Analysis of Infectious Disease Incidence and Complex Survey Data in Space and Time

Qi Dong

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

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Jon Wakefield, Chair

M. Elizabeth Halloran

Lurdes Inoue

Program Authorized to Offer Degree:
Biostatistics - Public Health

University of Washington

Abstract

Analysis of Infectious Disease Incidence and Complex Survey Data in Space and Time

Qi Dong

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

Immunization is one of the most cost-effective health interventions for preventing infectious diseases. Despite the tremendous advances in global immunization coverage in the past few decades, challenges remain in many low- and middle-income countries (LMICs). Timely and comprehensive evaluation of immunization program performance, especially at the sub-national level, is essential for monitoring progress and planning future interventions. Since the administrative data in many LMICs is unreliable due to inadequate health information systems, other data sources, such as disease surveillance systems and household surveys are turned to, to obtain valid estimates for policy guidance. This dissertation develops statistical methods for analyzing infectious disease incidence and complex survey data in space and time to facilitate planning and assessing immunization programs in LMICs. In particular, we develop space-time smoothing models for estimating sub-national measles routine immunization coverage using survey data, and build time-series models for estimating efficacy of measles supplementary immunization activities using incidence data. We also consider vaccination coverage mapping via Bayesian geostatistical modeling of household survey data, demonstrate appropriate ways of accounting for the survey design in modeling, and develop a novel approach to coverage map presentation to allow comparison and control of the overall map uncertainty level.

TABLE OF CONTENTS

Page

List of Figures iii

List of Tables x

Glossary xii

Chapter 1: Introduction 1
1.1 Motivating examples 2
1.2 Organization of dissertation 4

Chapter 2: Background 6
2.1 Survey sampling 6
2.2 Gaussian Markov random fields (GMRF) 7
2.3 Models for infectious diseases 11
2.4 Bayesian analysis and computation 14
2.5 Model assessment metrics 17

Chapter 3: Space-time smoothing models for routine immunization coverage estimation with complex survey data 20
3.1 Introduction 20
3.2 Motivating data 22
3.3 Methods 27
3.4 Simulation study 33
3.5 Application to Nigeria data 37
3.6 Discussion 54

Chapter 4: Estimating efficacy of supplementary immunization activities with disease incidence time series 57

4.1 Introduction . . . 57
4.2 Methods . . . 59
4.3 Simulation study . . . 65
4.4 Analysis of the measles incidence data in Benin . . . 78
4.5 Discussion . . . 83

Chapter 5: Modeling and presentation of vaccination coverage estimates using data from household surveys . . . 86
5.1 Introduction . . . 86
5.2 Motivating example: the 2018 Nigeria DHS . . . 88
5.3 Acknowledging the survey design in modeling . . . 92
5.4 Presenting model-based estimates and their uncertainties . . . 100
5.5 Discussion . . . 126

Chapter 6: Discussion and Future Work . . . 132

Bibliography . . . 135

Appendix A: Appendix for Chapter 3 . . . 149
A.1 Autoregressive model of order 1 (AR1) for Gaussian random vectors . . . 149
A.2 Additional results from the sensitivity analysis using 12-month birth cohorts 150

Appendix B: Appendix for Chapter 4 . . . 155

Appendix C: Appendix for Chapter 5 . . . 166
C.1 Covariate data sources . . . 166
C.2 The approximated enumeration area (EA) map . . . 167

LIST OF FIGURES

Figure Number — Page

2.1 Susceptible-Infectious-Recovered (SIR) model representation. Solid arrows show the movement from S to I to R. 12

3.1 Illustration of the data generating mechanism for a generic area. We let b index birth cohort and consider two birth cohorts — cohort $b = 1$ represented by the blue circles, and cohort $b = 2$ represented by the orange circles. Each circle indicates the underlying MCV1 coverage of the birth cohort at a specific time. The blue/orange shade represents the proportion of children covered by RI, i.e., the RI-specific MCV1 coverage; and the yellow shade represent the proportion covered by SIA. Because all children within the target age group can receive vaccination during an SIA regardless of whether they have been vaccinated before, some children in the birth cohorts might be covered by both RI and SIA. They are represented by the intersection of the blue/orange shade and the yellow shade. 29

3.2 Summary results of the simulation study, including bias, root mean square error (RMSE), the nominal coverage and the average widths of 95% posterior credible intervals (CIs) of the estimated RI-specific MCV1 coverage $p_{\text{RI},ib}$ across all areas and birth cohorts. 36

3.3 Design-based HT estimates of the overall MCV1 coverage across birth cohorts from each survey for Nigeria's 37 states. 38

3.4 The posterior medians and 95% CIs of the survey-specific effects ϵ_s. 43

3.5 The posterior medians of the ICAR spatial random effects α_i (left) and the IID spatial random effects γ_i (right) from the fitted *ICAR-AR1 Model*. . . . 44

3.6 The posterior medians and 95% CIs of the RW2 temporal random effects δ_b (left) and the IID temporal random effects τ_b (right) from the fitted *ICAR-AR1 Model*. 44

3.7 The posterior medians of the space-time interactions ϕ_{ib} across space over selected birth cohorts from the fitted *ICAR-AR1 Model*. Specifically, each map shows the estimated interaction for the cohorts born in the first 6 months of the corresponding year. 45

3.8 The posterior medians of the space-time interactions ϕ_{ib} over time in all states from the fitted *ICAR-AR1 Model*. 46

3.9 The posterior medians of the state-level RI-specific MCV1 coverage (dark grey lines) and the associated 95% CIs (light grey ribbons) for the birth cohorts born between 2000 and 2018 based on the fitted *ICAR-AR1 Model*. The design-based overall MCV1 coverage estimates were also shown for references. 47

3.10 Design-based HT estimates of the overall MCV1 coverage across birth cohorts from each survey for Nigeria's 37 states in the sensitivity analysis with 12-month birth cohort data. 51

3.11 The posterior medians of the state-level RI-specific MCV1 coverage (dark grey lines) and the associated 95% CIs (light grey ribbons) for the birth cohorts born between 2000 and 2018 based on the fitted *ICAR-AR1 Model*. 53

4.1 The simulated semi-monthly time series of total population (left) and adjusted births entering the susceptible population (right). 66

4.2 The simulated semi-monthly time series of underlying incidence, without (left) and with (right) a "planned" SIA at the beginning of the forecast period. . . 66

4.3 The simulated semi-monthly time series of underlying susceptible population, without (left) and with (right) a "planned" SIA at the beginning of the forecast period. 67

4.4 The point estimates and 95% confidence intervals of the reporting rate parameter ρ obtained by fitting OLS regression models with robust standard errors to the cumulative monthly adjusted births and the cumulative monthly reported incidence before the first SIA. The estimates are calculated by drawing 10000 random samples from the distribution $\mathrm{N}\left(\hat{\kappa}_{\mathrm{OLS}}, \hat{\sigma}^2_{\kappa}\right)$ and taking the inverse of the samples. When $\rho = 0.01$, the point estimate is 0.0102 and the 95% confidence interval is $(0.0095, 0.0111)$. When $\rho = 0.1$, the point estimate is 0.107 and the 95% confidence interval is $(0.099, 0.117)$. 71

4.5 The posterior medians and the 95% posterior credible intervals (CIs) of γ_1, γ_2, γ_3 and γ_4 for when the reporting rate is 0.01, 0.1, 0.3, 0.5 and 0.7 (along the x-axis), computed with (black) and without (blue) uncertainty propagation from the reporting rate estimation in the first stage. 72

4.6 The posterior medians and the 95% posterior credible intervals (CIs) of β^{EN}, ϕ, θ and p for when the reporting rate is 0.01, 0.1, 0.3, 0.5 and 0.7 (along the x-axis), computed with (black) and without (blue) uncertainty propagation from the reporting rate estimation in the first stage. 73

4.7 The simulated true values, posterior medians and 95% posterior CIs/predictive intervals of the underlying measles incidence computed with uncertainty propagation for when the reporting rate is 0.01, 0.1, 0.3, 0.5 and 0.7 (from top to bottom), without (left) and with (right) a "planned" SIA at the beginning of the forecast period, respectively. 74

4.8 The simulated true values, posterior medians and 95% posterior CIs/predictive intervals of the underlying measles incidence computed without uncertainty propagation (i.e., true ρ plugged in) for when the reporting rate is 0.01, 0.1, 0.3, 0.5 and 0.7 (from top to bottom), without (left) and with (right) a "planned" SIA at the beginning of the forecast period, respectively. 75

4.9 The simulated true values, posterior medians and 95% posterior CIs/predictive intervals of the underlying susceptible population computed with uncertainty propagation for when the reporting rate is 0.01, 0.1, 0.3, 0.5 and 0.7 (from top to bottom), without (left) and with (right) a "planned" SIA at the beginning of the forecast period, respectively. 76

4.10 The simulated true values, posterior medians and 95% posterior CIs/predictive intervals of the underlying susceptible population computed without uncertainty propagation (i.e., true ρ plugged in) for when the reporting rate is 0.01, 0.1, 0.3, 0.5 and 0.7 (from top to bottom), without (left) and with (right) a "planned" SIA at the beginning of the forecast period, respectively. 77

4.11 The root mean squared error of the underlying incidence in the forecast period, RMSE_f, calculated using the posterior medians of the semi-monthly incidence estimated with (black) and without (blue) uncertainty propagation from the reporting rate estimation in the first stage. 78

4.12 The monthly reported measles incidence in Benin between January 2012 and December 2018. 80

4.13 The posterior medians and the 95% posterior CIs/predictive intervals of (1) Top: the underlying monthly measles incidence scaled by the estimated reporting rate (with the true reported incidence time series); (2) Middle: the underlying semi-monthly true incidence; and (3) Bottom: the underlying semi-monthly susceptible population in Benin between 2012 and 2018. 82

5.1 Left: state-level Horvitz-Thompson direct estimates of MCV1 coverage among children aged 12–23 months. Right: the widths of the associated 90% confidence intervals. 90

5.2 Spatial distribution of the observed MCV1 coverage among children aged 12–23 months as recorded at the 2018 NDHS cluster level. The cluster-level observed MCV1 coverage is calculated as the proportion of children sampled in a survey cluster who have had at least one dose of MCV at the time of interview, based on evidence from either vaccination cards or caregiver recall. 91

5.3 The empirical semi-variogram for MCV1 coverage in Nigeria based on the posterior medians of the random effects C_{ic} in the non-spatial binomial mixed model (5.1). 92

5.4 Posterior medians of the spatial fields based on the *Binomial NN* (top left), *Beta-Binomial OD* (top right), *Lono-Binomial OD* (bottom left) and *Binomial TS* (bottom right) models that include the urban/rural strata variable. 103

5.5 Maps of the posterior medians (top row) and the widths of 90% credible intervals (botton row) for the estimated MCV1 coverage at the 1×1 km pixel (left), LGA (middle) and state (right) levels, based on the *Binomial NN* model that includes the urban/rural strata variable. 105

5.6 Maps of the posterior medians (top row) and the widths of 90% credible intervals (botton row) for the estimated MCV1 coverage at the 1×1 km pixel (left), LGA (middle) and state (right) levels, based on the *Beta-Binomial OD* model that includes the urban/rural strata variable. 106

5.7 Maps of the posterior medians (top row) and the widths of 90% credible intervals (botton row) for the estimated MCV1 coverage at the 1×1 km pixel (left), LGA (middle) and state (right) levels, based on the *Lono-Binomial OD* model that includes the urban/rural strata variable. 107

5.8 Maps of the posterior medians (top row) and the widths of 90% credible intervals (botton row) for the estimated MCV1 coverage at the 1×1 km pixel (left), LGA (middle) and state (right) levels, based on the *Binomial TS* model that includes the urban/rural strata variable. 108

5.9 Map of the posterior medians of the MCV1 coverage estimates for Nigeria's 37 states, with hatching density proportional to the posterior coefficient of variation (CV) of the vaccination odds. The results are based on 1000 posterior samples of the MCV1 coverage estimates from the *Lono-Binomial OD* model that includes the urban/rural strata variable. 110

5.10 The posterior distributions of the MCV1 coverage estimates for Nigeria's 37 states, arranged into a grid that preserved some of the geographical orientation, based on 1000 posterior samples from the *Lono-Binomial OD* model that includes the urban/rural strata variable. 112

5.11 Ridgeline plots of the posterior distributions of the MCV1 coverage estimates for Nigeria's 37 states, ordered by posterior median, based on 1000 posterior samples from the *Lono-Binomial OD* model that includes the urban/rural strata variable. 114

5.12 Histograms of the posterior ranking distributions of the 5 states with the lowest (left column) and 5 with the highest (right column) expected ranks (ERs), based on 1000 posterior samples of the MCV1 coverage estimates of Nigeria's 37 states from the *Lono-Binomial OD* model that includes the urban/rural strata variable. 116

5.13 Posterior ranking distributions of all 37 states in Nigeria with their expected ranks (ER) based on 1000 posterior samples of the MCV1 coverage estimates from the *Lono-Binomial OD* model that includes the urban/rural strata variable. 117

5.14 The range of the posterior rankings of all 37 states in Nigeria with their expected ranks (ERs) based on 1000 posterior samples of the MCV1 coverage estimates from the *Lono-Binomial OD* model that includes the urban/rural strata variable. The points indicate the ERs of the states, and the vertical line segments mark the complete range of the posterior rankings of the states. 118

5.15 Posterior ranking distributions of the 5 LGAs with the lowest (left column) and 5 with the highest (right column) expected ranks (ERs) based on 1000 posterior samples of the MCV1 coverage estimates of Nigeria's 774 LGAs from the *Lono-Binomial OD* model that includes the urban/rural strata variable. 119

5.16 The range of the posterior rankings of all 774 LGAs in Nigeria with their expected ranks (ER) based on 1000 posterior samples of the MCV1 coverage estimates from the *Lono-Binomial OD* model that includes the urban/rural strata variable. The points indicate the expected ranks (ER) of the LGAs, and the vertical line segments mark the complete range of the posterior rankings of the LGAs. 120

5.17 Maps of the posterior probabilities of LGA-level MCV1 coverage estimates exceeding the 20%, 50% and 80% thresholds respectively, based on the *Lono-Binomial OD* model that includes the urban/rural strata variable. 122

5.18 Top: Maps of MCV1 coverage estimates at the 1×1 km pixel, LGA and state levels using the discrete color scale composed of the *extremely low* $[0\%, 20\%)$, *low* $[20\%, 50\%)$, *medium* $[50\%, 80\%)$ and *high* $[80\%, 100\%]$ coverage intervals. Bottom: The corresponding histograms of the *true classification probabilities* (TCPs) with the *average true classification probability* (ATCP) highlighted by the blue vertical line. 124

5.19 Top: Maps of MCV1 coverage estimates at the 1×1 km pixel level using discrete color scales with $K = 2, 3$ and 4 quantile intervals. For each K, the color scale is formed by creating the intervals $[L_0, L_1), \ldots, [L_{K-1}, L_K]$, where $L_0 = 0\%$, $L_K = 100\%$, and L_k equals the $100 \times k/K$ quantile of the pooled posterior samples of the pixel-level coverage estimates based on the *Lono-Binomial OD* model that includes the urban/rural strata variable. Bottom: The corresponding histograms of the *true classification probabilities* (TCPs) of the pixels with the *average true classification probability* (ATCP) highlighted by the blue vertical line. 125

5.20 Maps of MCV1 coverage estimates at the LGA level (top row) and the state level (bottom row) using a discrete color scale with $K = 3, 4$ and 5 quantile intervals. The color scales are formed by creating the intervals $[L_0, L_1), \ldots, [L_{K-1}, L_K]$, where $L_0 = 0\%$, $L_K = 100\%$, and L_k equals the $100 \times k/K$ quantile of the pooled posterior samples of the coverage estimates at the LGA and state level respectively, based on the *Lono-Binomial OD* model that includes the urban/rural strata variable. 127

5.21 The average ture classification probabilities (ATCP) corresponding to the state, LGA and 1×1 km pixel maps with discrete color scales defined by various number of quantile intervals K. For each K within each spatial resolution (state, LGA or pixel), the quantile intervals are formed by creating the intervals $[L_0, L_1), \ldots, [L_{K-1}, L_K]$, where $L_0 = 0\%$, $L_K = 100\%$, and L_k equals the $100 \times k/K$ quantile of the pooled posterior samples of the coverage estimates at that spatial scale based on the *Lono-Binomial OD* model that includes the urban/rural strata variable. 128

A.1 The posterior medians and 95% CIs of the independent survey-specific effects ϵ_s in the sensitivity analysis with 12-month birth cohort data.. 150

A.2 The posterior medians of the ICAR spatial random effects α_i (left) and the IID random effects γ_i (right) from the fitted *ICAR-AR1 Model* in the sensitivity analysis with 12-month birth cohort data. 151

A.3 The posterior medians and 95% CIs of the RW2 temporal random effects δ_b (left) and the IID random effects τ_b (right) from the fitted *ICAR-AR1 Model* in the sensitivity analysis with 12-month birth cohort data. 151

A.4 The posterior medians of the space-time interactions ϕ_{ib} across space over selected 12-month birth cohorts from the fitted *ICAR-AR1 Model* in the sensitivity analysis. Specifically, each map shows the estimated interaction for the cohorts born in the corresponding year. 152

A.5 The posterior medians of the space-time interactions ϕ_{ib} over time in all states from the fitted *ICAR-AR1 Model* in the sensitivity analysis with 12-month birth cohort data. 153

A.6 The posterior medians of the state-level RI-specific MCV1 coverage (dark grey lines) and the associated 95% CIs (light grey ribbons) for the 12-month birth cohorts born between 2000 and 2018 based on the fitted *ICAR-AR1 Model* in the sensitivity analysis. The design-based overall MCV1 coverage estimates were also shown for references. 154

B.1 Traceplots of posterior samples of model parameters in the simulation study computed with uncertainty propagation for when the reporting rate $\rho = 0.01$. 155

B.2 Traceplots of posterior samples of model parameters in the simulation study computed with uncertainty propagation for when the reporting rate $\rho = 0.1$. 156

B.3 Traceplots of posterior samples of model parameters in the simulation study computed with uncertainty propagation for when the reporting rate $\rho = 0.3$. 157

B.4 Traceplots of posterior samples of model parameters in the simulation study computed with uncertainty propagation for when the reporting rate $\rho = 0.5$. 158

B.5 Traceplots of posterior samples of model parameters in the simulation study computed with uncertainty propagation for when the reporting rate $\rho = 0.7$. 159

B.6 Traceplots of posterior samples of model parameters in the simulation study computed without uncertainty propagation (i.e., true ρ plugged-in) where reporting rate $\rho = 0.01$. 160

B.7 Traceplots of posterior samples of model parameters in the simulation study computed without uncertainty propagation (i.e., true ρ plugged-in) where reporting rate $\rho = 0.1$. 161

B.8 Traceplots of posterior samples of model parameters in the simulation study computed without uncertainty propagation (i.e., true ρ plugged-in) where reporting rate $\rho = 0.3$. 162

B.9 Traceplots of posterior samples of model parameters in the simulation study computed without uncertainty propagation (i.e., true ρ plugged-in) where reporting rate $\rho = 0.5$. 163

B.10 Traceplots of posterior samples of model parameters in the simulation study computed without uncertainty propagation (i.e., true ρ plugged-in) where reporting rate $\rho = 0.7$. 164

B.11 Traceplots of posterior samples of model parameters in the Benin analysis. . 165

LIST OF TABLES

Table Number Page

3.1 Summary information of survey sampling design, including source of sampling frame, survey strata, cluster and household sample sizes, age group with vaccination data, and features of weight calculation. 25

3.2 The start and end dates, target age groups, and geographical areas of the SIAs in Nigeria we use for the analysis. There are 20 northern states and 17 southern states in Nigeria. 26

3.3 The space-time interaction assumptions of the models. 39

3.4 Summary of model assessment metrics and parameter estimates. Bold figures represent the "best" models according to the relevant criteria. The posterior medians are taken as the point estimates and the corresponding posterior 95% CIs are shown in the parentheses. 40

3.5 Proportions of variance explained by each random effect component (percent). 41

3.6 Summary of model assessment metrics and parameter estimates in the sensitivity analysis with 12-month birth cohort data. Bold figures represent the "best" models according to the relevant criteria. The posterior medians are taken as the point estimates and the corresponding posterior 95% CIs are shown in the parentheses. 52

4.1 The posterior medians and the 95% posterior CIs of model parameters from the Benin analysis. The estimates for ρ were obtained by fitting an OLS regression model with robust standard errors to the cumulative monthly adjusted births and the cumulative monthly reported incidence before December 2014. The estimates are calculated by drawing 10000 random samples from the distribution $\mathrm{N}\left(\hat{\kappa}_{\mathrm{OLS}}, \hat{\sigma}^2_{\kappa}\right)$ and taking the inverse of the samples. 81

5.1 Estimates of parameters and model validation results. Reported are the posterior medians and 2.5% and 97.5% quantiles of the regression coefficients and the parameters of the spatial field and cluster-level excess variation. The spatial range parameter ρ is on the longitude-latitude degree scale, which, given the geographical location of Nigeria, equals an average of 111 km per degree. The widely applicable information criterion (WAIC) is calculated based on all the data. The bias, mean absolute error (MAE) and root mean squared error (RMSE) are calculated based on the cross-validation exercise described in the supplementary materials. Bold figures represent the "best" models according to the relevant criteria. 101

5.2 Estimates of parameters for each model. Reported are the posterior medians and 2.5% and 97.5% quantiles of the regression coefficients for covariates. Note that the estimates for the coefficients of covariates are not directly comparable because the covariates are measured on different scales. 102

A.1 Proportions of variance explained by each random effect component (percent) in the sensitivity analysis with 12-month birth cohort data. 150

GLOSSARY

DHS: Demographic and Health Surveys

EA: Enumeration area

GMRF: Gaussian Markov random field

ICAR: Intrinsic conditional autoregressive

INLA: Integrated nested Laplace approximation

LGA: Local government area

LMIC: Low- and middle-income country

MCMC: Markov chain Monte Carlo

MCV: Measles-containing-vaccines

MICS: Multiple Indicator Cluster Surveys

NNHS: National Nutrition and Health Surveys

RI: Routine immunization

RW2: Second-order random walk

SIA: Supplementary immunization activity

SPDE: Stochastic partial differential equation

TSIR: Time-series susceptible-infected-recovered

WHO: World Health Organization

ACKNOWLEDGMENTS

First of all, I would like to thank my advisor, Jon Wakefield, for being so patient and encouraging throughout my graduate career. His passion for statistics and dedication to both work and life are inspiring. I have learned so much from him and I am truly grateful for his support and guidance.

I would also like to thank my committee members, Betz Halloran, Lurdes Inoue, Laina Mercer, and Kenny Sherr, for providing valuable insight into this dissertation and for being my role models. In additional, thank you to my research collaborators, Kevin McCarthy, Niket Thakkar, Kurt Frey, and Dale Rhoda, for sharing your knowledge in measles vaccination campaigns and household surveys.

From the University of Washington and Fred Hutch, I would like to thank my research assistantship advisors, Sharon Browning, Brian Browning, and Elizabeth Brown, for their funding support and mentorship. In addition, I want to thank Scott Emerson, Katie Kerr, Marco Carone, Adam Szpiro, Lyn Brumback, Lloyd Mancl, Patrick Heagerty, Tim Thornton, Ken Rice, Thomas Fleming, Jim Hughes, Amy Willis, Gitana Garofalo, Sandra Coke, and Stephanie Shadbolt, for supporting me as instructors, mentors, and administrators during the course of my graduate studies. I would also like to extend my gratitude to the members of the Space Time Analysis Bayes (STAB) working group, the 2015 biostatistics cohort, and other students in the department for their company and friendship.

Finally, I would like to thank my family and friends. Thank you to my parents, Hongchang Dong and Suben Chen, for always believing in me and giving me the freedom to make my own decisions. You are the greatest parents and I love you so much! Thank you to my husband, Scott Coggeshall, for being such a loving and supportive partner. Whenever I felt down and

discouraged, you were always there to cheer me up. I could not have done this without you! Last, but not least, thank you to my parents-in-law, Jack and Donna Coggeshall, and my sisters-in-law, Sarah Shipley and Abby Cardenaz, for making me feel so loved. I am so lucky to have you as my families!

DEDICATION

To my family and friends.

Chapter 1

INTRODUCTION

Immunization is one of the most successful and cost-effective health interventions for controlling and preventing life-threatening infectious diseases (World Health Organization, 2020a). Between 2010 and 2018, an estimated 23 million deaths were averted with measles vaccine alone (Patel et al., 2019). Despite the tremendous advances in immunization coverage in the past two decades, challenges remain in many low- and middle-income countries (LMICs), where disease burdens are the highest (World Health Organization, 2020b). The Immunization Agenda 2030 sets an ambitious global strategy "to leave no one behind", expanding the vision of the Global Vaccine Action Plan to "extend the full benefits of immunization to all people, regardless of where they are born, who they are or where they live" (World Health Organization, 2012, 2020b). Timely and comprehensive evaluation of immunization program performance, especially at the sub-national level, is essential for monitoring progress and planning future interventions (Strategic Advisory Group of Experts on Immunization, 2019). Unfortunately, administrative data in many LMICs is unreliable due to inadequate health information systems (Dunkle et al., 2014). Therefore, other data sources, such as disease surveillance systems and household surveys are turned to, to obtain valid estimates for policy guidance (Cutts et al., 2013).

This dissertation focuses on developing statistical methods for analyzing infectious disease incidence and complex survey data to facilitate assessing and planning immunization programs in LMICs. In particular, we consider estimation of vaccination coverage and immunization campaign effectiveness across space and time.

1.1 Motivating examples

1.1.1 Measles vaccination coverage in Nigeria

Measles is a vaccine-preventable, highly contagious respiratory infection that affects mostly young children (World Health Organization, 2019). According to a 2017 World Health Organization (WHO) report, an estimated 20.8 million children are still missing the first dose of measles-containing-vaccines (MCV), and Nigeria is the country with the most unvaccinated children (3.3 million) (World Health Organization, 2017b). In high-burden settings, a key strategy for increasing MCV coverage is to conduct supplementary immunization activities (SIAs) in the form of vaccination campaigns, in addition to delivering scheduled vaccination through routine immunization (RI) programs (World Health Organization, 2017a; The Measles and Rubella Initiative, 2012).

Tracking vaccination coverage in LMICs relies heavily on data from household surveys (Cutts et al., 2016; Dolan and MacNeil, 2017). In Nigeria, there are three major survey programs measuring vaccination rates: the Demographic and Health Surveys (DHS) supported by the United States Agency for International Development (ICF, 2020a), the Multiple Indicator Cluster Surveys (MICS) supported by the United Nations Children's Fund (UNICEF, 2020), and the National Nutrition and Health Surveys (NNHS) based on the Standardized Monitoring and Assessment of Relief and Transitions (SMART) method supported by the National Bureau of Statistics of Nigeria (National Bureau of Statistics, 2016; SMART Technical Advisory Group, 2006, 2017). These surveys use similar, but nonidentical stratified cluster sampling designs to provide national and sub-national coverage estimates for the 36 states and the Federal Capital Territory of Nigeria, i.e., the 37 administrative level 1 (admin-1) areas, and hereafter referred to as Nigeria's 37 states. Specifically, the DHS samples were stratified by urban and rural areas within each state, whereas MICS and NNHS were stratified only by states.

In Chapter 3, we develop a space-time smoothing model for estimating sub-national RI-specific coverage of the first dose of measles-containing-vaccines (MCV1) using complex

survey data. Sub-national RI-specific coverage estimates are crucial for identifying areas where the routine vaccine delivery systems are the weakest, and for building measles epidemiological models to estimate underlying susceptible population dynamics and forecast disease incidence trends. However, measuring RI-specific coverage through surveys in the presence of SIAs is challenging because the questionnaires do not differentiate the source of vaccination. Our method combines data from multiple surveys with potentially different sampling designs, and accounts for the impact of SIAs by incorporating information from the WHO SIA calendar (World Health Organization, 2020c). This approach is applied to four DHS, three MICS and two NNHS surveys conducted in Nigeria between 2003 and 2018 to generate reliable estimates and associated uncertainty intervals for the RI-specific MCV1 coverage at the state level.

1.1.2 Efficacy of measles vaccination campaigns in Benin

A common metric used for evaluating SIA effectiveness is the campaign coverage. It is calculated as the ratio of the number of MCV doses administered during a campaign to the size of the target population of the campaign. While SIAs usually have high reported campaign coverage, it is unclear how many people are effectively removed from the susceptible population (Mbabazi et al., 2009). A more relevant metric is SIA efficacy, defined as the fraction of the susceptible population immunized after a measles SIA. However, estimation of SIA efficacy has been a programmatic challenge due to difficulties in estimating the underlying susceptible population.

