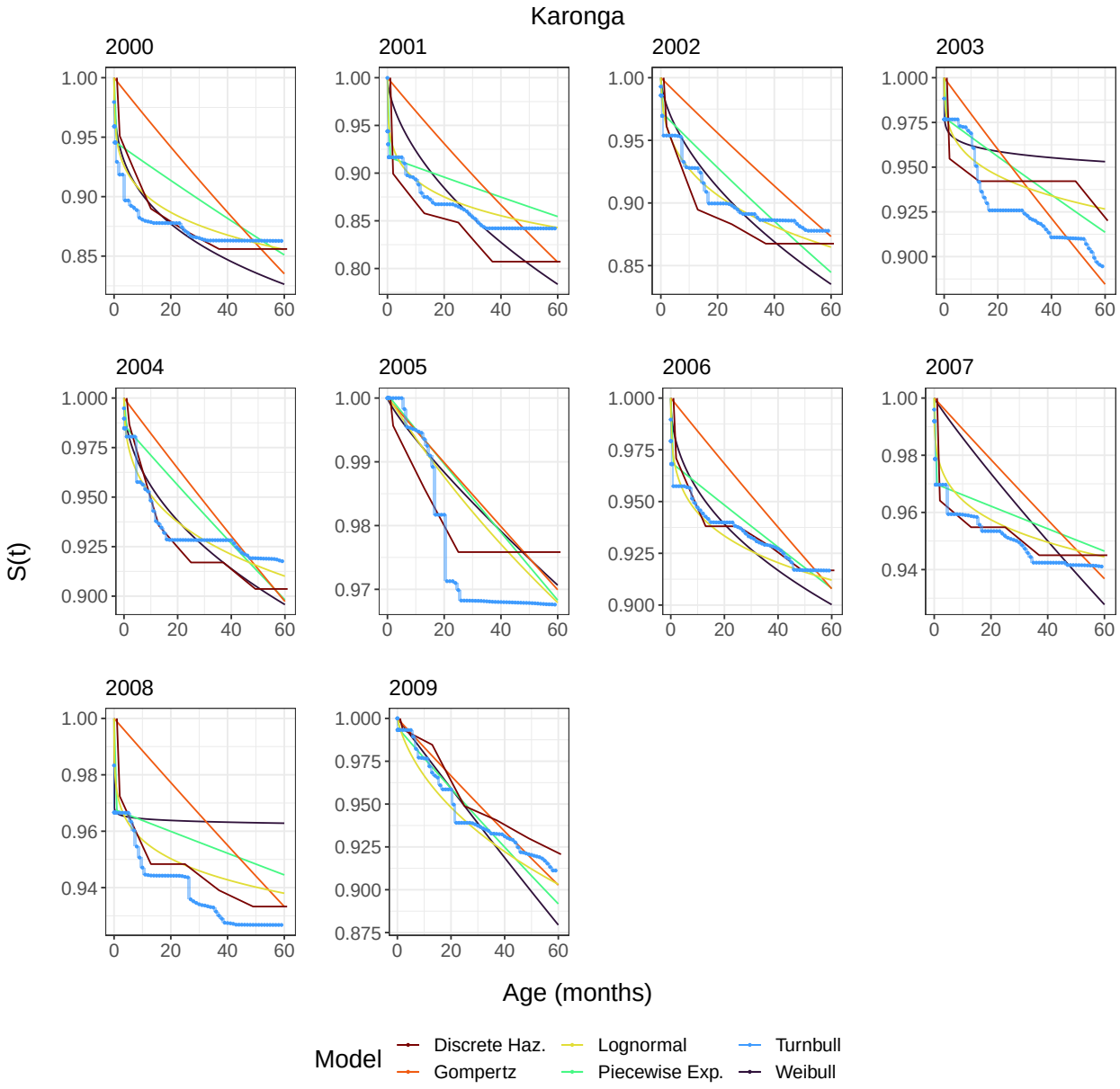


Figure C.21: Survey-weighted survival curves for the Karonga district of Malawi from 2000-2009.

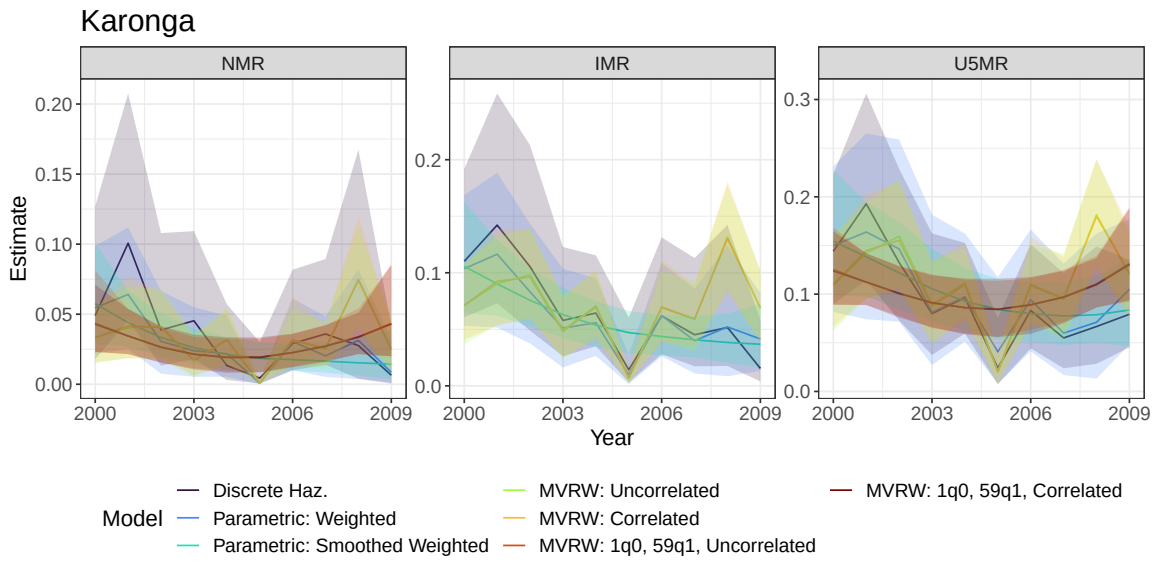


Figure C.22: Estimates of NMR, IMR, U5MR from direct estimates and smoothing models across time for the Karonga district of Malawi from 2000-2009.

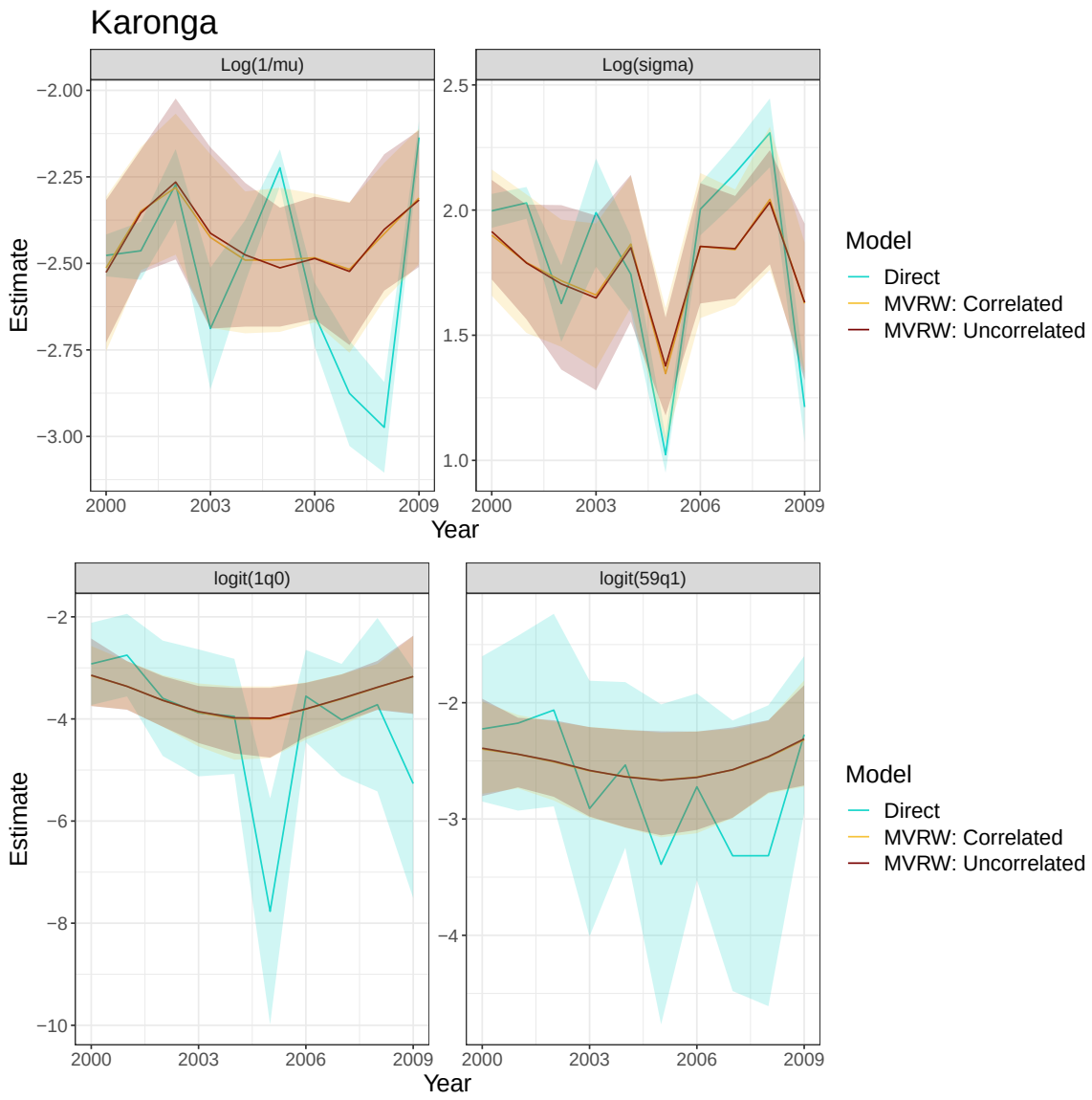


Figure C.23: Smoothed lognormal parameters (Top) and smoothed, transformed lognormal parameters (Bottom) for the Karonga district of Malawi from 2000-2009.

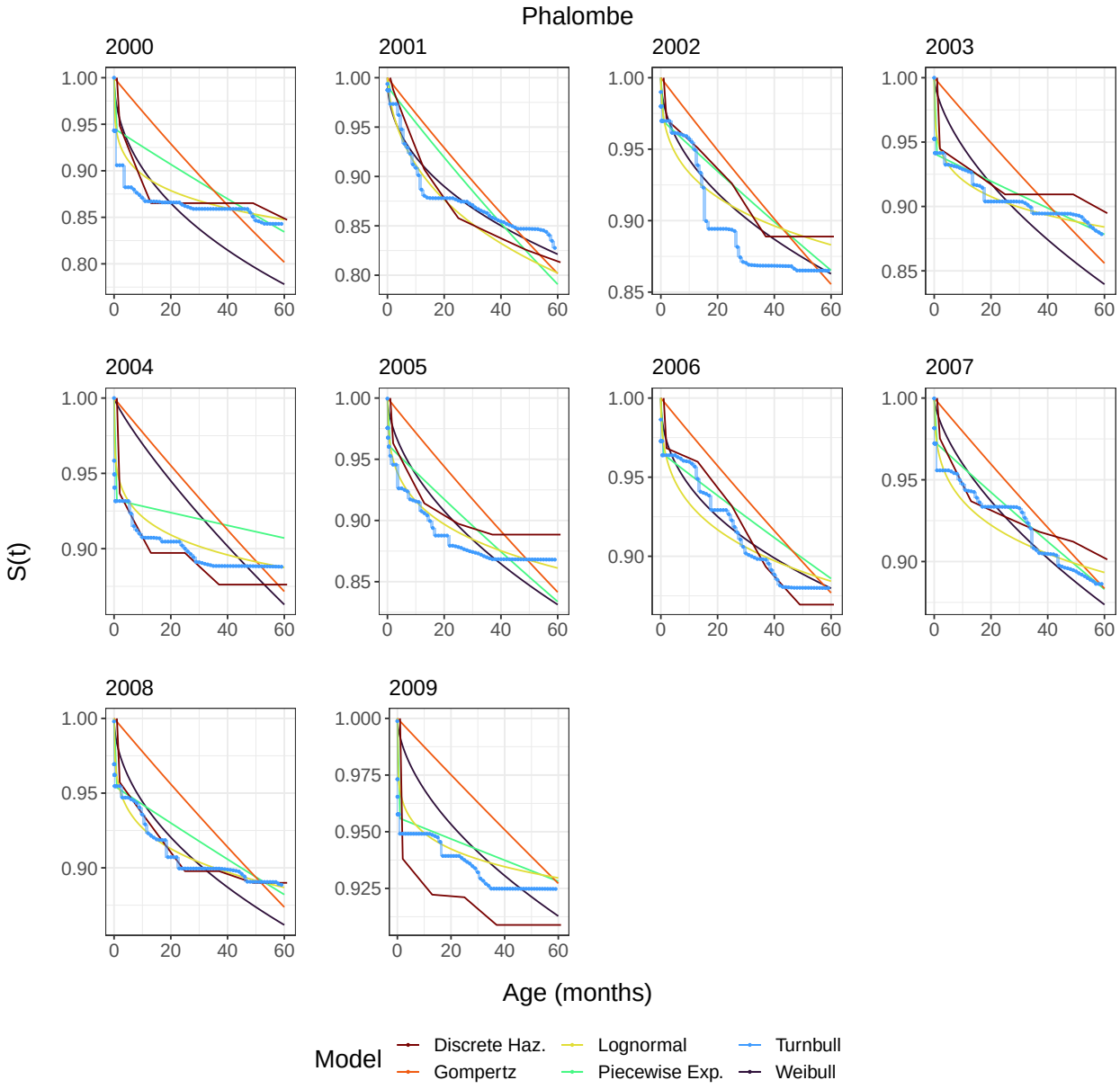


Figure C.24: Survey-weighted survival curves for the Phalombe district of Malawi from 2000-2009.

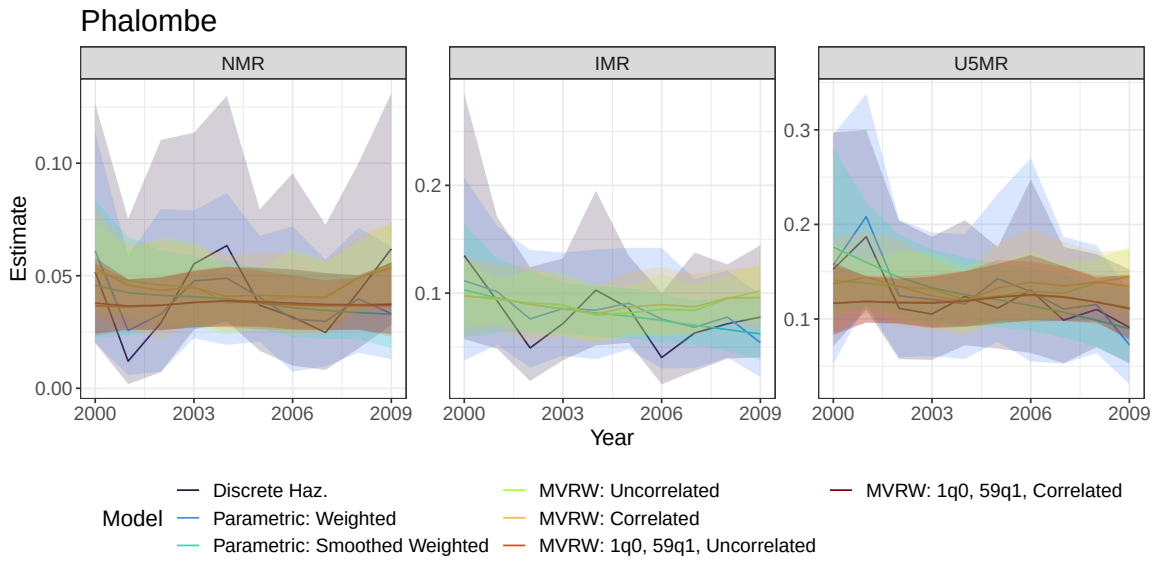


Figure C.25: Estimates of NMR, IMR, U5MR from direct estimates and smoothing models across time for the Phalombe district of Malawi from 2000-2009.

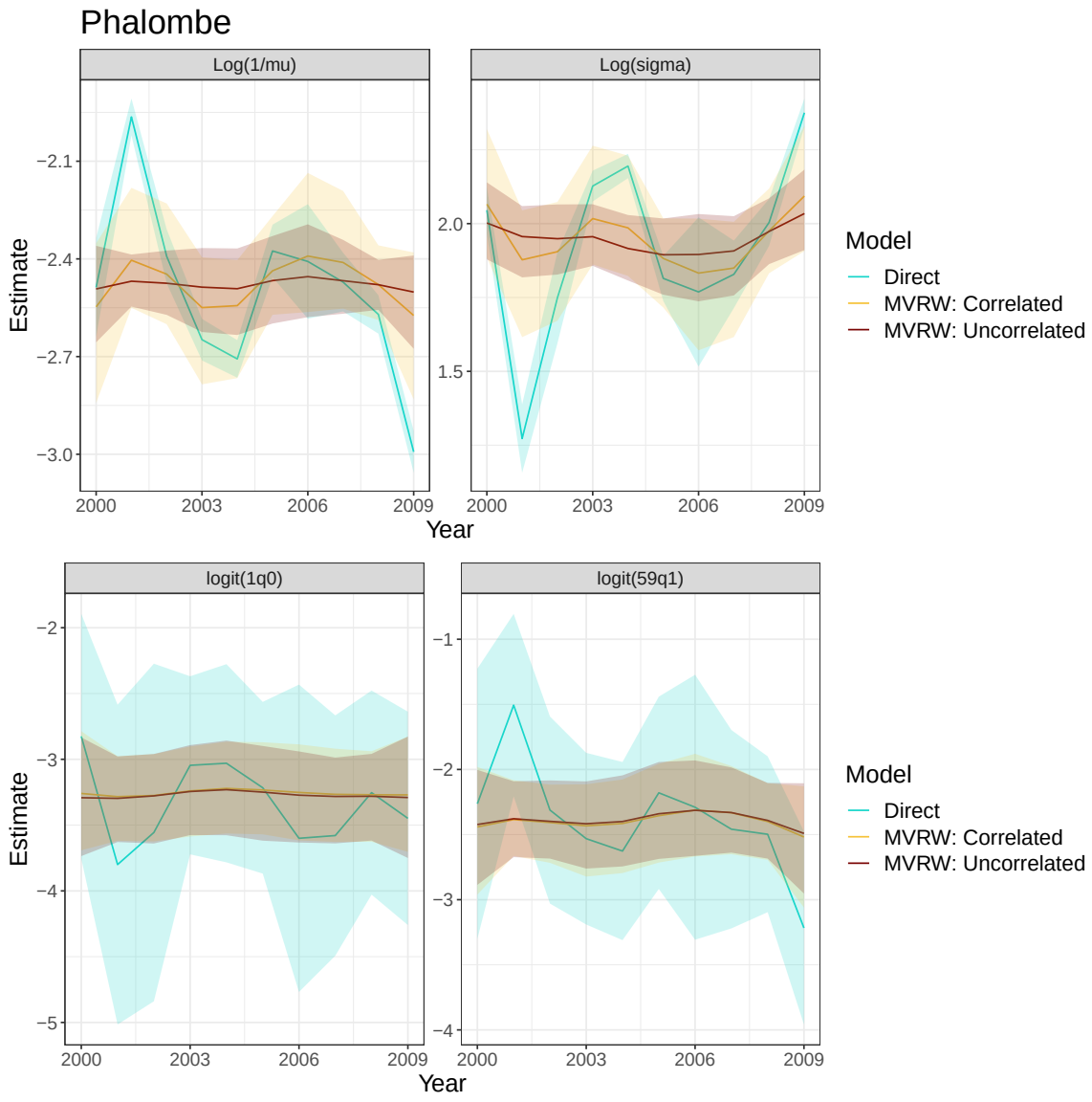


Figure C.26: Smoothed lognormal parameters (Top) and smoothed, transformed lognormal parameters (Bottom) for the Phalombe district of Malawi from 2000-2009.

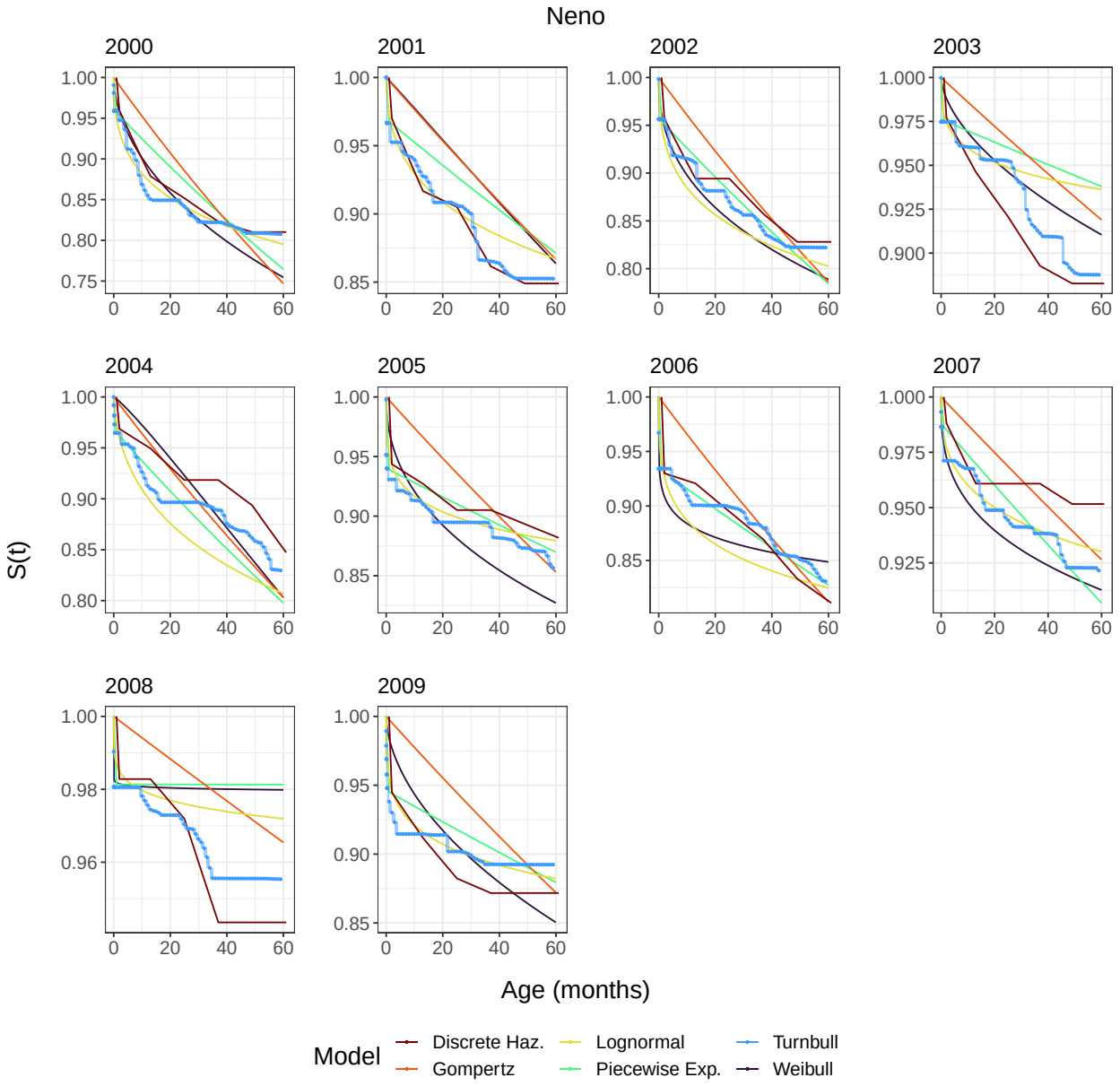


Figure C.27: Survey-weighted survival curves for the Neno district of Malawi from 2000-2009.

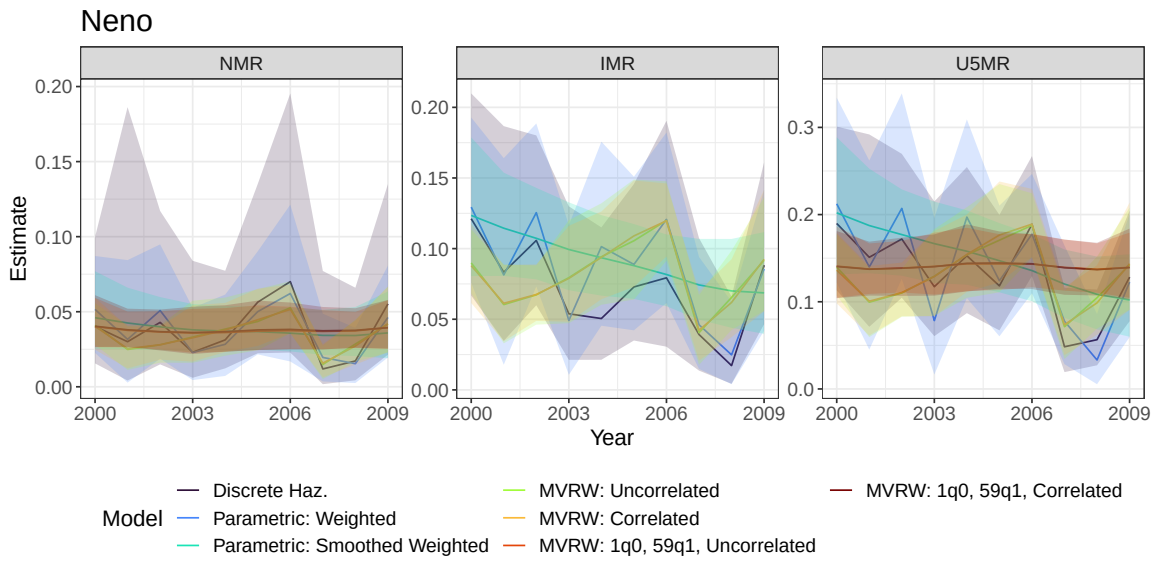


Figure C.28: Estimates of NMR, IMR, U5MR from direct estimates and smoothing models across time for the Neno district of Malawi from 2000-2009.