In Chapter 4, we develop a discrete-time hidden Markov model under the time-series susceptible-infected-recovered (TSIR) framework (Finkenstädt and Grenfell, 2000) to estimate SIA efficacy. Our approach accounts for under-reporting and seasonality of measles transmission, and accommodates monthly reported incidence data that are publicly available from the WHO Measles and Rubella Surveillance Database for most member countries (World Health Organization, 2020e). In particular, we apply our model to analyze the reported measles incidence in Benin between 2012 and 2018, as an illustrative example. We

estimate the underlying susceptible population dynamics, assess how many susceptible people were immunized by past SIAs, and forecast incidence trends in the future under various hypothetical SIA scenarios.

1.1.3 Vaccination coverage mapping with household survey data

It is becoming increasingly popular to produce high-resolution maps of vaccination coverage by fitting Bayesian geostatistical models to data from household surveys. The WorldPop project (WorldPop, 2020a) and the Institute for Health Metrics and Evaluation (IHME) (IHME, 2020a) are two major producers of vaccination coverage surfaces at fine spatial scales. Their estimates are used by researchers and policy makers from organizations across the globe (WorldPop, 2020b; IHME, 2020b).

In Chapter 5, we discuss a number of crucial choices with respect to two key aspects of the map production process adopted by WorldPop and IHME: the acknowledgement of the survey design in modeling, and the appropriate presentation of estimates and their uncertainties. Specifically, we consider the importance of accounting for survey stratification and cluster-level non-spatial excess variation in survey outcomes when fitting geostatistical models. We also discuss the trade-off between the geographical scale and precision of model-based estimates, and demonstrate visualization methods for mapping and ranking that emphasize the probabilistic interpretation of results. A novel approach to coverage map presentation is proposed to allow comparison and control of the overall map uncertainty level. We illustrate the various issues using the measles vaccination data from the 2018 Nigeria DHS survey.

1.2 Organization of dissertation

In Chapter 2, we provide a review of the statistical concepts that are heavily relied upon in this dissertation. We begin with a review of survey sampling, followed by a description of Gaussian Markov random field (GMRF) models. We give more details on intrinsic Gaussian Markov random field (IGMRF) models as well as the stochastic partial differential equation (SPDE) approach, which are commonly used for modeling temporal and spatial components

in Bayesian hierarchical models. Next, we briefly introduce the susceptible-infected-recovered (SIR) models and the time-series SIR (TSIR) framework for infectious disease modeling. We also discuss a number of Bayesian issues used in this dissertation, including penalized complexity (PC) priors, Markov chain Monte Carlo (MCMC) and the integrated nested Laplace approximation (INLA). Last, we describe three measures of model fit for model assessment and comparison.

The next three chapters detail the core methodological contributions of this dissertation. In Chapter 3, we develop a space-time smoothing model for estimating sub-national RI-specific MCV1 coverage using complex survey data. In Chapter 4, we extend the TSIR framework and develop a discrete time-series model for estimating SIA efficacy using monthly measles incidence data. In Chapter 5, we focus on vaccination coverage mapping via Bayesian geostatistical modeling of household survey data, and describe methods for acknowledging the survey design in modeling and presenting estimates and their uncertainties appropriately. Chapter 6 contains concluding remarks.

Chapter 2

BACKGROUND

In this chapter, we describe some of the statistical and computational tools that we will use throughout this dissertation.

2.1 Survey sampling

We focus on binary outcomes and let Y_{ik} be the indicator for the outcome of interest associated to individual $k = 1, \ldots, N_i$ in area i, $i = 1, \ldots, I$. For a given area i, a common characteristic of interest, especially in the context of vaccination coverage, is the true proportion of cases $P_i = \frac{Y_i}{N_i}$, where $Y_i = \sum_{k=1}^{N_i} Y_{ik}$ is the total number of cases in the population. To obtain an estimate for this quantity, a survey is conducted to sample a subset of individuals from the total population. Let I_{ik} denote the indicator for whether individual k in area i is sampled in the survey, and $\pi_{ik} = \Pr(I_{ik} = 1)$ denote the probability of selection. Then, the sample size in area i is $n_i = \sum_{k=1}^{N_i} I_{ik}$, and the design weight associated with individual k in area i is $w_{ik} = \pi_{ik}^{-1}$.

A widely-used estimator of the population proportion P_i is the Horvitz-Thompson estimator (Horvitz and Thompson, 1952), denoted by $\hat{P}_i^{HT}$ and defined by

$$\hat{P}_i^{HT} = \frac{1}{N_i} \sum_{k=1}^{N_i} \frac{Y_{ik} I_{ik}}{\pi_{ik}} = \frac{1}{N_i} \sum_{k=1}^{n_i} \frac{y_{ik}}{\pi_{ik}}, \tag{2.1}$$

where y_{ik} is the observed outcome from sample k in area i. The estimated design variance associated with $\hat{P}_i^{HT}$ is

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

Population totals N_i are often unknown in household surveys, as the design weights are commonly normalized during weight calculation. In such cases and in general, we can use the Hájek estimator (Hájek, 1971) for population proportion estimation

$$\hat{P}_i = \frac{\sum_{k=1}^{n_i} w_{ik} y_{ik}}{\sum_{k=1}^{n_i} w_{ik}}. \tag{2.3}$$

The estimated design variance of the Hájek estimator can be obtained via various approximations (Lohr, 2009).

2.2 Gaussian Markov random fields (GMRF)

Let $\boldsymbol{x} = [x_1, \ldots, x_n]^\intercal$ be an $n-$dimensional random vector following a multivariate Gaussian distribution with $\boldsymbol{x} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. $\boldsymbol{x}$ is a Gaussian Markov random field (GMRF) with mean $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$ if and only if its density $\pi(\boldsymbol{x})$ can be written in the following form

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

and for all $i \neq j$,

$$\boldsymbol{Q}_{ij} = 0 \text{ if and only if } x_i \perp\!\!\!\perp x_j \mid \boldsymbol{x}_{-\{i,j\}}, \tag{2.4}$$

where $\boldsymbol{x}_{-\{i,j\}}$ denotes all elements of $\boldsymbol{x}$ except x_i and x_j. The Markov property of GMRF is apparent in the *conditional independence* condition described in (2.4). It implies that given $\boldsymbol{x}_{-\{i,j\}}$, the value of x_i is independent of the value of x_j.

The conditional mean and precision of an element x_i of a GMRF $\boldsymbol{x}$ are

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

$$\mathrm{Prec}\left[x_i | \boldsymbol{x}_{-i}\right] = Q_{ii}$$

where $j \sim i$ means the jth element is a neighbor of the ith element, i.e., $Q_{ij} \neq 0$.

A GMRF $\boldsymbol{x} = [x_1, \ldots, x_n]^\intercal$ is an intrinsic Gaussian Markov random field (IGMRF) with mean $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{Q}$ if and only if its density $\pi(\boldsymbol{x})$ can be written in the following form

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

where $\boldsymbol{Q}$ has rank $n-k$ with $k \geq 1$, and $|\cdot|^*$ denotes a generalized determinant. An IGMRF of kth order is an improper GMRF of rank $n-k$ and its density is invariant to the addition of any constant. Assuming $\boldsymbol{\mu} = \boldsymbol{0}$, the conditional mean and precision of an element x_i of an IGMRF $\boldsymbol{x}$ are

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

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

2.2.1 Second-order random walk (RW2) model

The second-order random walk (RW2) model of regular locations is an IGMRF of second order with density

$$\pi(\boldsymbol{x}|\kappa) \propto \kappa^{(n-2)/2} \exp\left(-\frac{1}{2}\boldsymbol{x}^\intercal \boldsymbol{Q}\boldsymbol{x}\right),$$

where κ is the precision parameter, and $\boldsymbol{Q} = \kappa \boldsymbol{R}$ is the precision matrix with a structure matrix $\boldsymbol{R}$ of the following form:

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

The RW2 model is commonly used for temporal random effects in Bayesian hierarchical models. It is constructed assuming independent second-order increments

$$\triangle^2 x_i = (x_{i+2} - x_{i+1}) - (x_{i+1} - x_i)|\kappa \sim \mathrm{N}(0, \kappa^{-1}).$$

2.2.2 Intrinsic conditional autoregressive (ICAR) model

The intrinsic conditional autoregressive (ICAR) model is an IGMRF with the density

$$\pi(\boldsymbol{x}|\kappa) \propto \kappa^{(rank(\boldsymbol{Q})-1)/2} \exp\left(-\frac{1}{2}\boldsymbol{x}^\intercal \boldsymbol{Q}\boldsymbol{x}\right),$$

where κ is the precision parameter, and $\boldsymbol{Q} = \kappa\boldsymbol{R}$ is the precision matrix with a structure matrix $\boldsymbol{R}$ with the following elements:

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

where m_i is the number of neighbors of the ith element, and $j \sim i$ means the jth element is a neighbor of the ith element.

The ICAR model is commonly used for spatial random effects in Bayesian hierarchical models. It can be specified conditionally via

$$x_i|\boldsymbol{x}_{-i}, \kappa \sim \mathrm{N}\left(\frac{1}{m_i}\sum_{j:j\sim i} x_j, \frac{1}{m_i\kappa}\right).$$

It should be noted that the marginal variance of an RW2 or ICAR model depends on both the dimension and the rank of its structure matrix $\boldsymbol{R}$. Following Sørbye and Rue (2014), we can scale an $n-$dimensional RW2 or ICAR model by obtaining the generalized inverse of $\boldsymbol{R}$ and setting the generalized variance of $\boldsymbol{x}$ to be 1. Specifically, let $\boldsymbol{\Sigma}^* = \boldsymbol{R}^{-1}$ and $\sigma^2(x_i) = \Sigma^*_{ii}$, we calculate

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

The scaled model is then defined to have precision $\kappa\boldsymbol{R}_{GV}$ where $\boldsymbol{R}_{GV} = \sigma^2_{GV}\boldsymbol{R}$.

2.2.3 Stochastic partial differential equation (SPDE) approach

Rue and Tjelmeland (2002) first proposed to use GMRFs to approximate continuous Gaussian field (GF) and presented good "proof-of-concept" results for a number of common covariance functions. However, their approach requires all fits to be pre-computed on a regular grid only, hence is less desirable for practical purposes. Lindgren et al. (2011) proposed a new approach and considered a stochastic partial differential equation (SPDE) whose solution is a continuous Gaussian random field (GRF) with Matérn correlation. The Finite Element Method (FEM) was then used to build a GMRF representation of the GRF, which can be computed explicitly. The GMRF representation provides a sparse representation of the spatial effect through a sparse precision matrix, and hence enables many favorable computational properties of the GMRFs. We now briefly describe the SPDE approach proposed by Lindgren et al. (2011) with a focus on the case of 2-dimensional spatial field.

Consider a GRF $S(\boldsymbol{s})$ with the Matérn covariance function within a domain $\mathcal{D} \in \mathbb{R}^2$. That is, for any collection of points $(\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n)^\intercal$,

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

where $S_i = S(\boldsymbol{s}_i)$, $\mu_i = \mu(\boldsymbol{s}_i)$ for some mean function $\mu(\cdot)$, and $\Sigma_{ij} = C(S(\boldsymbol{s}_i), S(\boldsymbol{s}_j))$ for the Matérn covariance function $C(\cdot,\cdot)$ with a scale parameter $\kappa > 0$ and a smoothness parameter $\nu > 0$:

$$C(S(\boldsymbol{s}_i), S(\boldsymbol{s}_j)) = \sigma^2 \cdot \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\kappa||\boldsymbol{s}_i - \boldsymbol{s}_j||\right)^\nu K_\nu \left(\kappa||\boldsymbol{s}_i - \boldsymbol{s}_j||\right),$$

with $||\cdot||$ denoting the Euclidean distance in $\mathbb{R}^2$ and K_ν being the modified Bessel function of the second kind. Now, consider the SPDE

$$(\kappa^2 - \triangle)^{(\nu+1)/2} S(\boldsymbol{s}) = \lambda W(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathbb{R}^2$$

where $\triangle = (\partial^2/\partial s_1^2) + (\partial^2/\partial s_s^2)$ is the Laplacian operator on $\mathbb{R}^2$ and $W(\boldsymbol{s})$ is a spatial white noise Gaussian stochastic process with unit variance. Lindgren et al. (2011) show that the

solution to the SPDE is a GRF with Matérn covariance:

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

Now, a representation of the solution to the SPDE is constructed over a triangulation of the domain (i.e., a mesh) using FEM. Specifically, the solution is approximated by a weighted sum of basis functions

$$S(\boldsymbol{s}) \approx \sum_{k=1}^{m} \phi_k(\boldsymbol{s}) w_k$$

where ϕ_k are basis functions, w_k Gaussian distributed weights, and m is the number of vertices in the triangulation. In particular, the weights $\boldsymbol{w} = [w_1, w_2, \ldots, w_m]^\intercal$ are jointly Gaussian with mean $\mathbf{0}$ and precision matrix $\boldsymbol{Q}$. It can be shown that the sparse structure of $\boldsymbol{Q}$ can be preserved by carefully choosing the basis functions, and the joint distribution for the weights is a GMRF which determines the full distribution in the continuous domain. More information can be found in Bakka (2018). As such, the spatial process $S(\boldsymbol{s})$ can be approximated at any point inside the triangulated domain.

2.3 Models for infectious diseases

2.3.1 The susceptible-infected-recovered (SIR) model

Historically, infectious disease data were analyzed using deterministic models based on differential equations (Anderson et al., 1992). Models are set up based on a set of compartments in which individuals undergo homogeneous mixing.

A widely used compartmental model is the susceptible-infected-recovered (SIR) model, depicted in Figure 2.1. This approach is typically used when the number of disease counts is large, and the integer numbers in the constituent S, I and R compartments are taken to be continuous. Let $x(t)$, $y(t)$, $z(t)$ be the number of *susceptibles*, *infected*, *recovered* individuals at time t in a *closed population* of size N. The *hazard rate (force of infection)* is

$$\underbrace{\lambda^\dagger(t)}_{\text{Hazard}} = \underbrace{c(N)}_{\text{Contact Rate}} \times \underbrace{\frac{y(t)}{N}}_{\text{Prevalence}} \times \underbrace{p_I}_{\text{Infection Prob}} .$$

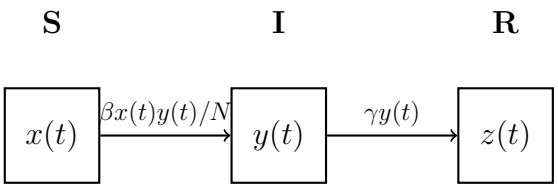

Figure 2.1: Susceptible-Infectious-Recovered (SIR) model representation. Solid arrows show the movement from S to I to R.

Two common forms for the contact rate (Begon et al., 2002) are

$$c(N) = \begin{cases} c_{\text{FD}} & \text{Frequency Dependent,} \\ Nc_{\text{DD}} & \text{Density Dependent.} \end{cases}$$

The *frequency-dependent* model is often used, particularly for childhood infections when the most relevant contact group is the classroom, whose size will be of the same order, regardless of the population size. Under *frequency dependency*,

$$\lambda^{\dagger}(t) = \beta y(t)/N,$$

where $\beta = c_{\text{FD}} \times p_{\text{I}}$.

The deterministic SIR model is defined through classic *mass-action* (Anderson et al., 1992). With *frequency dependent* transmission, we have the following set of ordinary differential equations:

$$\begin{aligned} \frac{dx(t)}{dt} &= -\frac{\beta x(t)y(t)}{N}, \\ \frac{dy(t)}{dt} &= \frac{\beta x(t)y(t)}{N} - \gamma y(t), \\ \frac{dz(t)}{dt} &= \gamma y(t), \end{aligned}$$

with *infection rate* β and *recovery rate* γ.

The discrete-time stochastic SIR model treats the current disease counts as a function of previous counts on a regular time scale. One may choose the time scale to be the transmission

dynamics time (i.e., latency plus infectious periods) or the generation time (i.e., time from infection of a primary case to infection of a secondary case infected by the primary case). For example, often, but not always, 2 weeks is used for measles.

2.3.2 The time-series SIR (TSIR) framework

The time-series SIR (TSIR) model is a discrete-time stochastic SIR model first described in Finkenstädt and Grenfell (2000). It was extended in a series of subsequent papers (Bjørnstad et al., 2002; Grenfell et al., 2002; Glass et al., 2003; Morton and Finkenstädt, 2005) and has been used to understand measles and rubella transmission in a variety of settings (Ferrari et al., 2008; Metcalf et al., 2011; Mahmud et al., 2017; Metcalf et al., 2013). Under this framework, the infected individuals are assumed to be infectious for one time unit before becoming removed. Therefore, we lose the recovery rate parameter and incidence is assumed equal to prevalence.

A simple model for the susceptibles at time t, denoted X_t, in the context of measles without vaccination is

$$X_t = X_{t-1} - Y_t + B_{t-d}$$

where Y_t is the number of infected individuals, and B_{t-d} is the number of births d time units previously, with d chosen to be the number of time units for which maternally derived immunity lasts.

The TSIR framework usually uses a negative binomial model for the number of infected individuals. It arises as the distribution of the population size in a linear birth process (Feller, 1950; Cox and Miller, 1977). Details of this derivation can be found in Wakefield et al. (2019). A classic form of the model for the number of infected individuals Y_t is

$$Y_t | Y_{t-1} = y_{t-1}, X_{t-1} = x_{t-1} \sim \text{NegBin}\left(\mu_t, y_{t-1}\right)$$

with

$$\mu_t = \frac{\beta y_{t-1}^{\alpha} x_{t-1}}{N},$$

where β is the infection rate parameter. The power parameter α is included to allow for deviations from mass action and to account for the discrete-time approximation to the continuous time model (Glass et al., 2003).

2.4 Bayesian analysis and computation

2.4.1 Bayesian hierarchical modeling

Bayesian hierarchical models in their simplest form comprise three levels:

$$\begin{aligned} &\text{The data model:} && \boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\phi} \sim p\left(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\phi}\right) \\ &\text{The latent model:} && \boldsymbol{\theta}|\boldsymbol{\phi} \sim p\left(\boldsymbol{\theta}|\boldsymbol{\phi}\right) \\ &\text{The hyperparameter model:} && \boldsymbol{\phi} \sim p\left(\boldsymbol{\phi}\right) \end{aligned}$$

where $\boldsymbol{y}$ denotes the data, $\boldsymbol{\theta}$ denotes the latent parameters, and $\boldsymbol{\phi}$ denotes the hyperparameters. The joint density of all model quantities takes the following form:

$$p\left(\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\phi}\right) = p\left(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\phi}\right) p\left(\boldsymbol{\theta}|\boldsymbol{\phi}\right) p\left(\boldsymbol{\phi}\right)$$

By Bayes rule, the posterior distribution of the latent parameters and the hyperparameters is

$$p\left(\boldsymbol{\theta},\boldsymbol{\phi}|\boldsymbol{y}\right) = \frac{p\left(\boldsymbol{y},\boldsymbol{\theta},\boldsymbol{\phi}\right)}{p\left(\boldsymbol{y}\right)} = \frac{p\left(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{\phi}\right) p\left(\boldsymbol{\theta}|\boldsymbol{\phi}\right) p\left(\boldsymbol{\phi}\right)}{p\left(\boldsymbol{y}\right)}$$

This represents the distribution of the unknown parameters given the observed data, which is usually the target of inference in Bayesian analysis.

2.4.2 Penalized complexity (PC) priors

Simpson et al. (2017) described a new approach for constructing prior distributions in nested models called penalized complexity (PC) priors. A key principle of the PC prior approach is Occam's razor: each prior should favor a simpler model, called the *base model*, unless the data indicate otherwise. Users can determine the level of shrinkage towards the *base model*

by specifying an intuitive statement about the parameter in terms of its median or other quantiles. For example, a PC prior for the standard deviation σ of a latent effect can be set by defining parameters (u, α) so that

$$\text{Prob}\,(\sigma > u) = \alpha,\ u > 0,\ 0 < \alpha < 1.$$

In the context of this dissertation, PC priors are mainly used for precision parameters in Gaussian random effect models. Consider $\boldsymbol{x} \sim \text{N}\,(\boldsymbol{0}, \boldsymbol{Q}^{-1})$, where $\boldsymbol{Q} = \kappa \boldsymbol{R}$ is the precision matrix. The simplest *base model* would be the absence of this random effect with $\kappa^{-1} = 0$. To derive a prior for κ, one can characterize the 'distance" between the base model and a more complex model using a function of the Kullback-Leibler (KL) distance. Using a constant rate penalization, we will have an exponential prior for this distance and consequently a PC prior for κ in the form

$$p(\kappa) = \frac{\lambda}{2}\kappa^{-3/2}\exp(-\lambda\kappa^{-1/2}).$$

If we specify the penalty parameter λ as $\lambda = -\frac{\log(\alpha)}{u}$, we will obtain an equivalent form of this prior as

$$\text{Prob}\left(\kappa^{-1/2} > u\right) = \alpha,$$

i.e., the probability of the standard deviation $\sigma = \kappa^{-1/2}$ being greater than u equals α.

2.4.3 Markov chain Monte Carlo (MCMC)

Markov chain Monte Carlo (MCMC) is a common Bayesian computing approach for estimating posterior distributions. The basic idea is that a Markov chain whose stationary distribution is the posterior distribution $p\,(\boldsymbol{\theta}|\boldsymbol{y})$ can be constructed so that inference for each unknown parameter can be obtained by drawing posterior samples from the converged Markov chain.

The Metropolis-Hastings algorithm is a widely-applicable MCMC method for obtaining random samples from posterior distributions. The algorithm proceeds as follows:

1. Choose an initial value, $\boldsymbol{\theta}^{(0)}$

2. For iteration $m = 1$, sample $\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(m-1)}$ from a proposal distribution $q_\sigma(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(m-1)})$, where σ is a parameter controlling the spread of the proposal distribution.

3. Compute the acceptance probability r
$$r = \min\left\{\frac{p(\boldsymbol{\theta}^*|\boldsymbol{y})/q_\sigma(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(m-1)})}{p(\boldsymbol{\theta}^{(m-1)}|\boldsymbol{y})/q_\sigma(\boldsymbol{\theta}^{(m-1)}|\boldsymbol{\theta}^*)}, 1\right\}$$

4. Generate random sample u from Unif$(0, 1)$.

5. Set the value for the current iteration
$$\boldsymbol{\theta}^{(m)} = \begin{cases} \boldsymbol{\theta}^* & \text{if } u < r, \\ \boldsymbol{\theta}^{(m-1)} & \text{otherwise.} \end{cases}$$

6. Repeat steps 2 through 5 for M iterations.

Before the Markov chain eventually converges to the stationary distribution, an initial number of samples will be discarded as *burn-in*. In addition, the σ parameter is typically selected such that the acceptance rate across all iterations is about 20–30% (Roberts and Rosenthal, 2009).

Another common MCMC algorithm is the Gibbs sampler, where the posterior distribution of a parameter is known and is easy to sample from directly. It is essentially a special case of the Metropolis-Hastings algorithm and is especially useful for parameters with conjugate priors. The two algorithms can be easily combined to form a hybrid approach, and that is what we will use in Chapter 4 for inference.

2.4.4 Integrated nested Laplace approximation (INLA)

When the posterior is high-dimensional (e.g., there are many latent parameters in a Bayesian hierarchical model) or the parameters in the posterior are correlated (e.g., the latent parameters are spatially and/or temporally correlated), the MCMC algorithm can be computationally expensive because convergence is slow and so a large number of MCMC samples are

required to explore the posterior distribution. The integrated nested Laplace approximation (INLA) approach proposed by Rue et al. (2009) provides a fast alternative to MCMC when the latent effects $p(\boldsymbol{\theta}|\boldsymbol{\phi})$ is a GMRF. Specifically, the INLA approach provides an accurate approximation to the posterior marginal distributions using a combination of Laplace approximations and numerical integration, such that the posterior marginals

$$p(\theta_i|\boldsymbol{y}) = \int p(\theta_i|\boldsymbol{\phi}, \boldsymbol{y})\, p(\boldsymbol{\phi}|\boldsymbol{y})\, d\boldsymbol{\phi}$$
$$p(\phi_i|\boldsymbol{y}) = \int p(\boldsymbol{\phi}|\boldsymbol{y})\, d\boldsymbol{\phi}_{-i}$$

are approximated by

$$\tilde{p}(\theta_i|\boldsymbol{y}) = \int \tilde{p}(\theta_i|\boldsymbol{\phi}, \boldsymbol{y})\, \tilde{p}(\boldsymbol{\phi}|\boldsymbol{y})\, d\boldsymbol{\phi}$$
$$\tilde{p}(\phi_i|\boldsymbol{y}) = \int \tilde{p}(\boldsymbol{\phi}|\boldsymbol{y})\, d\boldsymbol{\phi}_{-i}$$

where

$$\tilde{p}(\boldsymbol{\phi}|\boldsymbol{y}) \propto \frac{p(\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\phi})}{p_G(\boldsymbol{\theta}|\boldsymbol{\phi}, \boldsymbol{y})} \text{ evaluated at } \boldsymbol{\theta} = \boldsymbol{\theta}^*(\boldsymbol{\phi})$$
$$\tilde{p}(\theta_i|\boldsymbol{\phi}, \boldsymbol{y}) = \sum_g \tilde{p}(\theta_i|\boldsymbol{\phi}_g, \boldsymbol{y})\, \tilde{p}(\boldsymbol{\phi}_g|\boldsymbol{y})\, w_g.$$

with $p_G(\boldsymbol{\theta}|\boldsymbol{\phi}, \boldsymbol{y})$ denoting the Gaussian approximation of $p(\boldsymbol{\theta}|\boldsymbol{\phi}, \boldsymbol{y})$, $\boldsymbol{\theta}^*(\boldsymbol{\phi})$ denoting the model of $p(\boldsymbol{\theta}|\boldsymbol{\phi}, \boldsymbol{y})$ for a given $\boldsymbol{\phi}$, $\boldsymbol{\phi}_g$ denoting a grid of hyperparameter $\boldsymbol{\phi}$ and w_g denoting the associated weights.

2.5 Model assessment metrics

We briefly describe three model assessment metrics that allow for comparison across a variety of fitted models: the deviance information criteria (DIC), the Watanabe-Akaike information criteria (WAIC) and the log-score conditional predictive ordinate (LCPO).

2.5.1 The deviance information criteria (DIC)

Let $\boldsymbol{\theta}$ be a vector of parameters and $\boldsymbol{y}$ be the observed data, so that $p(\boldsymbol{y}|\boldsymbol{\theta})$ is the data model, $p(\boldsymbol{\theta})$ is the prior and $p(\boldsymbol{\theta}|\boldsymbol{y})$ is the posterior. The *deviance*, which provides a measure

of model "mis-fit", is defined as

$$D = -2 \log p(\boldsymbol{y}|\boldsymbol{\theta}).$$

Let $\bar{\boldsymbol{\theta}} = \mathbb{E}[\boldsymbol{\theta}|\boldsymbol{y}]$ be the posterior mean. The *effective number of parameters*, which provides a measure of model complexity, is defined as

$$p_{DIC} = \mathbb{E}_{\boldsymbol{\theta}}[D|\boldsymbol{y}] - D(\bar{\boldsymbol{\theta}}),$$

where $\mathbb{E}_{\boldsymbol{\theta}}[D|\boldsymbol{y}]$ represents the posterior mean of the *deviance*, and $D(\bar{\boldsymbol{\theta}})$ represents the *deviance* evaluated at the posterior mean $\bar{\boldsymbol{\theta}}$. The DIC is then defined as

$$\begin{aligned} DIC &= p_{DIC} + \mathbb{E}_{\boldsymbol{\theta}}[D|\boldsymbol{y}] \\ &= 2\mathbb{E}_{\boldsymbol{\theta}}[D|\boldsymbol{y}] - D(\bar{\boldsymbol{\theta}}) \\ &= 2p_{DIC} + D(\bar{\boldsymbol{\theta}}) \end{aligned}$$

Therefore, a lower DIC value generally indicates a better model fit.

The DIC can be viewed as a hierarchical modeling generalization of the Akaike information criterion (AIC) and Bayesian information criterion (BIC). It is based on asymptotic arguments and may under-penalize complex models with many random effects (Spiegelhalter et al., 2014).

2.5.2 The Watanabe-Akaike information criteria (WAIC)

The WAIC, also known as the *widely applicable information criterion*, was first described in Watanabe (2010) and has been receiving increasing popularity. It is a fully Bayesian approach for estimating the out-of-sample expectation, starting with the computed *log point-wise posterior predictive density* and then adding a correction for *effective number of parameters* to adjust for overfitting (Gelman et al., 2014). Specifically, the *log point-wise posterior predictive density* is defined as

$$\sum_i \log p(y_i|\boldsymbol{y}) = \sum_i \log \int p(y_i|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta},$$

and the *effective number of parameters* is defined as

$$p_{WAIC} = \sum_i \log \mathbb{E}\left[p(y_i|\boldsymbol{\theta})|\boldsymbol{y}\right] - \mathbb{E}\left[\log p(y_i|\boldsymbol{\theta})|\boldsymbol{y}\right].$$

The WAIC is then defined as

$$WAIC = 2p_{WAIC} - 2\sum_i \log p(y_i|\boldsymbol{y})$$

Therefore, a lower WAIC value generally indicates a better model fit.