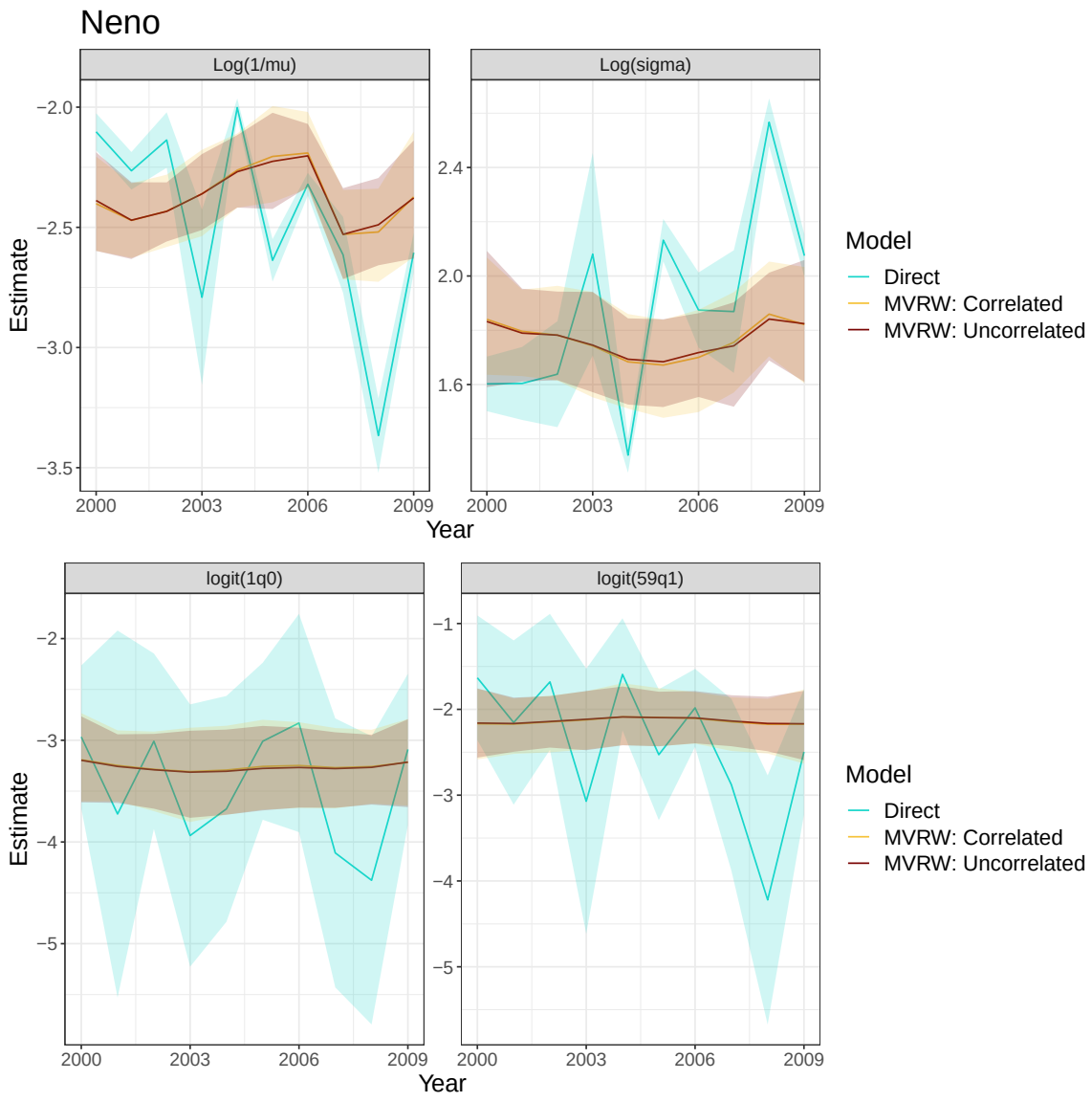


Figure C.29: Smoothed lognormal parameters (Top) and smoothed, transformed lognormal parameters (Bottom) for the Neno district of Malawi from 2000-2009.

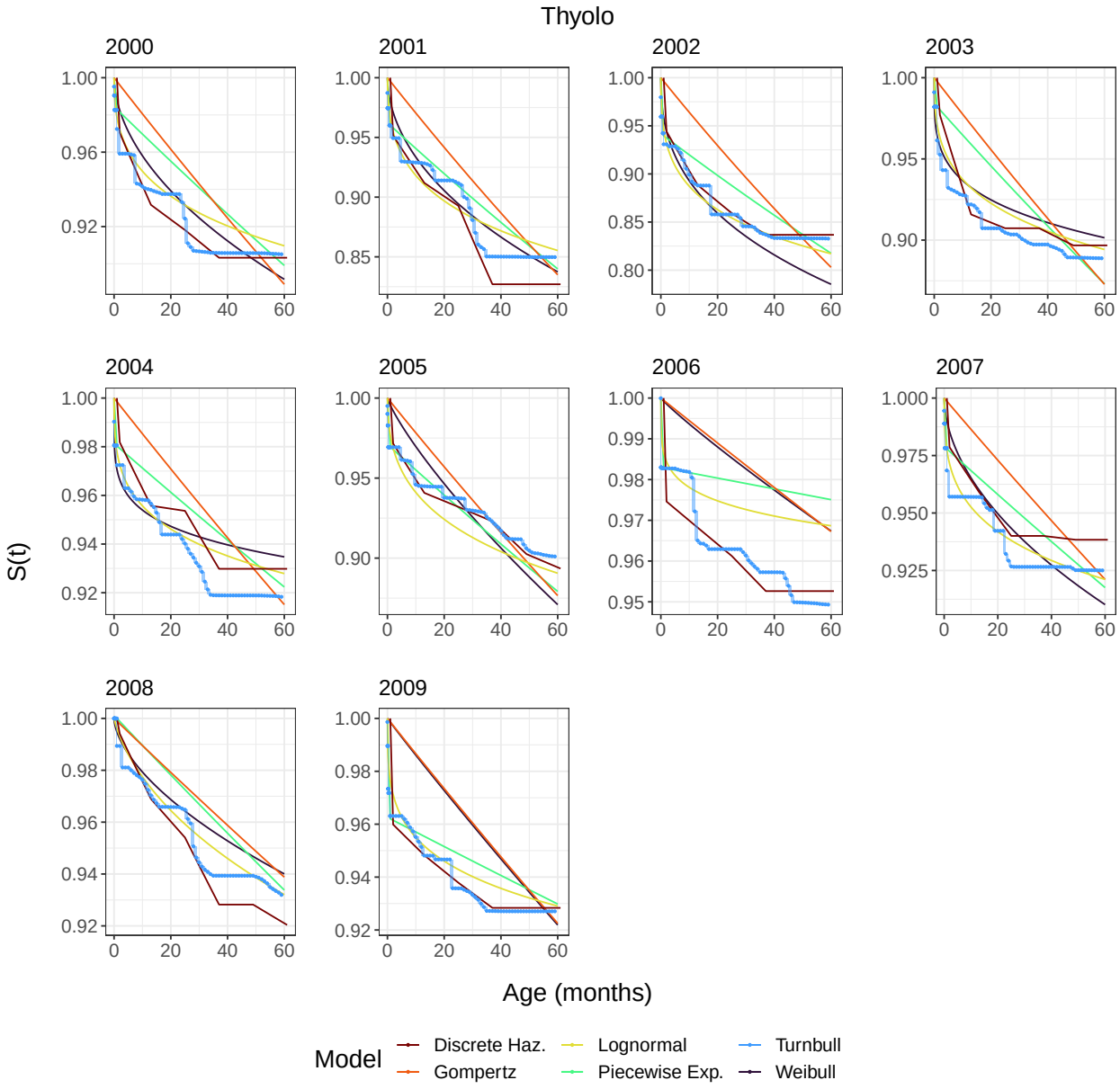


Figure C.30: Survey-weighted survival curves for the Thyolo district of Malawi from 2000-2009.

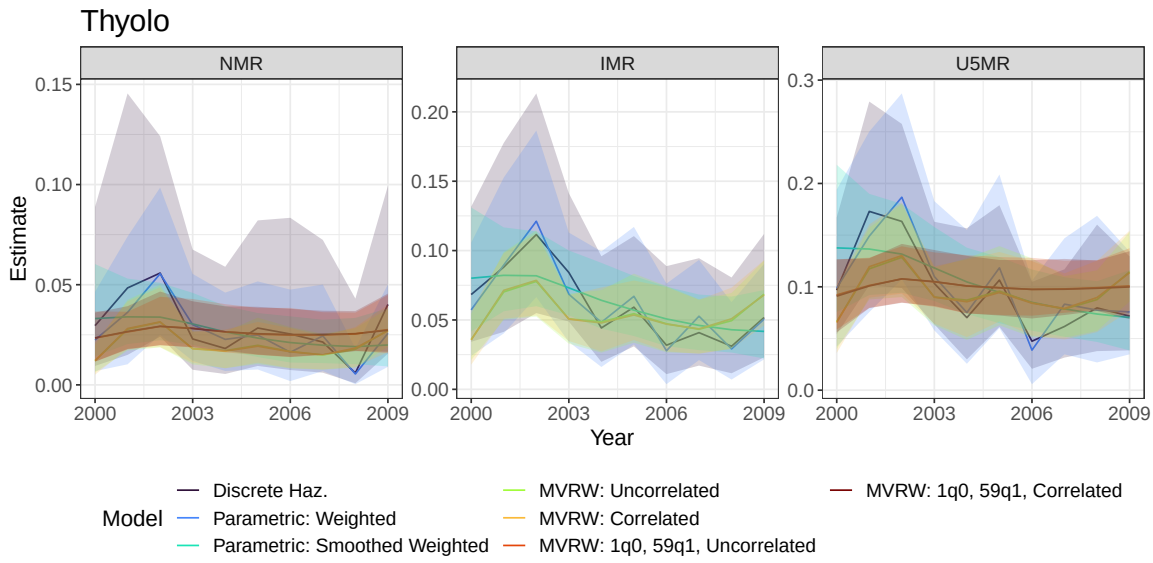


Figure C.31: Estimates of NMR, IMR, U5MR from direct estimates and smoothing models across time for the Thyolo district of Malawi from 2000-2009.

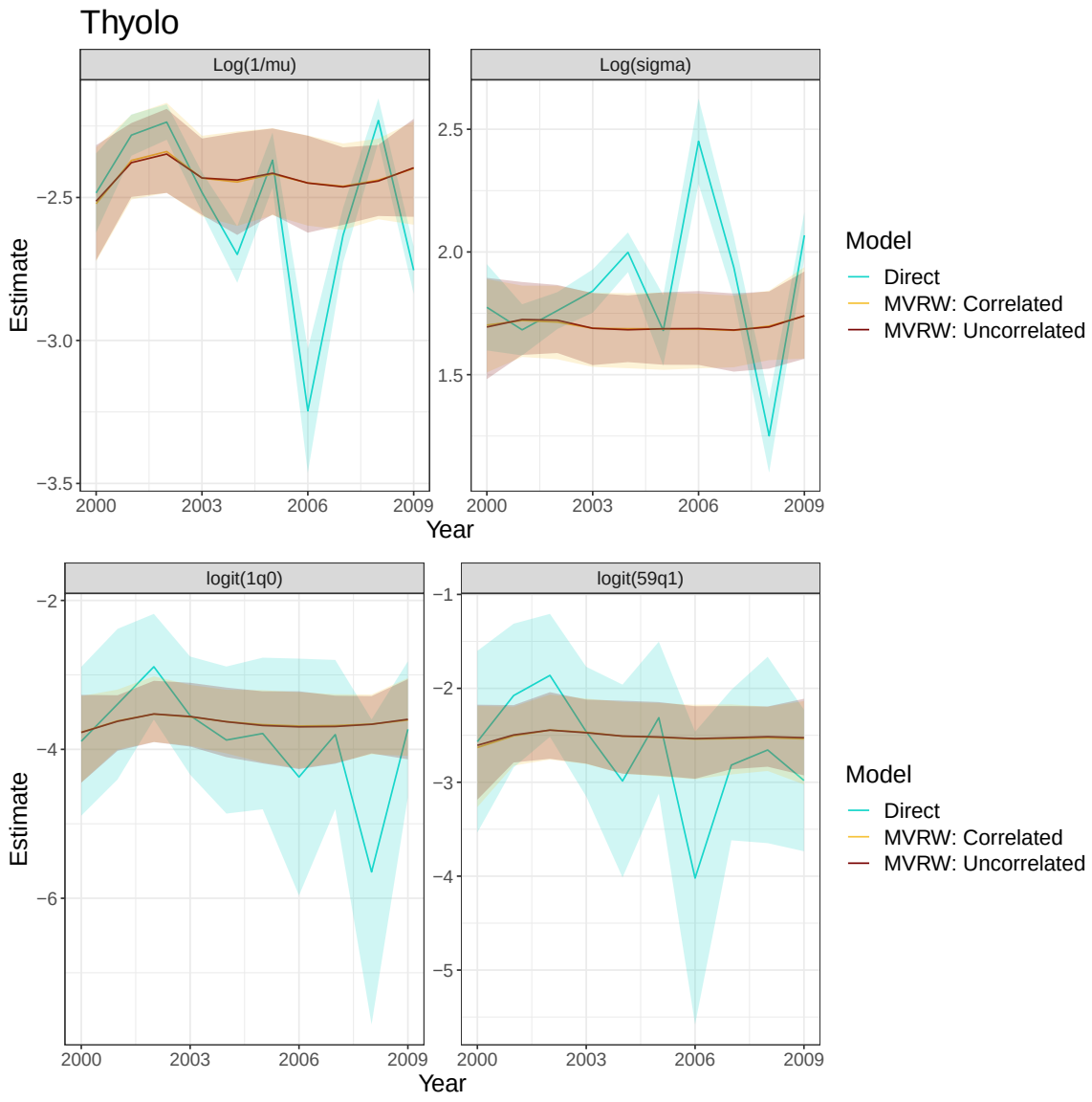


Figure C.32: Smoothed lognormal parameters (Top) and smoothed, transformed lognormal parameters (Bottom) for the Thyolo district of Malawi from 2000-2009.

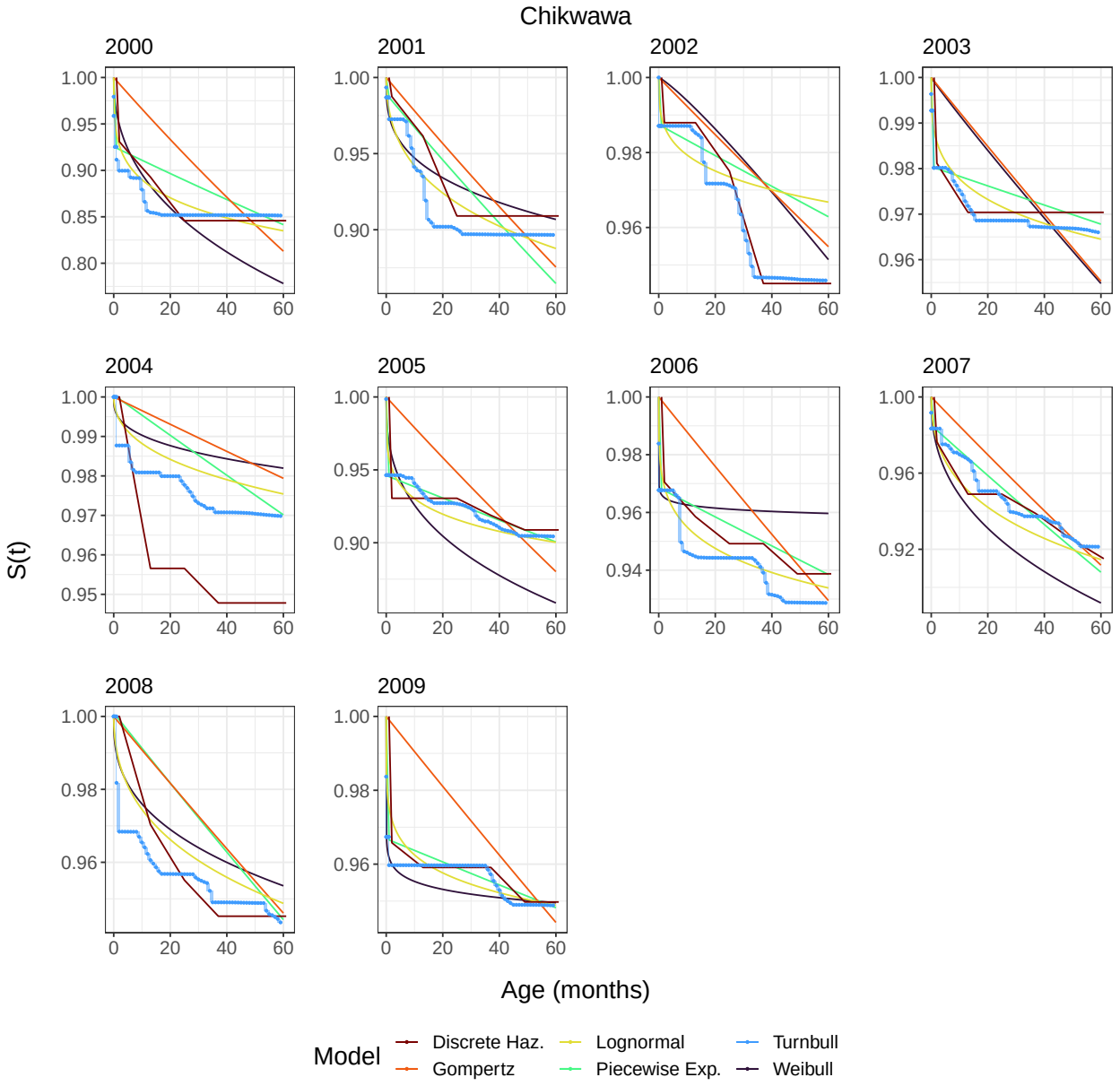


Figure C.33: Survey-weighted survival curves for the Chikwawa district of Malawi from 2000-2009.

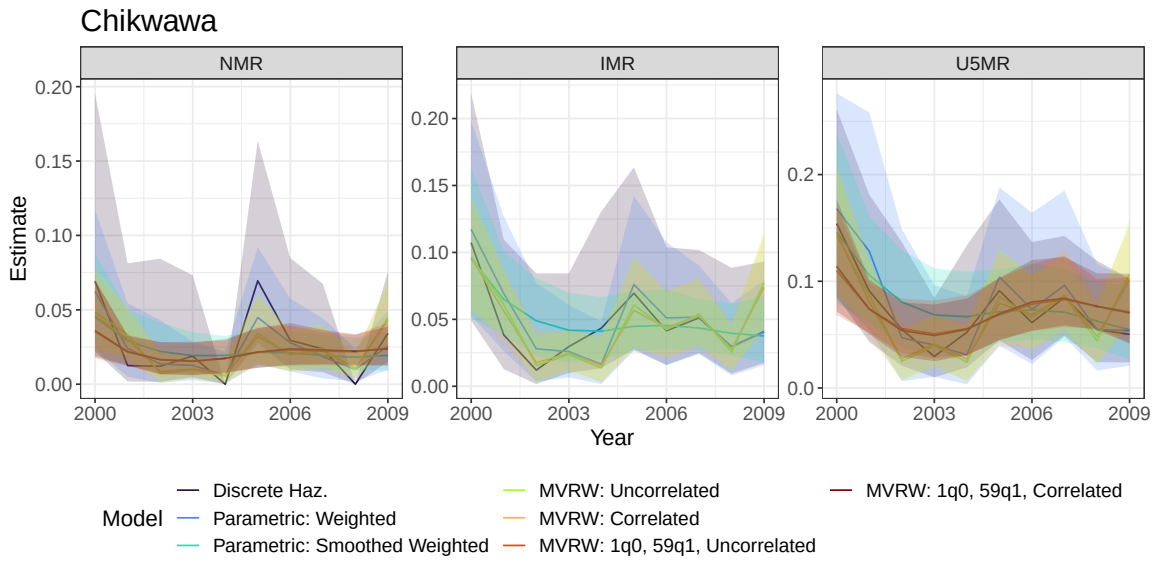


Figure C.34: Estimates of NMR, IMR, U5MR from direct estimates and smoothing models across time for the Chikwawa district of Malawi from 2000-2009.

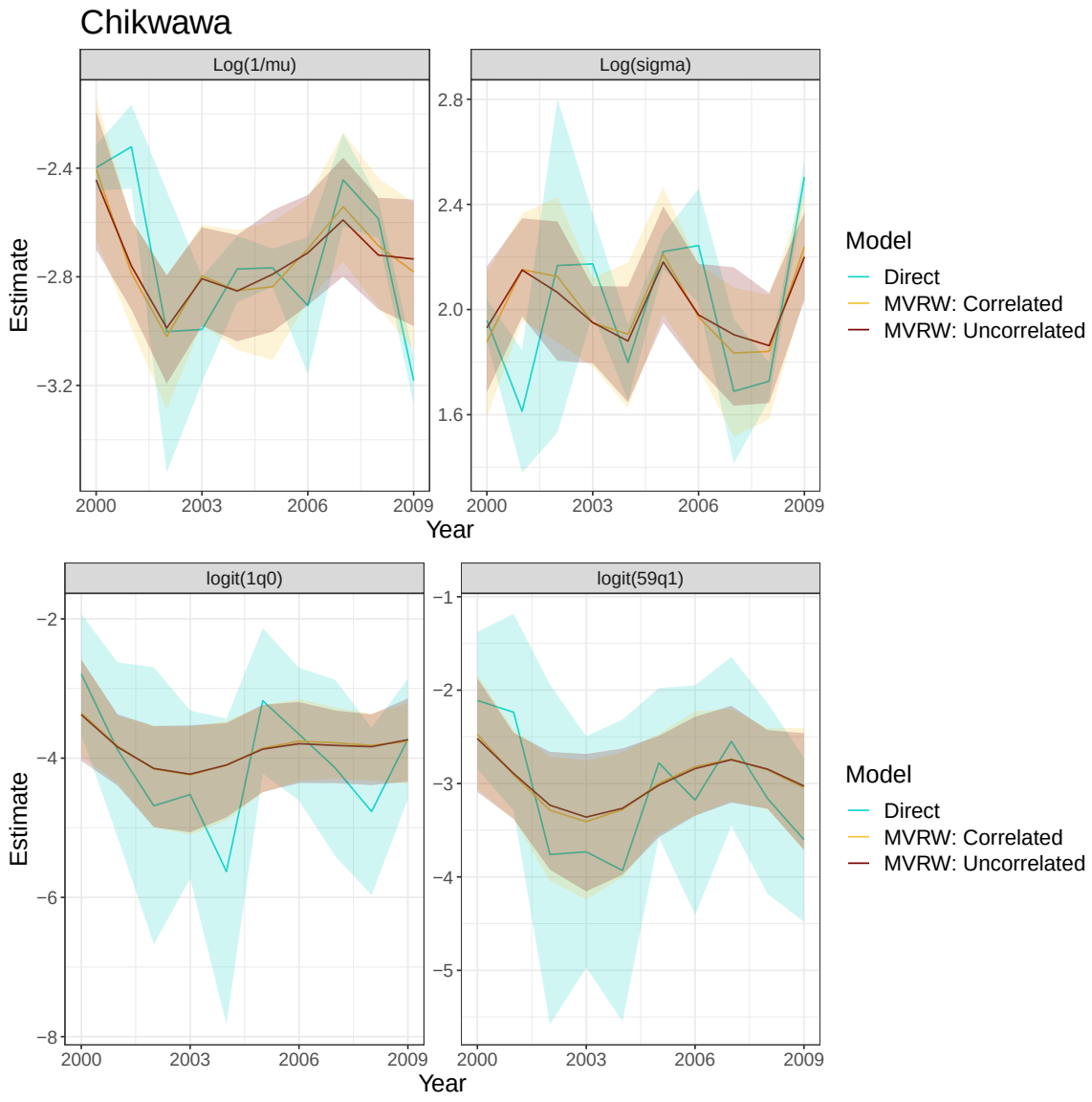


Figure C.35: Smoothed lognormal parameters (Top) and smoothed, transformed lognormal parameters (Bottom) for the Chikwawa district of Malawi from 2000-2009.

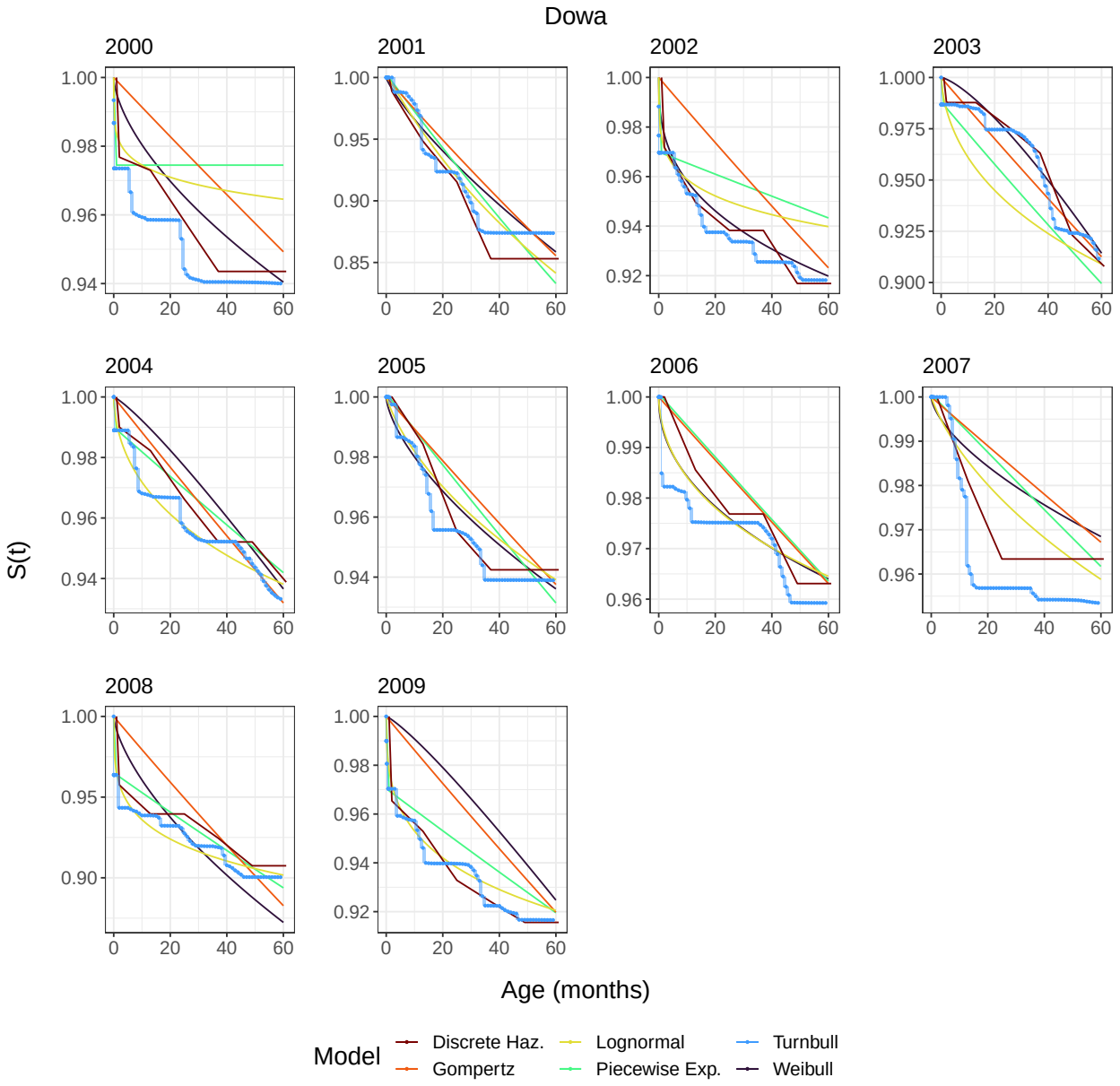


Figure C.36: Survey-weighted survival curves for the Dowa district of Malawi from 2000-2009.