Compared to the DIC, the WAIC utilizes the whole posterior distribution, instead of just the posterior mean. It can be computed using posterior samples.

2.5.3 The log-score conditional predictive ordinate (LCPO)

The conditional predictive ordinate (CPO) is a leave-one-out cross-validation score. It is defined as

$$\text{CPO}_i = p(y_i|\boldsymbol{y}_{-i}),$$

where $\boldsymbol{y}_{-i}$ denotes the observations $\boldsymbol{y}$ with the ith component removed. It represents the posterior probability of observing y_i when the model is fitted to all other data. Therefore, a higher value of CPO generally indicates a better fit of the model to y_i, and a low value suggest that y_i is an outlier or influential observation.

The log-score CPO (LCPO) is defined as

$$\text{LCPO} = -\frac{1}{n}\sum_i^n \log \text{CPO}_i$$

The negative sign is added here so that a smaller value of the LCPO indicates a better prediction quality of the model, which is consistent with DIC and WAIC.

Chapter 3

SPACE-TIME SMOOTHING MODELS FOR ROUTINE IMMUNIZATION COVERAGE ESTIMATION WITH COMPLEX SURVEY DATA

3.1 Introduction

Measles is a highly contagious viral disease. Before the introduction of measles vaccine in 1963 and widespread vaccination, major epidemics occurred approximately every 2 to 3 years and measles caused an estimated 2.6 million deaths each year (World Health Organization, 2019). Accelerated immunization activities have had a major impact on reducing global measles deaths. Between 2010 and 2018, measles vaccination prevented an estimated 23.2 million deaths (Patel et al., 2019). However, disease burden is still high in many low- and middle-income countries (LMICs), particularly in parts of Africa and Asia, where vaccination remains a challenge due to poor health care infrastructure and access (World Health Organization, 2017a). A key public health strategy in such high-burden settings is the combination of routine immunization (RI) and supplementary immunization activities (SIAs) (The Measles and Rubella Initiative, 2012). Although the World Health Organization (WHO) recommends two doses of measles-containing-vaccine (MCV) for all children — first at the age of 9 months and second at the age of 15 months, most LMICs only administer one dose in their RI schedule (World Health Organization, 2017a). In order to increase measles vaccination coverage, SIAs are carried out in the form of mass immunization campaigns every 2–4 years. During these campaigns, health workers run fixed-post vaccination sites and provide MCV to all children in a pre-specified target age group, regardless of whether they have been vaccinated previously. The aim is to immunize children missed by RI and reduce the susceptible population.

To achieve balanced implementations of RI and SIAs against measles, robust measurement of sub-national RI-specific coverage is crucial. Program officers rely on accurate estimation of local RI coverage to identify areas where the routine vaccine delivery systems are the weakest (Biellik and Orenstein, 2018; The Measles and Rubella Initiative, 2019). In addition, RI-specific MCV coverage is an essential input to measles epidemiological models, which estimate underlying susceptible population dynamics and inform optimal SIA schedules (Verguet et al., 2015; Thakkar et al., 2019; Dong et al., 2020b).

Despite ongoing efforts to strengthen administrative data quality (BID Initiative, 2016), the health management information systems in many LMICs are too weak to provide data of adequate quality for assessing and guiding health programs (Hancioglu and Arnold, 2013). Therefore, household surveys are the primary data sources for vaccination coverage estimates (Cutts et al., 2016; Dolan and MacNeil, 2017). The Demographic and Health Surveys (DHS) (ICF, 2020a) and the Multiple Indicator Cluster Surveys (MICS) (UNICEF, 2020) are two major survey programs that report sub-national measles vaccination coverage estimates. In addition, geospatial modeling approaches have been applied to the household survey data to generate MCV coverage estimates at various geographical scales (Utazi et al., 2018, 2019a,b, 2020). However, all the above-mentioned estimates are not RI-specific — they reflect the overall coverage of measles vaccination received via RI, SIAs or both sources. These coverage estimates can be influenced by past vaccination campaigns, and hence are not suitable for the aforementioned RI-specific purposes. For example, using the overall MCV coverage in epidemiological models may lead to under-estimation of the increase of susceptible pool, resulting in inaccurate estimation of measles transmission dynamics. To obtain RI-specific coverage estimates, the current method relies on only using data from birth cohorts that have not had any SIA opportunities yet. This restriction may result in some or all data from a household survey not being eligible for analysis — an inefficient use of already sparse data.

In this chapter, we propose a space-time smoothing model for estimating sub-national RI-specific coverage of first dose of MCV (MCV1) using data from complex surveys. Our model is inspired by the small area estimation methods developed by Mercer et al. (2015). The

proposed approach is able to combine data from multiple surveys with different sampling designs and construct sub-national coverage estimates with uncertainty that reflects the various data collection schemes. Our method allows more efficient use of data by modeling survey outcome not just from children who only had RI, but also from those who had RI and an SIA opportunity. We account for the impact of SIAs by incorporating information from the WHO SIA calendar (World Health Organization, 2020c). As a motivating example, we consider 8 household surveys conducted in Nigeria between 2003 and 2018 and generate state-level estimates of RI-specific coverage of MCV1 over time with associated uncertainty intervals.

The chapter is structured as follows. In Section 3.2, we describe the Nigeria data upon which estimation will be based. The details of our proposed method are provided in Section 3.3. A simulation study in Section 3.4 considers a number of scenarios with surveys of various sizes and demonstrates how our method improves estimation through more efficient use of data. The results of our modeling efforts of RI-specific MCV1 coverage in Nigeria from 2003 to 2018 are presented in Section 3.5. Section 3.6 contains concluding remarks.

3.2 Motivating data

We focus on the state-level RI-specific MCV1 coverage in Nigeria using data from four DHS surveys conducted in 2003, 2008, 2013 and 2018, three MICS surveys conducted in 2007, 2011 and 2016–17[1], and two National Nutrition and Health Surveys (NNHS) conducted in 2014 and 2015. In addition, we obtain the SIA calendar from the WHO website, which contains the start and end dates, target age groups and geographic areas of all SIAs implemented in Nigeria. More details about the data are provided below.

[1]The 2016–17 MICS is combined with the National Immunization Coverage Survey (NICS).

3.2.1 Household surveys

The DHS, MICS and NNHS used similar, but nonidentical stratified cluster sampling designs to provide vaccination coverage estimates at the national level and in each of Nigeria's 36 states and the Federal Capital Territory (FCT) (Dong et al., 2020a). Hereafter we refer to these as *Nigeria's 37 states*. The DHS samples were stratified by urban and rural areas within each state, whereas MICS and NNHS were stratified only by state. Respondents were selected in three stages: First, primary sampling units (PSUs or *clusters*) were selected from a census frame of enumeration areas (EAs) within each stratum. The 2003 DHS and 2007 MICS used frames based on the 1991 census and the other surveys used frames based on the 2006 census. All sampling frames cover the whole population. Second, a pre-determined number of households were randomly selected in each cluster from a recently-updated frame listing every residential household. Third, field teams attempted to interview all women aged 15–49 years and collect vaccination data for children under age 5 or 3 years from every selected household.

Every child in the survey has an associated survey weight that quantifies the relative number of children he or she represents in the population. Generally, DHS and MICS surveys calculate a base weight for each child as the inverse of the product of the probabilities of selection from each sampling stage. The base weights are then adjusted to reflect household or individual response rates. In contrast, the NNHS were conducted based on the Standardized Monitoring and Assessment of Relief and Transitions (SMART) method (SMART Technical Advisory Group, 2006, 2017), and each child was assigned the same base weight across all clusters under the standard assumption of equal selection probability. The base weights were then post-stratified to match a set of administrative population figures for the survey strata without further adjustment for non-response rate. Table 3.1 summarizes key information regarding the sampling design of each survey, including source of sampling frame, survey strata, cluster and household sample sizes, age group with vaccination data, and features of weight calculation.

For each child sampled in a survey, the care taker was asked to provide information about specific vaccines that the child had received up until the time of interview. Specifically, care takers were first asked to present any home-based records (HBR), such as vaccination cards, to the survey interviewer. If an HBR is available, the interviewer would populate a designated section of the survey questionnaire that detail the types and doses of vaccines received by the child. After this step is complete, the interviewer would ask the care giver a follow-up question of whether additional vaccine doses had been administered besides the ones listed on the HBR. This would include vaccines received at health care facilities via RI but had not been recorded on paper, or those received via SIAs. If yes, the interviewer will go back to the same section and record any additional doses based on the care giver's recall. If no HBR is available, the interviewer would skip the HBR section and ask the care giver to recall whether the child has received each specific vaccine at the time of survey interview, regardless of whether the vaccination is administered through RI or SIAs. As such, the surveys were designed to measure the percentage of children in various birth cohorts who received specific vaccines at any time before the survey through any source, and it is extremely difficult, if not impossible, to differentiate whether a child received MCV1 through RI or SIAs based on the answers to the survey questionnaires alone.

3.2.2 SIA calendar

The WHO SIA calendar is a publicly available data source where key information of the completed and planned SIAs in member states is provided (World Health Organization, 2020c). For each SIA, the calendar records the start and end dates, target age group and geographical areas in which the campaign is carried out. Table 3.2 shows the details of the SIAs in Nigeria conducted before the end of 2018 that we consider for this analysis. Excluded from the analysis are two mop-up campaigns implemented in 2007 and one outbreak response campaigns implemented in 2013. These campaigns' scales were very small in terms of target population and geographic area, and therefore were assumed to have negligible impact on the state-level MCV1 coverage.

Survey Time	Survey Type	Census from which Sampling Frame is derived	Strata	#Clusters per State	#Households per Cluster	Age Group with Vaccination Data	Base Weight Calculated using Stage 1 & 2 Selection Probabilities	Non-response Adjustment	Post-stratification
Mar. – Aug. 2003	DHS	1991	Urban/Rural within State	2–21	19–24	0 – 59 months	✓	✓	
Mar. – Apr. 2007	MICS	1991	State	30	25	0 – 59 months	✓	✓	
Jun. – Oct. 2008	DHS	2006	Urban/Rural within State	23–32	41	0 – 59 months	✓	✓	
Feb. – Mar. 2011	MICS	2006	State	40	20	0 – 59 months	✓	✓	
Feb. – Jun. 2013	DHS	2006	Urban/Rural within State	25–40	45	0 – 59 months	✓	✓	
Feb. – May 2014	NNHS	2006	State	30–35	22	0 – 59 months			✓
Jul. – Sep. 2015	NNHS	2006	State	32	22	0 – 59 months			✓
Sep. 2016 – Feb. 2017	MICS/NICS	2006	State	60–120	16	0 – 35 months	✓	✓	
Aug. – Dec. 2018	DHS	2006	Urban/Rural within State	35–53	30	0 – 35 months	✓	✓	

Table 3.1: Summary information of survey sampling design, including source of sampling frame, survey strata, cluster and household sample sizes, age group with vaccination data, and features of weight calculation.

Start Date	End Date	Target Age Group	Geographic Areas
12/06/2005	12/10/2005	9M–15Y	Northern States
10/03/2006	10/09/2006	9M–15Y	Southern States
11/26/2008	12/15/2008	9–59M	All States
01/26/2011	01/30/2011	9–59M	Northern States
02/23/2011	02/27/2011	9–59M	Southern States
10/05/2013	10/09/2013	9–59M	Northern States
11/02/2013	11/25/2013	9–59M	Southern States
12/21/2015	11/25/2015	9M–10Y	Northern States
01/28/2016	02/01/2016	9–59M	Southern States
11/09/2017	11/30/2017	9–59M	Northern States
03/01/2018	03/30/2018	9–59M	Southern States

Table 3.2: The start and end dates, target age groups, and geographical areas of the SIAs in Nigeria we use for the analysis. There are 20 northern states and 17 southern states in Nigeria.

3.3 Methods

The basic idea of our method is that we consider the RI-specific MCV1 coverage as the percentage of children in a birth cohort who received a first dose of measles vaccine through the routine vaccine delivery system according to the RI schedule, i.e., at the age of 9 months. Each cross-sectional survey serves as a snap-shot measurement of the MCV1 coverage within a birth cohort at the time of survey interview. We can carefully align the birth time, RI schedule, SIA schedule and survey time for each child to determine whether a birth cohort has had the RI opportunity or any additional SIA opportunity when the survey data is collected.

Figure 3.1 illustrates the data generating mechanism we assume for a generic area. We discretize time into 6-month intervals and let the beginning of 2000 be the reference starting point. We let b index birth cohort, such that the children born between Jan. and Jun. in 2000 belong to birth cohort $b = 1$, and the children born between Jul. and Dec. in 2000 belong to birth cohort $b = 2$, and so on. We let t index survey time, such that a survey conducted between Jan. and Jun. in 2001 has $t = 3$, and a survey conducted between Jan. and Jun. in 2002 has $t = 5$, and so on. We assume each child is scheduled to receive MCV1 through RI at the age of 9 months.

As an example, we let there be an SIA targeting children aged 9–59M implemented between Jun. and Dec. 2001, and consider two birth cohorts in detail — cohort $b = 1$ represented by the blue circles, and cohort $b = 2$ represented by the orange circles. Each circle indicates the underlying MCV1 coverage of the birth cohort at a specific time. The blue/orange shade represents the proportion of children covered by RI, i.e., the RI-specific MCV1 coverage; and the yellow shade represent the proportion covered by SIA. Because all children within the target age group can receive vaccination during an SIA regardless of whether they have been vaccinated before, some children in the birth cohorts might be covered by both RI and SIA. They are represented by the intersection of the blue/orange shade and the yellow shade.

Now we consider three annual cross-sectional surveys conducted at time $t = 3, 5$ and 7 respectively. Let $p_{b,t}$ denote the proportion of children in birth cohort b who have received MCV1 at time t. At $t = 3$, children in cohort $b = 1$ should already have had an RI opportunity to receive MCV1, but those in cohort $b = 2$ are still too young to be vaccinated against measles. Therefore, we can use the survey data to obtain a design-based estimate of the MCV1 coverage for birth cohort $b = 1$ at survey time $t = 3$, which we denote $\hat{p}_{1,3}$. In Figure 3.1, $\hat{p}_{1,3}$ and its associated uncertainty interval are represented by a blue dot with a vertical line segment. Because the birth cohort has not been impacted by any SIA at the time of survey, $\hat{p}_{1,3}$ gives an estimate of the RI-specific MCV1 coverage. In the next survey at $t = 5$, both birth cohorts should have had their RI opportunities. In addition, both birth cohorts should also have been targeted during the SIA implemented in the previous time interval. Since survey data cannot differentiate the source of the vaccination, the design-based estimates of the MCV1 coverage for birth cohort $b = 1$ and 2 at time $t = 5$, denoted $\hat{p}_{1,5}$ and $\hat{p}_{2,5}$, would reflect the overall MCV1 coverage in these birth cohorts instead of the RI-specific coverage. In Figure 3.1, the overall MCV1 coverage is represented by the union of the blue/orange shade and the yellow shade. The same logic applies to the survey conducted at $t = 7$.

As mentioned in Section 3.1, our method uses survey data from children who only had RI and children who had RI and one SIA opportunity. Based on the data generating mechanism described above, our proposed approach takes three steps to construct sub-national RI-specific MCV1 coverage over time: first, individual-level data from each survey is processed to identify eligible birth cohorts; second, design-based MCV1 coverage estimates and associated variances are calculated for all eligible birth cohorts identified in step one at the sub-national area level; last, a space-time smoothing model is applied to the design-based estimates from step two to predict RI-specific MCV1 coverage over time for all sub-national areas. We now describe each step in detail.

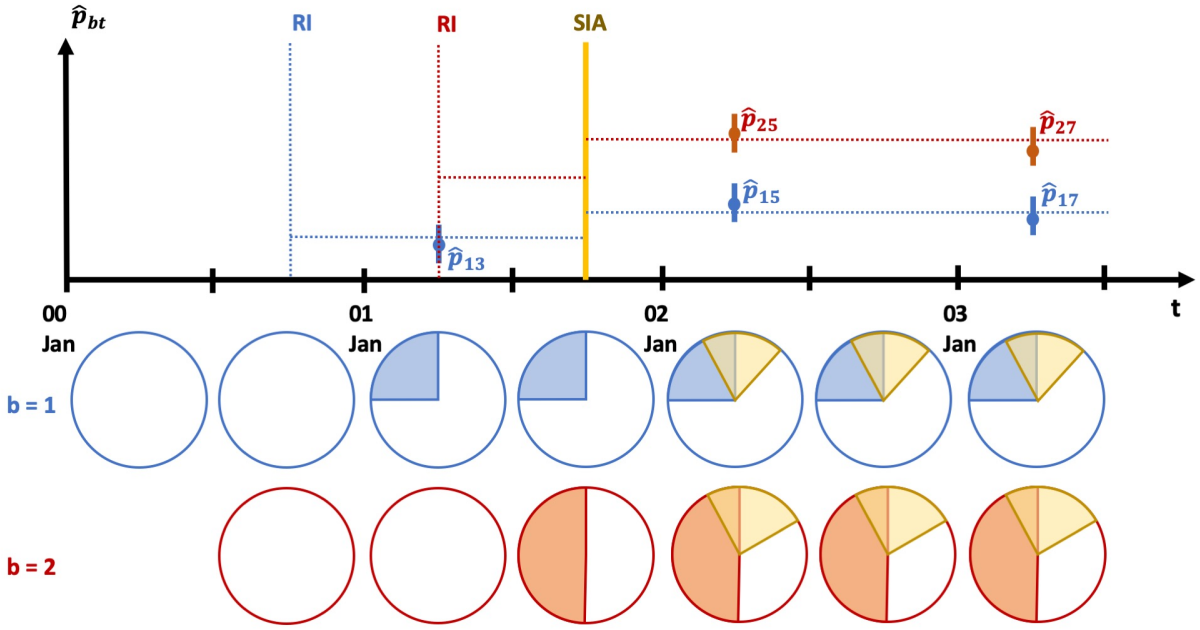

Figure 3.1: Illustration of the data generating mechanism for a generic area. We let b index birth cohort and consider two birth cohorts — cohort $b = 1$ represented by the blue circles, and cohort $b = 2$ represented by the orange circles. Each circle indicates the underlying MCV1 coverage of the birth cohort at a specific time. The blue/orange shade represents the proportion of children covered by RI, i.e., the RI-specific MCV1 coverage; and the yellow shade represent the proportion covered by SIA. Because all children within the target age group can receive vaccination during an SIA regardless of whether they have been vaccinated before, some children in the birth cohorts might be covered by both RI and SIA. They are represented by the intersection of the blue/orange shade and the yellow shade.

3.3.1 Data processing

In each survey, the time of interview for each child is recorded in Century-Month Code (CMC) format. In addition, the child's age in months at the time of interview is also recorded. Combining the two pieces of information allows us to identify the birth month, and hence the birth cohort, for each child.

Assuming every child has their RI opportunity at the age of 9 months, we use the age information of each child to see whether he/she has had the RI opportunity at the time of the survey. We also align the SIA calendar with each child's birth month to identify how many SIA opportunities he/she has had before the survey interview.

With each child's birth cohort, RI and SIA information collected, we now identify, within each sub-national area, the birth cohorts in which all children only had their RI opportunity, or all children had RI plus one SIA opportunity in each survey. Data from these eligible birth cohorts will be used in the next two steps.

3.3.2 Design-based overall MCV1 coverage estimation

For each eligible birth cohort identified in each survey, the design-based overall MCV1 coverage estimate is calculated at the sub-national area level. Let i, b, s and k be the indices for area, birth cohort, survey and sampled child respectively. We obtain the Horvitz-Thompson (HT) direct estimate (Horvitz and Thompson, 1952) of the overall MCV1 coverage in birth cohort b in area i from survey s, denoted $\hat{p}_{ibs}$, using

$$\hat{p}_{ibs} = \frac{\sum_{k=1}^{n_{ibs}} w_{ibsk} y_{ibsk}}{\sum_{k=1}^{n_{ibs}} w_{ibsk}},$$

where y_{ibsk} is the 0/1 indicator of whether child k in birth cohort b in area i has ever received MCV when interviewed for survey s, w_{ibsk} is the survey weight associated to the child, and n_{ibs} is the total number of children in birth cohort b in area i who are sampled in survey s. In addition, we also obtain the design-based estimate of the variance associated with $\hat{p}_{ibs}$, which we denote $\hat{v}_{ibs}$. See Section 2.1 for details of the calculation.

All design-based estimates can be computed using functions from the `survey` package (Lumley, 2004) within the `R` computing environment (Team, R Core, 2020).

3.3.3 Space-time smoothing model for RI-specific MCV1 coverage estimation

We extend the pseudo-likelihood approach developed in Mercer et al. (2015) and specify a Bayesian hierarchical space-time smoothing model to construct sub-national RI-specific MCV1 coverage estimates over time with uncertainty that reflects the various survey sampling designs.

First, the design-based estimates for the logit-transformed overall coverage and its associated variance, denoted by $\hat{\theta}_{ibs}$ and $\hat{V}_{ibs}$, are obtained via

$$\hat{\theta}_{ibs} = \log\left(\frac{\hat{p}_{ibs}}{1-\hat{p}_{ibs}}\right),$$
$$\hat{V}_{ibs} = \frac{\hat{v}_{ibs}}{\hat{p}^2_{ibs}(1-\hat{p}^2_{ibs})}.$$

Next, we take the following asymptotic distribution as a working likelihood:

$$\hat{\theta}_{ibs}|\theta_{ibs} \sim \text{Normal}\left(\theta_{ibs}, \hat{V}_{ibs}\right) \tag{3.1}$$

where θ_{ibs} is the underlying overall coverage in cohort b in area i measured by survey s. To account for the potential impact of SIA and specific surveys on the overall coverage, we specify θ_{ibs} as

$$\theta_{ibs} = \mu_{ib} + \beta_1 \times x_{ibs} + \epsilon_s \tag{3.2}$$

$$\mu_{ib} = \beta_0 + \alpha_i + \gamma_i + \delta_b + \tau_b + \phi_{ib} \tag{3.3}$$

The potential impact of SIA on the overall MCV1 coverage is captured by $\beta_1 \times x_{ibs}$, where x_{ibs} is an 0/1 indicator of whether children in birth cohort b in area i have had one SIA opportunity in additional to RI in survey s, and β_1 is a parameter quantifying the impact. We would expect $\beta_1 > 0$. If a birth cohort only had RI at the time of a survey, x_{ibs} would be zero and the overall coverage measured in that survey would not be influenced by any SIA. In

addition, we use the survey random effects $\epsilon_s \sim \text{Normal}\,(0, \sigma_\epsilon)$ to capture any survey-specific biases, although this is relative to the average of all the surveys, and does not correct for any overall bias in the surveys combined.

The underlying RI-specific MCV1 coverage in birth cohort b in area i is captured by the RI term μ_{ib}. Note that this quantity remains unchanged across surveys, because we consider the RI-specific MCV1 coverage to be the percentage of children in a birth cohort who received a first dose of MCV through the routine immunization system at the age of 9 months. Therefore, the underlying RI-specific coverage of a birth cohort only depends on when and where the children are born.

Space-time smoothing is achieved by modeling the RI term μ_{ib} as a combination of six components, see Equation (3.3). Specifically, β_0 is a common intercept for all birth cohorts and areas. There are two spatial terms that correspond to the convolution model of Besag et al. (1991) as described in Section 2.2.2. In particular, $\alpha_i \sim \text{ICAR}(\sigma_\alpha)$ are the intrinsic conditional autoregressive (ICAR) terms for $i = 1, \ldots, , I$ sub-national areas (e.g., $I = 37$ for Nigeria's 37 states), and $\gamma_i \sim_{iid} \text{Normal}(0, \sigma_\gamma)$ are independent and identically distributed (IID) random effects. There are also two temporal terms, with $\delta_b \sim \text{RW2}(\sigma_\delta)$ following an (intrinsic) random walk of order 2 model (described in Section 2.2.1) to pick up structured temporal trends in RI-specific coverage, and $\tau_b \sim_{iid} \text{Normal}(0, \sigma_\tau)$ being IID to pick up short-term fluctuations with no structure. Finally, the space-time interaction term ϕ_{ib} can be specified as one of the four types of interactions described in Knorr-Held (2000) to capture any additional independent or structured variation in RI coverage across space over time.

To construct RI-specific MCV1 coverage estimates and associated uncertainty intervals for each sub-national area, we use the Bayesian posterior samples of the RI component of the space-time smoothing model. Specifically, the percentage of children in birth cohort b in area i who received a first dose of MCV via RI, $p_{\text{RI},ib}$, can be estimated by drawing posterior samples

$$p_{\text{RI},ib}^{(m)} = \text{expit}\left(\mu_{ib}^{(m)}\right) = \text{expit}\left(\beta_0^{(m)} + \alpha_i^{(m)} + \gamma_i^{(m)} + \delta_b^{(m)} + \tau_b^{(m)} + \phi_{ib}^{(m)}\right),$$

where the superscript (m) denotes the mth posterior sample of the respective component. The survey specific effects ϵ_s are not included because they are bias terms.

Computation in this step, including model fitting and prediction, can be efficiently carried out using the Integrated Nested Laplace Approximation (INLA) (Rue et al., 2009) as implemented in the `INLA` package in `R`. See Section 2.4.4 for details.

3.4 Simulation study

We conduct a simulation study to examine how our approach performs under a number of scenarios with surveys of various sizes. Under each scenario, we compare two models:

1. *RI+SIA Model*: this model is similar to that described by Equations (3.1) – (3.3) in Section 3.3.3, but omitting the survey-specific effects, i.e. $\theta_{ibs} = \mu_{ib} + \beta_1 \times x_{ibs}$. For RI-specific coverage estimation, this model will be applied to the simulated survey data from birth cohorts that either had RI only or had RI plus one SIA opportunity.

2. *RI Only Model*: this model is similar to the *RI+SIA Model*, but omitting the component accounting for the impact of SIA, i.e., $\theta_{ibs} = \mu_{ib}$. This model will be applied to the simulated survey data from birth cohorts that had RI only — the same as the current common practice for RI-specific coverage estimation.

3.4.1 Set up

We simulate complete enumeration areas (EAs) and households within each of Nigeria's 37 states based on the distribution of population and EAs summarized in Table B.1 in the 2013 Nigeria DHS report (National Population Commission - NPC and ICF, 2014). For faster computation, the number of simulated EAs in each state is set to be one tenth of the true number of EAs, and each EA has 80 simulated households.

A total of 34 birth cohorts of children born between 2000 and 2016 are simulated assuming 1 new live birth per EA per month — a rate similar to Nigeria's birth rate during that time

period (The World Bank, 2019). Each simulated child is randomly assigned to a household. The resultant population from this step is used for the entire simulation study.

Next, for each iteration, we simulate RI and SIA activities roughly based on what have been implemented in Nigeria between 2000 and 2016. All children have their RI opportunities at the age of 9 months. In addition, we take Jan. 2000 as month 1 and let there be five SIAs conducted in all states in months 72, 108, 134, 167 and 193 respectively. We simulate the true RI-specific MCV1 coverage for children in birth cohort b in state i, $p_{\text{RI},ib}$, and the increase in the overall MCV1 coverage by the first SIA opportunity, $q_{\text{SIA},ib}$, under the following model:

$$\begin{aligned} p_{\text{RI},ib} &= \text{expit}(\beta_0 + \alpha_i + \gamma_i + \delta_b + \tau_b + \phi_{ib}) \\ &= \text{expit}(\mu_{ib}) \\ q_{\text{SIA},ib} &= \text{expit}(\mu_{ib} + \beta_1) - \text{expit}(\mu_{ib}) \\ &= \text{expit}(\mu_{ib} + \beta_1) - p_{\text{RI},ib} \end{aligned}$$

$$\alpha_i \sim \text{ICAR}(\sigma_\alpha) \quad \gamma_i \sim_{iid} \text{Normal}(0, \sigma_\gamma)$$
$$\delta_b \sim \text{RW2}(\sigma_\delta) \quad \tau_b \sim_{iid} \text{Normal}(0, \sigma_\tau)$$
$$\phi_{ib} \sim_{iid} \text{Normal}(0, \sigma_\phi)$$

The true values for the parameters are specified as follows:

$$\beta_0 = 0 \quad \beta_1 = 0.8$$
$$\sigma_\alpha = 0.8 \quad \sigma_\gamma = 0.01$$
$$\sigma_\delta = 0.8 \quad \sigma_\tau = 0.01$$
$$\sigma_\phi = 0.2$$

The MCV1 status for each child over time is simulated based on the true RI and SIA coverage in the birth cohort and state that the child belongs to.