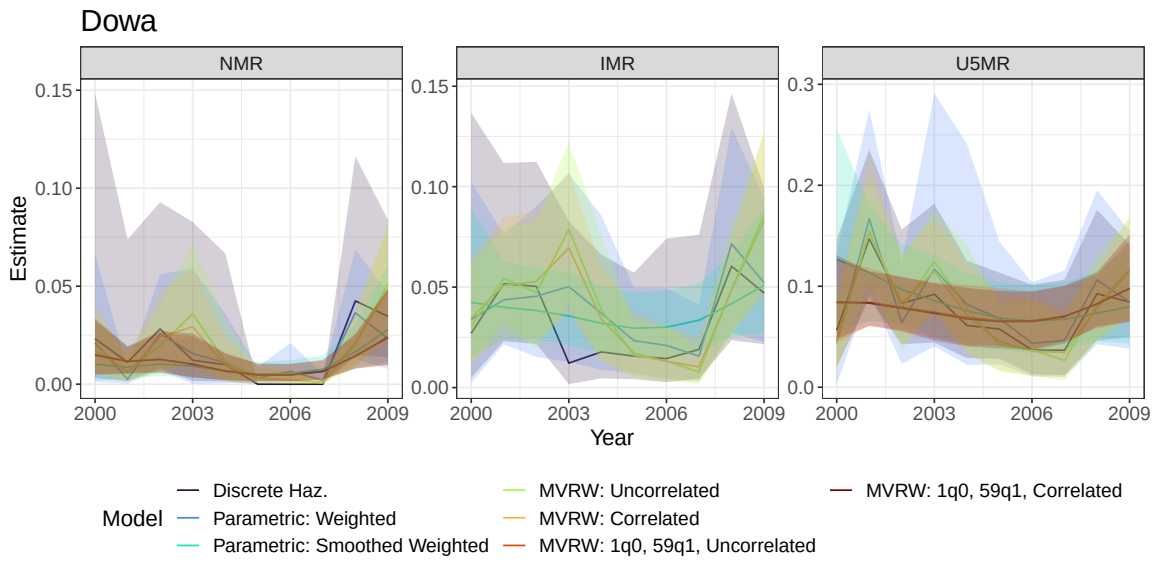


Figure C.37: Estimates of NMR, IMR, U5MR from direct estimates and smoothing models across time for the Dowa district of Malawi from 2000-2009.

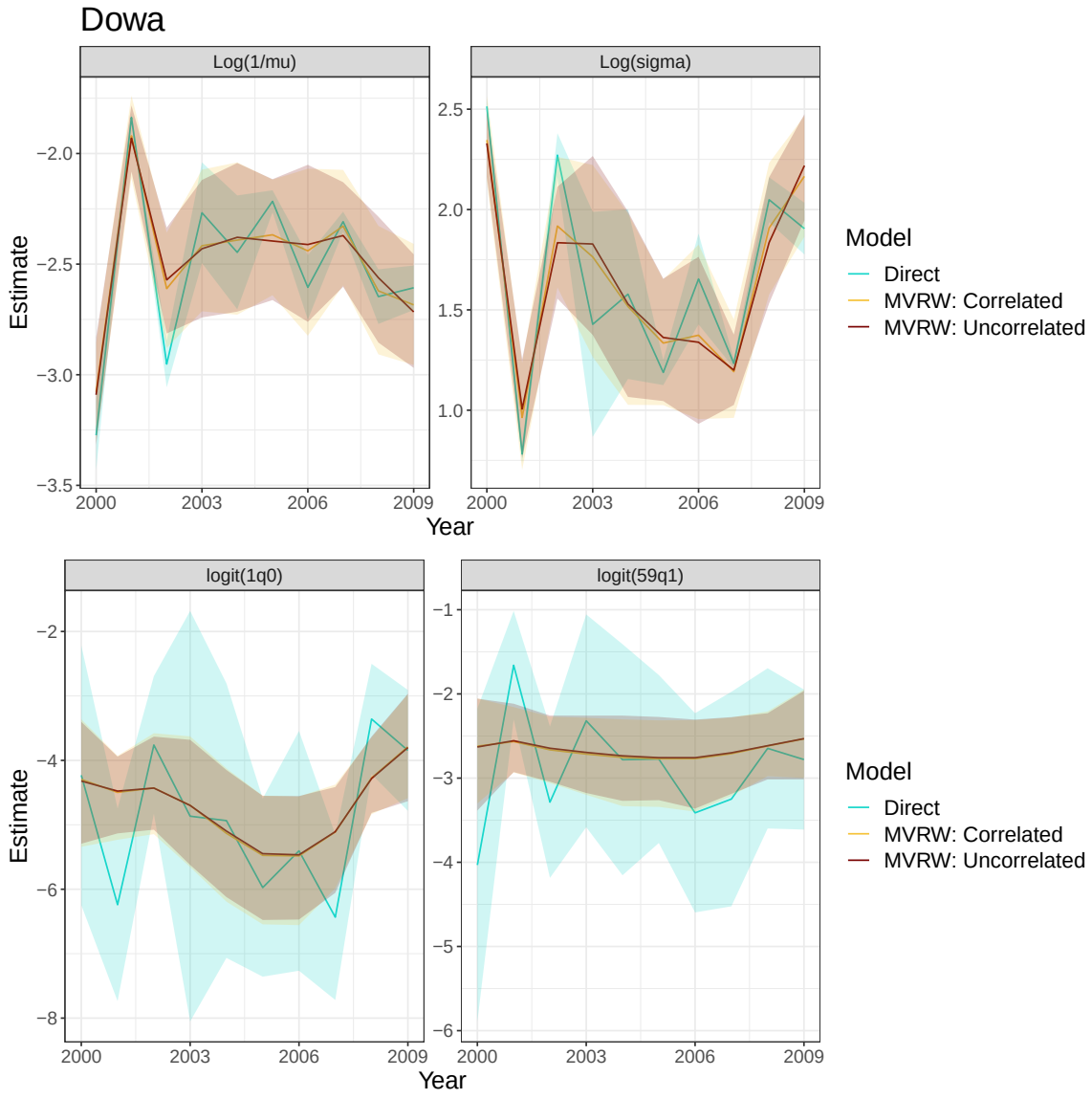


Figure C.38: Smoothed lognormal parameters (Top) and smoothed, transformed lognormal parameters (Bottom) for the Dowa district of Malawi from 2000-2009.

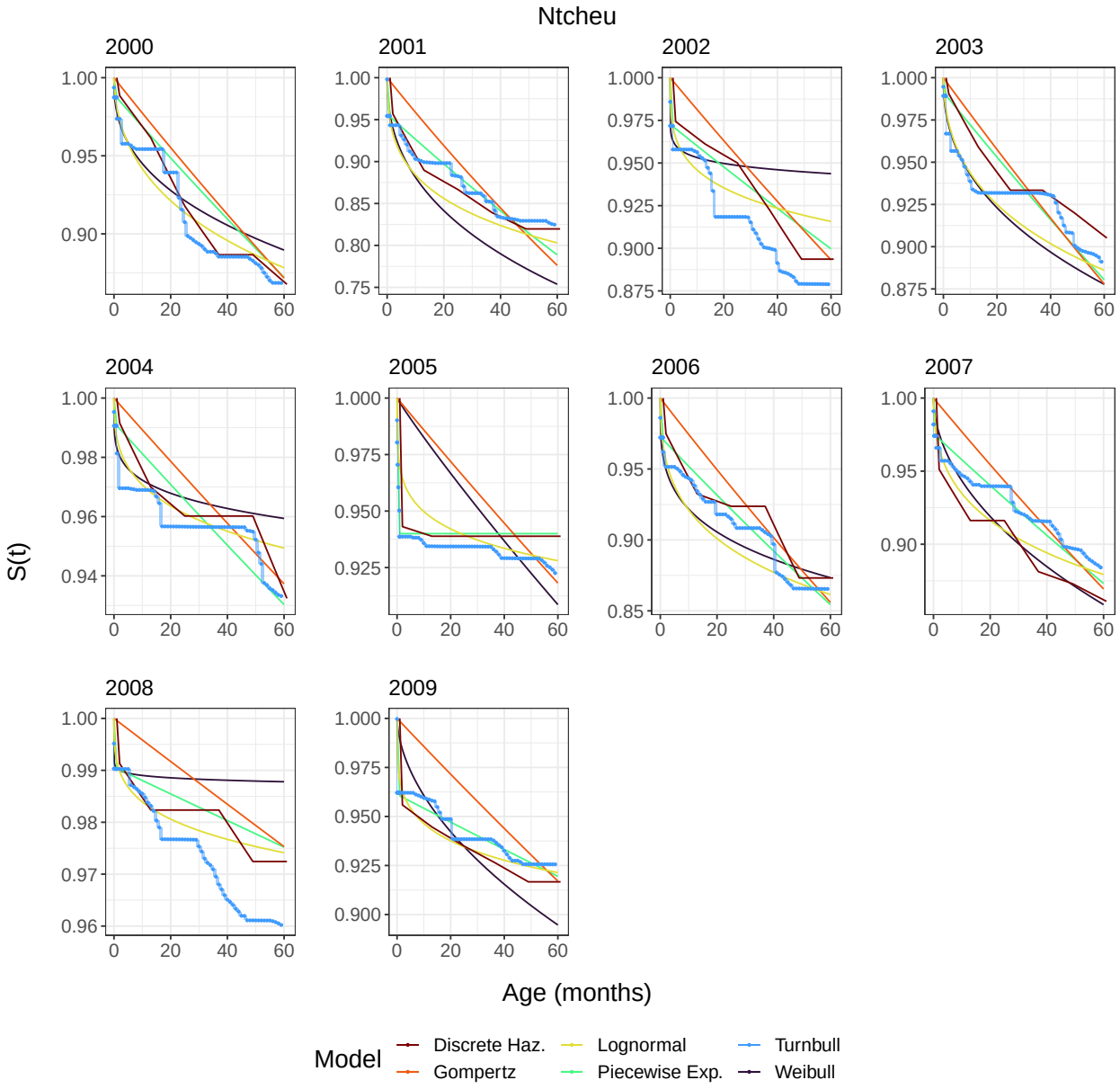


Figure C.39: Survey-weighted survival curves for the Ntcheu district of Malawi from 2000-2009.

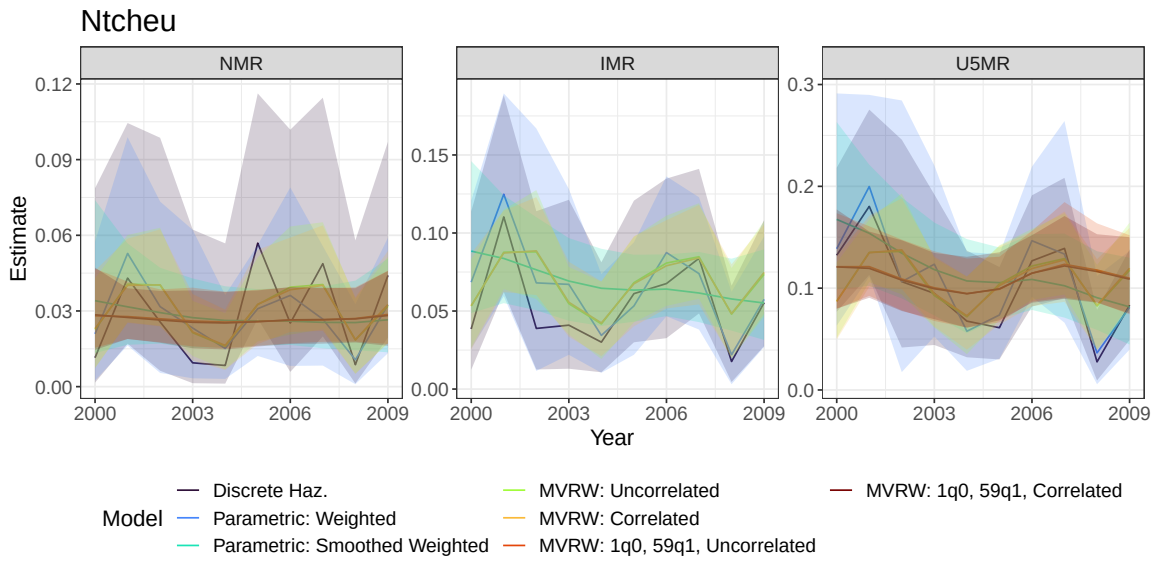


Figure C.40: Estimates of NMR, IMR, U5MR from direct estimates and smoothing models across time for the Ntcheu district of Malawi from 2000-2009.

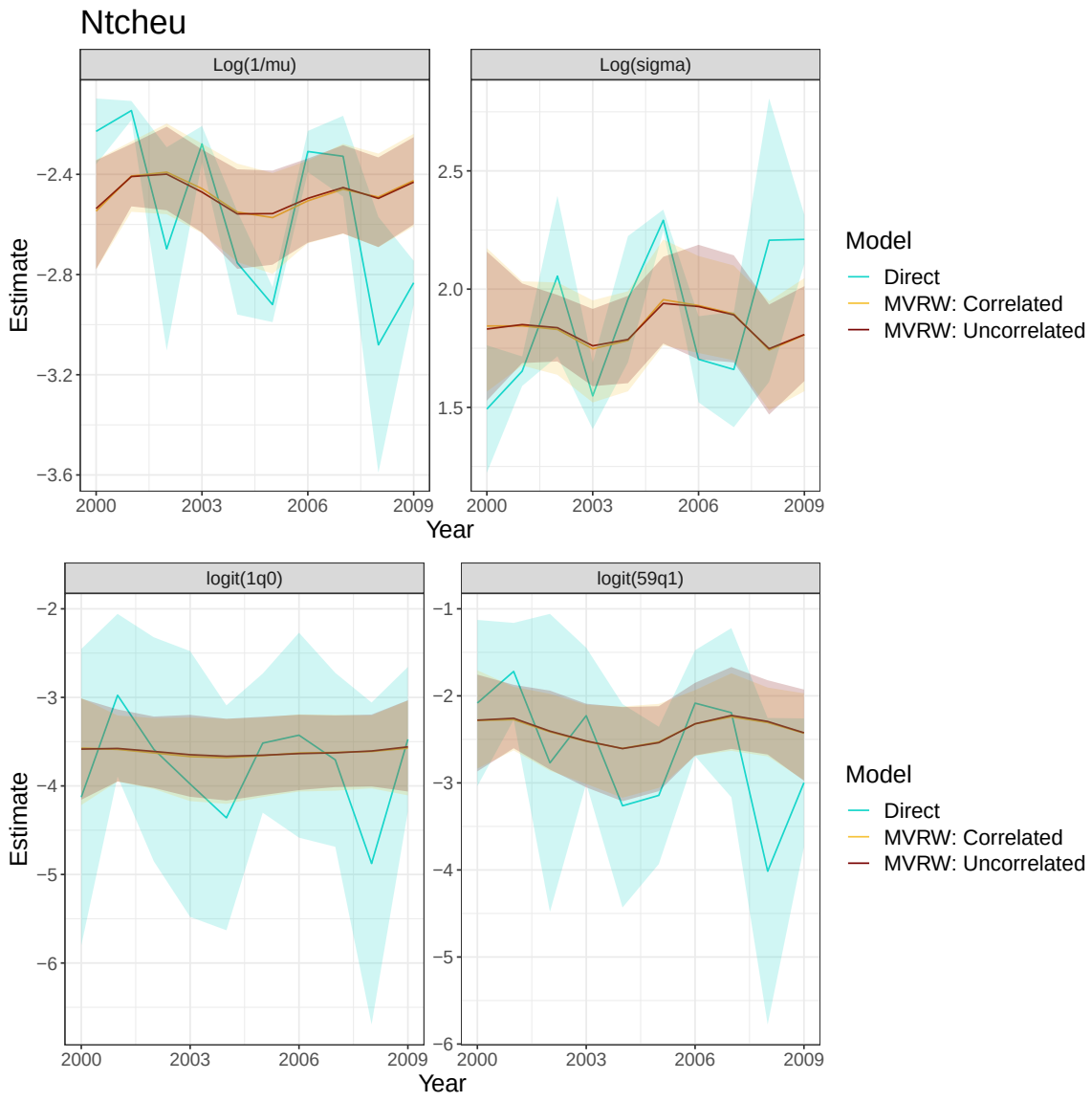


Figure C.41: Smoothed lognormal parameters (Top) and smoothed, transformed lognormal parameters (Bottom) for the Ntcheu district of Malawi from 2000-2009.

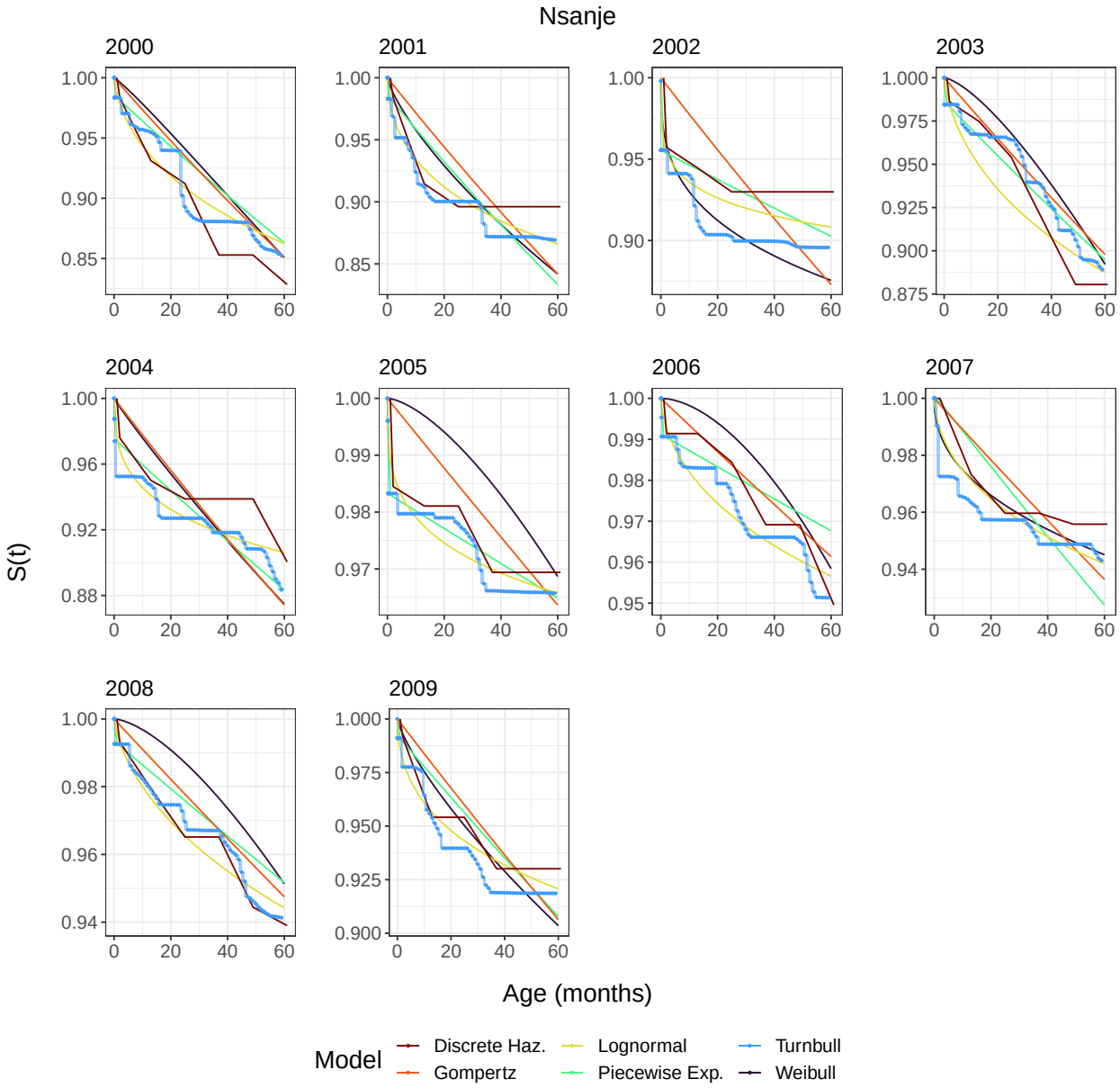


Figure C.42: Survey-weighted survival curves for the Nsanje district of Malawi from 2000-2009.

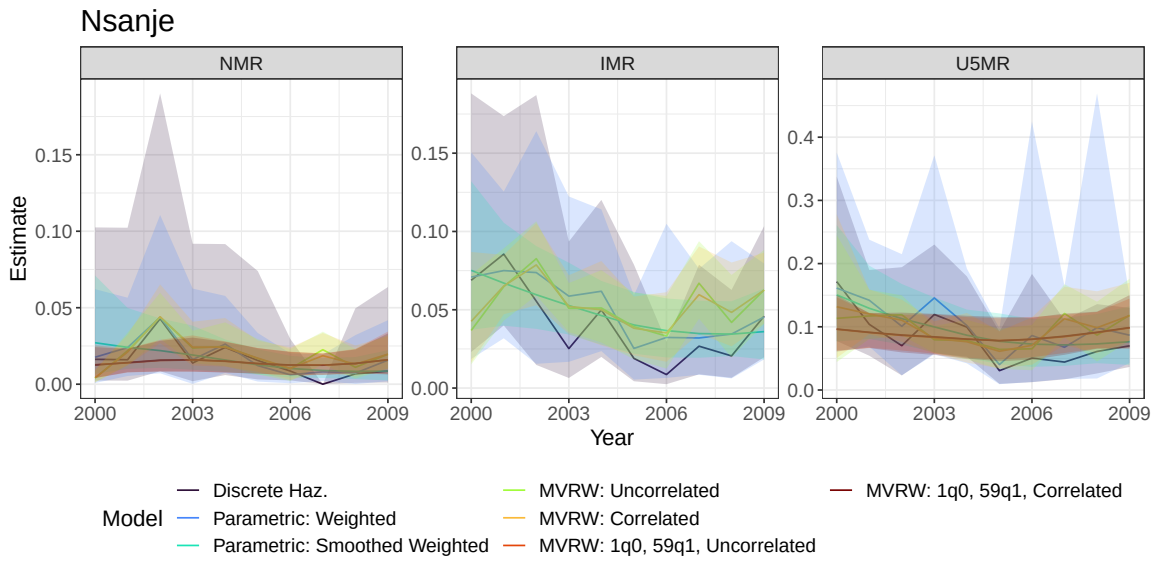


Figure C.43: Estimates of NMR, IMR, U5MR from direct estimates and smoothing models across time for the Nsanje district of Malawi from 2000-2009.

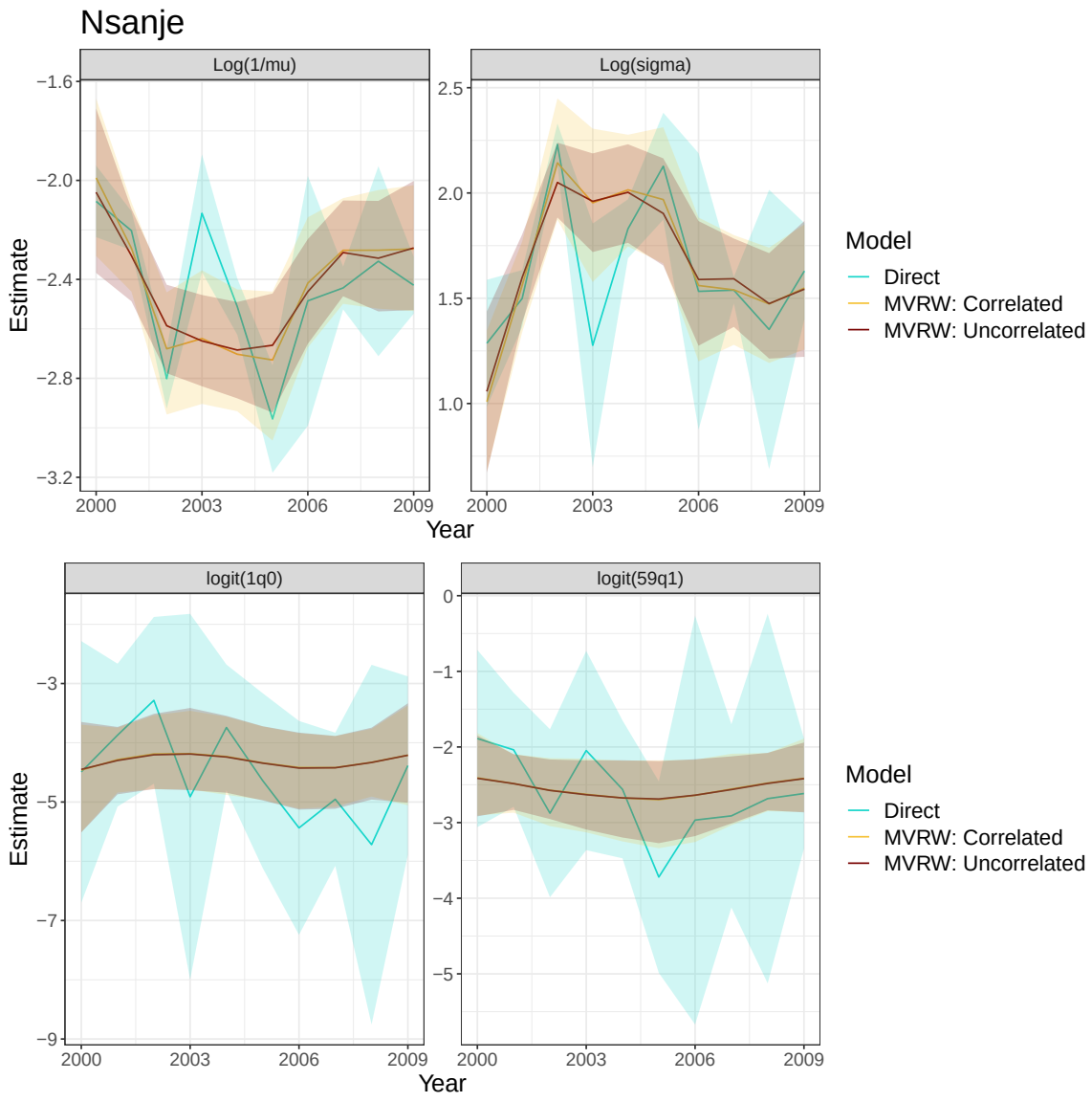


Figure C.44: Smoothed lognormal parameters (Top) and smoothed, transformed lognormal parameters (Bottom) for the Nsanje district of Malawi from 2000-2009.

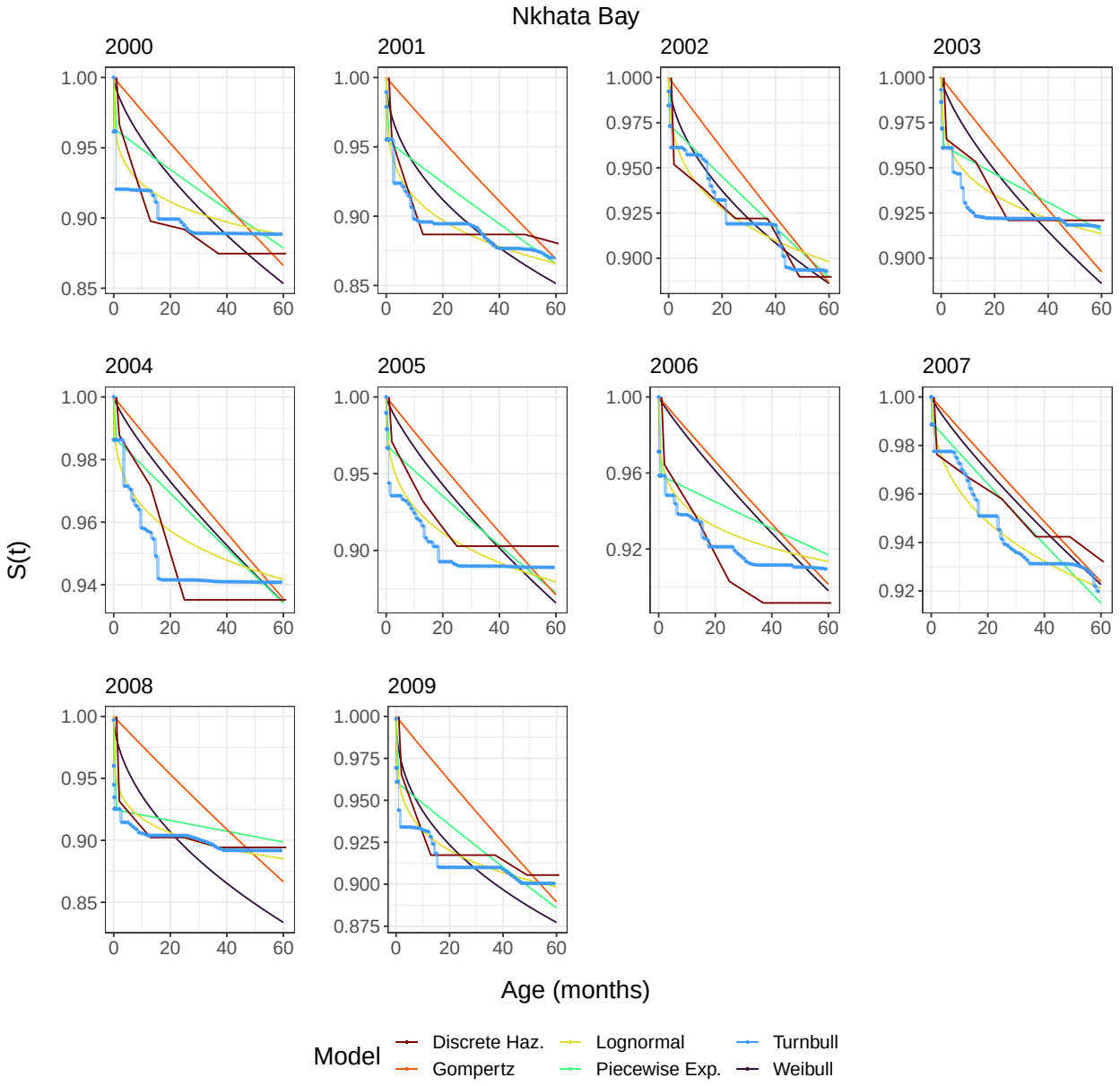


Figure C.45: Survey-weighted survival curves for the Nkhata Bay district of Malawi from 2000-2009.

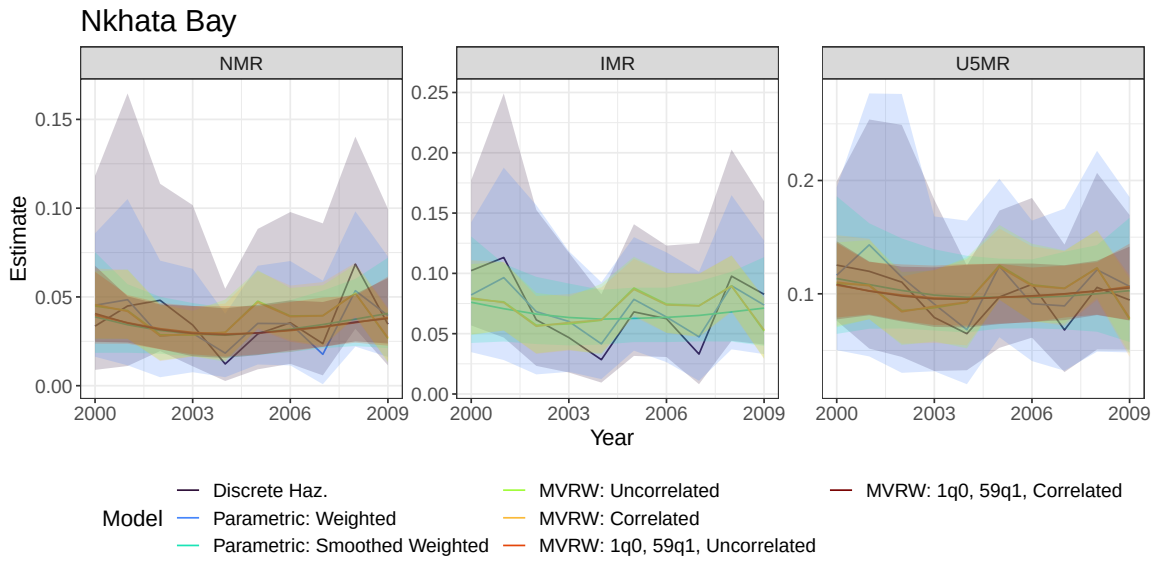


Figure C.46: Estimates of NMR, IMR, U5MR from direct estimates and smoothing models across time for the Nkhata Bay district of Malawi from 2000-2009.

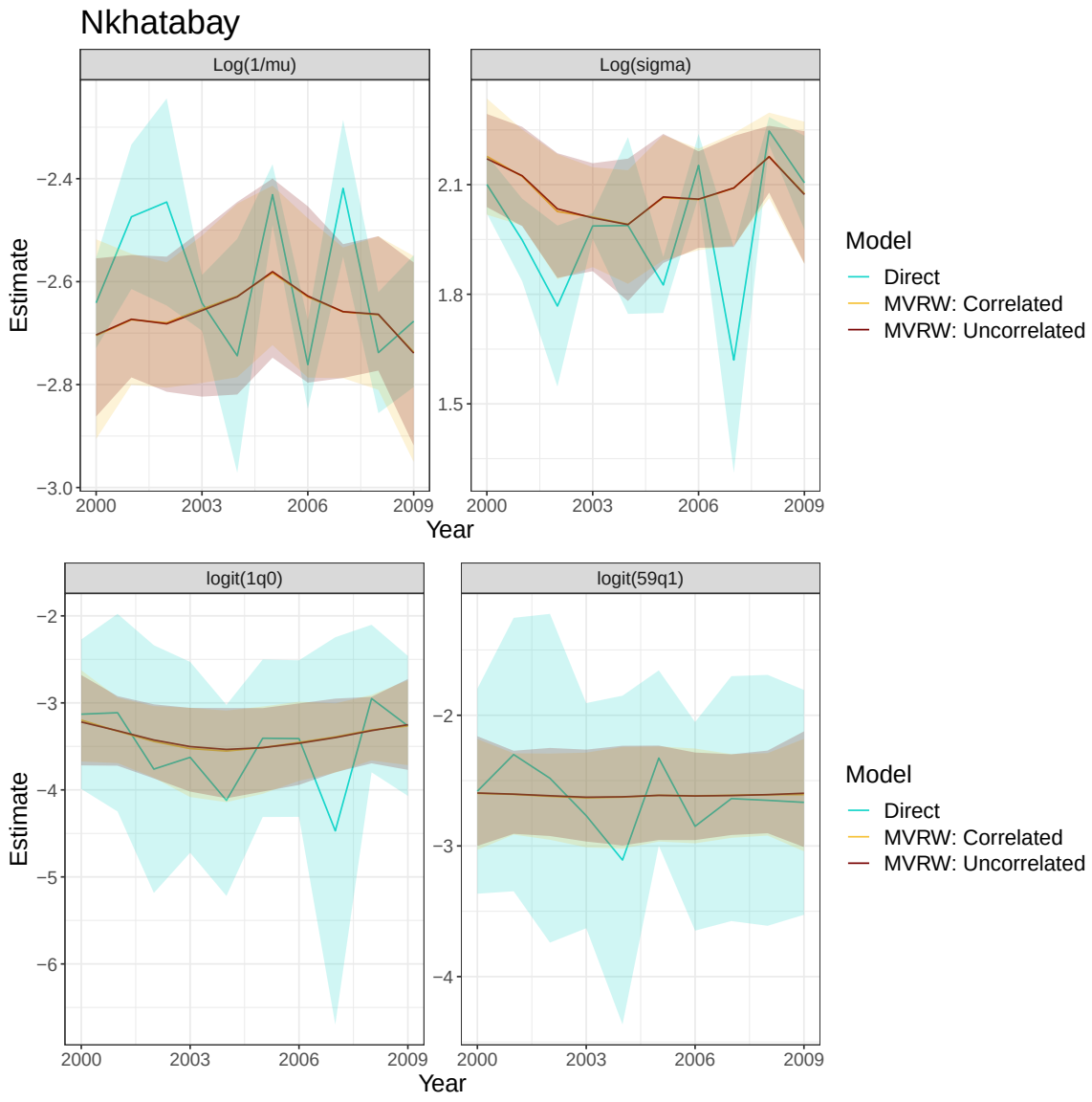


Figure C.47: Smoothed lognormal parameters (Top) and smoothed, transformed lognormal parameters (Bottom) for the Nkhata Bay district of Malawi from 2000-2009.

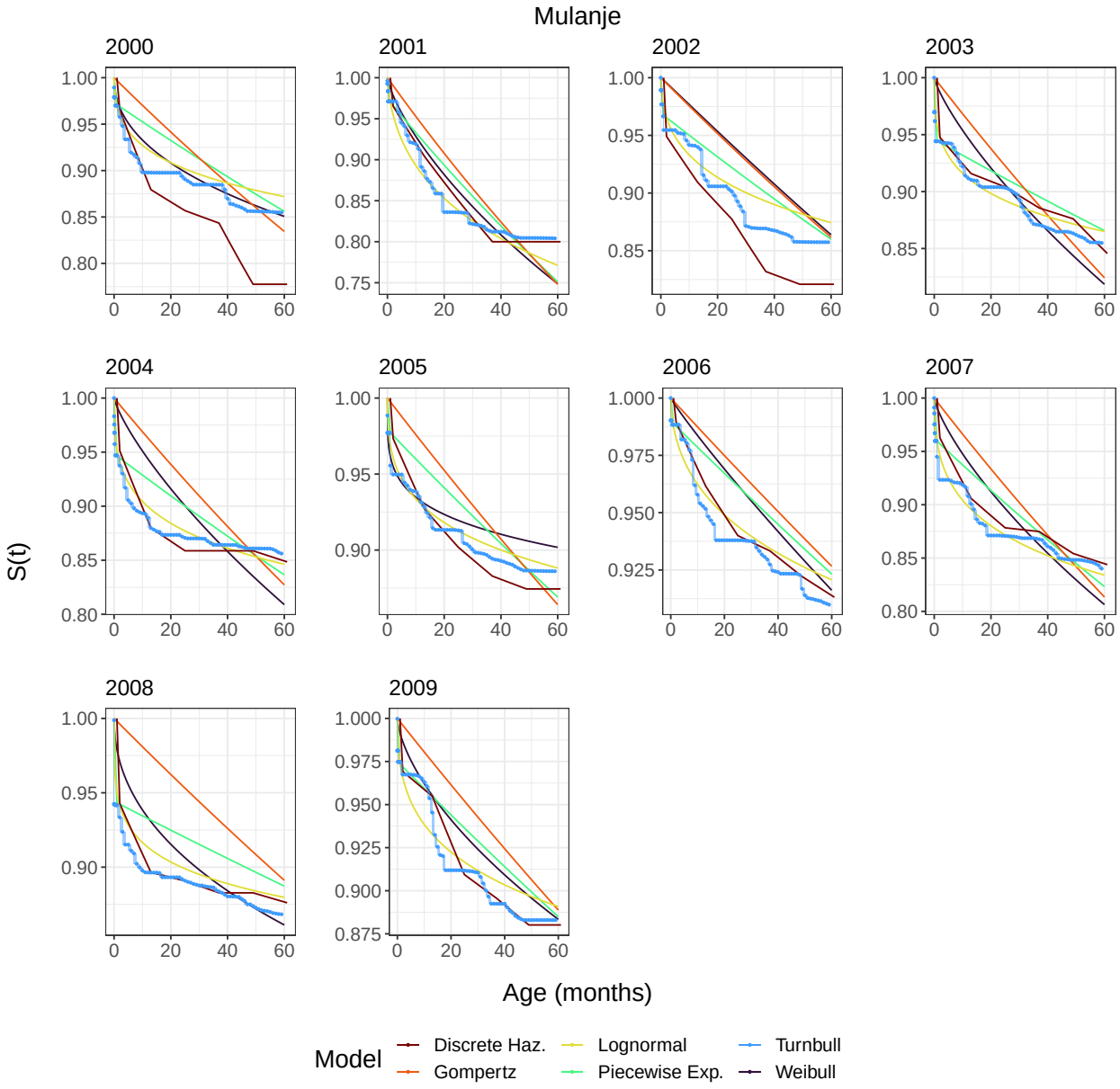


Figure C.48: Survey-weighted survival curves for the Mulanje district of Malawi from 2000-2009.

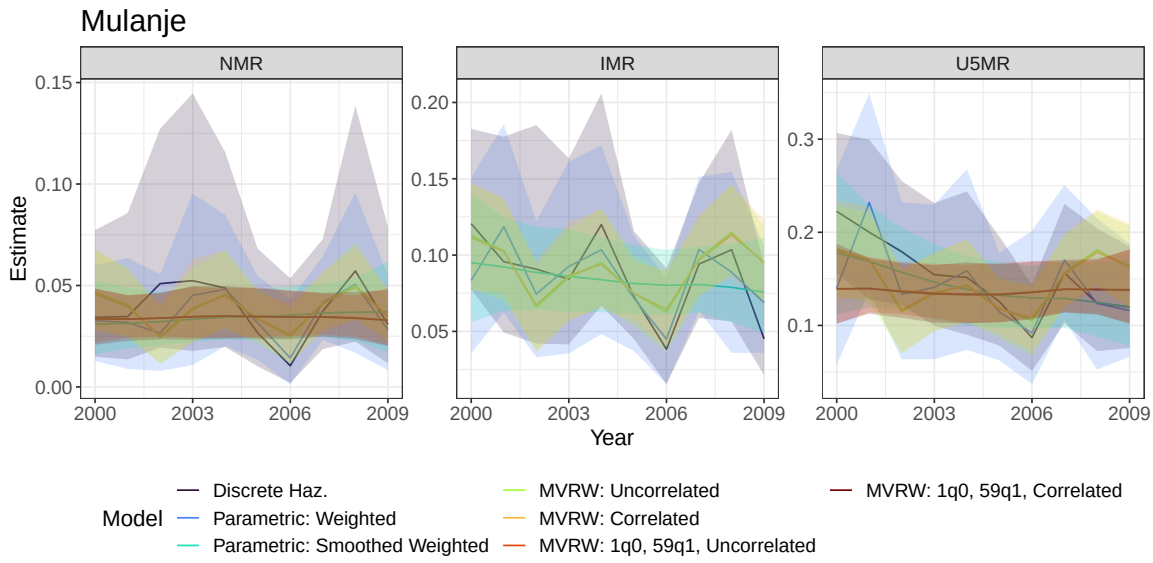


Figure C.49: Estimates of NMR, IMR, U5MR from direct estimates and smoothing models across time for the Mulanje district of Malawi from 2000-2009.

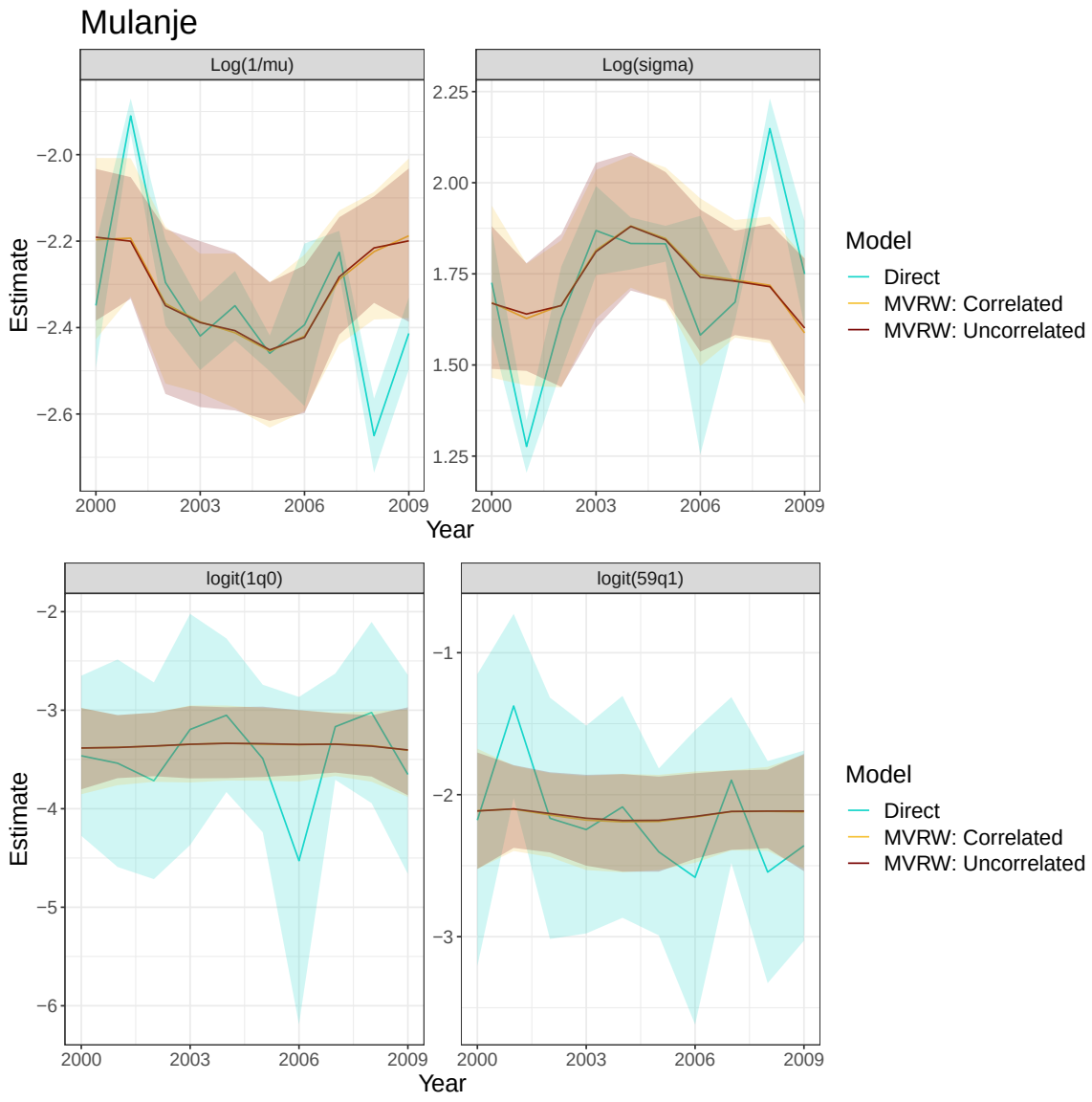


Figure C.50: Smoothed lognormal parameters (Top) and smoothed, transformed lognormal parameters (Bottom) for the Mulanje district of Malawi from 2000-2009.

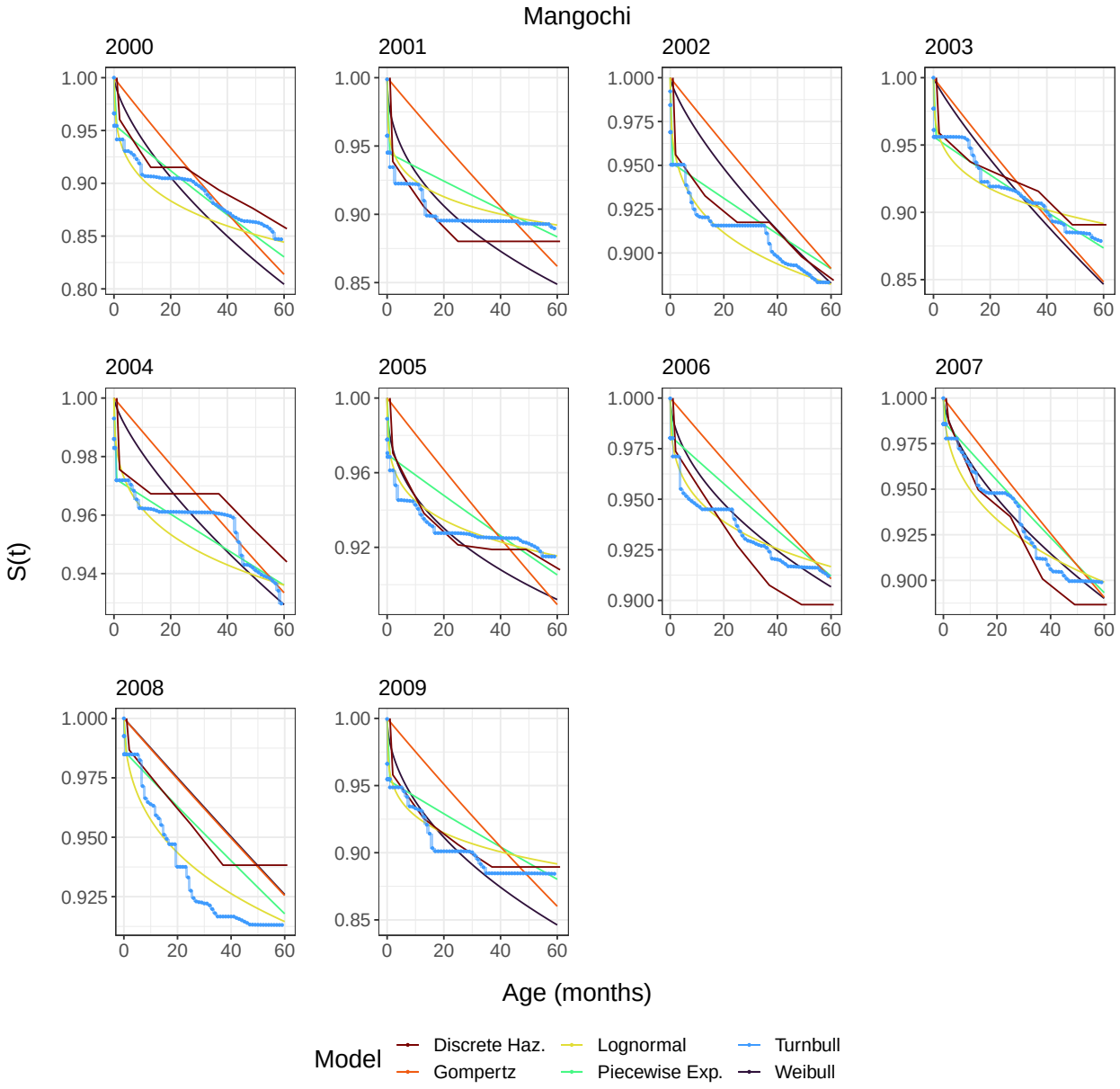


Figure C.51: Survey-weighted survival curves for the Mangochi district of Malawi from 2000-2009.

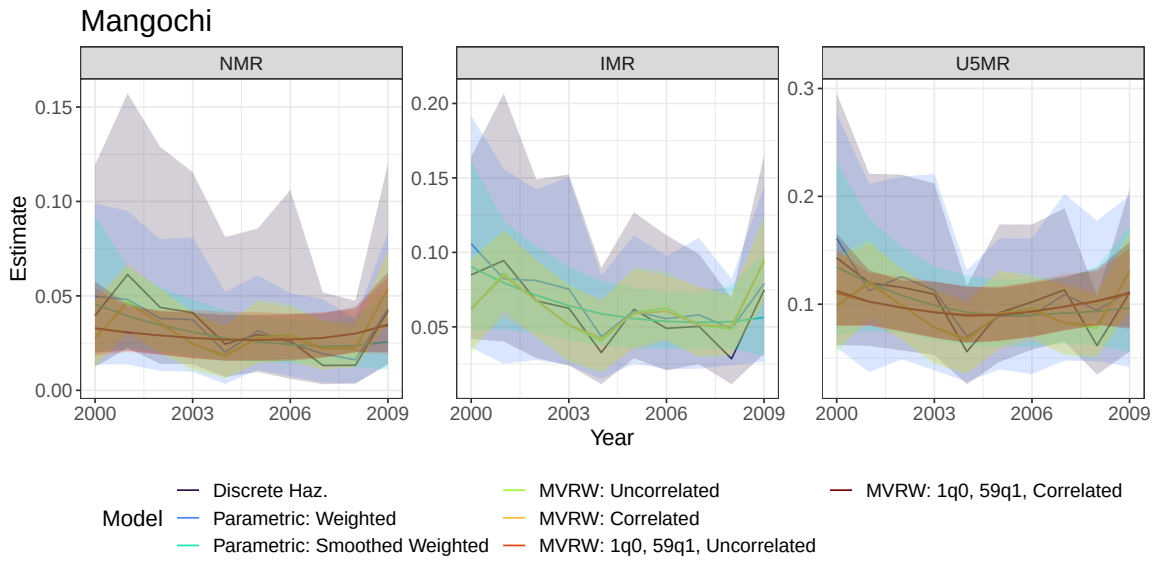


Figure C.52: Estimates of NMR, IMR, U5MR from direct estimates and smoothing models across time for the Mangochi district of Malawi from 2000-2009.