A total of eight surveys are simulated based on the survey schedule in Nigeria between 2000 and 2017. Each survey is stratified by state and adopts a two-stage cluster sampling

design. No stratification by urban/rural is carried out in this simulation study. Within each state i (i.e., stratum), c_i clusters are randomly sampled from the C_i EAs in the first stage, then within each selected cluster, h_{ic} households are randomly sampled from the H_{ic} households in the second stage. All children under 5 years of age in the selected households are sampled, and their MCV1 status are recorded.

Using the simulated survey data, we implement the data processing and design-based coverage estimation procedures described in Section 3.3. We then apply the *RI+SIA Model* and *RI Only Model* to the respective data sets to obtain the point estimates and associated 95% posterior credible intervals of the state-level RI-specific MCV1 coverage for each birth cohort.

The performance of our approach is examined under a number of scenarios with surveys of various sizes. Specifically, we conduct 300 iterations under each combination of $c_i = 20, 40, 60$ and $h_{ic} = 15, 30, 45$, with each iteration using a different simulated data set. The estimation accuracy and precision of the two models are evaluated based on four metrics: bias, RMSE, the nominal coverage and the average widths of 95% posterior CIs of the estimated RI-specific coverage $p_{\mathrm{RI},ib}$ across all areas and birth cohorts.

3.4.2 Results

The simulation results are summarized in Figure 3.2. In each plot, results from the *RI+SIA Model* and *RI Only Model* are represented by red and blue dots respectively, with darker shades corresponding to greater numbers of households sampled per cluster. Along the x-axis from left to right are increasing numbers of clusters sampled per state.

Both models have very small bias under all scenarios. Within each model, both the RMSE and the 95% nominal coverage improves as the number of clusters per state and the number of households per cluster increases. The average CI widths show a similar trend in general, except for when the *RI Only Model* is applied to very sparse data ($c_i = 20, h_{ic} = 15$).

Under each scenario, the *RI+SIA Model* always outperforms the *RI Only Model* in terms of RMSE, coverage and average CI widths. This demonstrates the improved accuracy and

precision of the RI-specific coverage estimation by the *RI+SIA Model*, which makes more efficient use of data by having a model component that accounts for the impact of SIAs and incorporating information from birth cohorts who had an SIA opportunity in addition to RI.

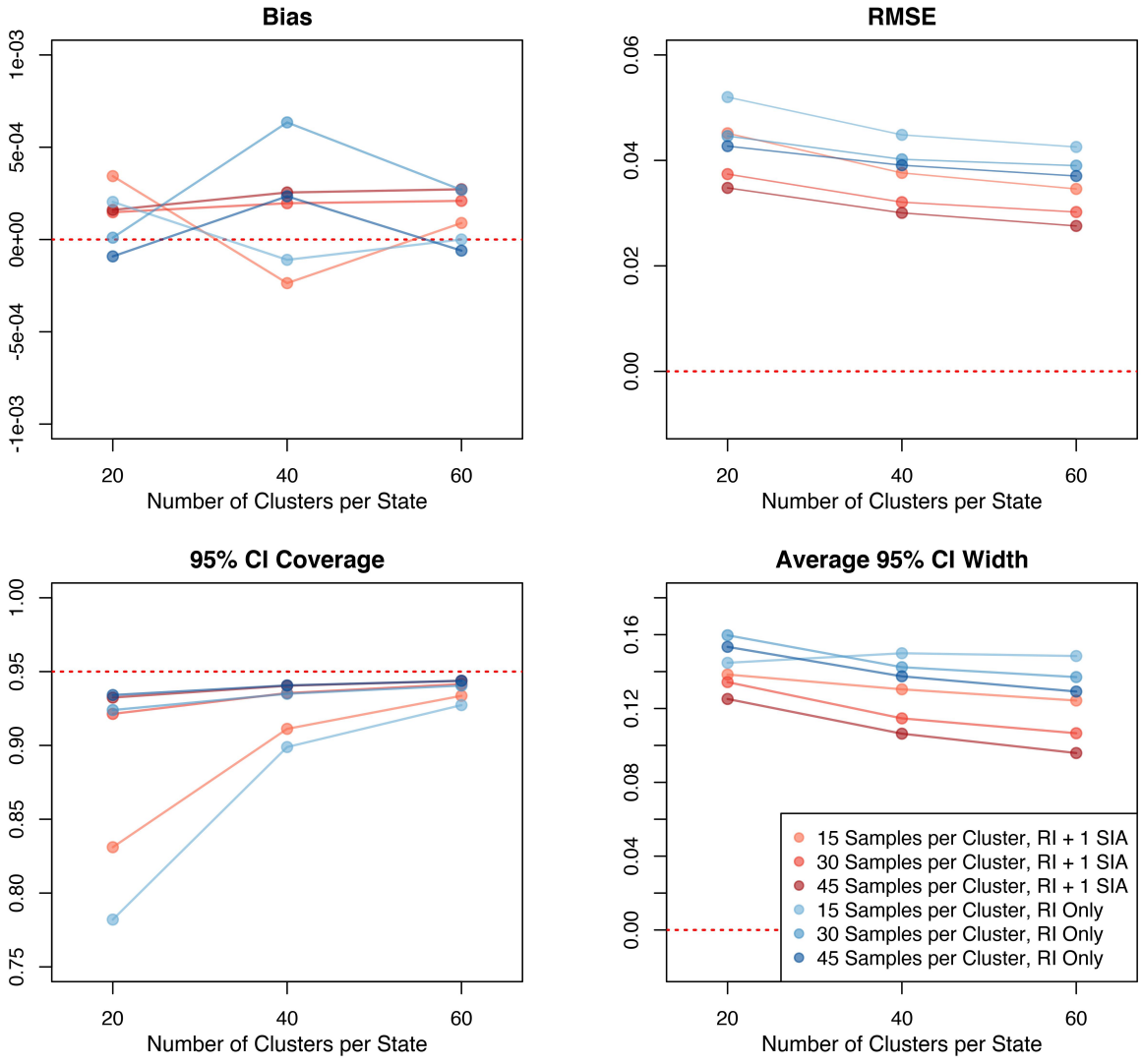

Figure 3.2: Summary results of the simulation study, including bias, root mean square error (RMSE), the nominal coverage and the average widths of 95% posterior credible intervals (CIs) of the estimated RI-specific MCV1 coverage $p_{\mathrm{RI},ib}$ across all areas and birth cohorts.

3.5 Application to Nigeria data

We apply our proposed method to estimate the state-level RI-specific MCV1 coverage in Nigeria between 2000 and 2018 using the household survey and SIA calendar data described in Section 3.2. We discretize time into 6-month intervals and let the beginning of 2000 be the reference starting point. During this time period, all children in Nigeria were scheduled to receive one dose of MCV via RI at the age of 9 months.

3.5.1 Data processing and design-based overall MCV1 coverage estimation

We carry out the data processing step outlined in Section 3.3.1 to identify, within each state, the birth cohorts in which all children only had their RI opportunity, or all children had RI plus one SIA opportunity in each survey. The design-based HT estimates of the state-level overall MCV1 coverage are calculated for each eligible birth cohort in each survey using the `survey` package (Lumley, 2004) within the `R` computing environment (Team, R Core, 2020). Figure 3.3 shows the resultant estimates with the associated confidence intervals across birth cohorts for Nigeria's 37 states. The plots are arranged to approximately match the relative geographical locations of the states. The different colors represent the surveys upon which the design-based estimates are calculated, and the shape of the dots indicates whether a birth cohort only had RI or had RI plus one SIA at the time of survey. The northern states generally have lower MCV1 coverage estimates than the southern states. As expected, within the same survey, the birth cohorts with an additional SIA opportunity tend to have slightly higher estimated MCV1 coverage than the RI-Only cohorts, although the differences are often small in magnitude. The coverage estimates of the same birth cohort from different surveys sometimes show large differences even after accounting for vaccination activity, suggesting that there might be some survey-specific effects. As a result of small sample sizes in some areas and surveys, the design-based coverage estimates for some birth cohorts are associated with very wide confidence intervals.

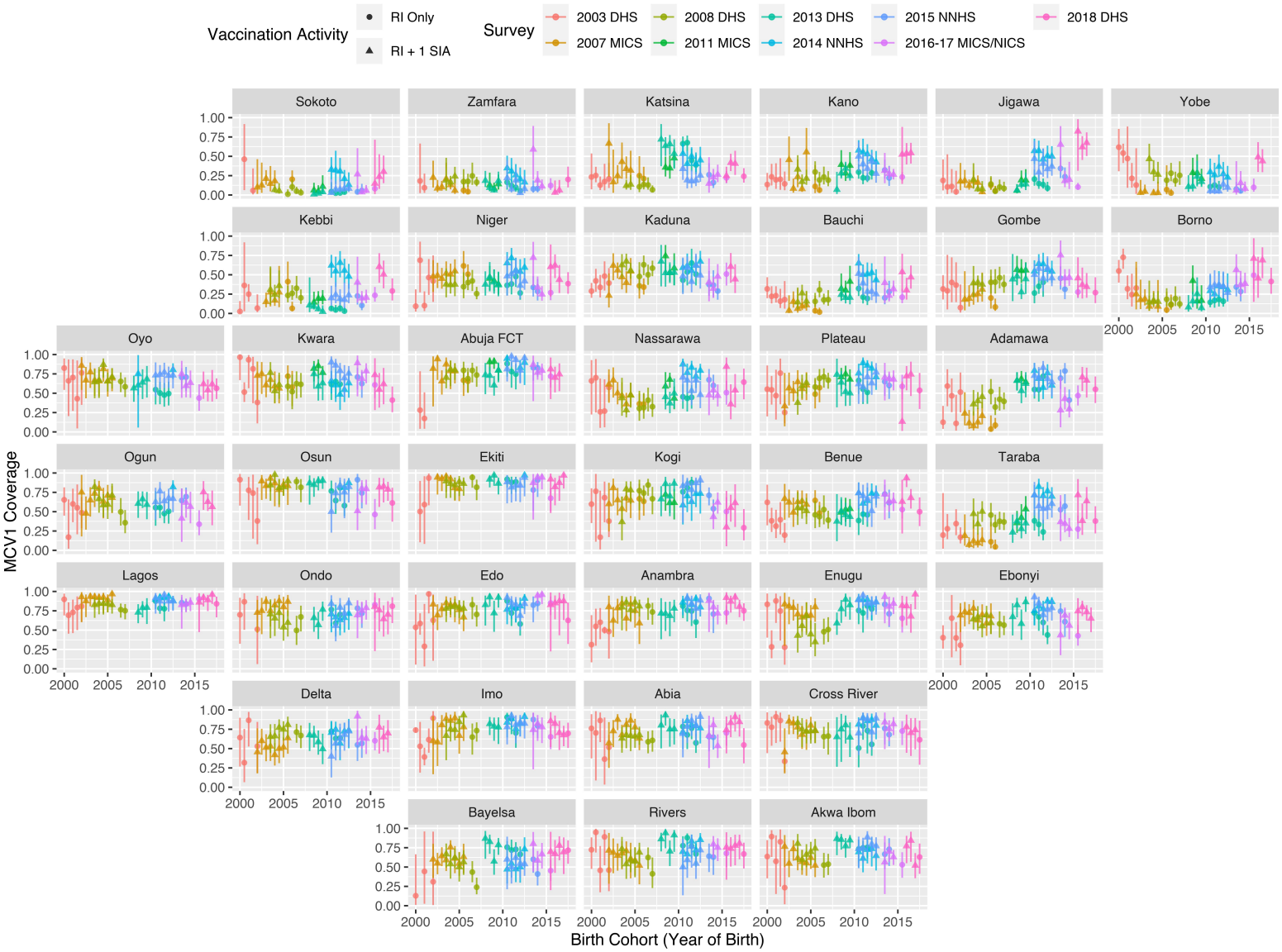

Figure 3.3: Design-based HT estimates of the overall MCV1 coverage across birth cohorts from each survey for Nigeria's 37 states.

3.5.2 RI-specific MCV1 coverage estimation via space-time smoothing

We fit six space-time smoothing models to the design-based estimates. Each model has the same working likelihood and model components as specified in Equations (3.1) – (3.3),

but a different specification of the space-time interaction term ϕ_{ib}. Specifically, we consider the *IID-IID*, *ICAR-IID*, *IID-RW2*, *ICAR-RW2*, *IID-AR1*, *ICAR-AR1* models, whose space-time interaction assumptions are detailed in Table 3.3 (Knorr-Held, 2000). In particular, *AR1* stands for the autoregressive (AR) model of order 1 for Gaussian random vectors. The definition of the AR1 model can found in Appendix A.

Model	Space-time Interaction ϕ_{ib}
IID-IID	IID spatial effects × IID temporal effects
ICAR-IID	ICAR spatial effects × IID temporal effects
IID-RW2	IID spatial effects × RW2 temporal effects
ICAR-RW2	ICAR spatial effects × RW2 temporal effects
IID-AR1	IID spatial effects × AR1 temporal effects
ICAR-AR1	ICAR spatial effects × AR1 temporal effects

Table 3.3: The space-time interaction assumptions of the models.

Implementation of the models was carried out in `R` using the `INLA` package. We used the Penalized Complexity (PC) priors introduced in Simpson et al. (2017) for all hyper-parameters (see Section 2.4.2). Specifically, we applied weak PC priors on the precision parameters for the random effects, such that $\Pr(\sigma > 2) = 0.01$ for the spatial and temporal terms, and $\Pr(\sigma > 1) = 0.01$ for the independent terms.

In Table 3.4, we report the posterior medians and 95% CIs of the model parameters. In addition, we also compute three model assessment metrics: the deviance information criterion (DIC), the Watanabe-Akaike information criterion (WAIC) and the log-score conditional predictive ordinate (LCPO) values, all of which are described in detail in Section 2.5. Based on the estimates for β_1, the odds of receiving at least one dose of MCV is 33.6% – 37.7% higher after an SIA campaign is conducted in a birth cohort. The *ICAR-AR1 Model* performs the best among the six models based on all three model metrics.

	Model Assessment Metric			Parameter Estimate								
Model	DIC	WAIC	LCPO	Intercept	SIA	ICAR	Area IID	RW2	Time IID	Space×Time	Survey	AR1
				β_0	β_1	σ_α	σ_γ	σ_δ	σ_τ	σ_ϕ	σ_ϵ	ρ_ϕ
IID-IID	3387	3616	1.097	-0.013 (-0.211, 0.203)	0.32 (0.19, 0.43)	0.67 (0.53, 0.87)	0.072 (0.010, 0.243)	0.25 (0.10, 0.57)	0.074 (0.038, 0.136)	0.34 (0.31, 0.37)	0.23 (0.13, 0.45)	
ICAR-IID	3344	3564	1.068	-0.007 (-0.208, 0.218)	0.32 (0.17, 0.43)	0.67 (0.53, 0.87)	0.074 (0.010, 0.245)	0.23 (0.09, 0.54)	0.091 (0.057, 0.139)	0.34 (0.31, 0.38)	0.25 (0.14, 0.53)	
IID-RW2	3250	3386	0.984	0.001 (-0.203, 0.205)	0.29 (0.19, 0.37)	0.25 (0.12, 0.46)	0.062 (0.009, 0.216)	0.19 (0.07, 0.49)	0.089 (0.057, 0.136)	0.51 (0.40, 0.63)	0.24 (0.14, 0.45)	
ICAR-RW2	3206	3341	0.972	-0.001 (-0.240, 0.264)	0.30 (0.15, 0.41)	0.28 (0.03, 1.45)	0.14 (0.01, 0.73)	0.24 (0.10, 0.55)	0.088 (0.055, 0.135)	0.60 (0.49, 0.74)	0.28 (0.15, 0.62)	
IID-AR1	3186	3347	0.983	-0.006 (-0.225, 0.235)	0.30 (0.17, 0.42)	0.59 (0.44, 0.78)	0.068 (0.009, 0.233)	0.24 (0.10, 0.57)	0.085 (0.051, 0.136)	0.43 (0.36, 0.51)	0.26 (0.14, 0.56)	0.88 (0.81, 0.92)
ICAR-AR1	**3154**	**3302**	**0.967**	-0.007 (-0.208, 0.220)	0.32 (0.18, 0.43)	0.51 (0.34, 0.71	0.072 (0.010, 0.235)	0.23 (0.10, 0.54)	0.089 (0.056, 0.136)	0.51 (0.41, 0.69)	0.26 (0.14, 0.54)	0.92 (0.86, 0.96)

Table 3.4: Summary of model assessment metrics and parameter estimates. Bold figures represent the "best" models according to the relevant criteria. The posterior medians are taken as the point estimates and the corresponding posterior 95% CIs are shown in the parentheses.

Table 3.5 contains the proportions of variance explained by each of the random effects in the space-time smoothing model. The "ICAR" column corresponds to the structured spatial effects; "Space IID" to the unstructured spatial effects; "RW2" to the structured temporal trends; "Time IID" to the unstructured temporal effects; and "Space×Time" to the respective space-time interaction terms in each model. Across all models, the combined effect from the RW2 and Time IID terms accounts for an average of 4.3% of the variability, indicating that the overall trend in the RI-specific MCV1 coverage remains relatively stable during the time period of interest. The survey-specific effects and the spatial IID effects also account for relatively small proportions of the variability. The majority of the variation is explained by the the structured spatial ICAR terms and the space-time interactions: the combined effect of the two terms accounts for an average of 88.5% of the variability, and is relatively consistent across all models. The relative proportions explained by the space-time interaction term depend on the specification of the interaction structure. This suggests that the distribution of variability between the spatial ICAR term and the space-time interaction can be sensitive to model specification. It also implies that spatial variation plays an important role in explaining the RI-specific MCV1 coverage.

Model	ICAR σ_α^2	Space IID σ_γ^2	RW2 σ_δ^2	Time IID σ_τ^2	Space×Time σ_ϕ^2	Survey σ_ϵ^2
IID-IID	80.8	1.6	3.8	0.4	8.7	4.7
ICAR-IID	79.4	1.7	3.7	0.6	9.3	5.2
IID-RW2	12.5	1.4	3.4	0.8	75.6	6.3
ICAR-RW2	17.7	4.2	2.3	0.3	72.0	3.5
IID-AR1	72.5	1.4	4.5	0.6	15.2	5.8
ICAR-AR1	52.7	1.9	4.7	0.8	33.1	6.9

Table 3.5: Proportions of variance explained by each random effect component (percent).

We now focus on the *ICAR-AR1 Model* and visualize the estimated random effects in the model. The estimated survey-specific effects ϵ_s are shown in Figure 3.4. The estimates suggest that there are moderate survey effects — a result that is also reflected in Table 3.5, where the survey effects are the third largest contributor to the variance explained by the random effects (6.9%) in the *ICAR-AR1 Model* model.

The posterior medians of the ICAR spatial random effects α_i and the IID spatial random effects γ_i are shown in Figure 3.5. The ICAR random effects show apparent spatial dependencies across states, with the northern states having lower estimated values than the southern states. In contrast, the IID random effects show no apparent spatial pattern, and their magnitudes are much smaller compared to the ICAR random effects. This suggests that the overall spatial variation in MCV1 coverage is mostly structured.

The posterior medians of the RW2 temporal random effects δ_b and the IID random effects τ_b are shown in Figure 3.6, along with the 95% point-wise CI envelopes. The RW2 temporal effects show slight downward trends at the beginning and the end of the time period, and a slight upward trend in the middle. Due to data sparsity, the uncertainties associated with the estimates at the beginning and the end of the time period are relatively large compared to those in the middle of the time period.

In Figure 3.7 and Figure 3.8, we show the estimated space-time interactions across space and over time for all states. At each time point, there is moderate spatial dependencies across states, with northern states having lower values. The temporal trends in most states were relatively flat with moderate fluctuations, except for a few northern states in which there were large changes over time.

Based on the fitted *ICAR-AR1 Model*, we calculated the state-level RI-specific MCV1 coverage estimates and the associated 95% CIs for the birth cohorts born between 2000 and 2018 using the procedure described in Section 3.3.3. Figure 3.9 shows the model results, along with the design-based overall MCV1 coverage estimates from each survey for reference. There is considerable spatial variation in the estimated RI-specific MCV1 coverage, with southern states generally having higher coverage than the northern states. In most states, the RI-

specific coverage remained at a constant level without significant changes in trend between 2003 and 2015. The 95% CIs associated with the RI coverage estimates during this time period are relatively narrow. In contrast, the RI-specific coverage estimates in 2000–2002 and 2016–2018 show greater temporal variation, but they are associated with wider CIs. Therefore, although there seem to be a slight downward trend in many states, one should be mindful of the uncertainties associated with the trend, especially when considering prediction for the 2018 birth cohorts.

In summary, this analysis illustrates how our model is able to combine information from multiple household surveys to produce spatially- and temporally-smoothed RI-specific MCV1 coverage estimates while accounting for the impact of SIAs as well as survey-specific effects.

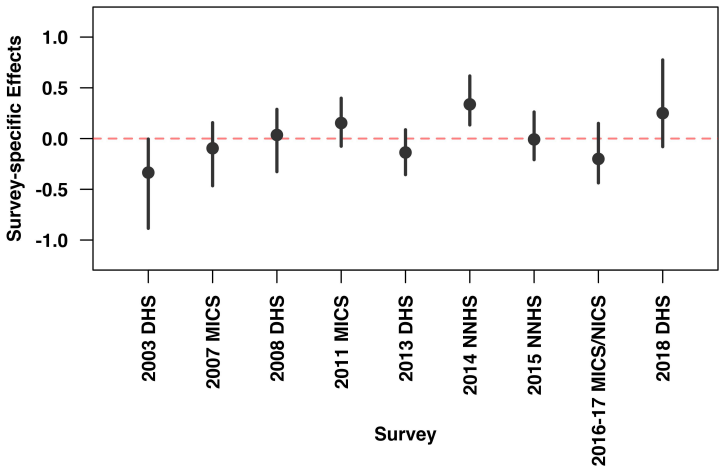

Figure 3.4: The posterior medians and 95% CIs of the survey-specific effects ϵ_s.

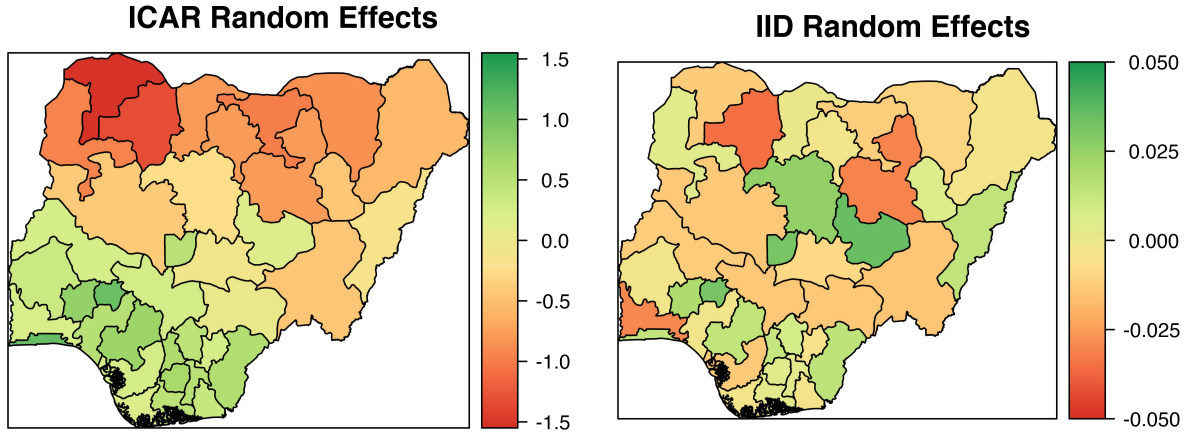

Figure 3.5: The posterior medians of the ICAR spatial random effects α_i (left) and the IID spatial random effects γ_i (right) from the fitted *ICAR-AR1 Model*.

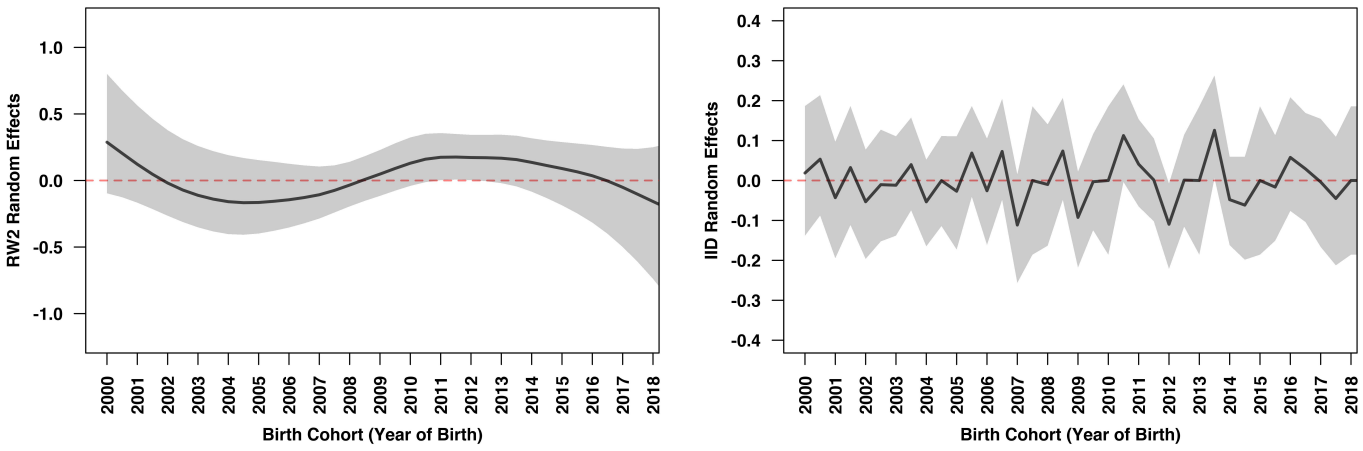

Figure 3.6: The posterior medians and 95% CIs of the RW2 temporal random effects δ_b (left) and the IID temporal random effects τ_b (right) from the fitted *ICAR-AR1 Model*.

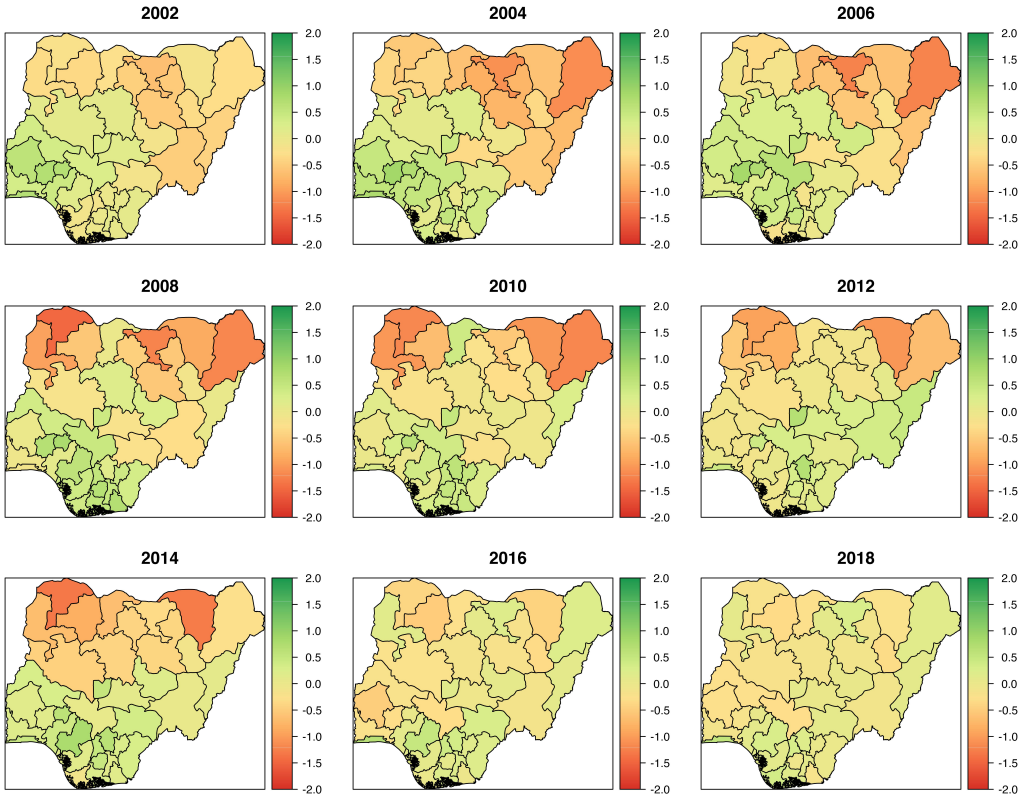

Figure 3.7: The posterior medians of the space-time interactions ϕ_{ib} across space over selected birth cohorts from the fitted *ICAR-AR1 Model*. Specifically, each map shows the estimated interaction for the cohorts born in the first 6 months of the corresponding year.