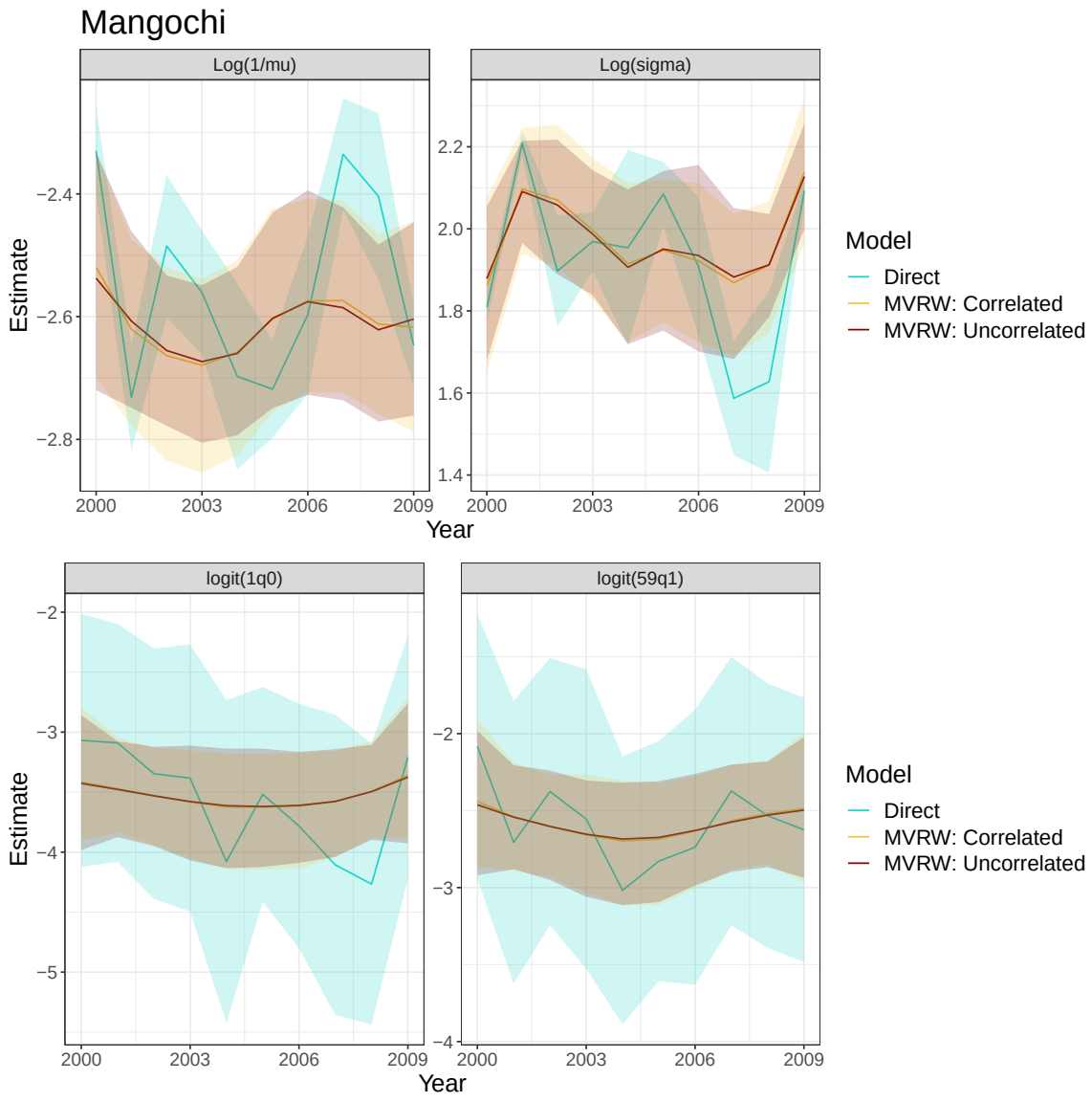


Figure C.53: Smoothed lognormal parameters (Top) and smoothed, transformed lognormal parameters (Bottom) for the Mangochi district of Malawi from 2000-2009.

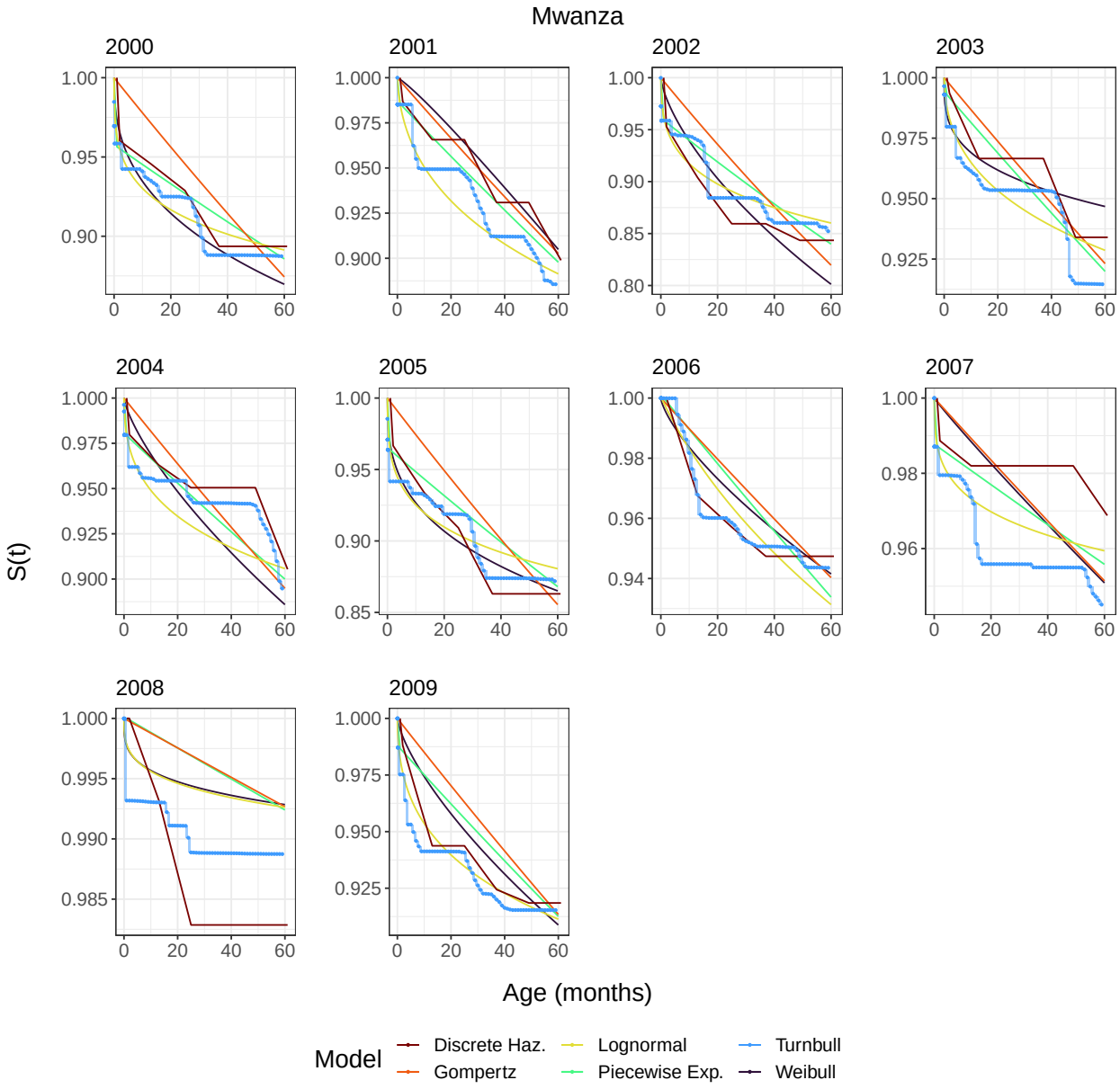


Figure C.54: Survey-weighted survival curves for the Mwanza district of Malawi from 2000-2009.

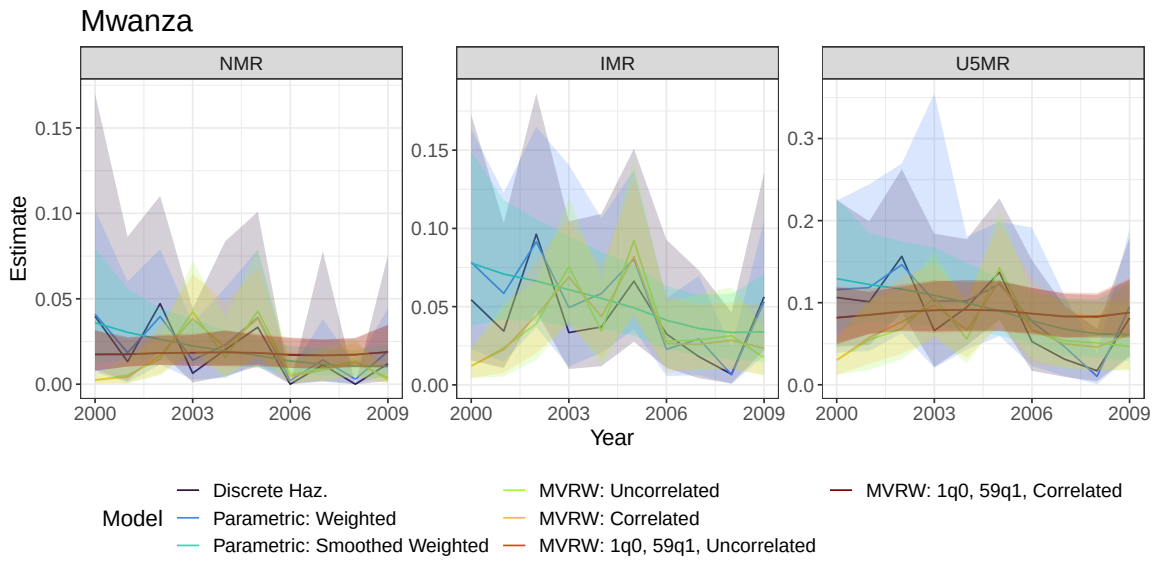


Figure C.55: Estimates of NMR, IMR, U5MR from direct estimates and smoothing models across time for the Mwanza district of Malawi from 2000-2009.

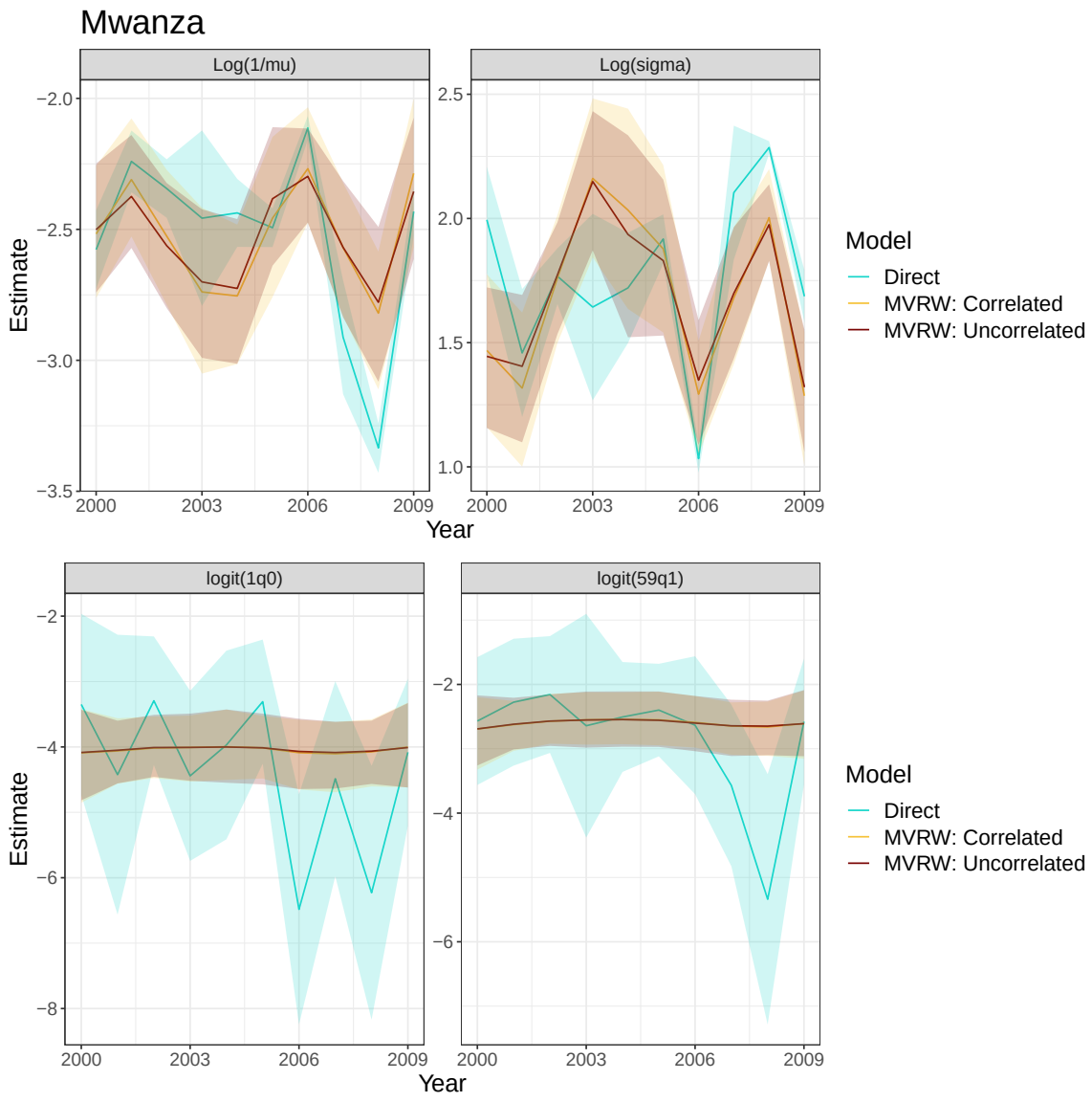


Figure C.56: Smoothed lognormal parameters (Top) and smoothed, transformed lognormal parameters (Bottom) for the Mwanza district of Malawi from 2000-2009.

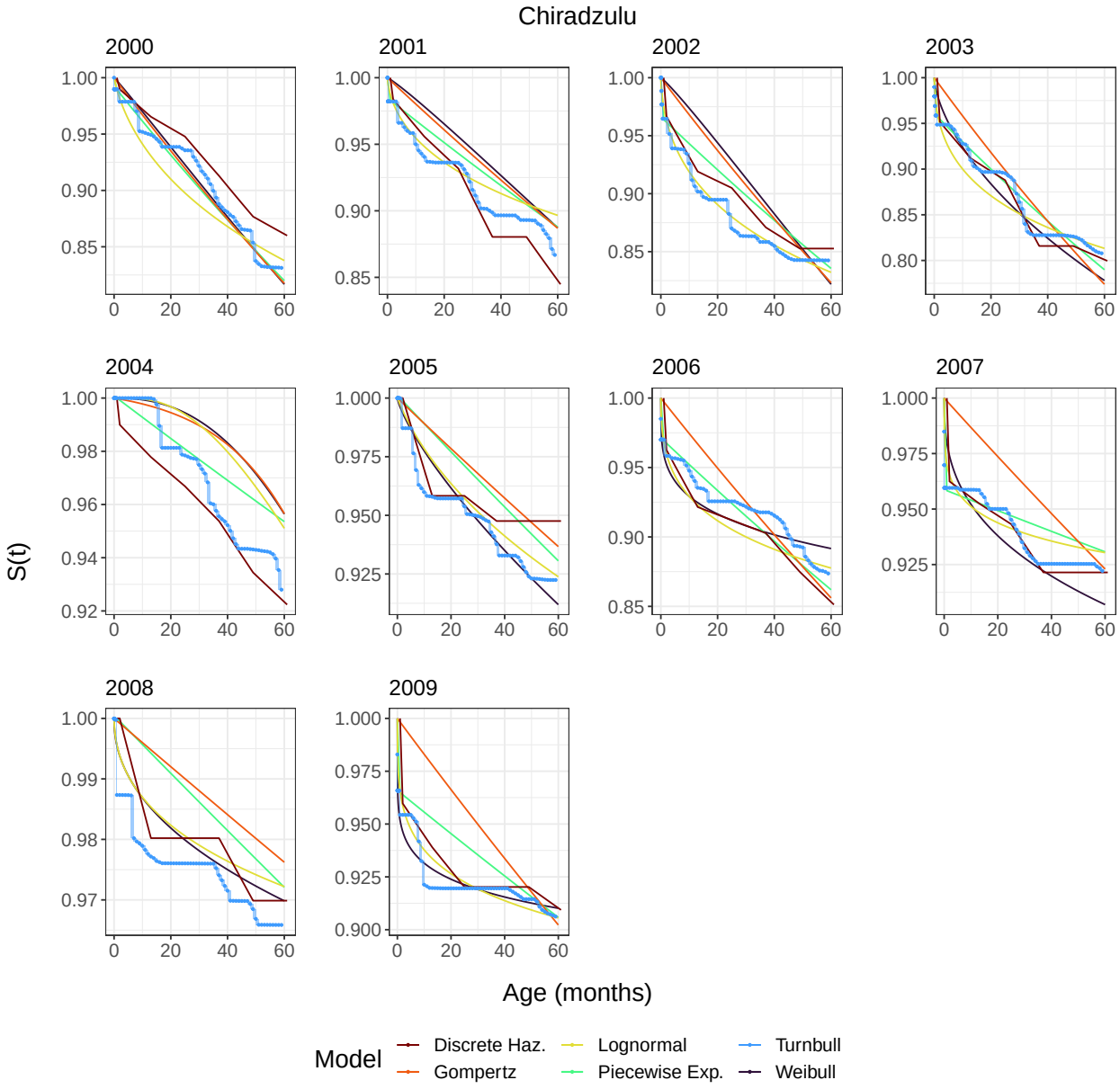


Figure C.57: Survey-weighted survival curves for the Chiradzulu district of Malawi from 2000-2009.

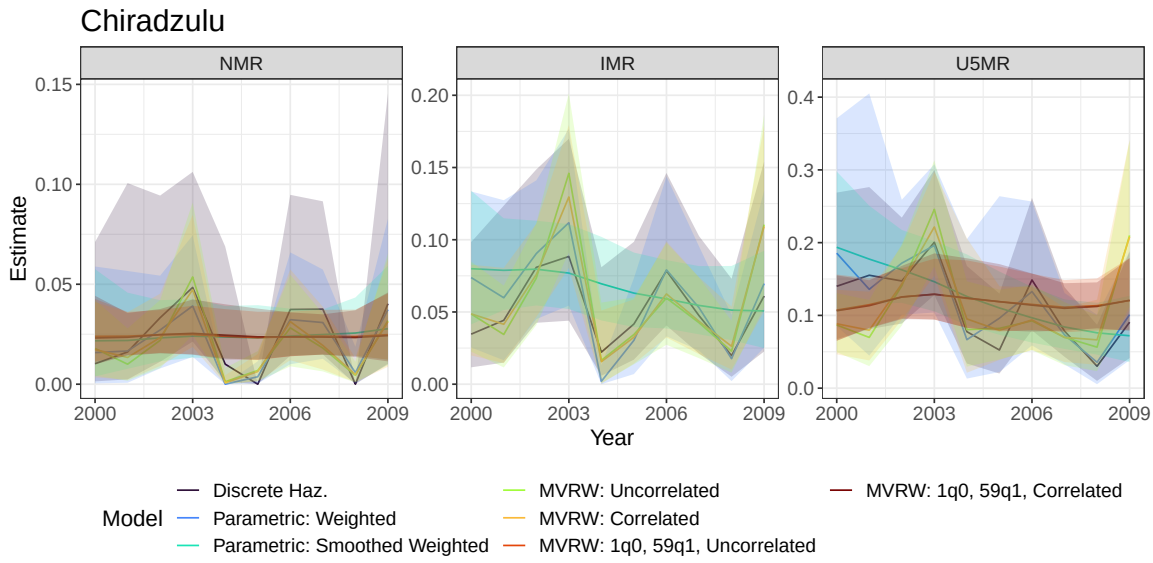


Figure C.58: Estimates of NMR, IMR, U5MR from direct estimates and smoothing models across time for the Chiradzulu district of Malawi from 2000-2009.

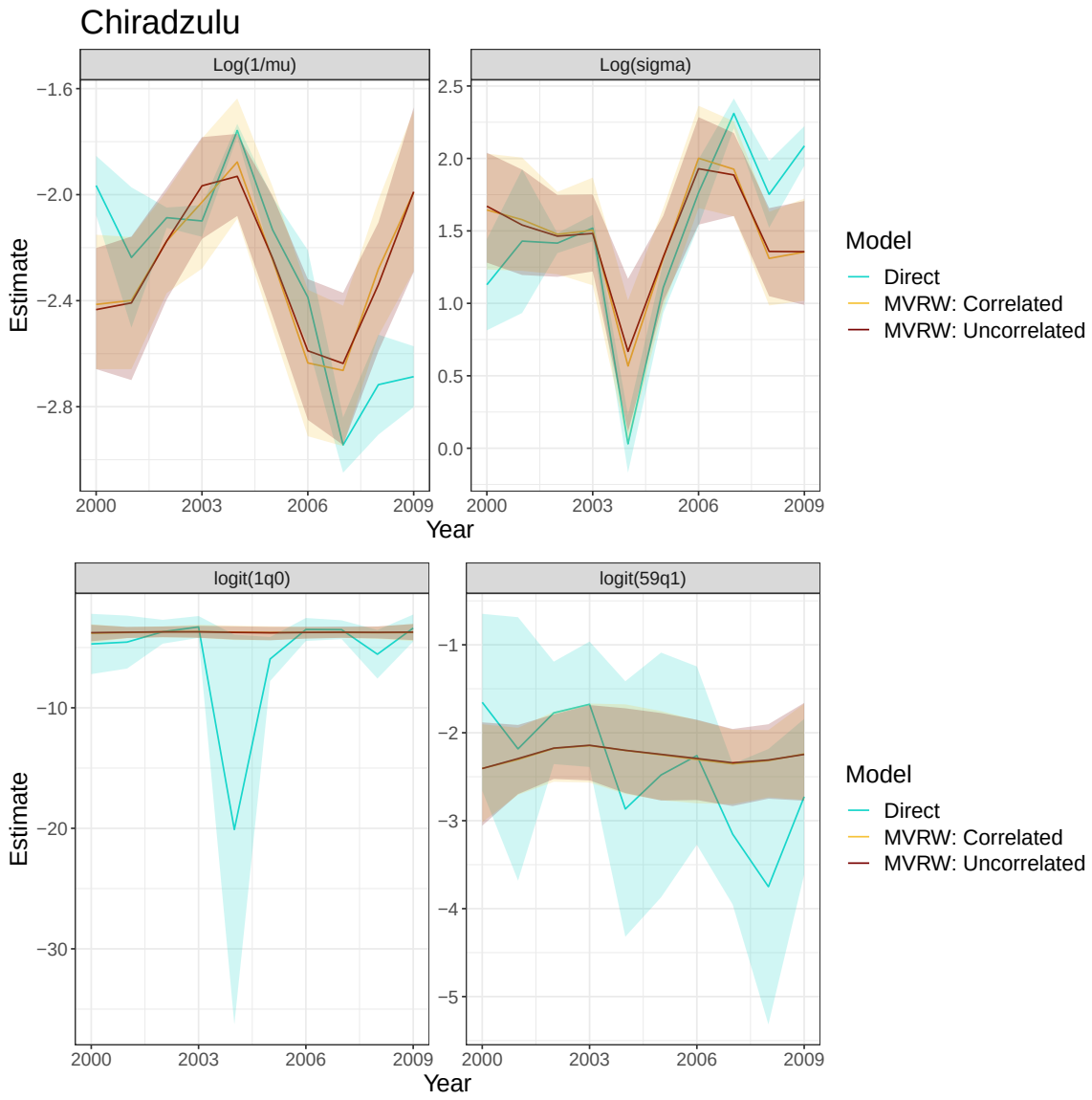


Figure C.59: Smoothed lognormal parameters (Top) and smoothed, transformed lognormal parameters (Bottom) for the Chiradzulu district of Malawi from 2000-2009.