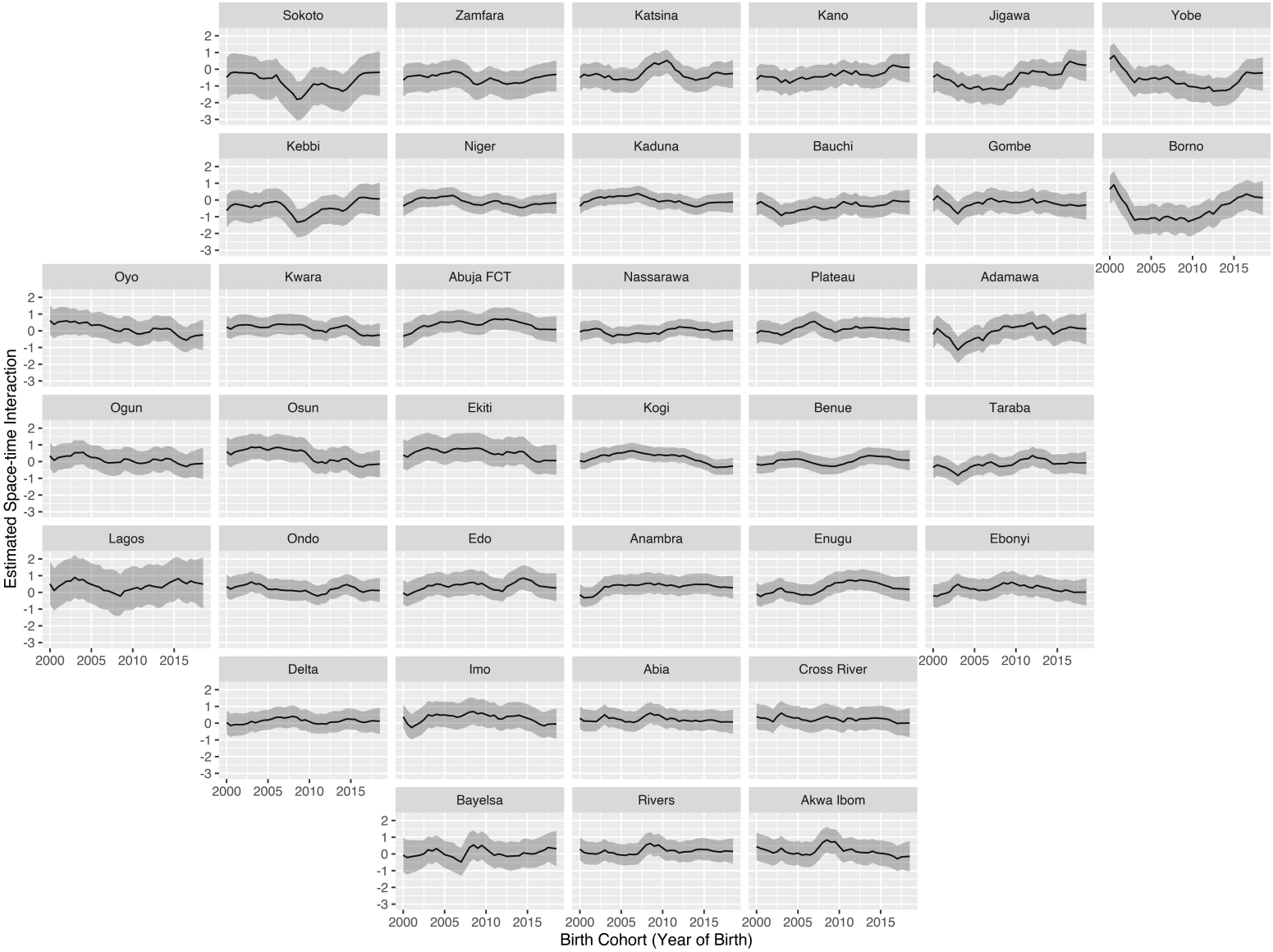

Figure 3.8: The posterior medians of the space-time interactions ϕ_{ib} over time in all states from the fitted *ICAR-AR1 Model*.

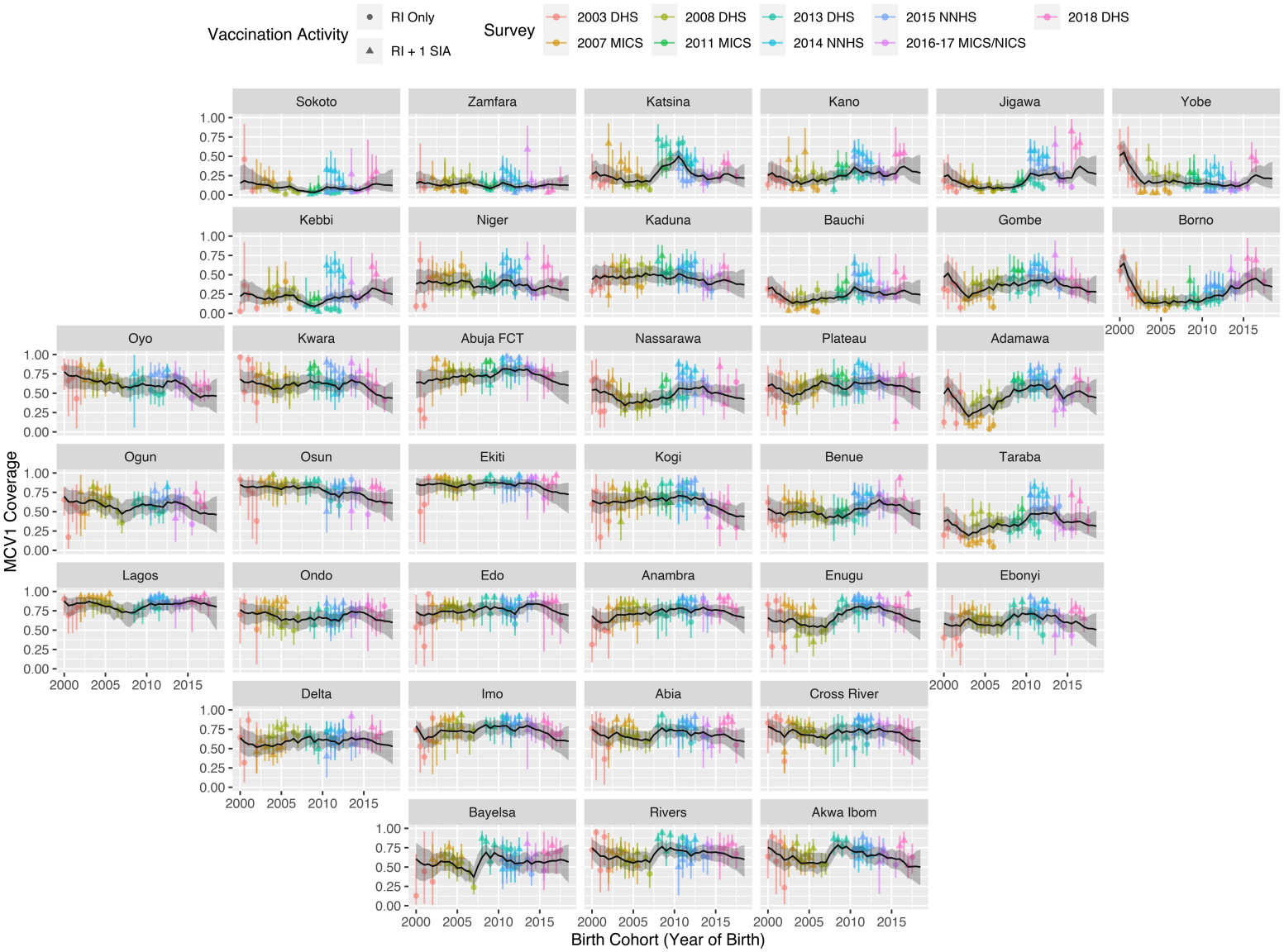

Figure 3.9: The posterior medians of the state-level RI-specific MCV1 coverage (dark grey lines) and the associated 95% CIs (light grey ribbons) for the birth cohorts born between 2000 and 2018 based on the fitted *ICAR-AR1 Model*. The design-based overall MCV1 coverage estimates were also shown for references.

3.5.3 Sensitivity analysis: using 12-month birth cohorts

We chose to use 6-month birth cohorts in our model as they gave a good balance between two competing objectives: on one hand, we want larger birth cohorts so that there would be enough samples from each sub-national area in each survey to provide design-based coverage estimates with reasonably high precision; on the other hand, smaller birth cohorts are more desirable because they are more likely to have children with the same RI and SIA history, and hence be eligible for analysis. Here we present a sensitivity analysis using 12-month birth cohorts.

We discretize time into 12-month intervals and let the beginning of 2000 be the reference starting point. We carry out the same data processing step outlined in Section 3.3.1 and calculate the design-based HT estimates of the state-level overall MCV1 coverage for each eligible birth cohort in each survey. Figure 3.10 shows the resultant estimates with the associated confidence intervals across birth cohorts for Nigeria's 37 states. As expected, the estimates for 12-month birth cohorts generally have narrower 95% CIs compared to the estimates for 6-month birth cohorts shown in Figure 3.3. However, only 61900 of the 86033 (71.9%) survey samples that were eligible for the previous analysis remain eligible for this analysis, and the proportion of the *RI Only* design-based coverage estimates (among the *RI Only* and *RI + 1 SIA* design-based coverage estimates) decreased from 608/1802 (33.7%) to 133/760 (17.5%). In particular, 35 of the 37 states do not have any *RI Only* design-based coverage estimate for any birth cohort born after 2012. As a result, we see that many southern states do not have any eligible data for birth cohorts born between 2006 and 2008, and a majority of the design-based overall MCV1 coverage estimates were influenced by a previous SIA.

The six space-time smoothing models specified in the main analysis were fitted to the design-based overall MCV1 coverage estimates for the 12-month birth cohorts. The posterior medians and 95% CIs of the model parameters as well as the model assessment metrics are reported in Table 3.6. Compared to the results in the main analysis, the estimates of the

intercept β_0 are considerably higher and the estimates of the coefficient for the SIA indicator β_1 are lower. The 95% CIs for these two parameters are also wider across all models. These results are likely due to the fact that less data, especially *RI Only* design-based coverage data, is used in this analysis. Since less information is available for the model to draw inference on the parameters, the resultant estimates differ from those in the main analysis and have higher associated uncertainties.

Similar to the main analysis, the *ICAR-AR1 Model* performs the best among the six models based on all three model metrics considered. The proportions of variance explained by the various random effects also show similar patterns compare to the main analysis. The details are provided in Table A.1 in Appendix A.

We now focus on the *ICAR-AR1 Model* and compare the results from fitting this model to the 12-month birth cohort data with the *ICAR-AR1 Model* results in the main analysis. The estimated ICAR and IID spatial random effects and the survey specific effects show very similar patterns between the two analyses. The estimated space-time interactions also show similar general trends, but some small-scale fluctuations shown in the main analysis are missing in this analysis, especially during time periods where *RI Only* design-based coverage data are not available. The estimated RW2 temporal random effects in this analysis are following a slightly different trend compared to the corresponding estimates in the main analysis: instead of showing a slight downward trend after 2012, the estimated RW2 random effects follow a slight increasing trend. In addition, the IID temporal random effects have considerably wider uncertainty intervals than those in the main analysis. Plots of these estimates and the associated uncertainties can be found in Appendix A.

In Figure 3.11, we compare the state-level RI-specific MCV1 coverage estimates produced by fitting the *ICAR-AR1 Model* to the 6-month and 12-month birth cohort data. The estimates from the 12-month birth cohort analysis are generally higher than those from the 6-month analysis. In particular, the differences between the two analyses tend to be large during time periods in which (1) there is no design-based coverage data (e.g., between 2006 and 2008), or (2) there is no *RI Only* coverage data (e.g., after 2012) for fitting the space-

time smoothing model in the 12-month birth cohort analysis. The pattern we see in this comparison, as well as the difference in parameter and random effect estimates we showed earlier, are likely due to the fact that using 12-month birth cohorts resulted in considerably less survey data being eligible for the smoothing model, hence some temporal patterns that were captured in the 6-month birth cohort analysis were unable to be estimated in this analysis.

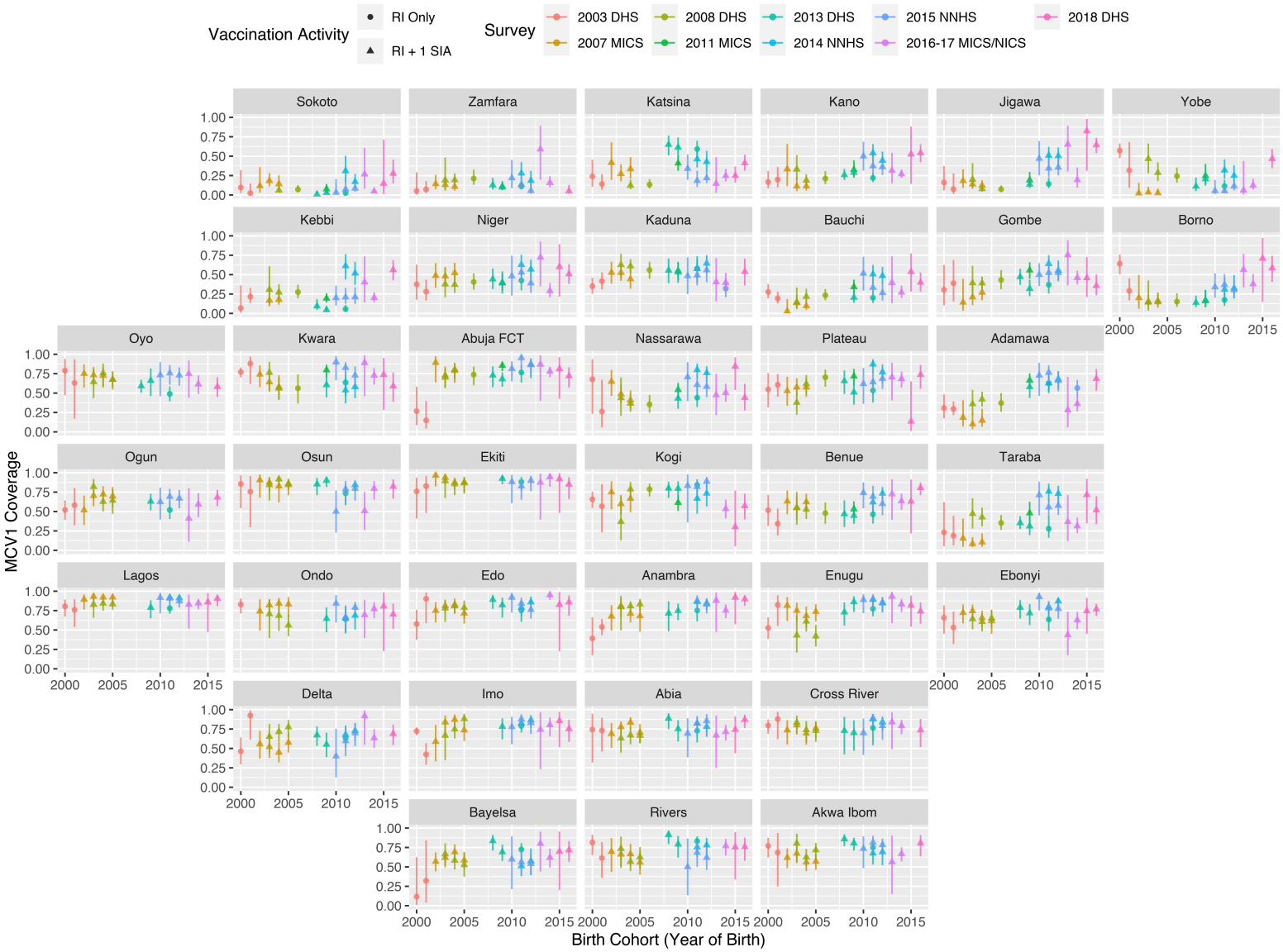

Figure 3.10: Design-based HT estimates of the overall MCV1 coverage across birth cohorts from each survey for Nigeria's 37 states in the sensitivity analysis with 12-month birth cohort data.

Model	Model Assessment Metric			Parameter Estimate								
	DIC	WAIC	LCPO	Intercept	SIA	ICAR	Area IID	RW2	Time IID	Space×Time	Survey	AR1
				β_0	β_1	σ_α	σ_γ	σ_δ	σ_τ	σ_ϕ	σ_ϵ	ρ_ϕ
IID-IID	1333	1463	1.112	0.211 (-0.019, 0.460)	0.195 (0.013, 0.368)	0.66 (0.52, 0.85)	0.071 (0.010, 0.238)	0.072 (0.007, 0.362)	0.081 (0.026, 0.193)	0.37 (0.33, 0.41)	0.21 (0.12, 0.40)	
ICAR-IID	1319	1452	1.078	0.206 (-0.029, 0.456)	0.198 (0.012, 0.376)	0.66 (0.52, 0.86)	0.075 (0.010, 0.245)	0.077 (0.008, 0.382)	0.105 (0.049, 0.212)	0.38 (0.33, 0.43)	0.20 (0.11, 0.39)	
IID-RW2	1294	1401	1.007	0.225 (-0.019, 0.468)	0.173 (-0.020, 0.358)	0.288 (0.028, 1.487)	0.184 (0.026, 0.964)	0.107 (0.014, 0.537)	0.126 (0.066, 0.235)	0.40 (0.31, 0.51)	0.20 (0.11, 0.39)	
ICAR-RW2	1281	1387	0.991	0.227 (-0.031, 0.503)	0.162 (-0.028, 0.344)	0.283 (0.027, 1.453)	0.144 (0.014, 0.736)	0.078 (0.008, 0.397)	0.125 (0.064, 0.235)	0.43 (0.34, 0.55)	0.20 (0.12, 0.40)	
IID-AR1	1273	1387	1.018	0.227 (-0.024, 0.496)	0.168 (-0.017, 0.347)	0.60 (0.45, 0.81)	0.068 (0.009, 0.238)	0.078 (0.008, 0.388)	0.104 (0.045, 0.211)	0.46 (0.38, 0.56)	0.21 (0.12, 0.40)	0.81 (0.70, 0.89)
ICAR-AR1	**1264**	**1377**	**0.994**	0.220 (-0.013, 0.474)	0.176 (-0.011, 0.353)	0.47 (0.28, 0.71	0.073 (0.010, 0.238)	0.077 (0.008, 0.385)	0.103 (0.047, 0.214)	0.58 (0.44, 0.83)	0.21 (0.12, 0.40)	0.88 (0.78, 0.95)

Table 3.6: Summary of model assessment metrics and parameter estimates in the sensitivity analysis with 12-month birth cohort data. Bold figures represent the "best" models according to the relevant criteria. The posterior medians are taken as the point estimates and the corresponding posterior 95% CIs are shown in the parentheses.

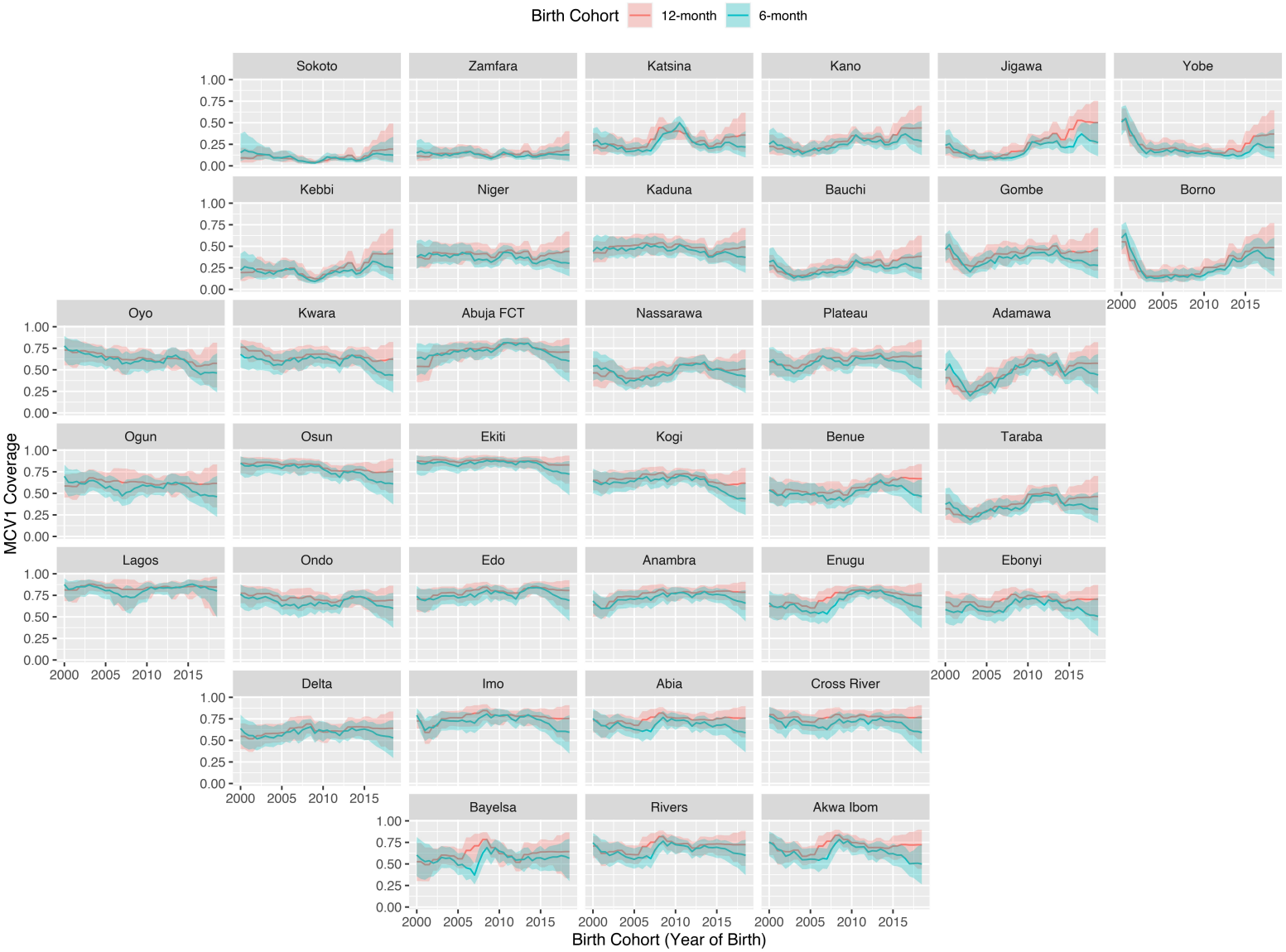

Figure 3.11: The posterior medians of the state-level RI-specific MCV1 coverage (dark grey lines) and the associated 95% CIs (light grey ribbons) for the birth cohorts born between 2000 and 2018 based on the fitted *ICAR-AR1 Model*.

3.6 Discussion

In this chapter, we developed a space-time smoothing model for estimating RI-specific MCV1 coverage at the sub-national level over time using data from household surveys. The proposed method is able to account for the impact of SIAs on the overall MCV1 coverage measured by cross-sectional surveys, and makes efficient use of data from not only birth cohorts who only had RI opportunities, but also those who had RI plus one SIA opportunity at the time of surveys. In addition, our model can be applied to data from multiple surveys with different sampling designs and construct coverage estimates with uncertainty that reflects the various data collection schemes. Any survey-specific effects such as sampling biases are also accounted for when predicting underlying RI coverage over time. We demonstrated how our method improves estimation through more efficient use of data via a simulation study that considered a number of scenarios with surveys of various sizes. We also applied our method to analyze data from nine household surveys conducted in Nigeria and constructed the RI-specific MCV1 coverage in each of Nigeria's 37 states from 2003 to 2018.

Since the DHS and MICS regularly conduct household surveys in most LMICs, our method is readily applicable to many countries with high measles burden to provide reliable RI-specific MCV1 coverage estimates over time to track routine vaccine delivery system strength. It helps identify areas in which the RI program is the weakest and facilitates balanced implementation of RI and SIAs to improve overall measles vaccination coverage. The output also provides important input data for measles epidemiological models, which estimate underlying susceptible population dynamics and inform optimal SIA schedules.

Under our framework, the RI-specific MCV1 coverage has a clear and sensible definition — the percentage of children in a birth cohort who receive a first MCV dose according to the RI schedule. This definition is particularly relevant in the context of building measles epidemiological models and evaluating SIA efficacy, where the RI-specific coverage is an essential input for calculating the number of children entering the susceptible poor over time (Verguet et al., 2015; Thakkar et al., 2019; Dong et al., 2020b). We emphasize here that the

denominator of this coverage only includes children who are still alive at their scheduled RI date, i.e., at the age of 9 months. In other words, our coverage definition excludes those who died before they become eligible to receive MCV1 through RI, and hence, does not require consideration of the mortality rate among those under the age of 9 months.

Our method does make assumptions about the mortality rates among children above the age of 9 months. Specifically, we assume that there is negligible difference in mortality rate between the children who are vaccinated against measles and the children who are not. This allows us to assume the underlying MCV1 coverage within each birth cohort remains constant as the children grow older, unless they have an SIA opportunity. As such, we can treat each cross-sectional survey as a snap-shot measurement of the same underlying MCV1 coverage within a birth cohort at the time of survey interview, and use survey data collected from children in older age groups for our model. However, if the unvaccinated children suffer significantly higher mortality than vaccinated children, the *proportion* of the surviving children in a birth cohort who are vaccinated would increase as the children grow older. In this case, the underlying MCV1 coverage within each birth cohort no long remains unchanged over time, and the coverage estimates from surveys implemented at different times among the same birth cohort are measuring different underlying coverage values. Therefore, a more complex model as well as additional mortality data are required to properly account for the impact of discrepancies in mortality rate between vaccinated and unvaccinated children, which would be an important topic for future research.

In addition, our model also assumes that children have their RI opportunity according to the stated schedule, and SIAs were carried out promptly as recorded in the calendar. These assumptions may not hold in reality, especially in settings where vaccine stock outs or disruption of services lead to delayed RI and SIAs. Their impact on RI-specific coverage estimation would depend on when and where SIAs and surveys were carried out. Further simulation studies could be done to investigate the issue under various scenarios.

Last, but not least, we assume there is negligible systematic recall bias across age groups in surveys. However, it is possible that care givers would under- or over-report doses for older

children due to memory recall. Previous studies gave mixed results regarding the presence, direction and magnitude of care giver recall bias in vaccination coverage surveys (Porth et al., 2019; Valadez and Weld, 1992; Ramakrishnan et al., 1999; Hu et al., 2019; Binyaruka and Borghi, 2018), and the results vary greatly by geographical regions. Future research can focus on how to detect and account for recall bias especially given specific context of application.

Chapter 4

ESTIMATING EFFICACY OF SUPPLEMENTARY IMMUNIZATION ACTIVITIES WITH DISEASE INCIDENCE TIME SERIES

4.1 Introduction

As mentioned in Chapter 3, a key public health strategy for measles prevention in high-burden settings is the combination of routine immunization (RI) and supplementary immunization activities (SIAs) (The Measles and Rubella Initiative, 2012). In this chapter, we develop epidemiological models for assessing SIA effectiveness and forecasting future incidence trend.

A common metric used for evaluating SIA effectiveness is the *administrative campaign coverage*. It is calculated as the ratio of the number of MCV doses administered during a campaign to the number of target population of the campaign. While SIAs usually have high reported campaign coverage, it is unclear how many people are effectively removed from the susceptible population (Mbabazi et al., 2009). Another popular indicator of SIA effectiveness is the *SIA coverage among MCV zero-dose children*, which is defined as the proportion of children in a specific age group with no history of receipt of MCV before the SIA who received a dose of MCV during the SIA. Portnoy et al. (2018) analyzed the Demographic and Health Survey (DHS) data from a few low- and middle-income countries (LMICs) to investigate how many children who have not previously received a measles vaccine dose are reached by SIAs. However, their methods require surveys being conducted within 2 years of implementation of SIAs and having a measles SIA indicator variable, which resulted in only 14 out of 111 countries having SIAs and DHS in the year 2000 onward to be eligible for the analysis. More recently, Utazi et al. (2020) analyzed the Nigeria 2017–18 post-campaign coverage survey

(PCCS) data to estimate the *SIA coverage among MCV zero-dose children* of the campaign. However, high quality PCCS are rare in most LMICs, and the resultant estimates cannot be easily applied to plan future campaigns.

An attractive measure of SIA effectiveness is the *SIA efficacy*, defined as the fraction of susceptible population immunized after a measles SIA. However, estimation of *SIA efficacy* has been a programmatic challenge due to difficulties in estimating the underlying susceptible population. Thakkar et al. (2019) proposed a linear regression approach that incorporates *SIA efficacy* to analyze measles incidence data in Pakistan to optimize campaign timing. While their method demonstrates effective use of *SIA efficacy* in campaign planning, it does not produce valid inference for the efficacy parameter. Specifically, the model does not account for data variability due to under-reporting, and is unable to construct valid uncertainty intervals for the model parameters and the predicted incidence.

In this chapter, we propose a discrete-time hidden Markov model under the time-series susceptible-infected-recovered (TSIR) framework (Finkenstädt and Grenfell, 2000) to estimate SIA efficacy using a hybrid of ordinary least squares (OLS) regression with robust standard errors and Markov chain Monte Carlo (MCMC) procedures. The TSIR model was first introduced by Finkenstädt and Grenfell (2000) to provide an approximation for measles time series in large communities where the disease is endemic. It was then extended in a series of subsequent papers (Bjørnstad et al., 2002; Grenfell et al., 2002; Glass et al., 2003; Morton and Finkenstädt, 2005) and has been used to understand measles and rubella transmission in a variety of settings (Ferrari et al., 2008; Metcalf et al., 2011; Mahmud et al., 2017; Metcalf et al., 2013). A recent review by Wakefield et al. (2019) details the theoretical motivation of the framework and compares it with other popular epidemic modelling approaches such as the Epidemic/Endemic (EE) models introduced by Held et al. (2005). More details regarding the TSIR framework are provided in Section 2.3. Our approach accounts for under-reporting and seasonality of measles transmission, and accommodates monthly reported incidence data that are publicly available from the WHO Measles and Rubella Surveillance Database for most member countries (World Health Organization, 2020c). The proposed model can be

used to estimate the underlying susceptible population dynamics, assess how many susceptible people were immunized by past SIAs, and forecast incidence trends in the future under various hypothetical SIA scenarios.

This chapter is structured as follows. In Section 4.2, we describe the model and the inference details. A simulation study in Section 4.3 considers various levels of under-reporting, where the observed incidence time series are obtained from extremely low to high reporting rates. In Section 4.4, we illustrate our model by using it to analyze the reported measles incidence in Benin from 2012 to 2018. Section 4.5 contains concluding remarks.

4.2 Methods

4.2.1 Model specification

In general, disease incidence (i.e. the number of new infections in a time period) is commonly modeled at time steps equal to the average generation time of the disease, which is about 14 days for measles (World Health Organization, 2020d). To better accommodate the monthly reported incidence data, we model the underlying true measles incidence in semi-monthly intervals. We let m and t index the monthly and the semi-monthly time points, and let $t(m)$ and $t(m)+1$ denote the first and second semi-monthly time points in month m. In addition, we use C and I to denote the reported and the underlying true measles incidence respectively. We account for under-reporting by letting the monthly reported incidence count C_m to follow a Binomial distribution with a sample size equal to the sum of the true semi-monthly incidence in that month, $I_{t(m)} + I_{t(m)+1}$, and a constant reporting probability ρ,

$$C_m | I_{t(m)}, I_{t(m)+1} \sim \text{Bin}\left(I_{t(m)} + I_{t(m)+1}, \rho\right). \tag{4.1}$$

For the underlying true incidence, we let the number of new infections in the current semi-month I_t follow a Negative Binomial distribution with a mean that depends on the incidence I_{t-1}, the susceptibles S_{t-1} and the known total population N_{t-1} from the previous

semi-month,

$$I_t|I_{t-1}, S_{t-1} \sim \text{NegBin}(\lambda_t, \phi) \quad \text{for } t \geq 2, \tag{4.2}$$

$$\lambda_t = \exp(\beta_t^{\text{AR}}) \frac{I_{t-1}^{\alpha} S_{t-1}}{N_{t-1}} + \exp(\beta^{\text{EN}}) N_{t-1}. \tag{4.3}$$

In particular, we assume a constant dispersion parameter ϕ and endemic parameter β^{EN} over time. The mixing parameter α is set to be 0.975 to account for the discrete-time approximation to the continuous time model (Glass et al., 2003). As for the auto-regressive parameter β_t^{AR}, we use the following formulation to capture the yearly seasonality of measles transmission,

$$\beta_t^{\text{AR}} = \gamma_1 + \gamma_2 t + \gamma_3 \sin\left(\frac{2\pi}{\omega} t\right) + \gamma_4 \cos\left(\frac{2\pi}{\omega} t\right), \tag{4.4}$$

where $\omega = 24$.

For the underlying dynamic of the susceptibles, we propose a balancing equation that incorporates the information from past SIA campaigns. Specifically, we model the number of susceptibles in the first semi-month S_1 as the product of the proportion of susceptibles in the total population θ (to be estimated) and the known size of the total population at that time point N_1. For semi-month $t \geq 2$, we let the number of susceptibles S_t be equal to the number of susceptibles in the previous semi-month S_{t-1}, plus the known number of births entering the susceptible pool B_t, minus the number of new infections I_t, and finally minus the number of susceptibles immunized in the most recent SIA campaign S_t^*,

$$S_1 = \theta \times N_1, \tag{4.5}$$

$$S_t = S_{t-1} + B_t - I_t - S_t^* \quad \text{for } t \geq 2. \tag{4.6}$$

The number of births entering the susceptible pool B_t can be calculated as the number of 9-month-old children who are losing maternal immunity and are not effectively vaccinated via RI. We call this quantity the *adjusted births*. Let L_t and R_t be the number of 9-month-old children in the population and the RI-specific coverage of the first dose of measles-containing-vaccine (MCV1) among the 9-month-olds at semi-month t. The adjusted births B_t can be

obtained by

$$B_t = L_t \times (1 - R_t \times 0.87), \tag{4.7}$$

assuming the efficacy of MCV1 in LMIC settings is 87% (World Health Organization, 2017a).

We formulate S_t^*, the number of susceptibles immunized in semi-month t in the most recent SIA campaign, as

$$S_t^* = p \times S_{k(t)} \times \delta_t \tag{4.8}$$

where p denotes the overall SIA efficacy parameter, representing the fraction of susceptible population immunized after a complete SIA campaign, $S_{k(t)}$ denotes the number of susceptibles right before the most recent campaign started, and δ_t denotes the fraction of the total target population to be covered in semi-month t in the most recent campaign. For any semi-month t in which no SIA activity is carried out, we set $\delta_t = 0$; otherwise, an estimate of δ_t can be obtained using the target population information from the SIA calendar on the World Health Organization (WHO) website (World Health Organization, 2020c). This formulation of S_t^* allows for flexible modeling of susceptible dynamics when an SIA campaign is implemented in several phases with part of the total target population covered in each phase (e.g., by geographic regions).

4.2.2 Computation

Morton and Finkenstädt (2005) described a Markov chain Monte Carlo (MCMC) algorithm for making inference about parameters in a TSIR model, but acknowledged that this approach can perform poorly in estimating the reporting rate parameter ρ. In preliminary work we also encounter serious unidentifibility issues when we estimate ρ jointly with all other model parameters. Therefore, we propose a two-stage hybrid approach that estimates the reporting rate parameter separately from the rest of the model parameters. Specifically, we estimate the reporting rate parameter in the first stage via an ordinary least squares (OLS) regression model with robust standard errors. The results are then used to estimate

the rest of the model parameters in the second stage via MCMC. We now describe each stage in detail.

Stage 1: reporting rate estimation via OLS regression with robust standard errors

The reporting rate estimation step is inspired by the susceptible reconstruction method developed by Finkenstädt and Grenfell (2000). We consider a time period during which **no SIA is implemented**. Let the first month of this time period be $m = 1$ and last month be $m = M$. Let C_m, I_m, S_m and B_m be the number of reported incidence, underlying true incidence, susceptible population and the adjusted births entering the susceptible pool in month m. Given a realization of the underlying monthly true incidence $\{I_m\}$ and the observed monthly reported incidence $\{C_m\}$, for $m = 1, \ldots, M$, we let ρ_m be the ratio

$$\rho_m = \frac{C_m}{I_m} \quad \text{such that} \quad I_m = \frac{1}{\rho_m} C_m = \kappa_m C_m$$

where κ_m is the inverse of ρ_m.

Now, we consider the underlying susceptible dynamics during this time period. Since no SIA is implemented, Equation (4.6) implies that the monthly susceptible time series follows

$$\begin{aligned} S_m &= S_{m-1} + B_m - I_m \\ &= S_{m-1} + B_m - \kappa_m C_m \\ &= (S_{m-2} + B_{m-1} - \kappa_{m-1} C_{m-1}) + B_m - \kappa_m C_m \\ &\vdots \\ &= S_0 + \sum_{i=1}^{m} B_i - \sum_{i=1}^{m} \kappa_i C_i \end{aligned}$$

where S_0 is the number of susceptibles right before month 1. Assuming the ratio between the reported incidence and the underlying true incidence ρ_m is roughly equal to ρ during the

entire time period, that is, $\rho_m \approx \rho$ and $\kappa_m = \frac{1}{\rho_m} \approx \frac{1}{\rho} = \kappa$, we have, approximately

$$S_m = S_0 + \sum_{i=1}^{m} B_i - \sum_{i=1}^{m} \kappa_i C_i \tag{4.9}$$

$$\approx S_0 + \sum_{i=1}^{m} B_i - \kappa \sum_{i=1}^{m} C_i \tag{4.10}$$

where $\kappa = \frac{1}{\rho}$. Let $\bar{S}$ be the average of $\{S_m\}$ during the time period $m = 1, \ldots, M$. We can write equation (4.10) as

$$\sum_{i=1}^{m} B_i = (S_m - S_0) + \kappa \sum_{i=1}^{m} C_i \tag{4.11}$$

$$= (\bar{S} - S_0) + \kappa \sum_{i=1}^{m} C_i + (S_m - \bar{S}) \tag{4.12}$$

Let $Y_m = \sum_{i=1}^{m} B_i$ be the cumulative adjusted births, $X_m = \sum_{i=1}^{m} C_i$ be the cumulative reported incidence, $\beta_0 = \bar{S} - S_0$ be the difference between S_0 and $\bar{S}$, and $U_m = S_m - \bar{S}$ be the deviation of S_m from the average $\bar{S}$. Then, we can rewrite equation (4.12) as

$$Y_m = \beta_0 + \kappa X_m + U_m \tag{4.13}$$

Assuming the deviations U_m follows a Normal distribution with mean 0 and variance σ_m^2, Equation (4.13) motivates an OLS regression model with robust standard errors (White, 1980) for estimating κ:

$$Y_m = \beta_0 + \kappa X_m + U_m \tag{4.14}$$

$$U_m \sim \text{Normal}(0, \sigma_m^2) \tag{4.15}$$

As such, we can obtain a point estimate $\hat{\kappa}_{\text{OLS}}$ and a robust standard error $\hat{\sigma}_\kappa$ for the inverse of the reporting rate $\kappa = \frac{1}{\rho}$ by fitting an OLS regression model with robust standard errors using the cumulative adjusted births as the outcome variable and the cumulative reported incidence as the predictor variable. We can acknowledge the uncertainty in the reporting rate parameter ρ by drawing random samples from the distribution $\text{N}\left(\hat{\kappa}_{\text{OLS}}, \hat{\sigma}_\kappa^2\right)$ and taking the inverse of the samples.

Stage 2: MCMC for Bayesian hierarchical model

In the second stage, we estimate the other model parameters and the latent variables I_t using an MCMC algorithm that is similar to the one outlined by Morton and Finkenstädt (2005). We use diffuse priors and derive the posterior conditional distribution for each model parameter given data and ρ. Note that the posterior conditional distribution of any I_t involves all future values of I and the other parameters, because I_t influences all S_j, $j > t$, from Equation (4.7). To propagate the uncertainty associated with the estimation of ρ from the first stage, we plug in a random value for ρ in each MCMC iteration by taking the inverse of a random sample drawn from the distribution $\mathrm{N}\left(\hat{\kappa}_{\mathrm{OLS}}, \hat{\sigma}^2_{\kappa}\right)$.

We use the Metropolis-Hasting algorithm to draw posterior samples and adopt the following guidelines:

1. We use independent proper diffuse priors, such as the Normal, Uniform or Beta distribution, depending on the range of the parameters.

2. The starting values for I_t and θ must be realizable, i.e. $C_m \leq I_{t(m)} + I_{t(m)+1}$ and $I_{t+1} \leq S_t \leq N_t$ for each m and t. Any proposed move such that these rules are violated is rejected.

3. We update each I_t individually using a discrete proposal distribution as outlined by Wakefield et al. (2011). After updating an I_t, all subsequent values in the susceptible series S_t are updated as well.

4. In preliminary work we saw that the posterior samples of θ, γ_1 and γ_2 tend to be highly correlated. We use the blocked Metropolis sampling technique (Wakefield, 2013) to update them together for more efficient convergence.

4.3 Simulation study

In this section, we illustrate our method and investigate the effect of uncertainty propagation via a simulation study considering a range of reporting rates, from extremely low to high ($\rho = 0.01, 0.1, 0.3, 0.5$ and 0.7).

4.3.1 Set up

We simulated semi-monthly population and live births time series assuming a yearly population growth rate of 2.5% and a yearly birth rate of 4%, a setting that is common among LMICs (United Nations, 2019). To simulate the time series of adjusted births entering the susceptible pool, we assumed an RI-specific MCV1 coverage of 70% and a first-dose efficacy of 87% (World Health Organization, 2017a). We set the starting values of the incidence and susceptibles times series based on a simulation using the Epidemiological MODeling (EMOD) software developed by the Institute of Disease Modeling (Insitutue for Disease Modeling, 2020), and generated the subsequent I_t and S_t values based on the model described in Section 4.2 using the following true parameter values:

$$\gamma_1 = 3, \quad \gamma_2 = 0, \quad \gamma_3 = 0.2, \quad \gamma_4 = 0.5, \quad \beta^{\text{EN}} = -12, \quad \phi = 10$$

We allowed the number of infected and susceptible individuals to stochastically change for one year before collecting eight years of semi-monthly time series data. This resulted in a 5.6% of the total population being susceptible to measles at the first time point, i.e., $\theta = 0.056$. We let there be one SIA with an overall efficacy of $p = 40\%$ implemented in two phases at the beginning of the fourth year, with 50% of the total target population covered in each phase. Finally, for each reporting rate, we simulated a monthly observed incidence C_m based on Equation (4.1).

We used the simulated data from the first six years as the training set for model fitting and the data from the last two years (i.e., the forecast period) for forecast validation. To evaluate the model's ability to estimate the potential effect of a "planned" SIA campaign, we

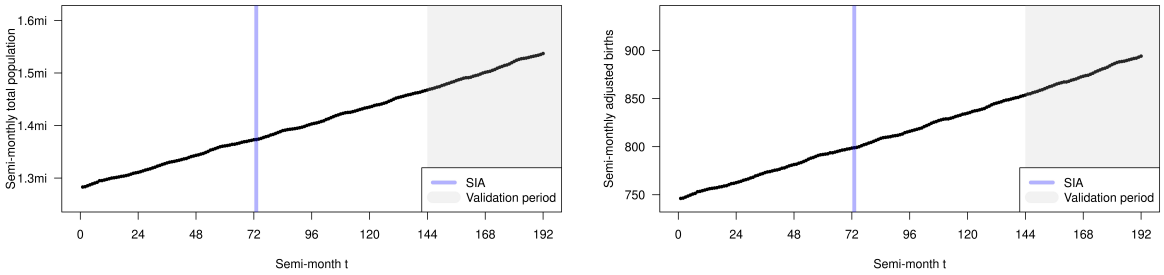

Figure 4.1: The simulated semi-monthly time series of total population (left) and adjusted births entering the susceptible population (right).

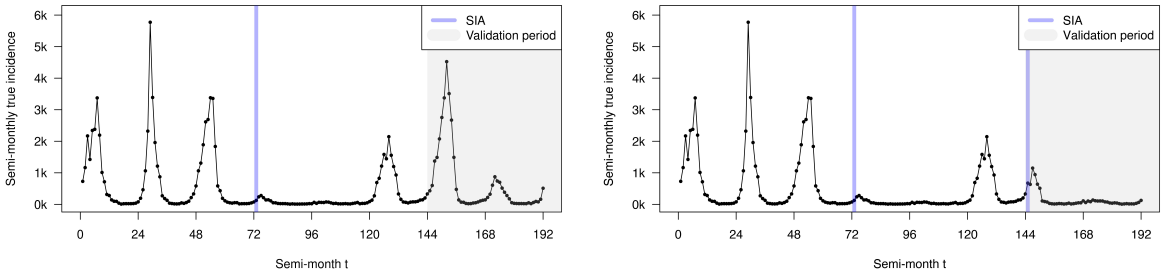

Figure 4.2: The simulated semi-monthly time series of underlying incidence, without (left) and with (right) a "planned" SIA at the beginning of the forecast period.

simulated an additional set of validation data assuming a second SIA with the same overall efficacy of $p = 40\%$ was implemented at the beginning of the forecast period.

The simulated total population and adjusted births entering the susceptible population are shown in Figure 4.1. The simulated time series of the underlying incidence and susceptible population, with and without a "planned" SIA at the beginning of the forecast period, are shown in Figures 4.2 and 4.3.

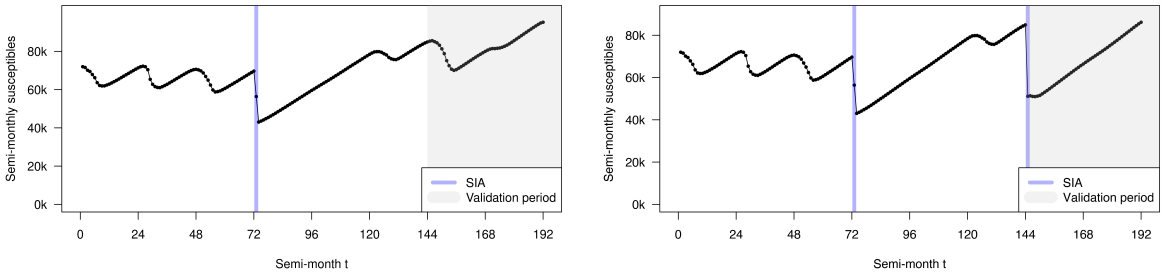

Figure 4.3: The simulated semi-monthly time series of underlying susceptible population, without (left) and with (right) a "planned" SIA at the beginning of the forecast period.

4.3.2 Computation and results

In each simulation corresponding to a reporting probability, we estimated the reporting rate ρ in the first stage using the training data from the time period before the first SIA. Figure 4.4 shows the point estimates and 95% confidence intervals of ρ obtained by fitting OLS regression models with robust standard errors to the cumulative monthly adjusted births and the cumulative monthly reported incidence. The estimates are calculated by drawing 10000 random samples from the distribution $\mathrm{N}\left(\hat{\kappa}_{\mathrm{OLS}}, \hat{\sigma}^2_{\kappa}\right)$ and taking the inverse of the samples. Under all reporting rate scenarios, the point estimate is higher than the true parameter value. Both the magnitude of the positive bias and the uncertainty associated with the estimation increases with the level of true reporting rate. A possible reason is the violation of the assumption that the ratio between the reported incidence and the underlying true incidence ρ_m is approximately equal to ρ throughout the entire time period, that is, $\rho_m \approx \rho$ and $\kappa_m = \frac{1}{\rho_m} \approx \frac{1}{\rho} = \kappa$. This would result in "errors" associated with the predictor variable (i.e., the cumulative reported incidence) that are not accounted for by the OLS regression model, and hence lead to "attenuation bias" in the estimates of κ. More discussion about this issue is provided in Section 4.5.

In the second stage of the estimation, we used all data from the first six years (i.e., the training set) to estimate the other model parameters and the underlying incidence and

susceptible dynamics. We drew 20000 posterior samples using the MCMC algorithm and the uncertainty propagation procedure described in Section 4.2.2, discarded the first 10000 samples as burn-in, and obtained the posterior medians and the 95% posterior credible intervals (CIs) of parameters. The traceplots of the posterior samples are shown in Figures B.1 to B.5 in Appendix B.

To investigate the impact of uncertainty propagation associated with the reporting rate estimation in the first stage, we carried out an additional round of model fitting using the same training data without the uncertainty propagation step. Specifically, instead of using random samples of ρ drawn from the distribution estimated from the first stage, we plugged in the true reporting rate in each MCMC iteration in the second stage. This would allow us to see how much uncertainty in the model estimation and prediction can be attributed to the uncertainty in reporting rate estimation. The traceplots of the posterior samples obtained from this round of computation are shown in Figures B.6 to B.10 in Appendix B.

Figures 4.5 and 4.6 show the posterior medians and the 95% posterior CIs of the model parameters computed with and without uncertainty propagation under various reporting rate scenarios. When we estimate the model parameters with the true ρ values plugged in (blue triangles and segments), most of the resultant point estimates are close to the true parameter values and all 95% posterior CIs cover the truth. The model tends to overestimate the SIA efficacy parameter p. Within each parameter, the uncertainties associated with the estimates are quite similar across the various levels of reporting rate, except for when the reporting rate is extremely low ($\rho = 0.01$), in which case the uncertainties tend to be considerably larger than the other scenarios.

When we estimate the model parameters with the first-stage uncertainty propagated (black triangles and segments), the resultant estimates tend to have wider 95% posterior CIs than the plugged-in results. The increases in CI widths tend to be larger when the reporting rate is higher. The impact of uncertainty propagation is especially considerable when $\rho = 0.7$, in which case the uncertainty associated with the first-stage estimation of ρ is the greatest compared to the other scenarios. One exemption we see is the estimation of the

dispersion parameter ϕ, in which case the uncertainty propagation resulted in considerable decreases in both the point estimates and the posterior CI widths. One possible reason is that some of the variability in the underlying incidence induced by the uncertainty of the ρ estimates was attributed to overdispersion, which lead to lower estimated values of the dispersion parameter ϕ.

To predict the underlying incidence and susceptible population dynamics in the forecast period, we used the posterior samples of the model parameters from each MCMC iteration and generated the potential realizations of the underlying dynamics according to Equations (4.3) through (4.6). For the simulations where a second SIA was implemented at the beginning of the forecast period, we carried out the prediction assuming the "planned" SIA has the same (estimated) efficacy of the previous campaign.

Figures 4.7 and 4.8 show the simulated true values, posterior medians and 95% CIs/predictive intervals of the underlying measles incidence under various reporting rate scenarios, computed with and without uncertainty propagation and/or a "planned" SIA at the beginning of the forecast period. The corresponding results for the underlying susceptible population dynamics are shown in Figures 4.9 and 4.10. Under all reporting rate scenarios, the model successfully predicted the general timing and the relative magnitudes of the two measles outbreaks in the forecast period when no SIA is implemented, and captured the effect of a "planned" SIA if it were implemented at the beginning of the forecast period. However, the model was not able to predict the peaks of the two outbreaks accurately. The results shown in Figures 4.7 and 4.8 also demonstrate the impact of uncertainty propagation in incidence forecasting. Under all reporting rate scenarios, we see that the predictive intervals became wider after the first-stage estimation uncertainty is accounted for in the second stage, and the magnitude of the change increases with the first-stage uncertainty.

Finally, we computed the root mean squared error (RMSE) of the predicted underlying

incidence during the forecast period, RMSE_f, to assess the model's forecast accuracy:

$$\text{RMSE}_f = \sqrt{\frac{1}{T_f}\sum_{t=1}^{T_f}\left(\hat{I}_t - I_t\right)^2},$$

where I_t and $\hat{I}_t$ are the "true" simulated incidence and the posterior median of the predicted incidence at time t, and $T_f = 48$ is the total number of semi-monthly time steps in the forecast period.

Figure 4.11 shows the results computed with (black) and without (blue) uncertainty propagation from the reporting rate estimation in the first stage. When the true ρ is plugged-in in the second stage estimation, the RMSE_f is the highest when the reporting rate is extremely low ($\rho = 0.01$). As reporting rate increased, the forecast accuracy first measured by the RMSE_f first improved sharply, and then remained relatively stable. The impact of uncertainty propagation on RMSE_f is small under under reporting rate scenarios, except for when the reporting rate is high ($\rho = 0.7$) and the fist-stage uncertainty and bias are large. In this case, the RMSE_f value computed with uncertainty propagation is much higher than the plugged-in value.

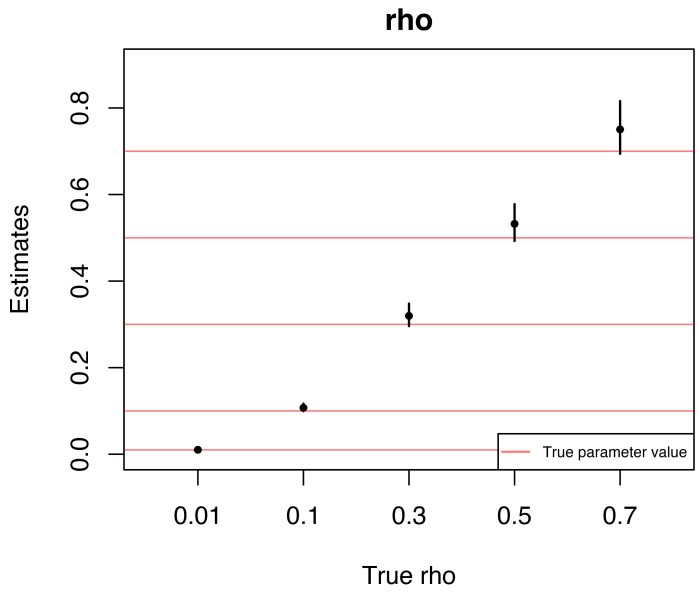

Figure 4.4: The point estimates and 95% confidence intervals of the reporting rate parameter ρ obtained by fitting OLS regression models with robust standard errors to the cumulative monthly adjusted births and the cumulative monthly reported incidence before the first SIA. The estimates are calculated by drawing 10000 random samples from the distribution $\mathrm{N}\left(\hat{\kappa}_{\mathrm{OLS}}, \hat{\sigma}_{\kappa}^{2}\right)$ and taking the inverse of the samples. When $\rho = 0.01$, the point estimate is 0.0102 and the 95% confidence interval is $(0.0095, 0.0111)$. When $\rho = 0.1$, the point estimate is 0.107 and the 95% confidence interval is $(0.099, 0.117)$.

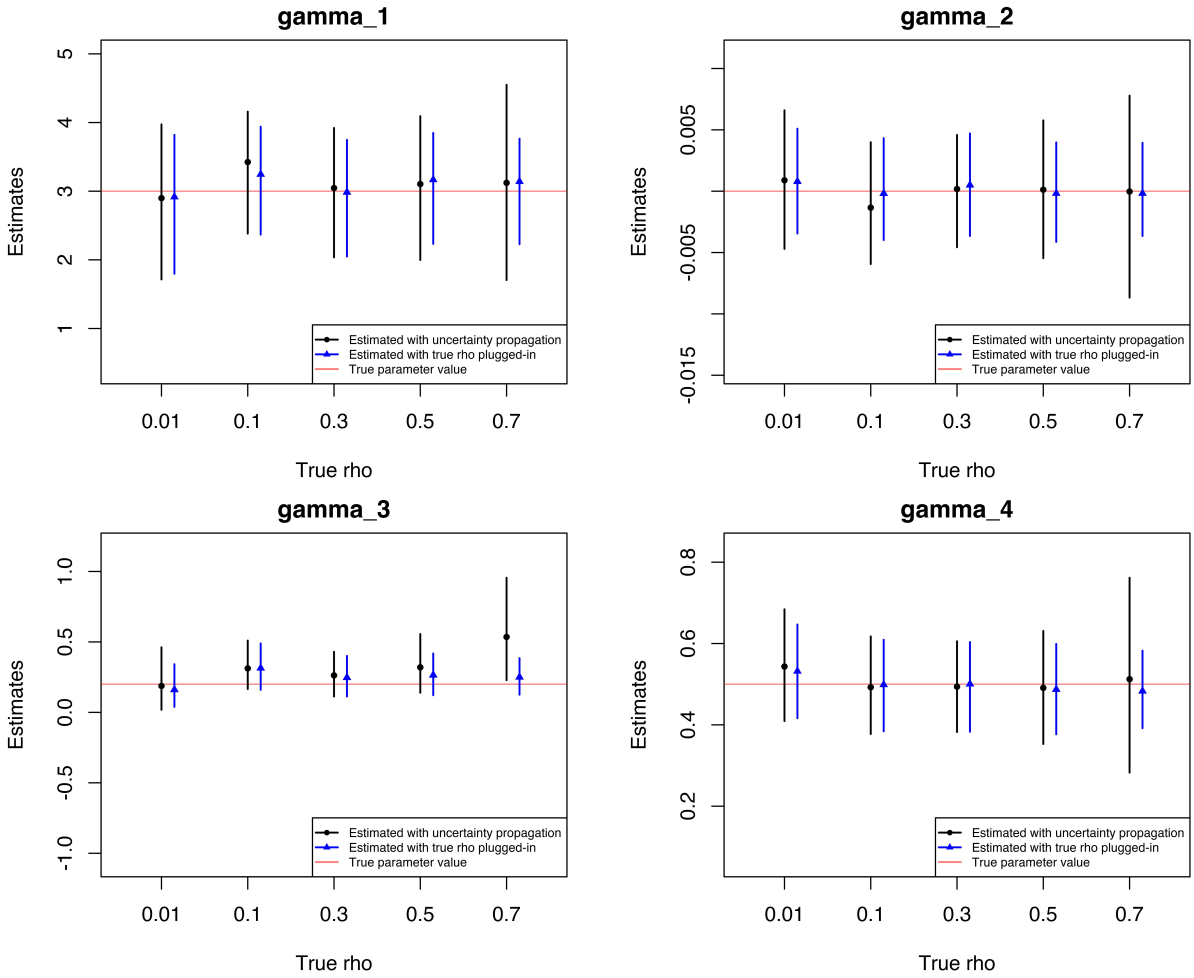

Figure 4.5: The posterior medians and the 95% posterior credible intervals (CIs) of γ_1, γ_2, γ_3 and γ_4 for when the reporting rate is 0.01, 0.1, 0.3, 0.5 and 0.7 (along the x-axis), computed with (black) and without (blue) uncertainty propagation from the reporting rate estimation in the first stage.