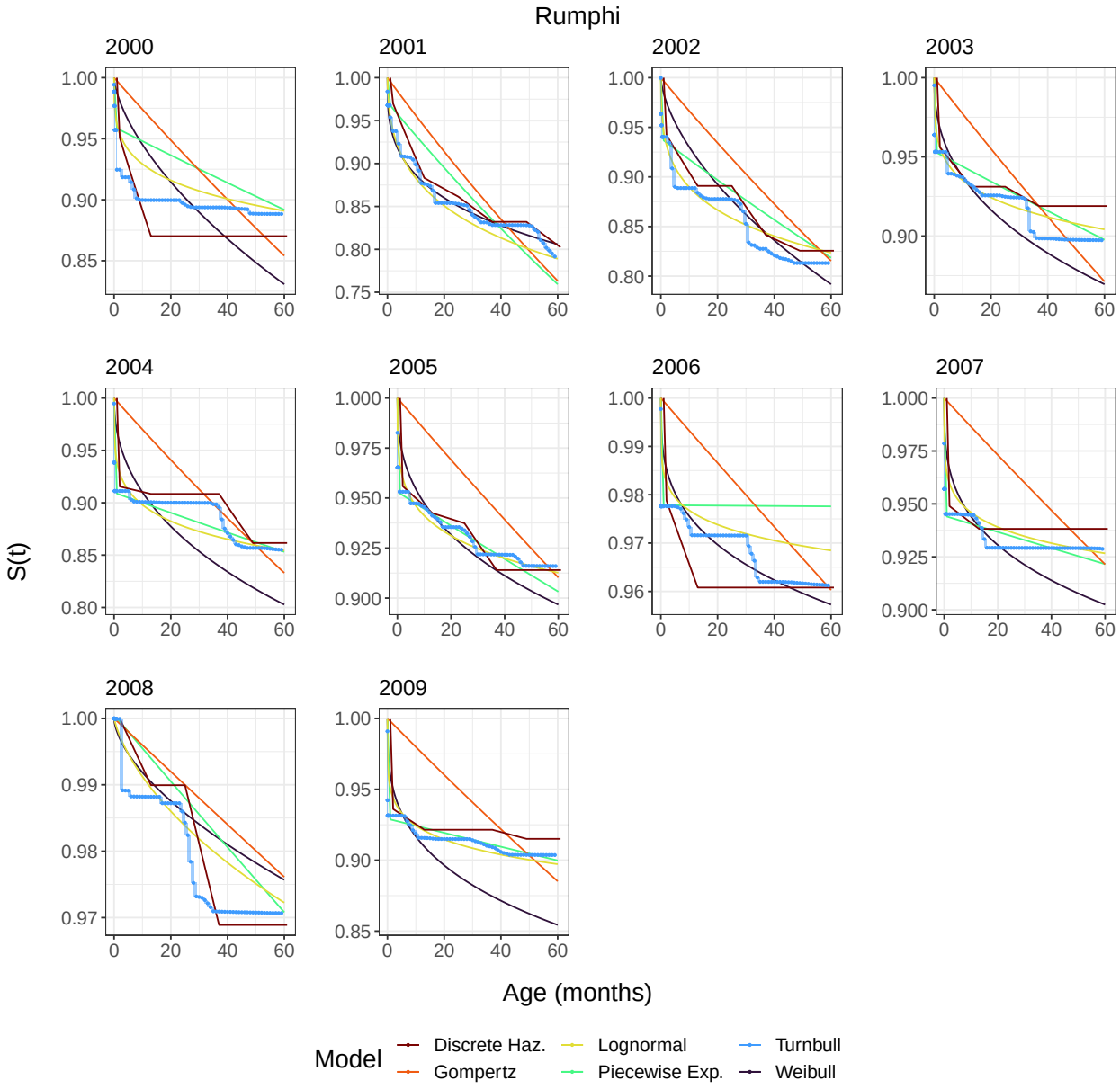


Figure C.60: Survey-weighted survival curves for the Rumphi district of Malawi from 2000-2009.

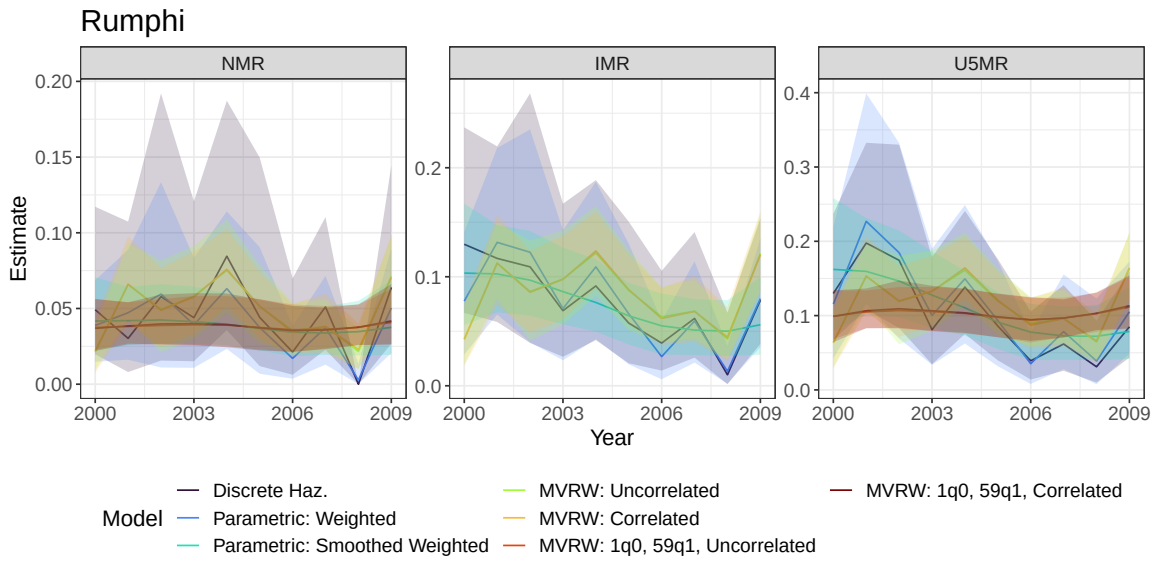


Figure C.61: Estimates of NMR, IMR, U5MR from direct estimates and smoothing models across time for the Rumphi district of Malawi from 2000-2009.

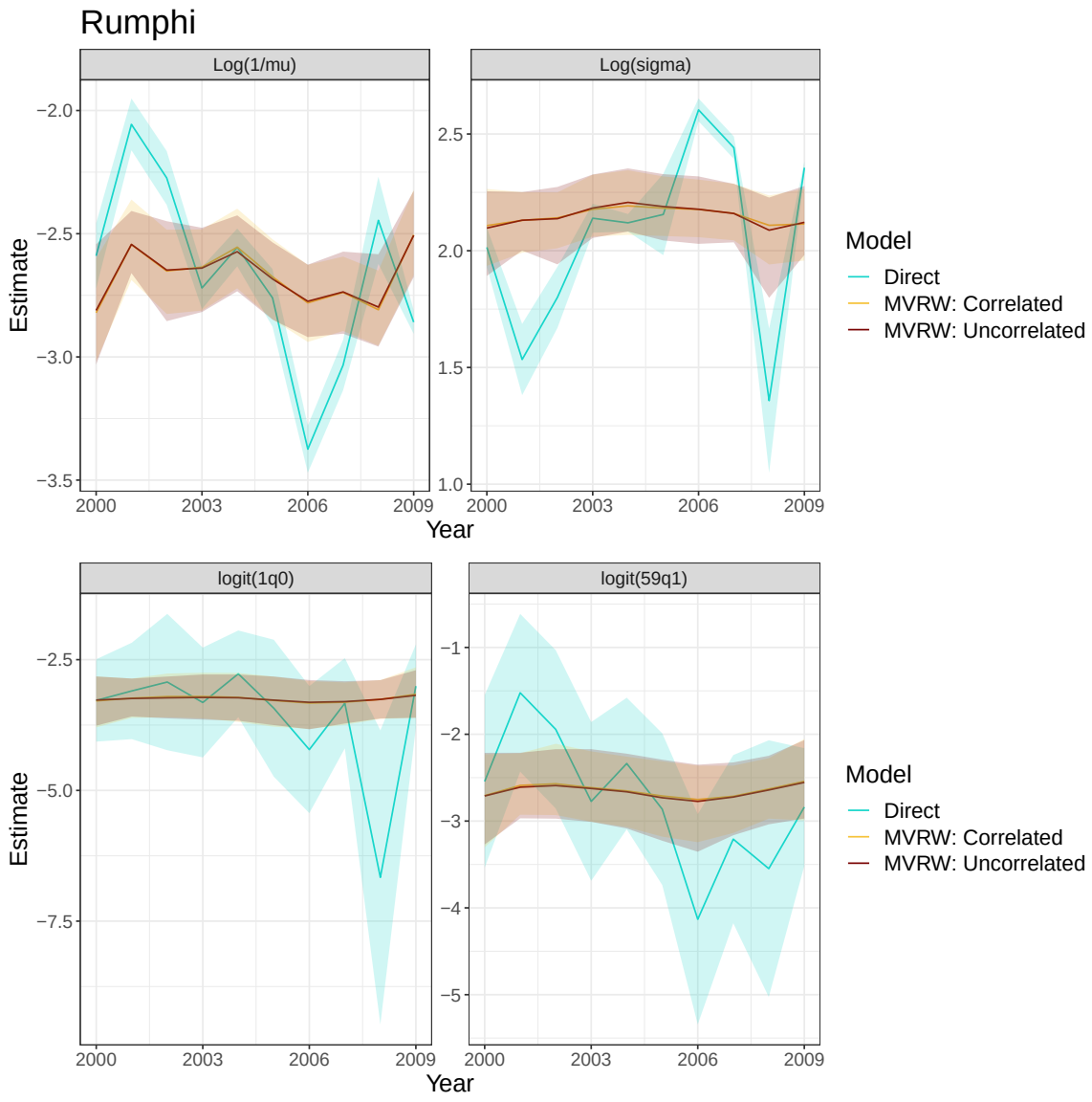


Figure C.62: Smoothed lognormal parameters (Top) and smoothed, transformed lognormal parameters (Bottom) for the Rumphi district of Malawi from 2000-2009.

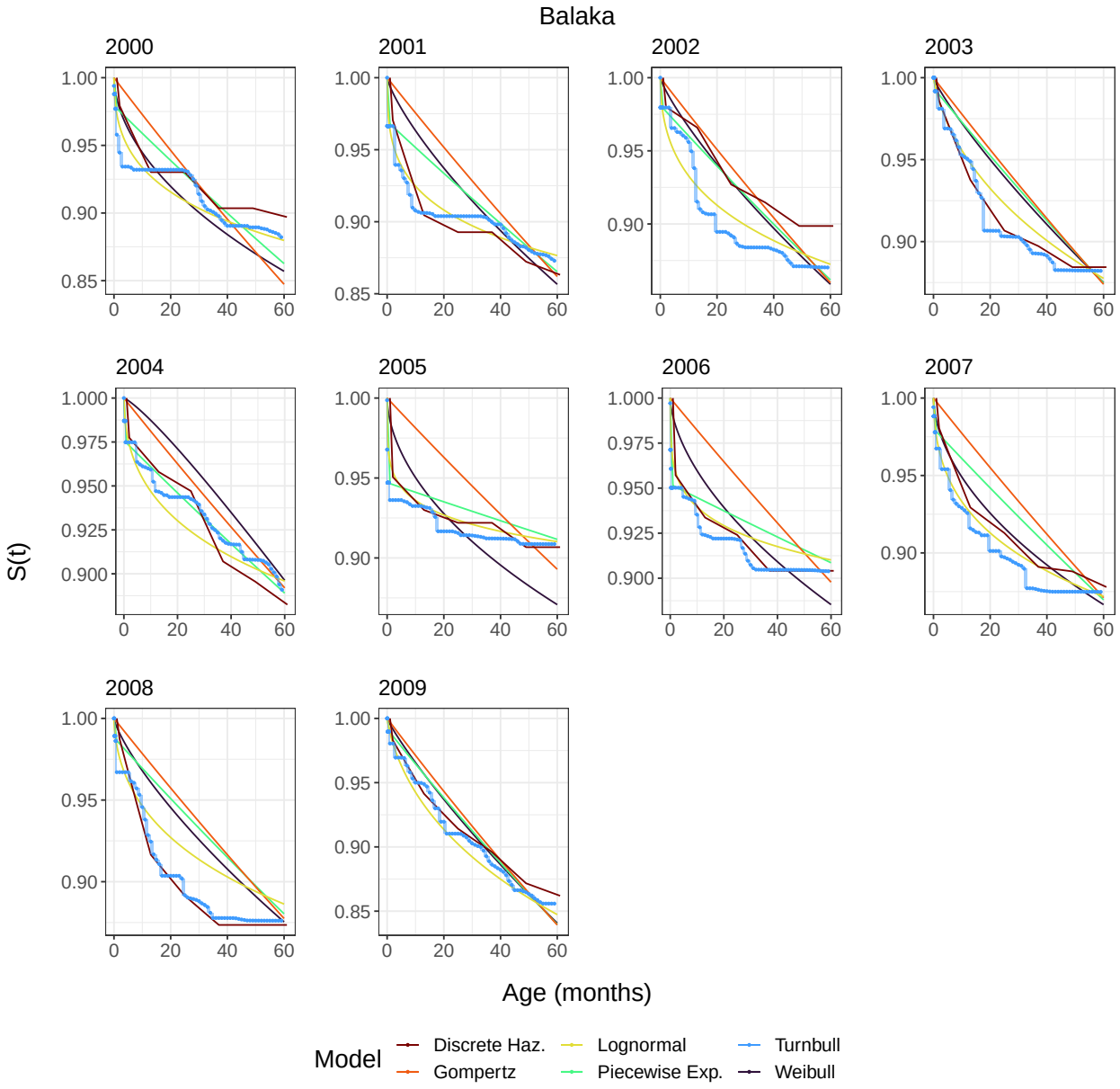


Figure C.63: Survey-weighted survival curves for the Balaka district of Malawi from 2000-2009.

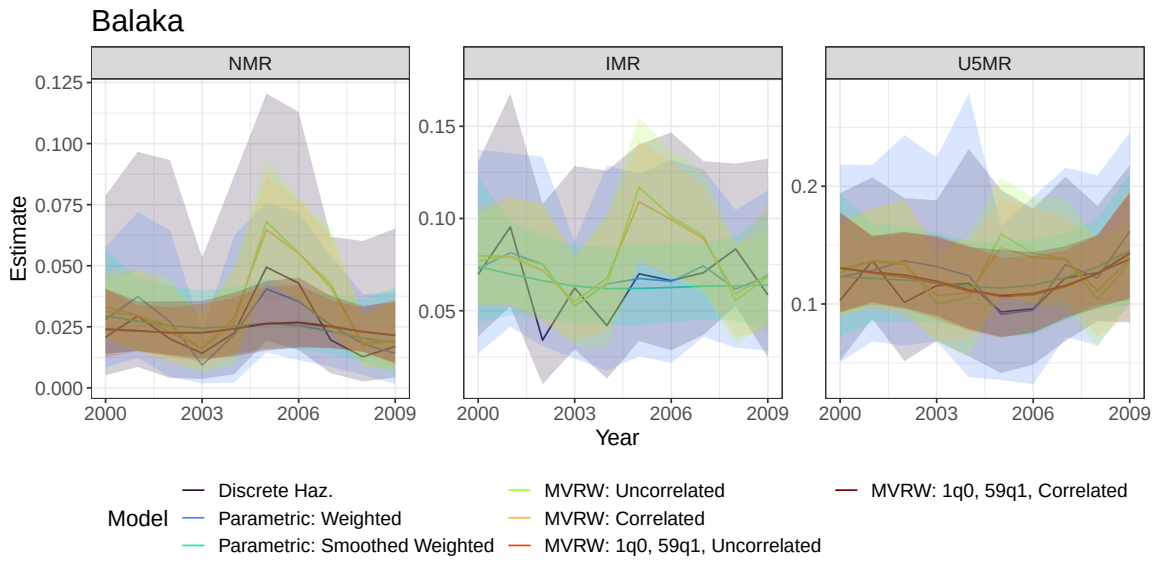


Figure C.64: Estimates of NMR, IMR, U5MR from direct estimates and smoothing models across time for the Balaka district of Malawi from 2000-2009.

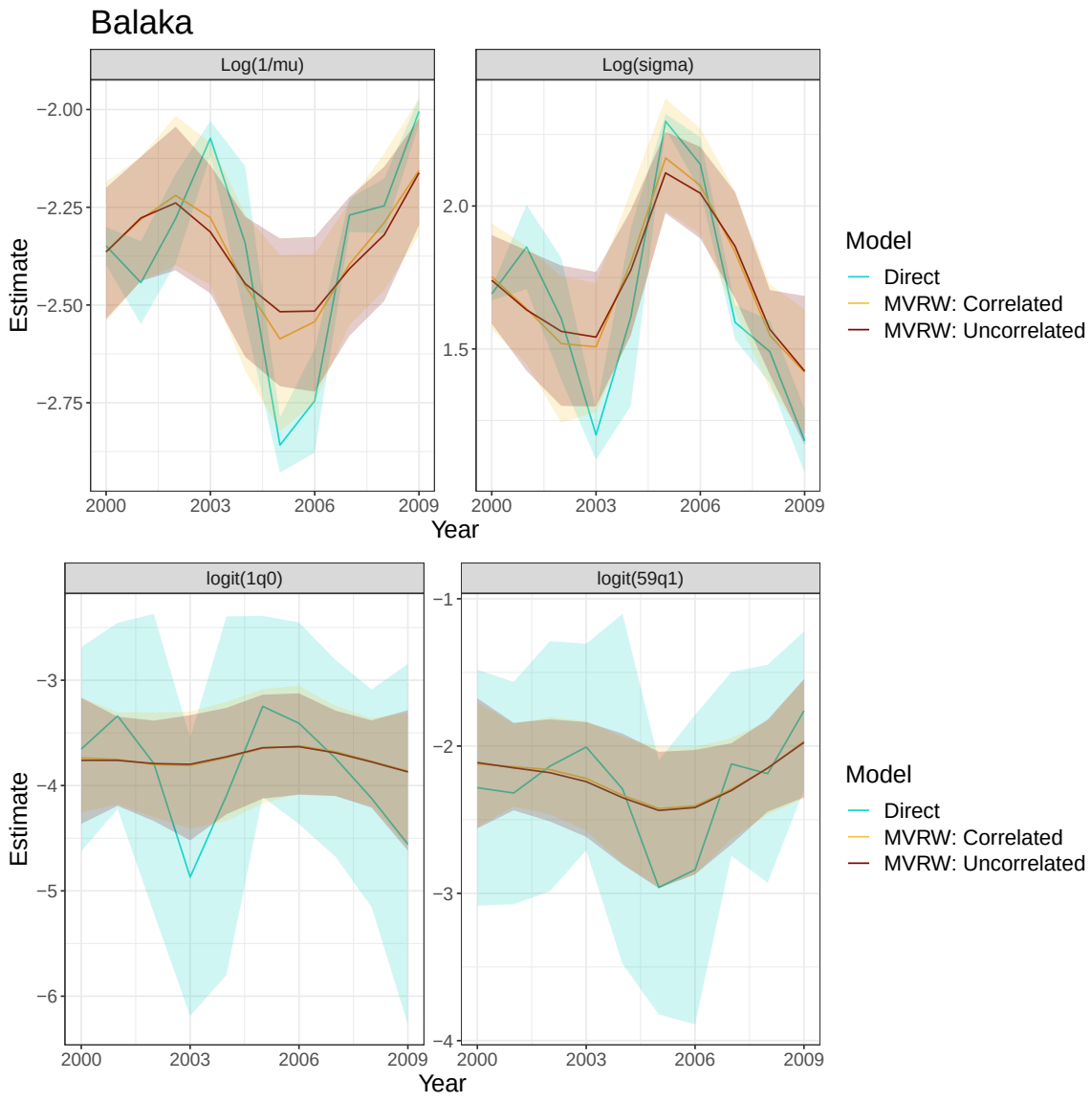


Figure C.65: Smoothed lognormal parameters (Top) and smoothed, transformed lognormal parameters (Bottom) for the Balaka district of Malawi from 2000-2009.

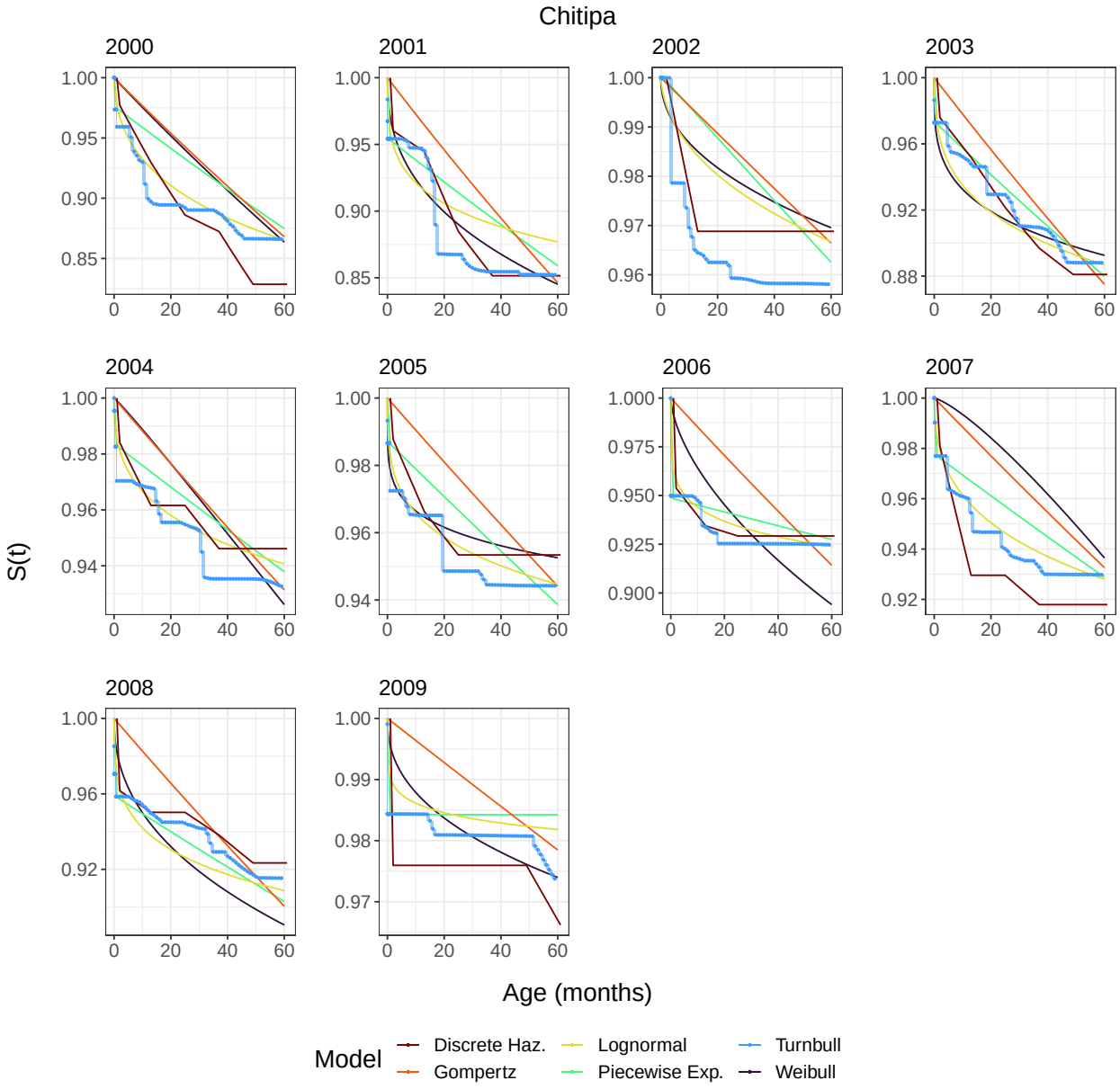


Figure C.66: Survey-weighted survival curves for the Chitipa district of Malawi from 2000-2009.

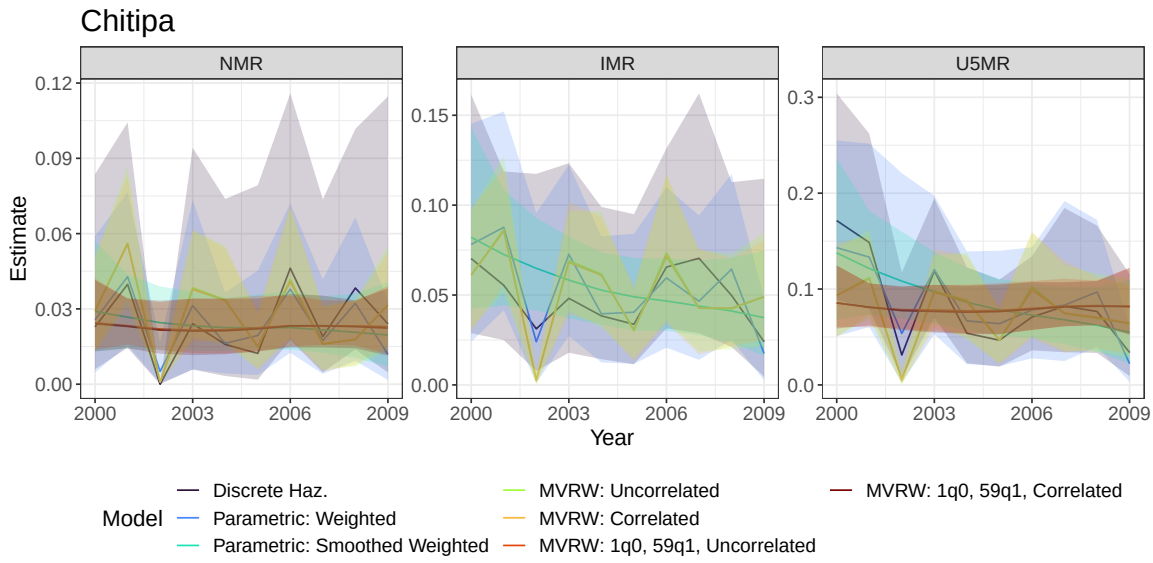


Figure C.67: Estimates of NMR, IMR, U5MR from direct estimates and smoothing models across time for the Chitipa district of Malawi from 2000-2009.