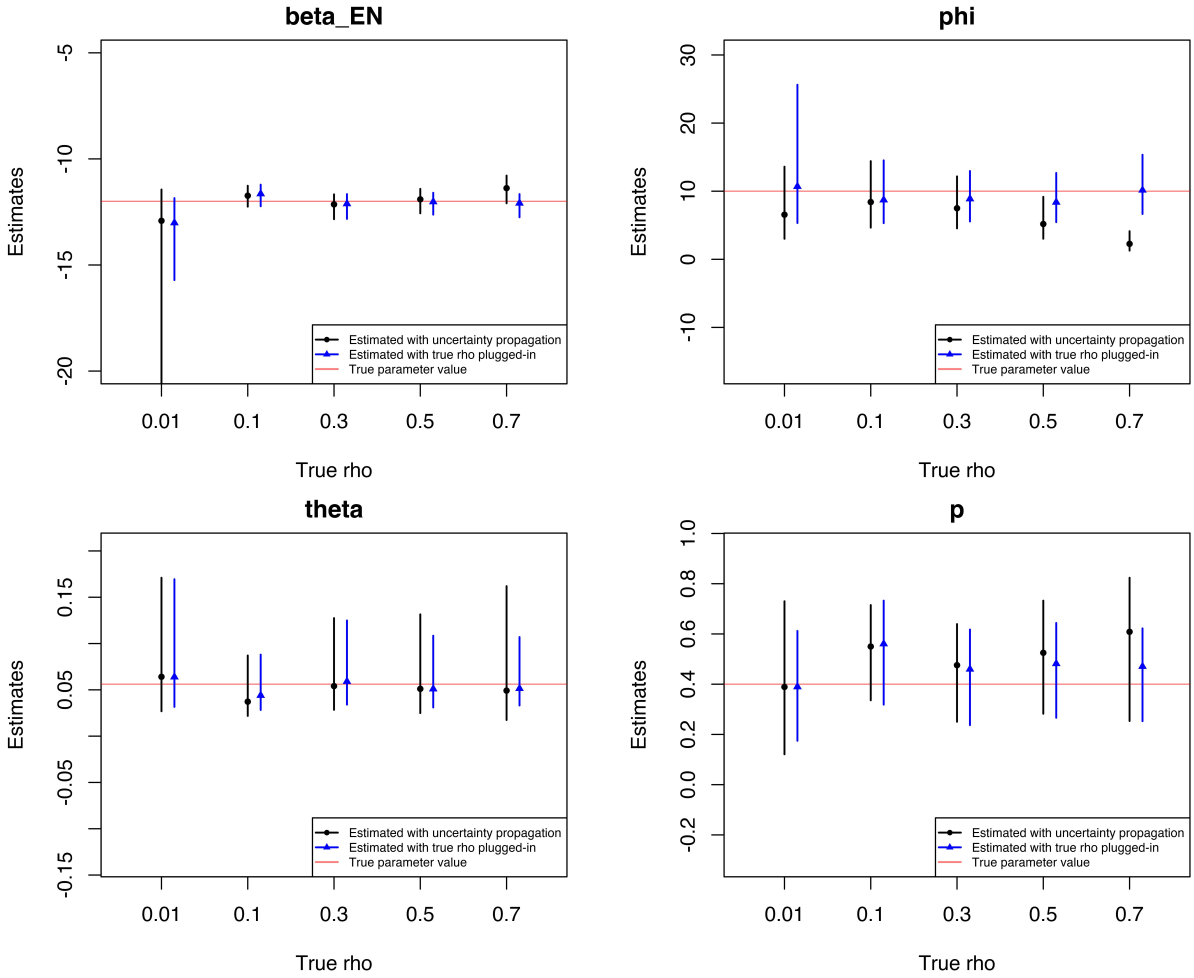

Figure 4.6: The posterior medians and the 95% posterior credible intervals (CIs) of β^{EN}, ϕ, θ and p for when the reporting rate is 0.01, 0.1, 0.3, 0.5 and 0.7 (along the x-axis), computed with (black) and without (blue) uncertainty propagation from the reporting rate estimation in the first stage.

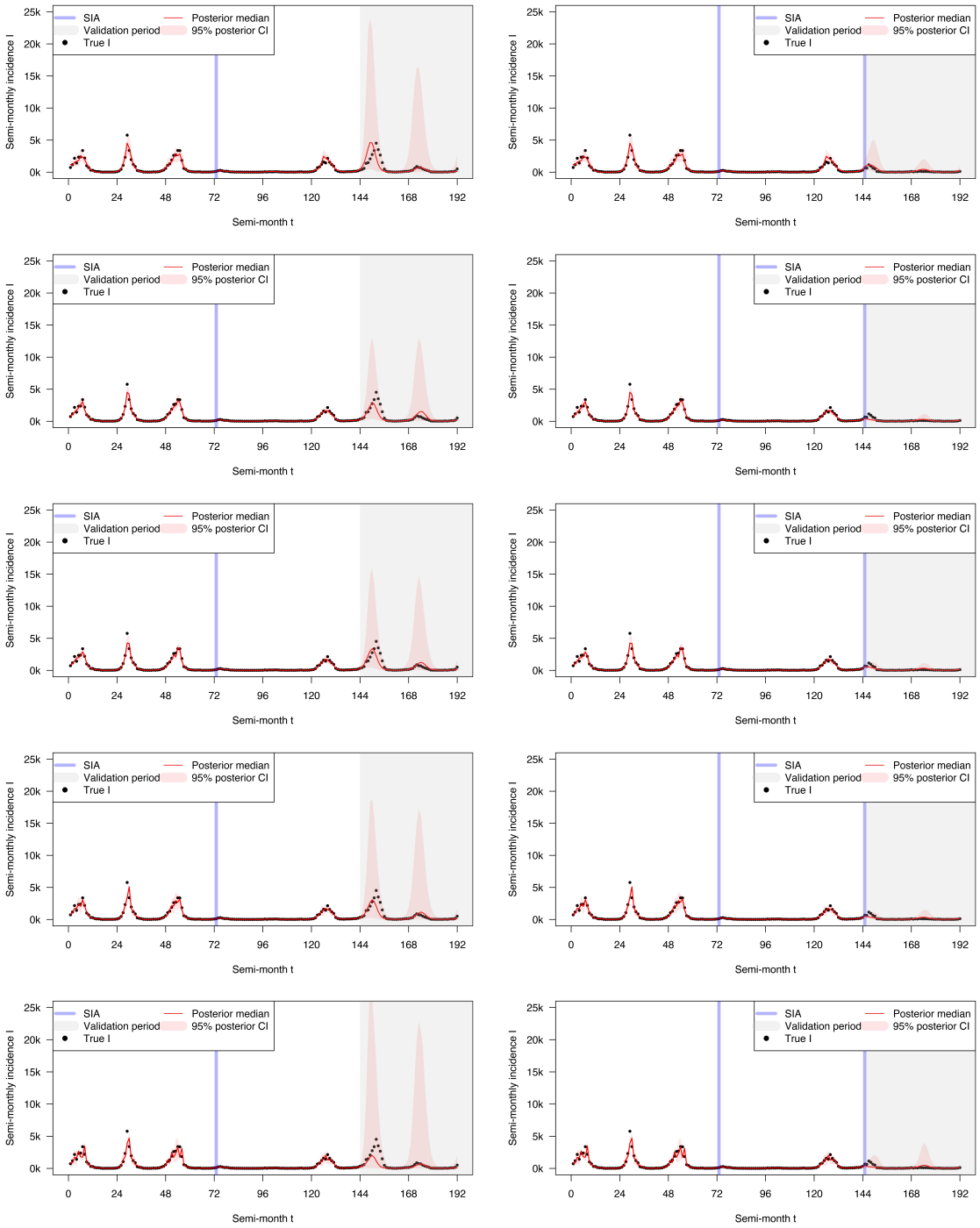

Figure 4.7: The simulated true values, posterior medians and 95% posterior CIs/predictive intervals of the underlying measles incidence computed with uncertainty propagation for when the reporting rate is 0.01, 0.1, 0.3, 0.5 and 0.7 (from top to bottom), without (left) and with (right) a "planned" SIA at the beginning of the forecast period, respectively.

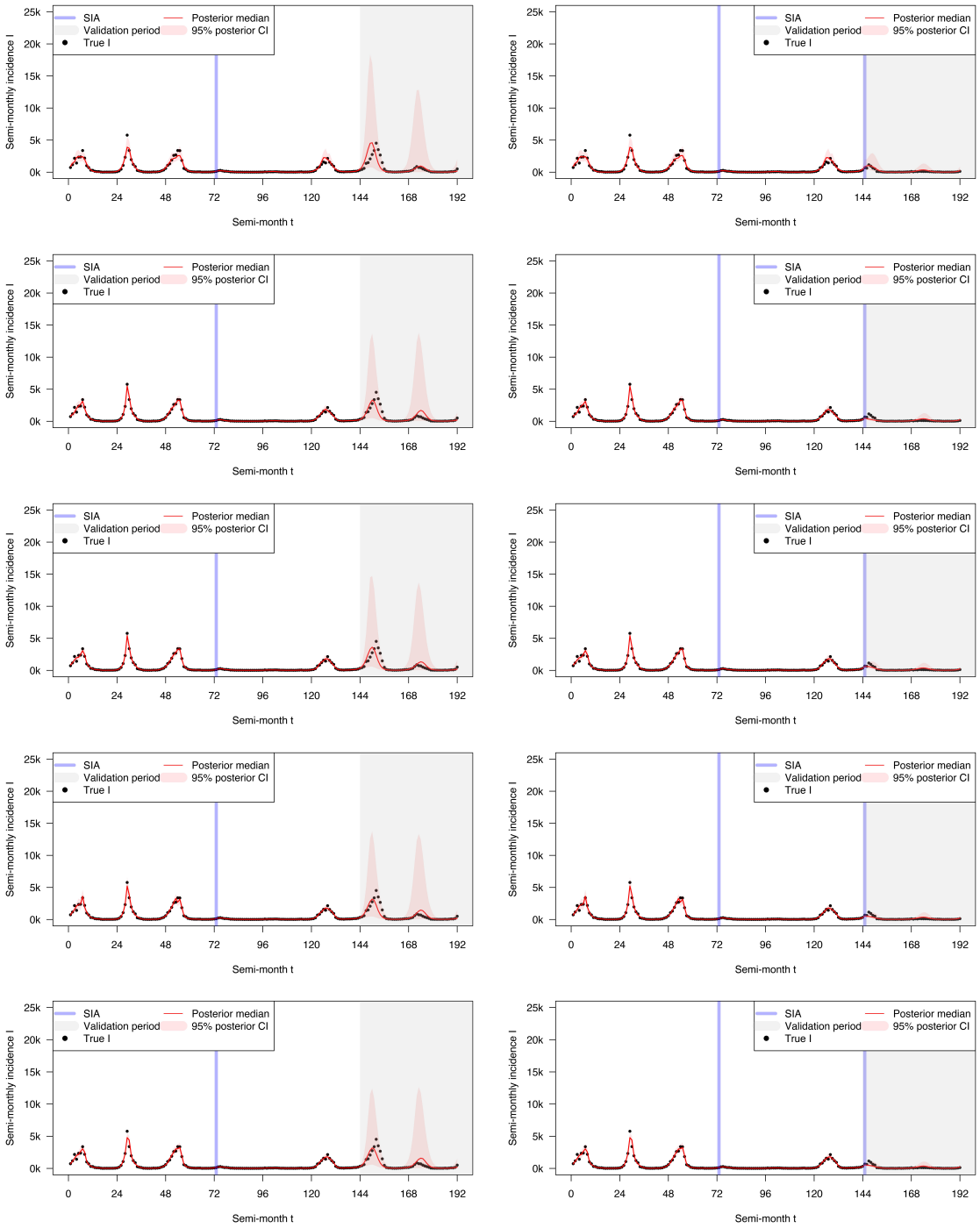

Figure 4.8: The simulated true values, posterior medians and 95% posterior CIs/predictive intervals of the underlying measles incidence computed without uncertainty propagation (i.e., true ρ plugged in) for when the reporting rate is 0.01, 0.1, 0.3, 0.5 and 0.7 (from top to bottom), without (left) and with (right) a "planned" SIA at the beginning of the forecast period, respectively.

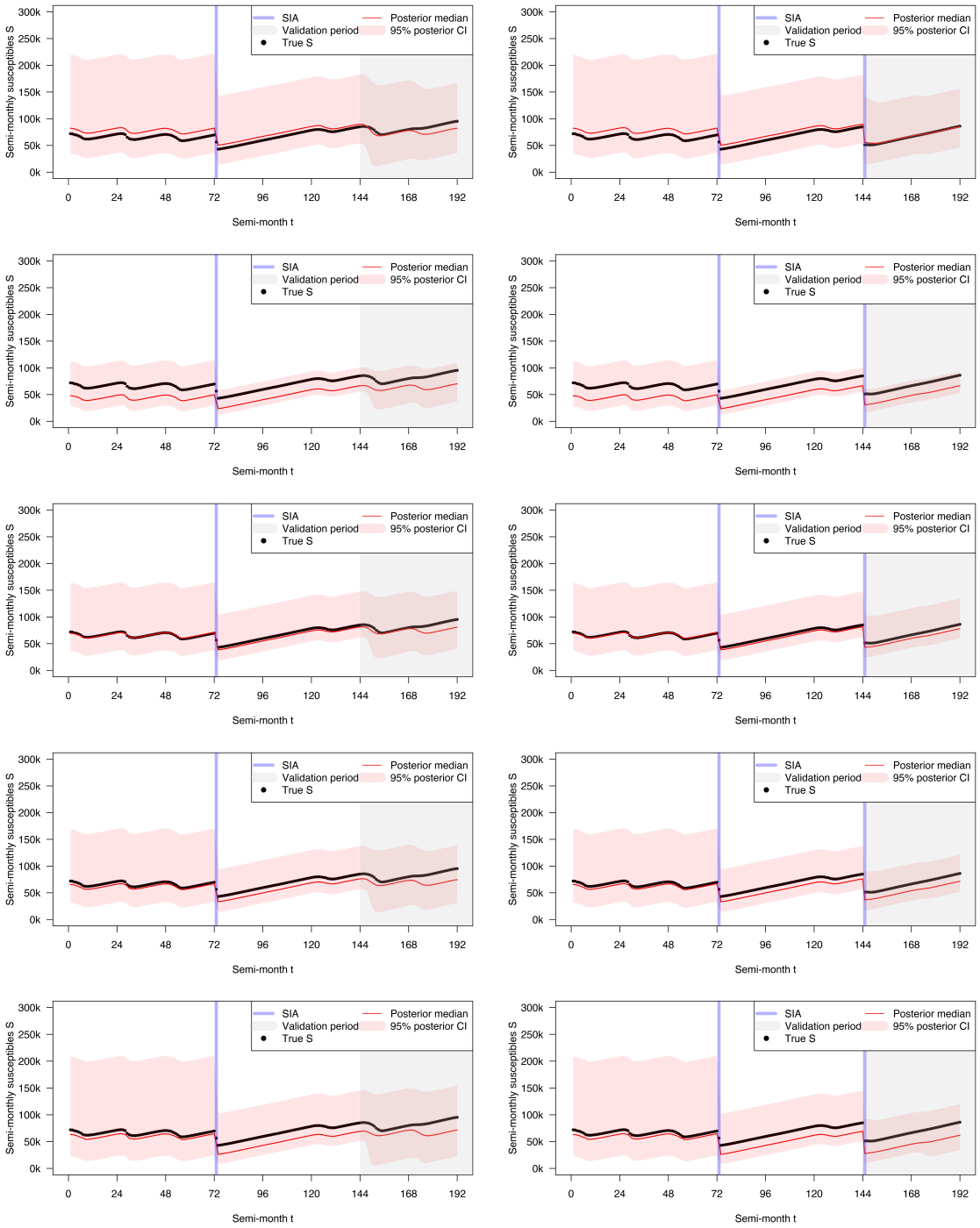

Figure 4.9: The simulated true values, posterior medians and 95% posterior CIs/predictive intervals of the underlying susceptible population computed with uncertainty propagation for when the reporting rate is 0.01, 0.1, 0.3, 0.5 and 0.7 (from top to bottom), without (left) and with (right) a "planned" SIA at the beginning of the forecast period, respectively.

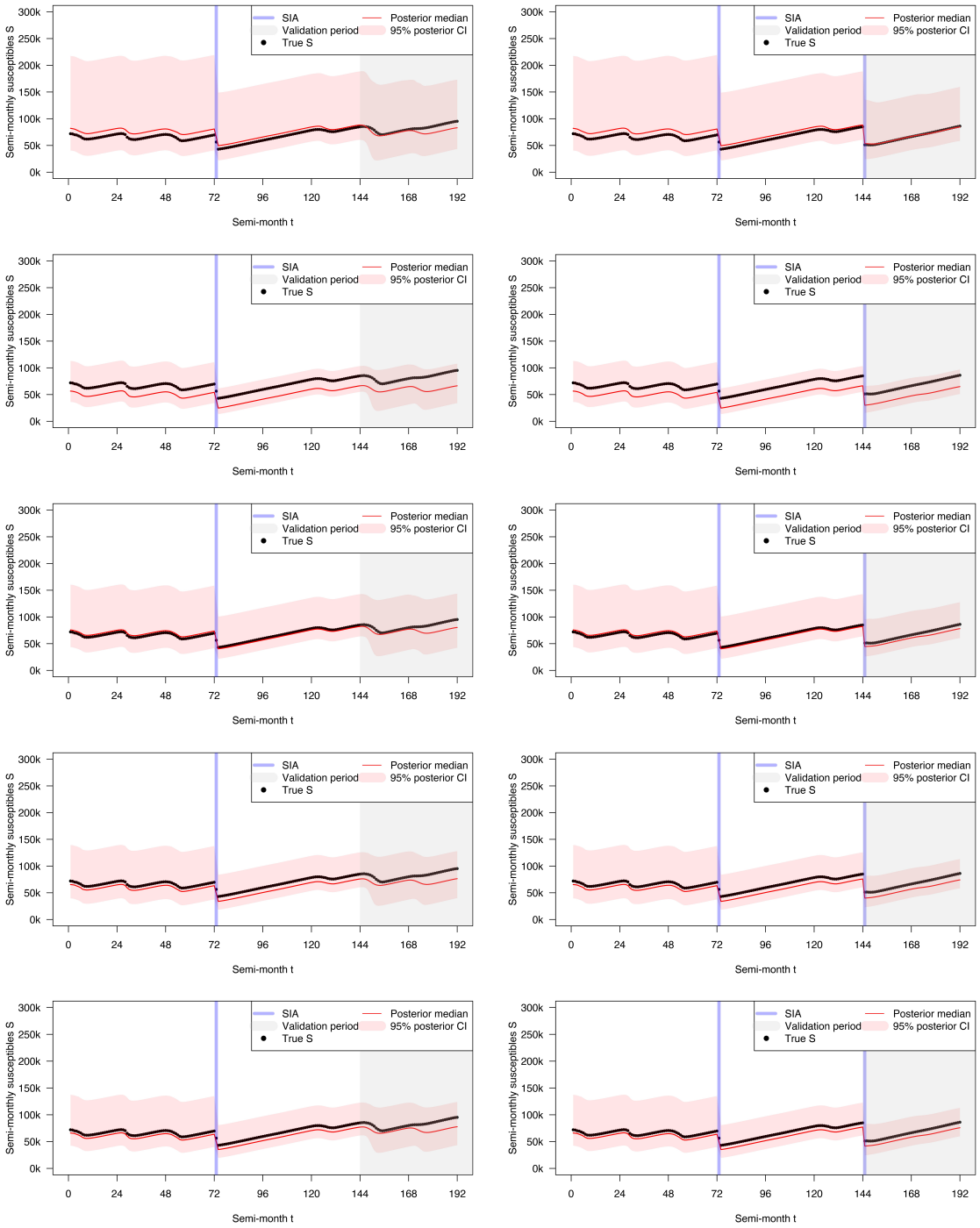

Figure 4.10: The simulated true values, posterior medians and 95% posterior CIs/predictive intervals of the underlying susceptible population computed without uncertainty propagation (i.e., true ρ plugged in) for when the reporting rate is 0.01, 0.1, 0.3, 0.5 and 0.7 (from top to bottom), without (left) and with (right) a "planned" SIA at the beginning of the forecast period, respectively.

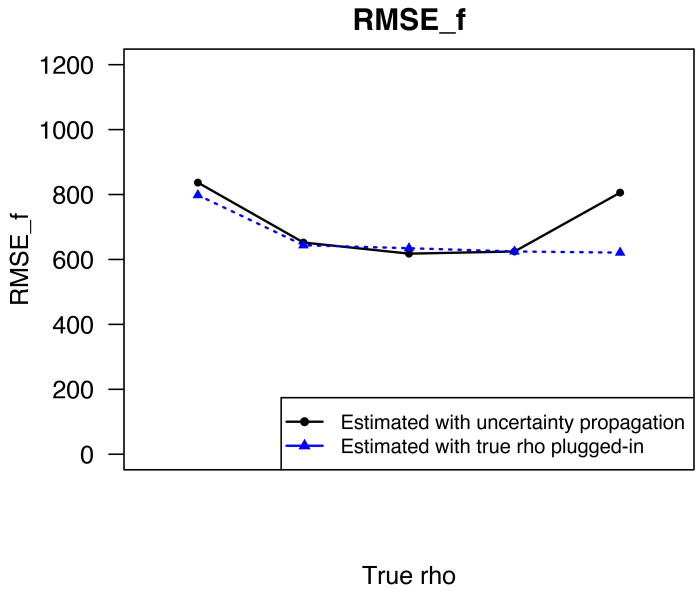

Figure 4.11: The root mean squared error of the underlying incidence in the forecast period, RMSE_f, calculated using the posterior medians of the semi-monthly incidence estimated with (black) and without (blue) uncertainty propagation from the reporting rate estimation in the first stage.

4.4 Analysis of the measles incidence data in Benin

In this section, we illustrate our method by analyzing the monthly reported measles incidence in Benin between 2012 and 2018. Specifically, we fit our model using the data collected between January 2012 and December 2016, estimate the efficacy of the one SIA campaign carried out in this time period, and forecast the reported incidence between January 2017 and December 2018 for model validation.

4.4.1 Data

We obtained the yearly national population estimates and crude birth rates of Benin from 2011 to 2018 from the WorldBank database (The World Bank, 2019). The estimates were linearly interpolated over time to calculate the semi-monthly total population and live-births

time series on the national level. To obtain the adjusted births entering the susceptible pool, we first calculated the department-level (administrative level 1 in Benin) live-births using a set of proportions derived from the WorldPop estimates (WorldPop, 2017). We used the infant mortality rate estimates from the Institute for Health Metrics and Evaluation (IHME) (Institute for Health Metrics and Evaluation, 2019) to approximate the under-9-month mortality rate in each department and calculated the number of 9-month-old children at each time point. Using the space-time smoothing model developed in Chapter 3 (Dong and Wakefield, 2020), we also obtained the department-level RI-specific MCV1 coverage with data from the Multiple Indicator Cluster Surveys (MICS) conducted in 2006, 2012 and 2017 (UNICEF, 2020) and the Demographic and Health Survey (DHS) conducted in 2014 (ICF, 2020a). Finally, we calculated the adjusted births time series for each of the 12 departments in Benin assuming the efficacy of the first dose of measles vaccine is 87% (World Health Organization, 2017a) based on Equation (4.7), and aggregated them to obtain the adjusted births on the national level.

The monthly reported measles incidence and the SIA calendar were downloaded from the WHO database (World Health Organization, 2020c). There was one national SIA campaign implemented in two phases during the time period of interest: 2.6 million children aged between 9 months to 9 years were first targeted at the beginning of December 2014, then another 0.4 million children of the same age group were targeted at the end of January 2015. The reported incidence time series is shown in Figure 4.12.

4.4.2 Computation and results

For reporting rate estimation, we fitted an OLS regression model with robust standard errors using the cumulative adjusted births and cumulative reported incidence data collected between January 2012 and December 2014. We drew 10000 random samples from the distribution $\mathrm{N}\left(\hat{\kappa}_{\mathrm{OLS}}, \hat{\sigma}^2_{\kappa}\right)$ and took the inverse of the samples to obtain estimates for the reporting rate parameter ρ. The resultant estimates for ρ are extremely low, with a point estimate of 0.0074 and a 95% confidence interval of (0.0069, 0.0081). This suggested that there was

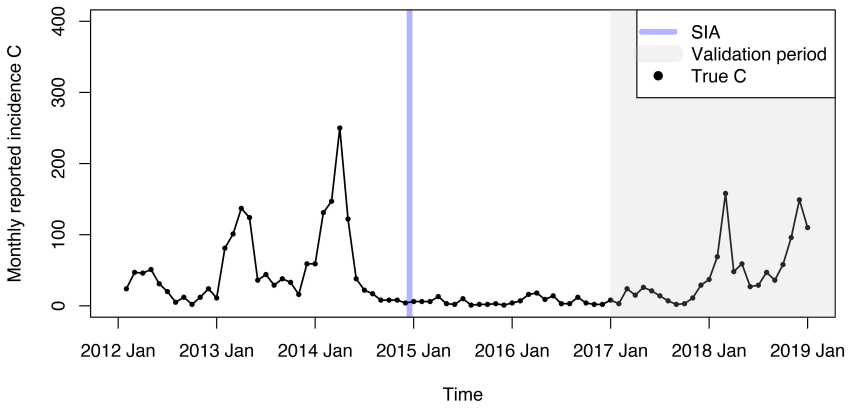

Figure 4.12: The monthly reported measles incidence in Benin between January 2012 and December 2018.

severe under-reporting in Benin's measles surveillance data during this time period.

We carried out the second stage of estimation described in Section 4.2.2 with uncertainty propagation using data collected between January 2012 and December 2016. We drew a total of 20000 posterior samples, discarded the first 10000 samples as burn-in, and obtained the posterior medians and the 95% posterior CIs of the model parameters and the underlying incidence time series. The traceplots of the posterior samples are shown in Figures B.1 to B.5 in Appendix B.

Table 4.1 shows the summary results for the parameter estimation. In particular, the estimates for θ implied that an estimated 2.8% (95% CI: 0.5%, 14.9%) of the total population was susceptible to measles at the beginning of year 2012. As for SIA efficacy, the point estimate of p suggested that an estimated 49.9% of the susceptible population were immunized after the SIA implemented in 2014–15. However, the 95% posterior CI of p is very wide (14.5%, 85.9%), indicating that the estimated efficacy was associated with huge uncertainty. This is expected considering the extremely low reporting rate estimates from the first stage and the relatively short time series we used for this analysis.

Finally, we predicted the underlying incidence and susceptible population dynamics in

the forecast period by generating the potential realizations of the underlying dynamics using the posterior samples of the model parameters from all MCMC iterations. In Figure 4.13, we compare the posterior medians and 95% CIs/predictive intervals of the underlying monthly measles incidence, scaled by the estimated reporting rate, to the reported incidence time series in Benin between 2012 and 2018. The model was able to predict the general timing of the first two outbreaks and their relative magnitudes in the forecast period, but failed to predict the occurrence or the magnitude of the third outbreak. The estimated underlying incidence and susceptible population dynamics are also shown in Figure 4.13 for reference.

Parameter	Estimate	95% CI
ρ	0.0074	(0.0069, 0.0081)
γ_1	3.08	(2.10, 5.34)
γ_2	-0.005	(-0.016, 0.003)
γ_3	0.173	(-0.002, 0.376)
γ_4	0.248	(0.080, 0.438)
β^{EN}	-11.2	(-12.5, -10.5)
θ	0.028	(0.005, 0.149)
p	0.499	(0.145, 0.859)
ϕ	4.89	(2.48, 9.87)

Table 4.1: The posterior medians and the 95% posterior CIs of model parameters from the Benin analysis. The estimates for ρ were obtained by fitting an OLS regression model with robust standard errors to the cumulative monthly adjusted births and the cumulative monthly reported incidence before December 2014. The estimates are calculated by drawing 10000 random samples from the distribution $\mathrm{N}\left(\hat{\kappa}_{\text{OLS}}, \hat{\sigma}^2_{\kappa}\right)$ and taking the inverse of the samples.