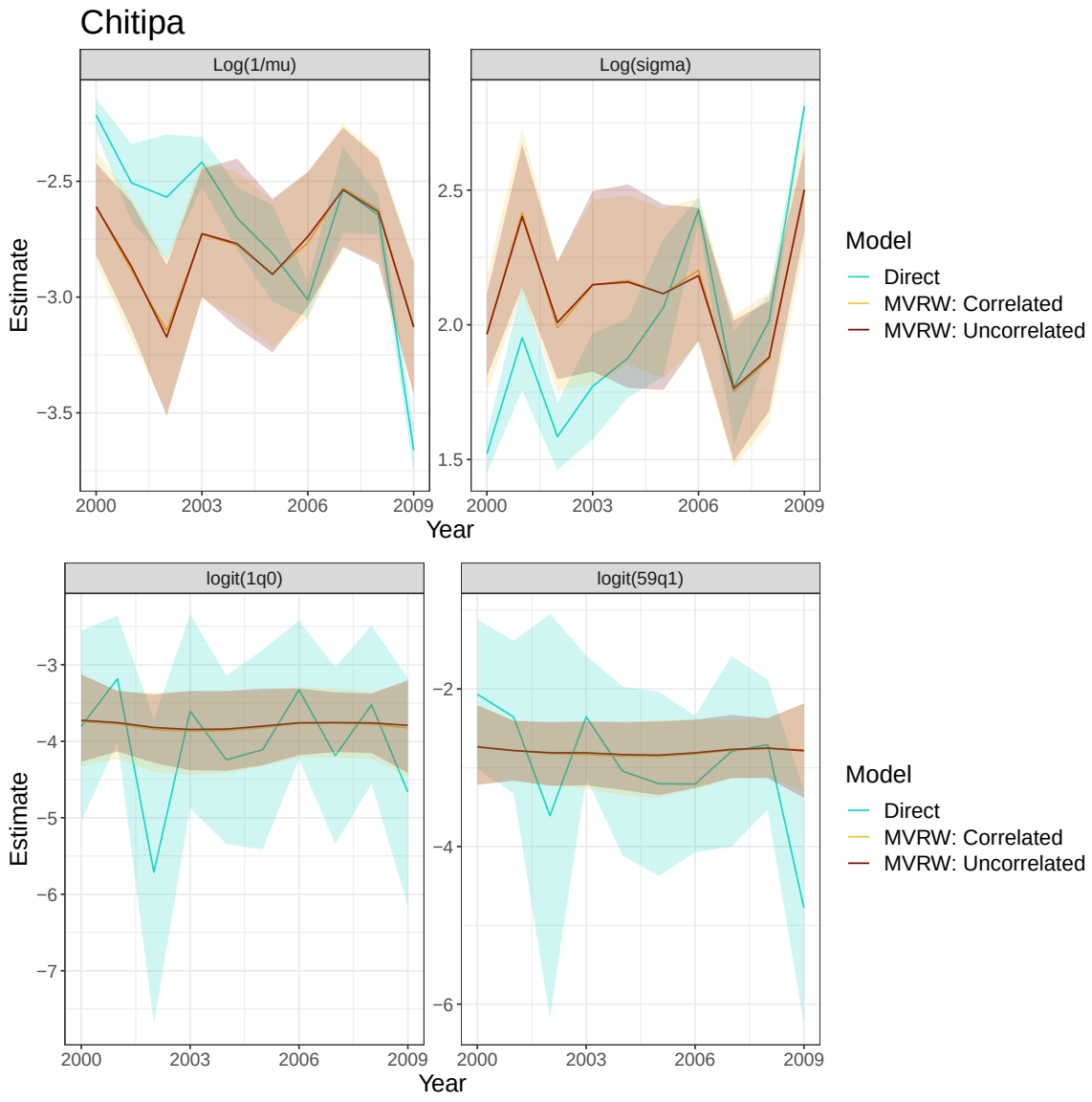


Figure C.68: Smoothed lognormal parameters (Top) and smoothed, transformed lognormal parameters (Bottom) for the Chitipa district of Malawi from 2000-2009.

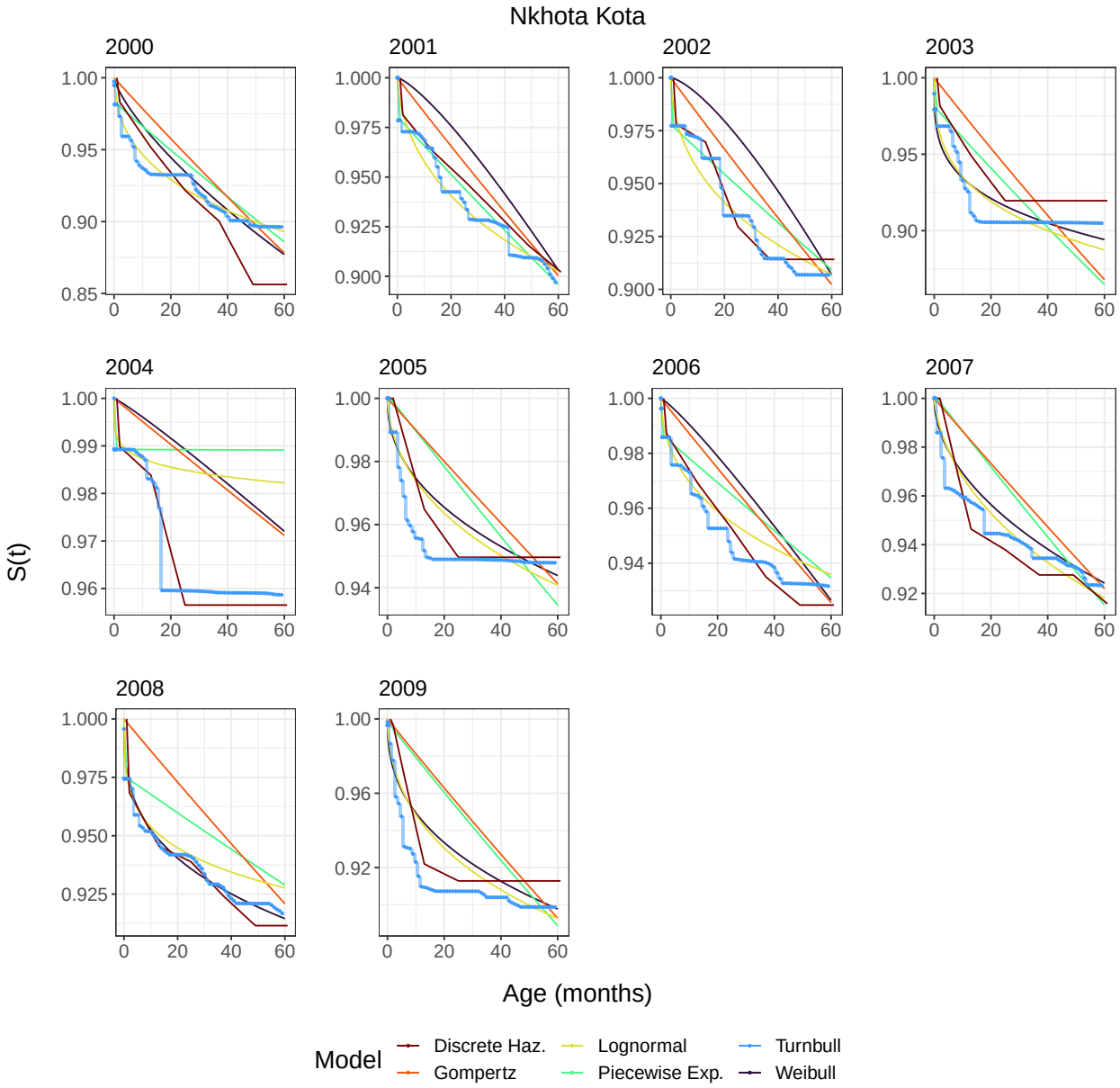


Figure C.69: Survey-weighted survival curves for the Nkhotakota district of Malawi from 2000-2009.

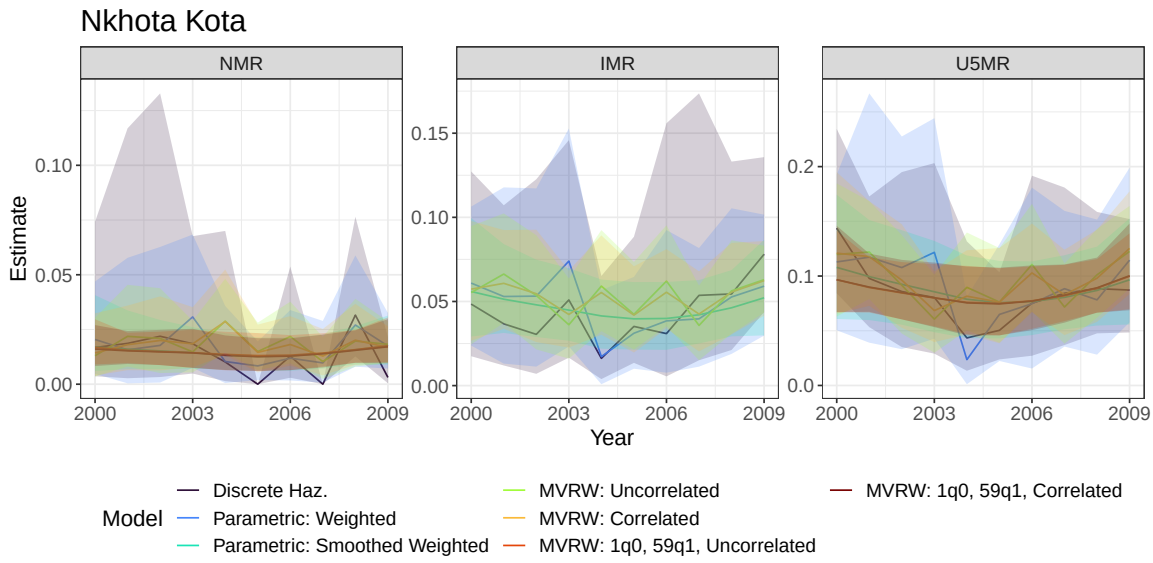


Figure C.70: Estimates of NMR, IMR, U5MR from direct estimates and smoothing models across time for the Nkhotakota district of Malawi from 2000-2009.

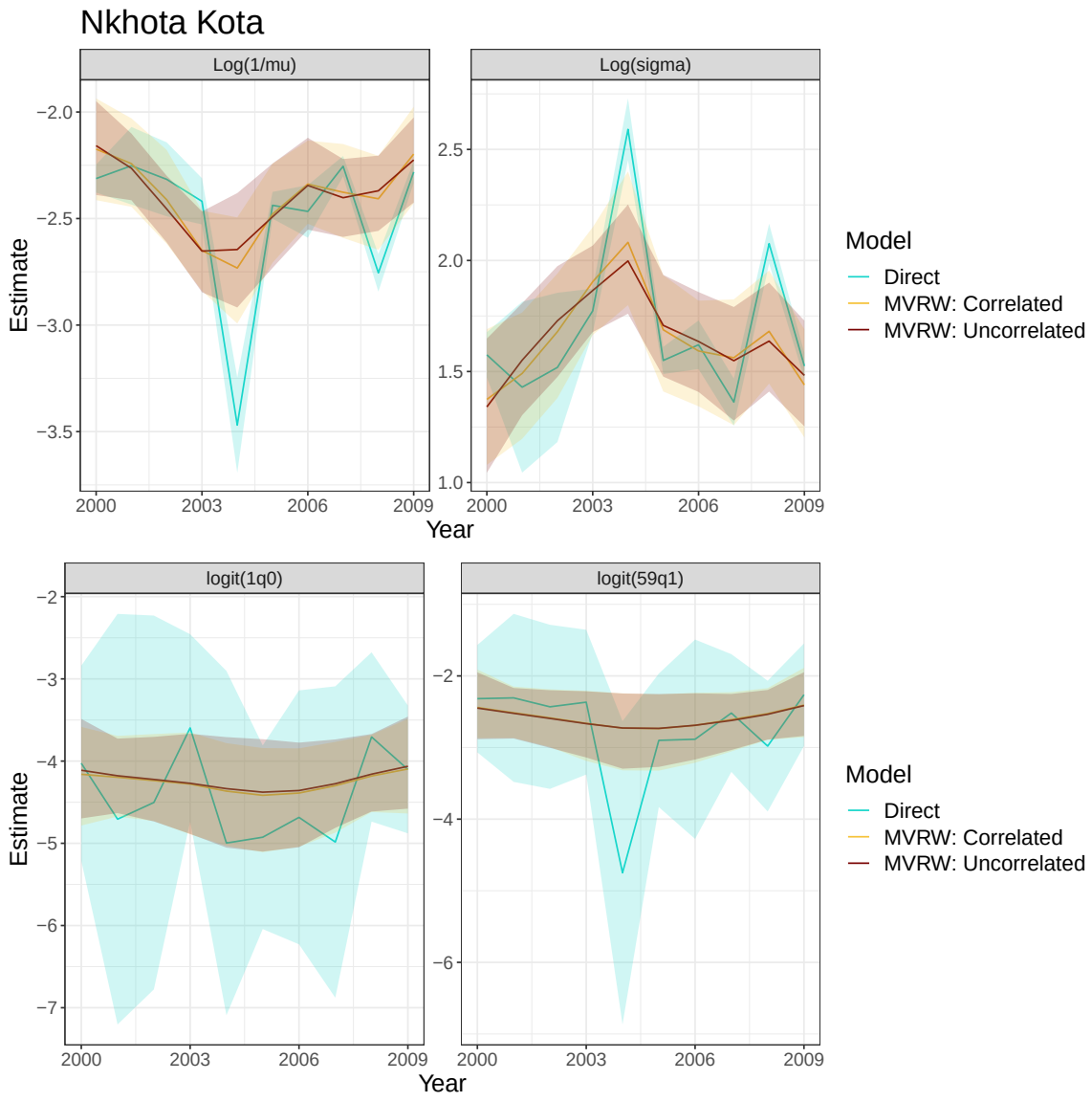


Figure C.71: Smoothed lognormal parameters (Top) and smoothed, transformed lognormal parameters (Bottom) for the Nkhotakota district of Malawi from 2000-2009.

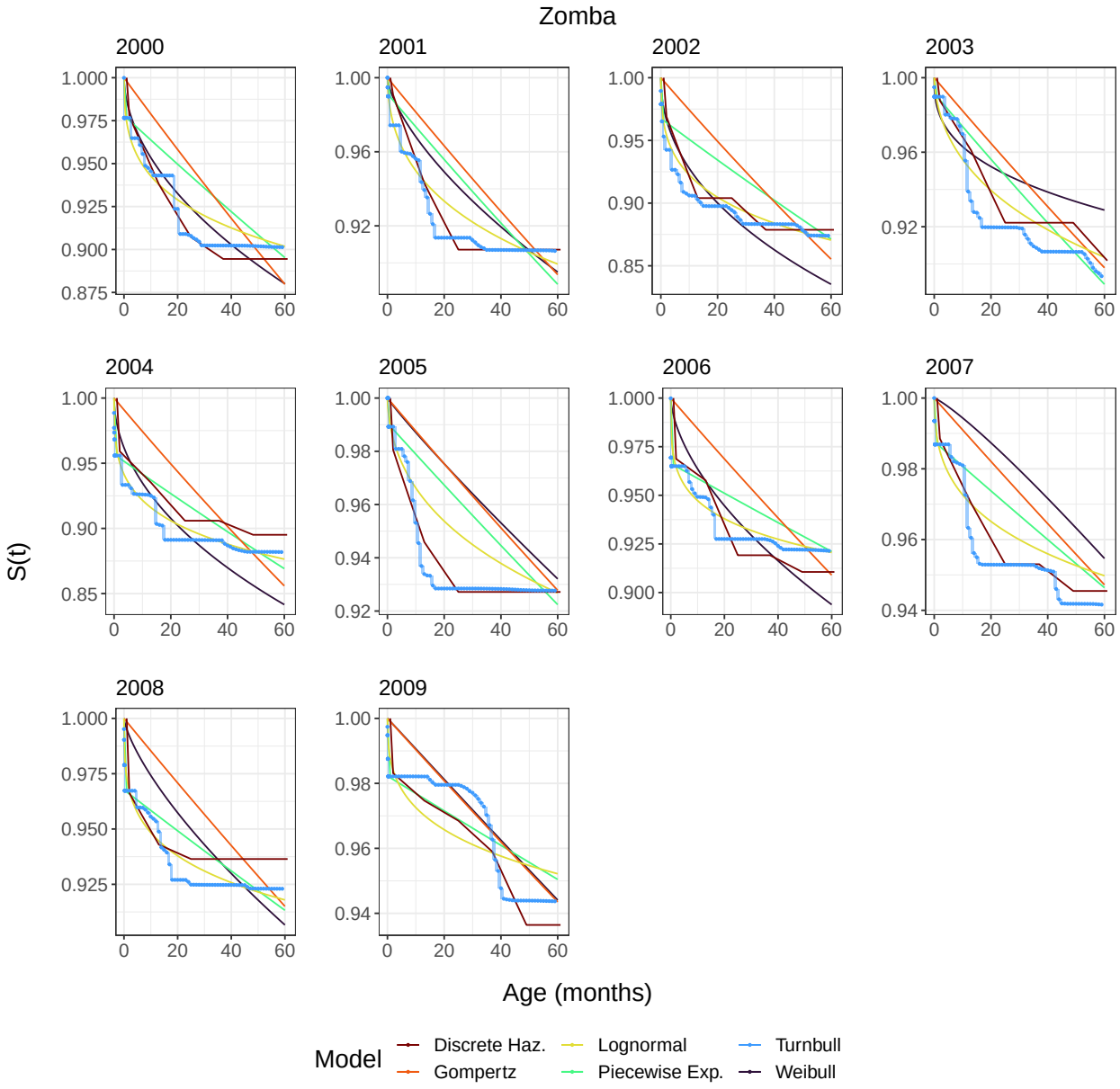


Figure C.72: Survey-weighted survival curves for the Zomba district of Malawi from 2000-2009.

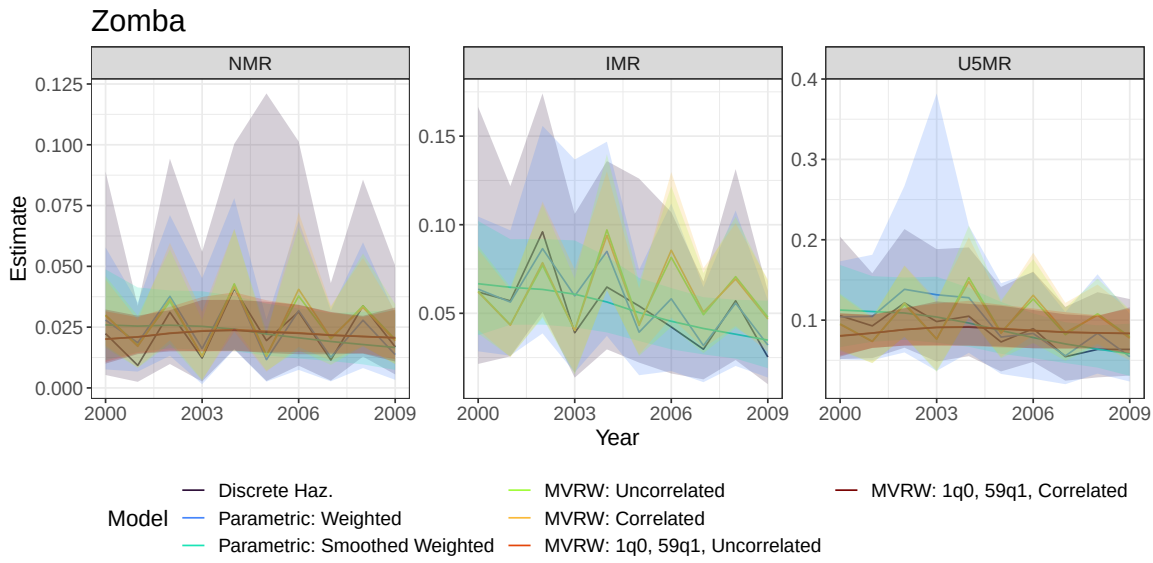


Figure C.73: Estimates of NMR, IMR, U5MR from direct estimates and smoothing models across time for the Zomba district of Malawi from 2000-2009.

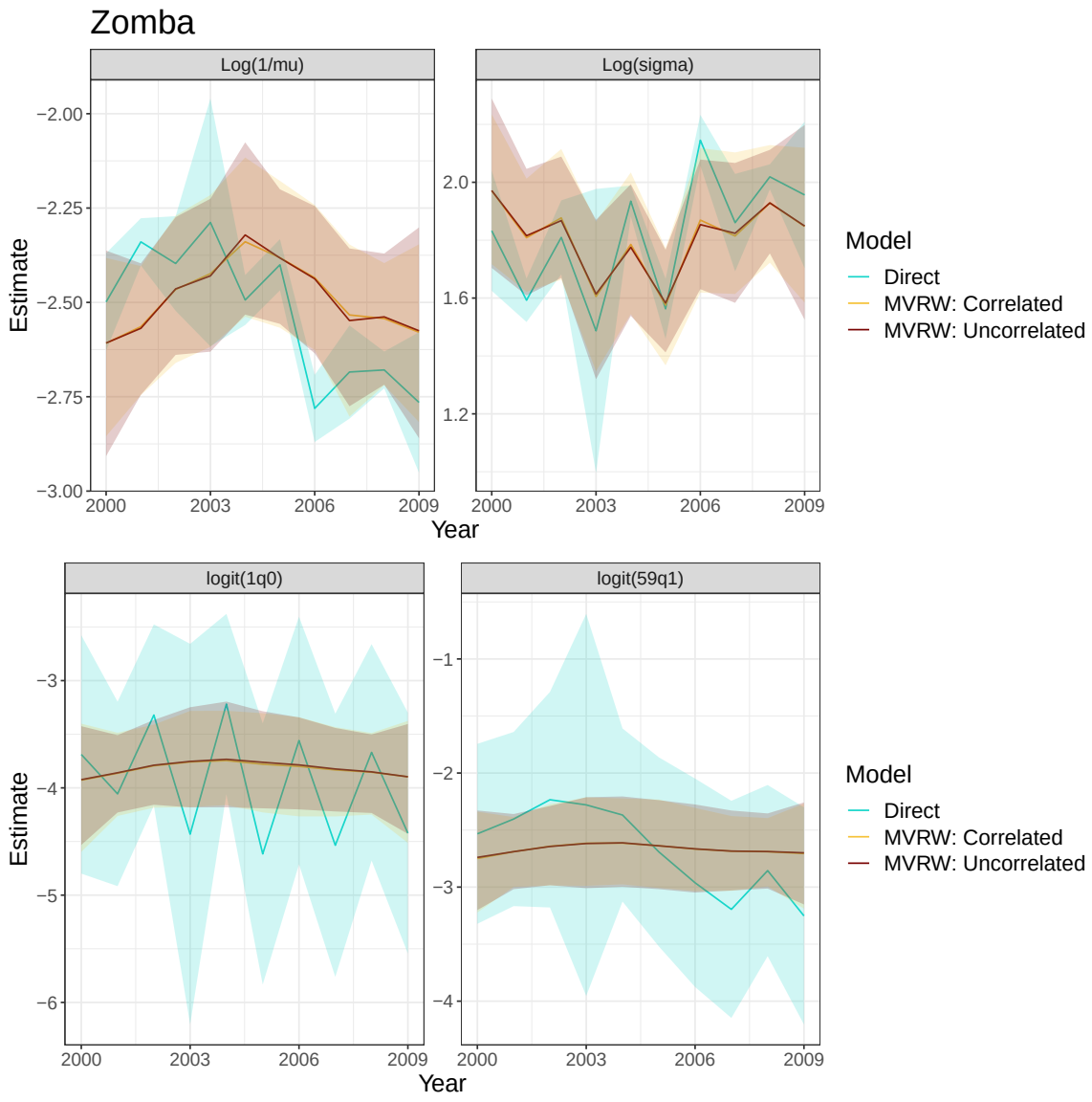


Figure C.74: Smoothed lognormal parameters (Top) and smoothed, transformed lognormal parameters (Bottom) for the Zomba district of Malawi from 2000-2009.

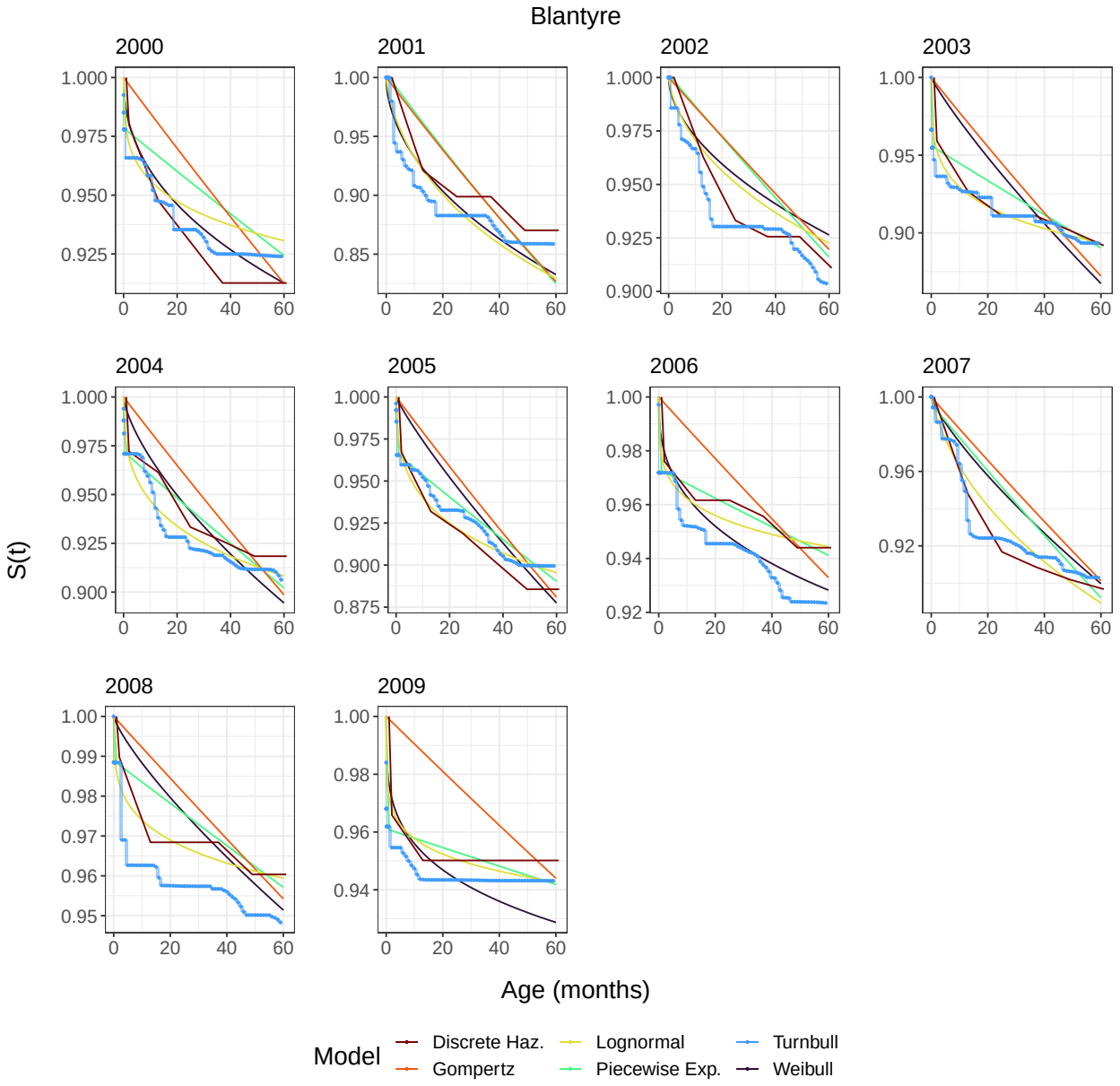


Figure C.75: Survey-weighted survival curves for the Blantyre district of Malawi from 2000-2009.

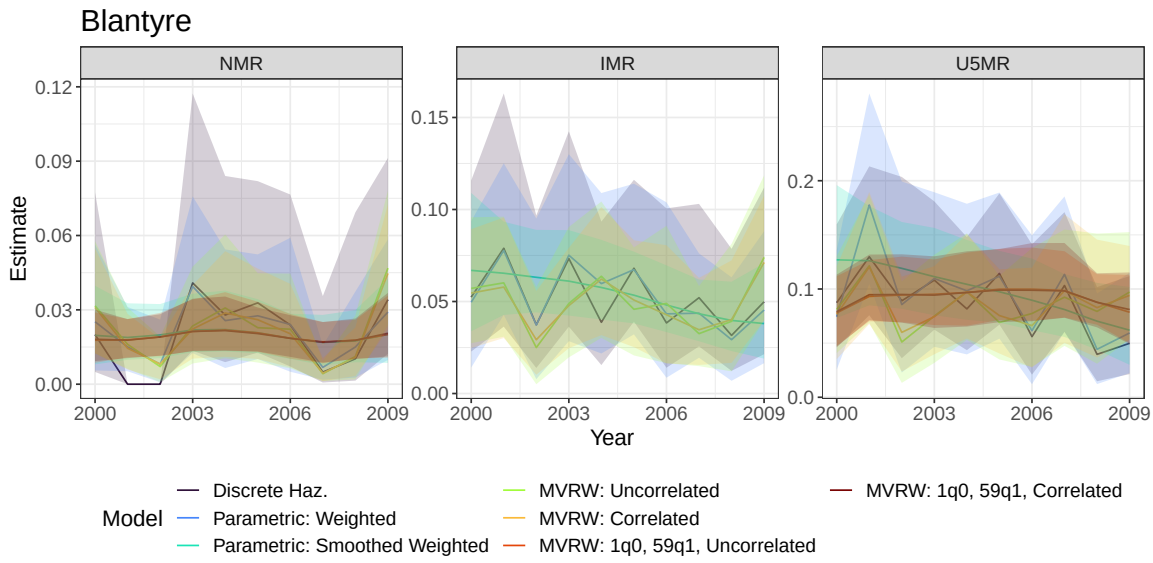


Figure C.76: Estimates of NMR, IMR, U5MR from direct estimates and smoothing models across time for the Blantyre district of Malawi from 2000-2009.

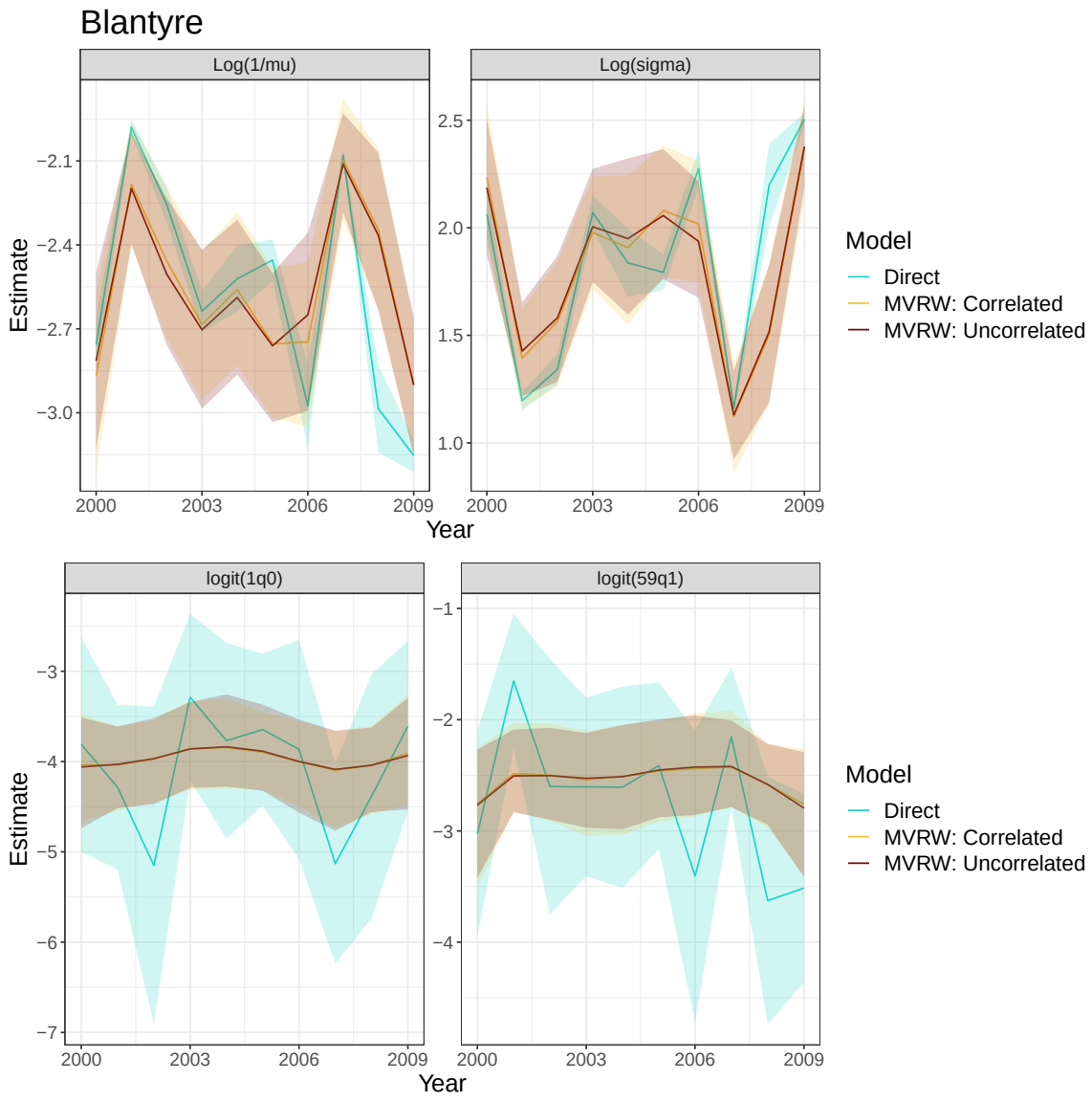


Figure C.77: Smoothed lognormal parameters (Top) and smoothed, transformed lognormal parameters (Bottom) for the Blantyre district of Malawi from 2000-2009.

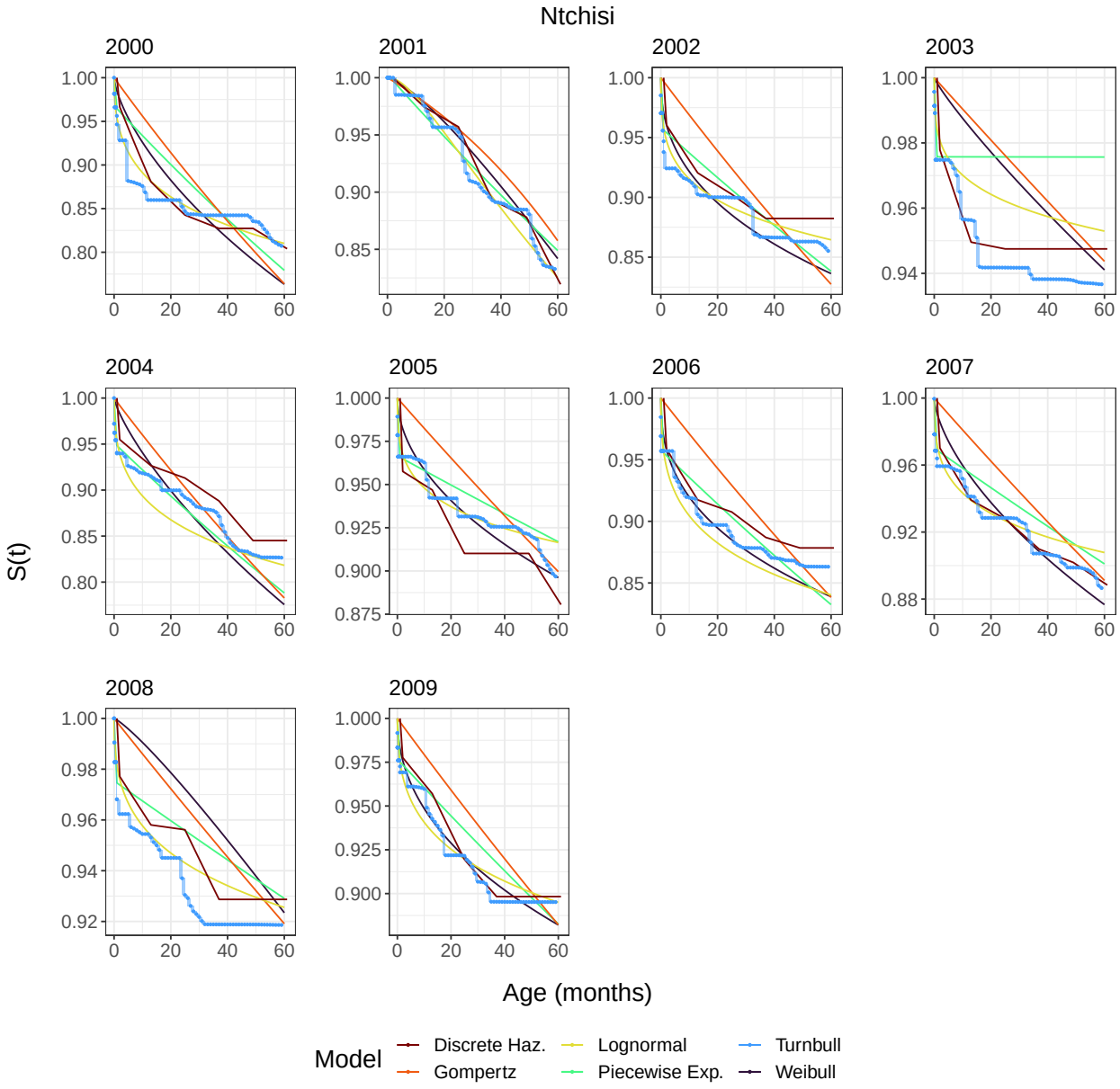


Figure C.78: Survey-weighted survival curves for the Ntchisi district of Malawi from 2000-2009.

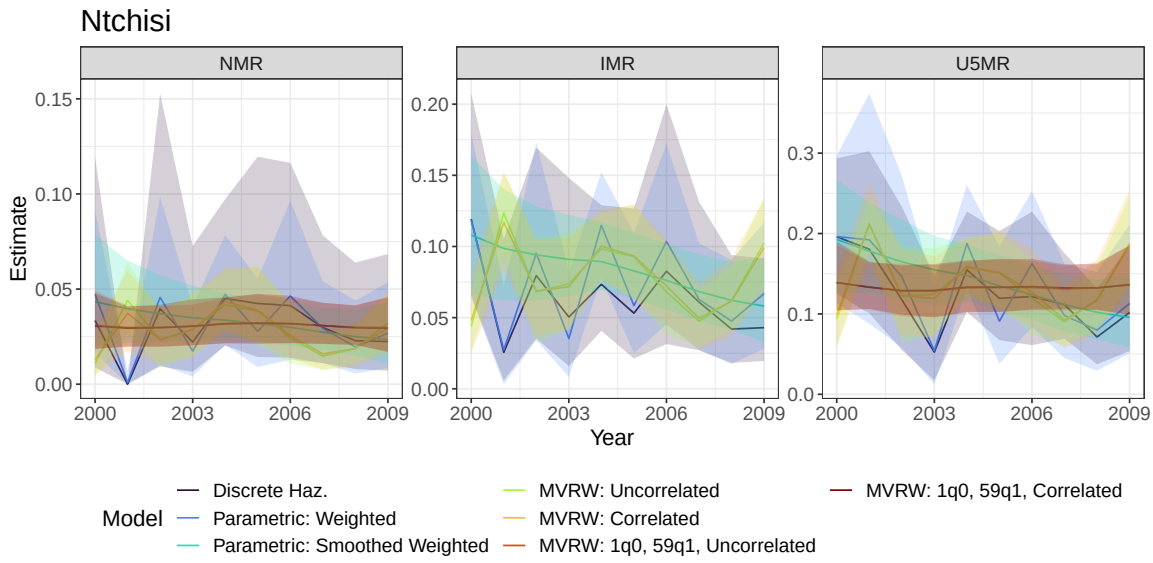


Figure C.79: Estimates of NMR, IMR, U5MR from direct estimates and smoothing models across time for the Nthchisi district of Malawi from 2000-2009.

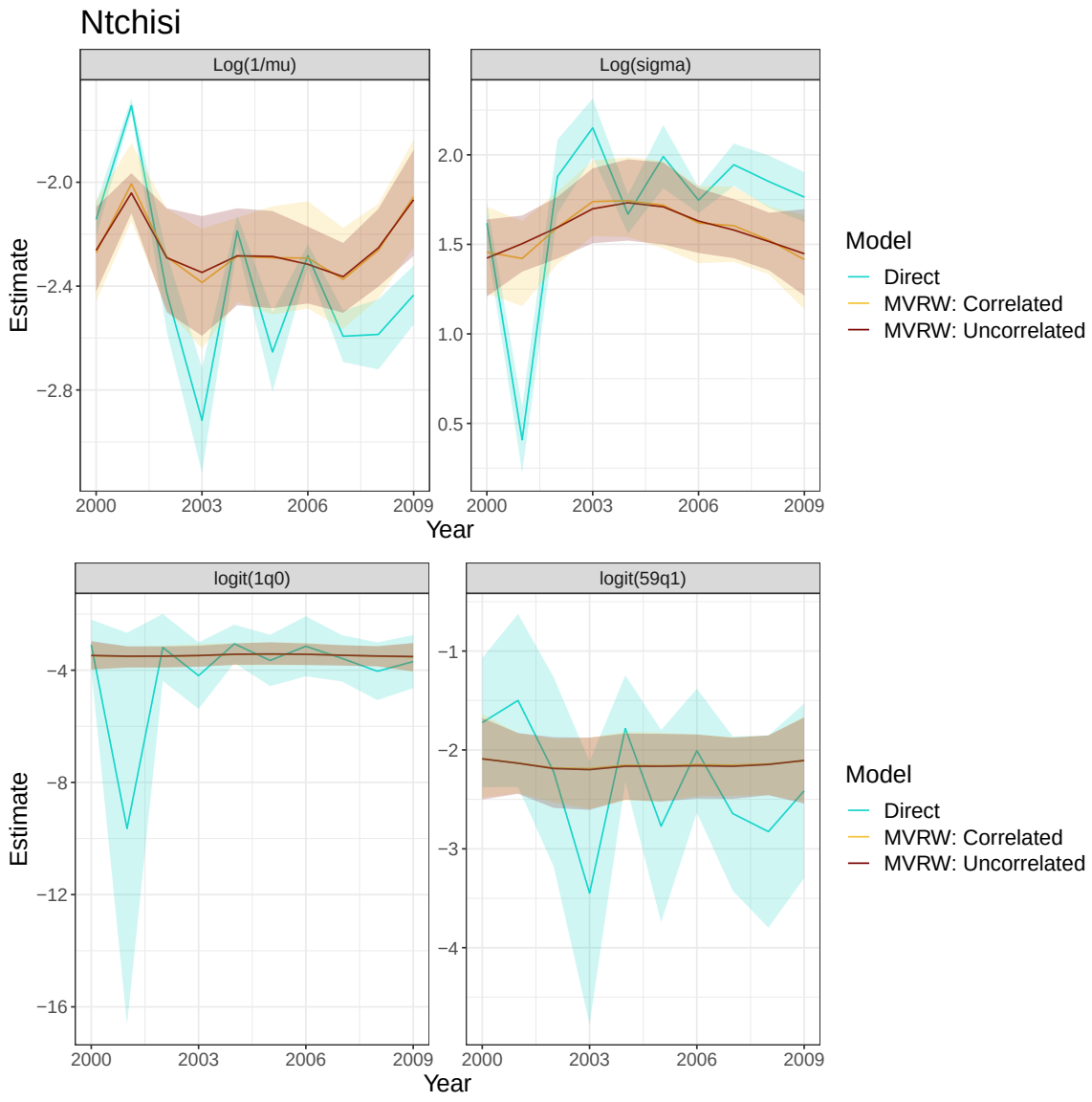


Figure C.80: Smoothed lognormal parameters (Top) and smoothed, transformed lognormal parameters (Bottom) for the Ntchisi district of Malawi from 2000-2009.

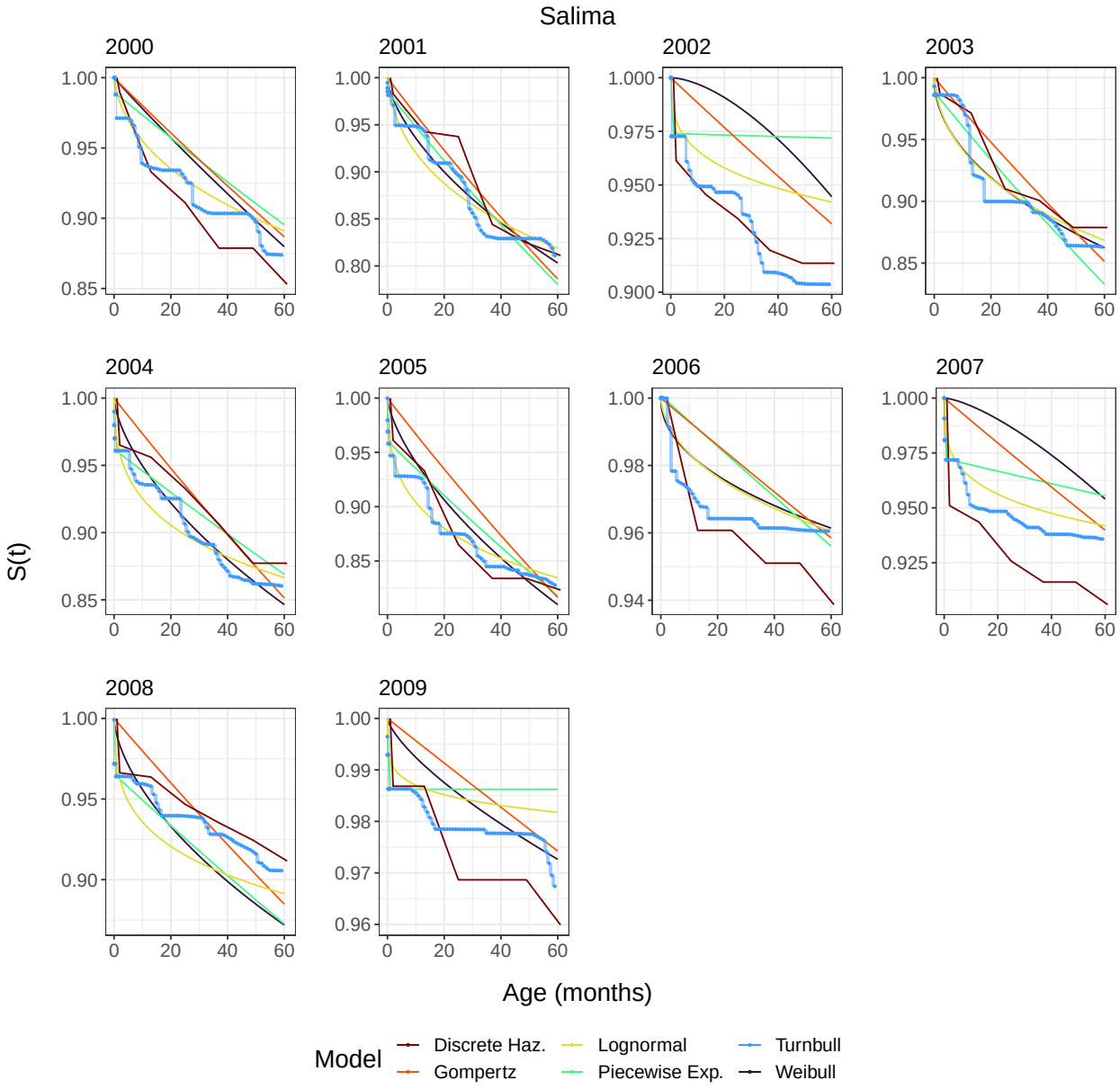


Figure C.81: Survey-weighted survival curves for the Salima district of Malawi from 2000-2009.

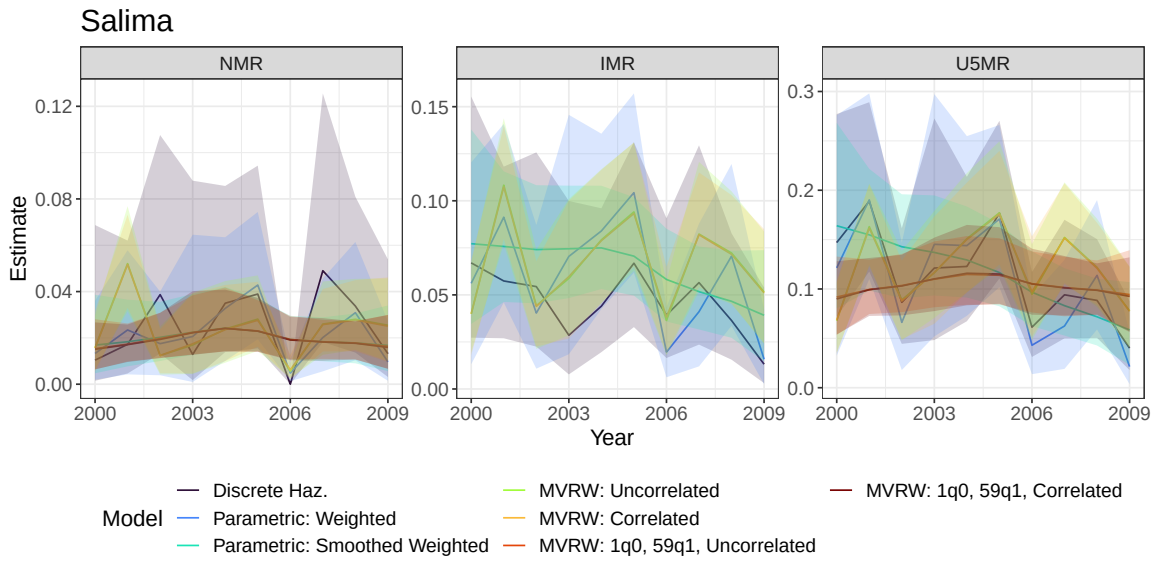


Figure C.82: Estimates of NMR, IMR, U5MR from direct estimates and smoothing models across time for the Salima district of Malawi from 2000-2009.

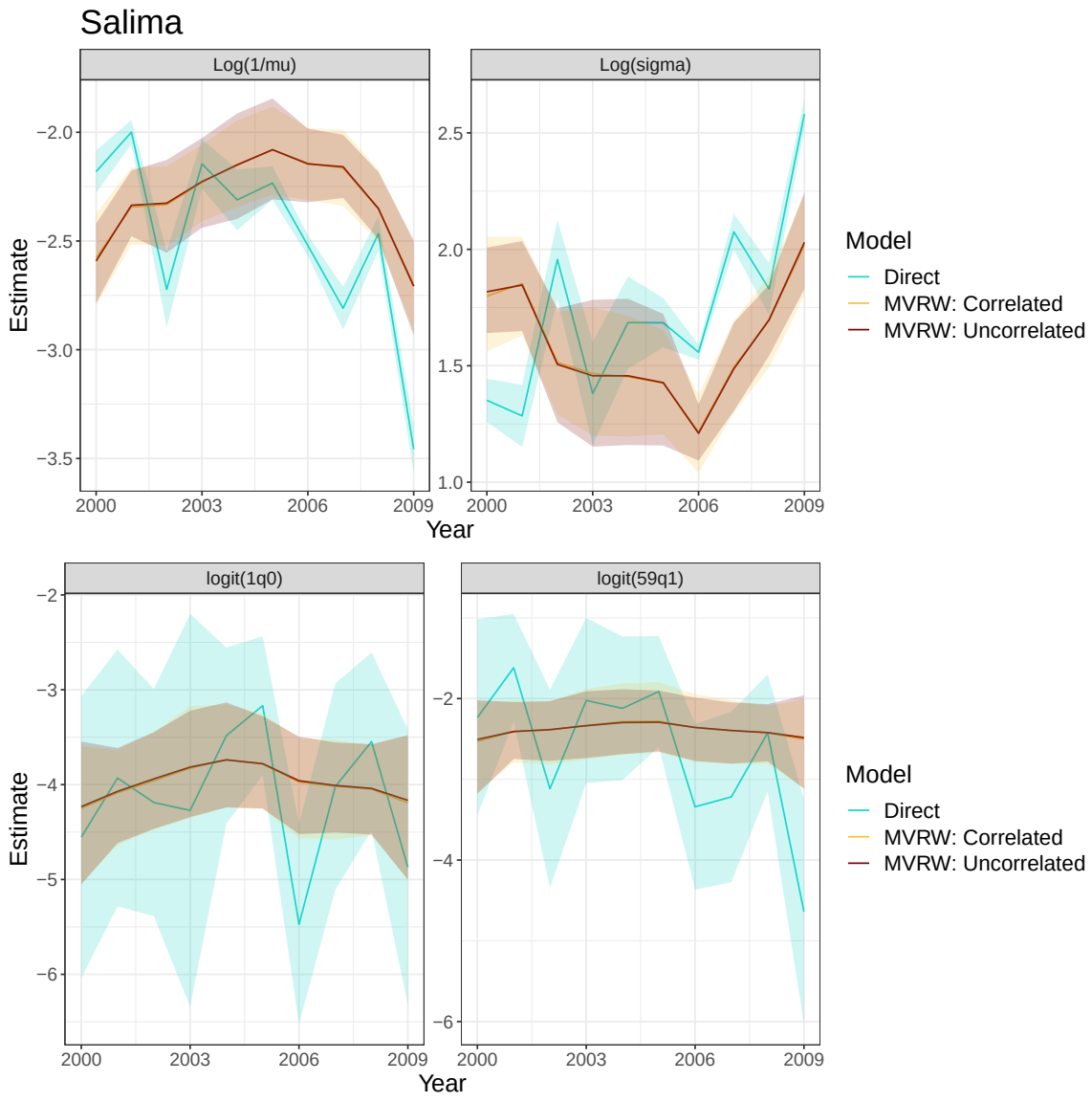


Figure C.83: Smoothed lognormal parameters (Top) and smoothed, transformed lognormal parameters (Bottom) for the Salima district of Malawi from 2000-2009.