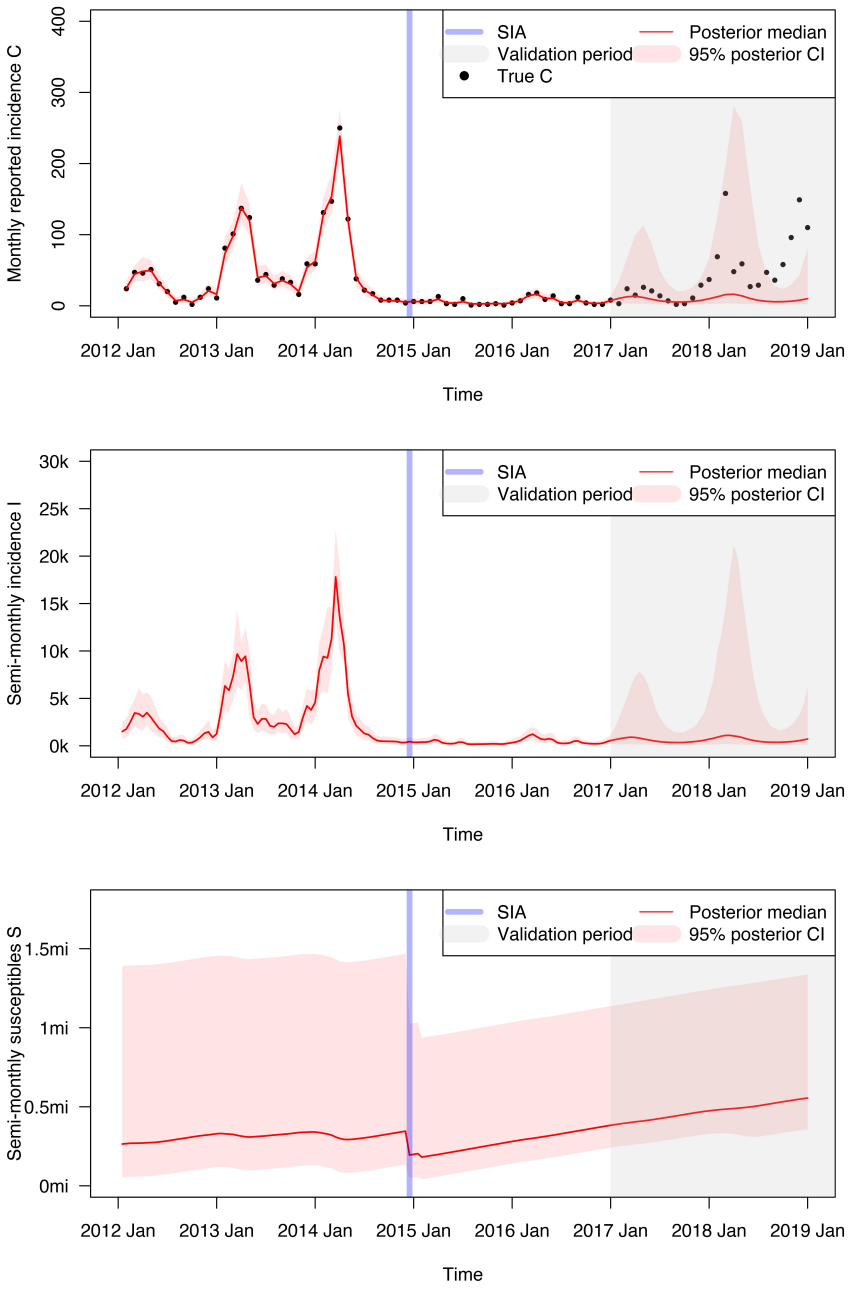

Figure 4.13: The posterior medians and the 95% posterior CIs/predictive intervals of (1) Top: the underlying monthly measles incidence scaled by the estimated reporting rate (with the true reported incidence time series); (2) Middle: the underlying semi-monthly true incidence; and (3) Bottom: the underlying semi-monthly susceptible population in Benin between 2012 and 2018.

4.5 Discussion

In this chapter, we developed a discrete-time hidden Markov model under the TSIR framework to estimate SIA efficacy using measles incidence time series data. Our approach accounts for under-reporting and seasonality of measles transmission, and accommodates monthly reported incidence data that is publicly available from the WHO database. We proposed a two-stage estimation procedure that first estimates the reporting rate through an OLS regression model with robust standard errors and then estimates the rest of the model parameters and the underlying incidence dynamics via an MCMC algorithm with uncertainty propagation. The proposed method can be used to estimate the fraction of susceptible people immunized by past SIAs and forecast incidence trends in the future under various hypothetical SIA scenarios.

We illustrated our method and investigated the impact of uncertainty propagation on parameter estimation and incidence prediction via a simulation study considering a range of reporting rates. In addition, we also applied our method to analyze the reported measles incidence data in Benin from 2012 to 2018. While both the simulation study and the Benin example demonstrate our model's ability to predict the general timing and relative magnitudes of measles outbreaks in the near future, they also revealed some of the limitations of our approach. As shown in Section 4.3.2, our reporting rate estimation method produces biased results, especially when the reporting rate is high. A possible cause of this issue is that the derivation of the OLS procedure outlined in Section 4.2.2 was based on a key approximation: the ratio of the reported cases to the underlying cases at each time point is approximately equal to the reporting rate parameter ρ, i.e., $\rho_m \approx \rho$ and $\kappa_m = \frac{1}{\rho_m} \approx \frac{1}{\rho} = \kappa$ for each month m. However, this approximation may not hold if the empirical ratios $\rho_m = \frac{C_m}{I_m}$ deviate from the parameter ρ at one or more time points. Such deviations may arise due to the discreteness of C_m and I_m, the binomial variation in Equation (4.1), or measurement

errors in the reported incidence C_m. In this case, we could write Equation (4.9) as

$$\begin{aligned} S_m &= S_0 + \sum_{i=1}^{m} B_i - \sum_{i=1}^{m} \kappa_i C_i \\ &= S_0 + \sum_{i=1}^{m} B_i - \sum_{i=1}^{m} \kappa(1+\epsilon_i) C_i \\ &= S_0 + \sum_{i=1}^{m} B_i - \kappa \sum_{i=1}^{m} (C_i + \epsilon_i C_i) \\ \implies \quad \sum_{i=1}^{m} B_i &= (S_m - S_0) + \kappa \sum_{i=1}^{m} (C_i + \epsilon_i C_i) \end{aligned}$$

where $\epsilon_i = \frac{\kappa_i}{\kappa} - 1$ quantify the deviations between κ_i and κ. Using the previously defined notation, the above can be written as

$$\begin{aligned} Y_m &= \beta_0 + \kappa \sum_{i=1}^{m} (C_i + \epsilon_i C_i) + U_m \\ &= \beta_0 + \kappa \sum_{i=1}^{m} C_i^* + U_m \\ &= \beta_0 + \kappa X_m^* + U_m \end{aligned}$$

where $C_i^* = C_i + \epsilon_i C_i$ and $X_m^* = \sum_{i=1}^{m} C_i^*$. If each ϵ_i value is known, we can use $X_m^* = \sum_{i=1}^{m} C_i^* = \sum_{i=1}^{m} (C_i + \epsilon_i C_i)$ as the predictor variable in the OLS regression model to obtain an unbiased estimator for κ. Unfortunately, ϵ_i is unknown and its variability is difficult to estimate with the available data. Our current procedure, described in Section 4.2.2, uses the cumulative observed counts $X_m = \sum_{i=1}^{m} C_i$ as the predictor variable instead. As such, there would be "errors" in the predictor variable X_m that are not accounted for by the OLS regression model, resulting in an underestimate of the coefficient κ, known as the "attenuation bias" (Chesher, 1991; Griliches and Ringstad, 1970; Frost and Thompson, 2000; Fuller, 2009). Since κ is defined to be the inverse of ρ, this would translate into a positive bias in the estimates of ρ, which has been seen in our simulation study. What makes our case even more complicated is the fact that the predictor variable X_m is a cumulative quantity. As a result, the "errors" associated with X_m increase with m, making correcting the bias a very difficult problem that requires more in-depth research.

Another aspect of our model that could be improved is to allow temporal fluctuations in the reporting rate that may be influenced by time-varying factors such as the state of the epidemic and reports in the media. Finkenstädt and Grenfell (2000) described a spline-based method for estimating time-varying reporting rates, but the approach uses an ad-hoc algorithm for deciding the bandwidth and gave poor performance in some of our preliminary work. As Angela (2019) has pointed out, under-reporting corrections should be geographical- and disease-specific, as well as age- and gender-dependent. We are currently exploring alternative model specifications that can potentially incorporate additional information, such as measures of availability and reliability of surveillance system, to detect and account for temporal trends in reporting rates.

In our model, the semi-monthly time series of total population and adjusted births are input data that are treated as the truth — we consider them to be deterministic quantities with fixed values. A potential extension of our current method is to account for potential uncertainties in the population and births data and propagate these uncertainties in parameter estimation and incidence prediction. In addition, the balancing equation characterizing the susceptible dynamic currently assumes negligible non-measles deaths among the susceptible people. To relax this assumption, one can potentially add another component to the balancing equation to explicitly account for the people exiting the susceptible pool due to non-measles deaths. The viability and data requirement of this extension still requires further investigation.

It should be noted that our model was derived based on some key underlying assumptions that the TSIR models require. These include homogeneous mixing of the human host population and *frequency-dependent* contact rate (see Section 2.3). Therefore, the model is not suitable for analyzing data from extremely large geographical areas in which there is considerable heterogeneity in measles transmission dynamics. An interesting future research topic is to extend the current method to jointly model incidence from multiple areas while allowing for spatially and temporally correlated transmission patterns.

Chapter 5

MODELING AND PRESENTATION OF VACCINATION COVERAGE ESTIMATES USING DATA FROM HOUSEHOLD SURVEYS

5.1 Introduction

There has been an explosion in high-resolution map production for health and demographic indicators (Gething et al., 2015, 2016; Golding et al., 2017; Steele et al., 2017; Bosco et al., 2017; Osgood-Zimmerman et al., 2018; Graetz et al., 2018), including childhood vaccination coverage (Takahashi et al., 2017; Utazi et al., 2018, 2019a,b; Mosser et al., 2019; Utazi et al., 2020). The WorldPop project (WorldPop, 2020a) and the Institute for Health Metrics and Evaluation (IHME) (IHME, 2020a) are two major producers of vaccination coverage surfaces at fine spatial scales. Their estimates are used by researchers and policy makers from organizations across the globe (WorldPop, 2020b; IHME, 2020b).

Household surveys, such as the Demographic and Health Surveys (DHS) (ICF, 2020a) and the Multiple Indicator Cluster Surveys (MICS) (UNICEF, 2020), are a main data source for mapping vaccination coverage (Utazi et al., 2018, 2019a,b; Mosser et al., 2019). In recent DHS surveys, the groupings of households, known as clusters, are geo-referenced via their Global Positioning System (GPS) coordinates. Both WorldPop and IHME use model-based geostatistics (Diggle and Giorgi, 2019) to analyze cluster-level survey data. The core of their methods involves the following steps: match survey cluster locations to a collection of geospatial covariates, fit Bayesian spatial regression models to the cluster-level data using a continuous Gaussian Process (GP) model, and produce pixel-level vaccination coverage estimates on a fine grid. Often, the pixel-level estimates are aggregated by population density to form administrative-area-level estimates which are usually more relevant in the

context of immunization program monitoring and intervention planning.

This map production framework has a number of advantages. It allows for utilization of geospatial covariates from different sources, avoids the arbitrariness of the definition of neighbors in a discrete spatial model, allows combination of data with different geographical resolutions (Wilson and Wakefield, 2018), and allows coverage estimation for small administrative areas in which there are few or no survey clusters. However, it also has many challenges that need to be carefully addressed. First, the survey cluster locations are often randomly displaced within a certain radius (known as jittering) to preserve respondent confidentiality (ICF, 2020b). This makes it difficult to accurately match the survey clusters to geospatial covariates as the exact cluster locations are unknown. In addition, household surveys in low- and middle-income countries (LMIC) use multi-stage stratified cluster designs, with stratification by geographical region crossed with urban/rural. Not accounting for the survey design is well known to lead to biased estimates when stratification is ignored and anticonservative uncertainty intervals when clustering is not acknowledged (Paige et al., 2020). The appropriate acknowledgement of survey design in a model-based framework is challenging. Last, but not least, most household surveys are powered to provide reliable estimates at a particular sub-national scale (e.g., administrative-1 areas which are one below the national level). This implies that pixel-level vaccination coverage estimates are often associated with high uncertainties due to the sparsity of survey data, and care needs to be taken to appropriately present the estimates with their uncertainties.

In this chapter, we focus on two key aspects of the map production process: the acknowledgement of the survey design in modeling, and the appropriate presentation of estimates and their uncertainties. We use the coverage of the first dose of measles-containing-vaccine (MCV1) among children aged 12–23 months in Nigeria as a motivating example and illustrate modeling and presentation using data from the 2018 Nigeria Demographic and Health Survey (NDHS) (National Population Commission - NPC and ICF, 2019). In Section 5.2, we introduce the 2018 NDHS and present descriptive summary of the data. We then compare common models in Section 5.3, including the model used in Utazi et al. (2018), to illus-

trate approaches to modeling survey stratification and cluster-level non-spatial variation. A discussion of the trade-off between the geographical scale and precision of model-based estimates is presented in Section 5.4, along with a demonstration of visualization methods for mapping and ranking that emphasize the probabilistic interpretation of results. We also propose a novel approach to coverage map presentation that allows comparison and control of the overall map uncertainty level. In Section 5.5, we conclude with a general discussion and guidelines for map production and presentation.

5.2 Motivating example: the 2018 Nigeria DHS

The 2018 NDHS used a stratified, two-stage cluster design and a sampling frame derived from the Nigeria census conducted in 2006. Stratification was achieved by separating each of the 36 states and the Federal Capital Territory (i.e., the 37 administrative-1 areas, and hereafter referred to as Nigeria's 37 states) into urban and rural areas. Samples were selected independently in each stratum via a two-stage process: first, a pre-specified number of primary sampling units (PSUs), referred to as survey clusters, were selected from the sampling frame of census enumeration areas (EAs) with probability proportional to size; then, a fixed number of 30 households in every cluster were selected through equal probability systematic sampling. Due to the non-proportional allocation of the sample to the different states and the possible differences in response rates, each survey respondent is assigned a sampling weight to approximate the relative number of people he or she represents in the total population.

A total of 1389 survey clusters were selected to provide results representative at the national level as well as the state level. During the household listing operation, about 11 clusters were dropped as they were found to be insecure or vacated due to high levels of insurgency. In addition, due to extreme security issues in Borno state, 11 local government areas (LGAs, which are administrative-2 areas) where about 39% of Borno households reside, were dropped from the survey. Clusters selected from the dropped LGAs were replaced with other clusters from the remaining 16 LGAs in Borno. Consequently, state-level estimates for Borno are not representative for the dropped LGAs. In this paper, we focus on the coverage

of MCV1 among children aged 12–23 months based on evidence from either vaccination cards or caregiver recall. We remove the clusters that had no GPS location information or no eligible child samples and analyze data collected from the 5886 children aged 12–23 months in the remaining 1301 survey clusters.

5.2.1 Descriptive data analysis

We let i, c and k index states, survey clusters and children respectively and let y_{ick} be the 0-1 indicator of whether child k, sampled at cluster c in state i, has received MCV1 at the time of interview. We obtain the design-based Horvitz-Thompson (HT) direct estimates (Horvitz and Thompson, 1952) of the MCV1 coverage $\hat{p}_i$ for state i using

$$\hat{p}_i^{\text{HT}} = \frac{\sum_{c=1}^{m_i} \sum_{k=1}^{n_{ic}} y_{ick} \times w_{ick}}{\sum_{c=1}^{m_i} \sum_{k=1}^{n_{ic}} w_{ick}},$$

where m_i is the number of survey clusters in state i, n_{ic} is the total number of children sampled at cluster c in state i, and w_{ick} is the survey weight assigned to child k sampled at cluster c in state i. The survey weights are the reciprocal of the sampling probabilities with a non-response adjustment. The 90% confidence intervals (CIs) associated with the direct estimates can also be calculated using design-based variance estimators that acknowledge the stratification and clustering. All design-based estimation is implemented using the `survey` package (Lumley, 2004) in the `R` computing environment (Team, R Core, 2020).

Figure 5.1 shows the maps for the state-level HT direct coverage estimates $\hat{p}_i^{\text{HT}}$ and the widths of the associated 90% CIs. In general, the northern states of Nigeria have lower estimated MCV1 coverage than the southern states. Most 90% CI widths are relatively narrow, indicating moderate confidence in the state-level direct estimates.

In Figure 5.2, we show the spatial distribution of the observed MCV1 coverage among children aged 12–23 months as recorded at the cluster level. In general, the survey clusters in northern Nigeria tend to have lower observed MCV1 coverage than those in the south, and clusters that are closer to each other tend to have similar observed coverage – a sign of spatial correlation. However, it is not uncommon to see clusters in close proximity having

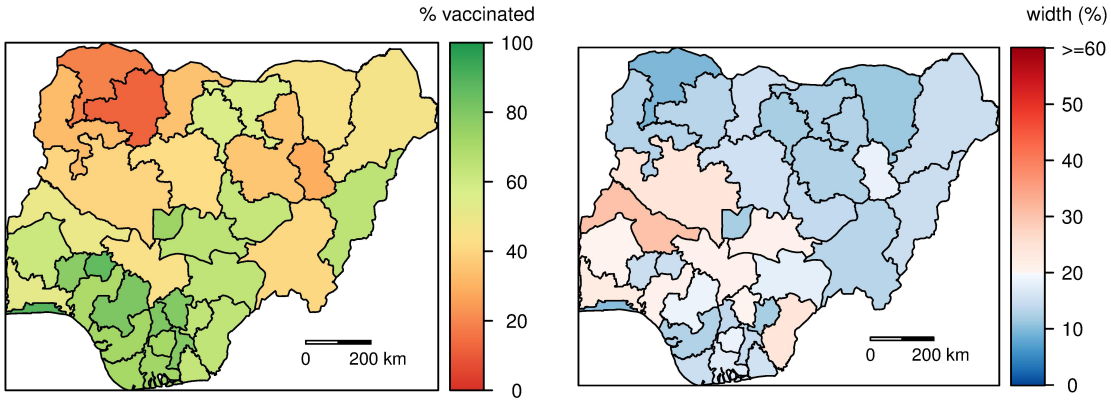

Figure 5.1: Left: state-level Horvitz-Thompson direct estimates of MCV1 coverage among children aged 12–23 months. Right: the widths of the associated 90% confidence intervals.

significantly different observed coverage, which could be outcome variation explained by factors other than spatial location and/or sampling variation.

5.2.2 Semi-variogram analysis

We take a closer look at the variation and spatial correlation in cluster-level observed coverage by examining the empirical semi-variogram (Diggle and Giorgi, 2019) based on the following non-spatial binomial mixed model:

$$\begin{aligned} Y_{ic}|p_{ic} &\sim \text{Binomial}\left(n_{ic}, p_{ic}\right) \\ \text{logit}(p_{ic}) &= \alpha + \boldsymbol{\beta}^{\top}\boldsymbol{x}_{ic} + C_{ic} \\ C_{ic} &\sim_{iid} \text{Normal}\left(0, \sigma_C^2\right), \end{aligned} \tag{5.1}$$

where $\sim_{iid}$ is short for "are independently and identically distributed as". Here, Y_{ic} is the random variable representing the number of vaccinated children out of n_{ic} who are sampled in cluster c of state i, p_{ic} is the probability of vaccination parameter in the binomial model that represents the cluster-level vaccination coverage, and $\boldsymbol{x}_{ic}$ is the vector of covariates associated with cluster c in state i. We use the same covariates selected by Utazi et al. (2018). These

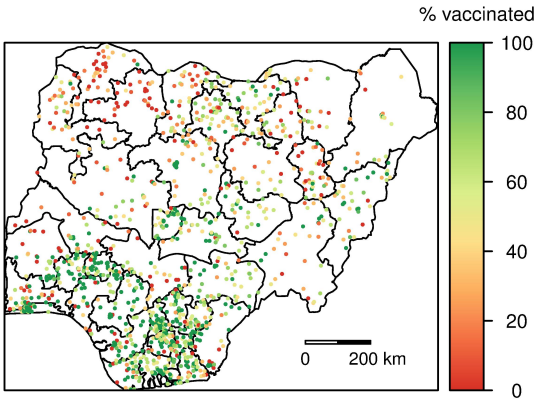

Figure 5.2: Spatial distribution of the observed MCV1 coverage among children aged 12–23 months as recorded at the 2018 NDHS cluster level. The cluster-level observed MCV1 coverage is calculated as the proportion of children sampled in a survey cluster who have had at least one dose of MCV at the time of interview, based on evidence from either vaccination cards or caregiver recall.

covariates are: poverty, aridity, log-transformed night-time lights, log-transformed travel time and enhanced vegetation index (EVI). Details of the data sources are provided in Appendix C. To extract the covariate values at each DHS cluster location and account for the jittering, we overlay the cluster GPS coordinates with covariate layers that have been standardized at a 1×1 km resolution and create 2 km and 5 km buffers around clusters in urban and rural areas respectively. The mean values of the grid cells within the buffers were extracted to be used as the covariate values for the cluster-level survey data. This procedure is adopted from the methods used by Utazi et al. (2018).

The semi-variogram is a standard tool for examining spatial dependence. Figure 5.3 shows the empirical semi-variogram based on the posterior medians of the random effects C_{ic}, estimated using the integrated nested Laplace approximation (INLA) approach (Rue et al., 2009), implemented in the `INLA` package in `R` (see Section 2.4.4). The semi-variogram shows an increasing trend with increasing distance between points, that flattens out at a range of

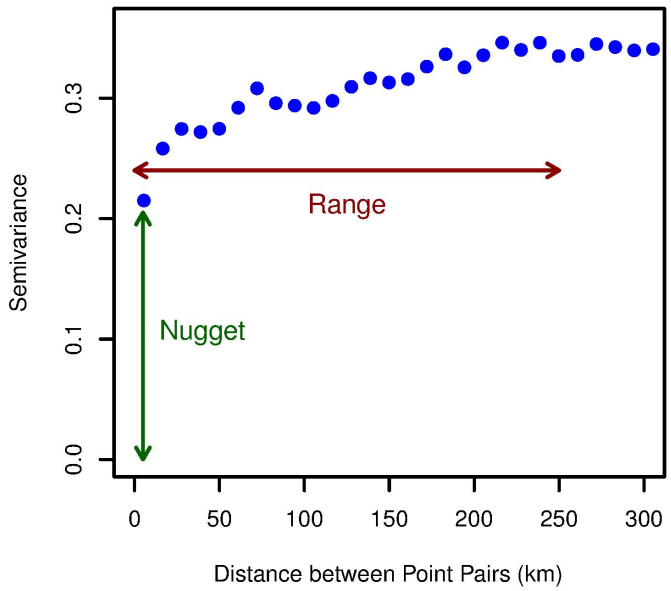

Figure 5.3: The empirical semi-variogram for MCV1 coverage in Nigeria based on the posterior medians of the random effects C_{ic} in the non-spatial binomial mixed model (5.1).

around 250 km, indicating that locations farther apart than that distance are not spatially correlated. In addition, the semi-variogram exhibits a considerable *nugget* effect when the spatial distance is very small. The *nugget* is typically associated with "measurement error" and small-scale variation (Diggle and Giorgi, 2019). This hints at the existence of non-spatial excess variation in the outcome at the cluster level, in addition to the variation explained by covariates and the spatial field.

5.3 Acknowledging the survey design in modeling

The stratified multi-stage cluster design adopted by the 2018 NDHS is ubiquitous among major household surveys that measure vaccination coverage. When fitting continuous spatial models to data from such surveys, it is essential to account for two key characteristics of the design: the stratification and clustering of the samples. A simulation study conducted by Paige et al. (2020) shows that explicitly accounting for survey stratification by including

urbanicity as a covariate in continuous spatial regression models improves predictions. If the response is associated with urbanicity then the improvement can be considerable. If any strong urban/rural association with the outcome is not properly modeled, large bias can result. WorldPop and IHME do not explicitly adjust for stratification in their models (Utazi et al., 2018, 2019a,b; Mosser et al., 2019; Utazi et al., 2020), but they do include extensive covariates which may, to some extent, implicitly adjust for the urban/rural stratification.

Acknowledging the clustering of survey samples is also challenging. WorldPop routinely ignore clustering in their models (Utazi et al., 2018, 2019a,b). IHME include an independent *nugget* error term in the linear predictor of cluster-level coverage estimate to capture the non-spatial excess variation in survey outcomes (Mosser et al., 2019). However, this approach has been used without explicit consideration of the mechanisms by which excess variation at the cluster level manifests itself. Traditionally, in the context of a continuous outcome, an estimated *nugget* effect has been attributed to measurement error or small-scale spatial variation (Diggle and Giorgi, 2019; Chiles and Delfiner, 2009). In the context of a vaccination coverage survey where Bernoulli sampling is carried out, measurement error would correspond to misclassification (willfully or by accident) of the binary outcomes. This undoubtedly occurs, but explicitly modeling the misclassification probabilities is difficult without gold standard data. In this paper, we focus on two other plausible forms of excess variation: (i) **within-cluster** variation that induces **overdispersion** at the cluster level, and (ii) **between-cluster** variation that represents **true signal** at the cluster level. The former is often referred to as excess-binomial variation, and the latter are "shocks" in the signal due to specific conditions at the cluster. One needs to carefully consider which form is adopted before choosing an estimation procedure that gives appropriate inference.

5.3.1 Model description

To illustrate and compare the various approaches of accounting for survey design in modeling, we consider variants of the following base model in which neither stratification nor clustering

is explicitly acknowledged:

$$Y_{ic}|p_{ic} \sim \text{Binomial}\left(n_{ic}, p_{ic}\right) \tag{5.2}$$

$$\text{logit}(p_{ic}) = \alpha + \boldsymbol{\beta}^{\top}\boldsymbol{x}_{ic} + S(\boldsymbol{s}_{ic}) \tag{5.3}$$

In addition to the notations defined previously, we let $\boldsymbol{s}_{ic}$ be the location associated with cluster c in state i and $S(\boldsymbol{s}_{ic})$ be a spatial random effect that follows a GP: $S(\cdot) \sim \text{GP}\left(0, \boldsymbol{\Sigma}_S\right)$ with $\boldsymbol{\Sigma}_S = [\sigma_S^2, \rho]$, where σ_S^2 is the marginal spatial variance and ρ is the spatial range (i.e., a distance at which the spatial correlation becomes negligible). Note that the GP we use is the solution to a stochastic partial differential equation (SPDE) which is approximated by a particular Gaussian Markov Random Field (GMRF) defined on a fine triangular mesh (Lindgren et al., 2011). This binomial model includes **no nugget** at the cluster level beyond the spatial field, hence we label this the *Binomial NN* model.

To obtain model-based vaccination coverage estimate for cluster c in state i, we can construct an approximation to its marginal posterior distribution by drawing posterior samples:

$$p_{ic}^{(m)} = \text{expit}\left(\alpha^{(m)} + \boldsymbol{\beta}^{\top(m)}\boldsymbol{x}_{ic} + S(\boldsymbol{s}_{ic})^{(m)}\right), \tag{5.4}$$

where the superscript (m) denotes the mth posterior sample of the respective parameter. Recall that a survey cluster is a census EA that has been sampled for the survey. Therefore, administrative-area-level coverage estimates can be obtained by aggregating the cluster-level estimates of all EAs in the area weighted by the population proportion in the EA. For example, the model-based vaccination coverage estimate for state i can be calculated using the posterior samples

$$p_i^{(m)} = \sum_{c=1}^{C_i} p_{ic}^{(m)} \times q_{ic},$$

where C_i is the total number of EAs in state i and q_{ic} is the proportion of the 12–23m population of state i that is in EA/cluster c. Almost always, however, the complete sampling frame of EAs is not available. In this case, we can approximate the state-level coverage by

aggregating cluster-level estimates over an approximated gridded EA map:

$$p_i^{(m)} \approx \sum_{g=1}^{G_i} p_{ig}^{(m)} \times q_{ig}, \tag{5.5}$$

where g indexes the gridded EA and $G_i = C_i$ is the total number of EAs in state i. The approximated EA map can be created using a gridded population density map and summary tables that are routinely available from survey reports. Details are provided in Appendix C.

To acknowledge the urban/rural stratification, we can extend (5.3) to include urbanicity as a covariate in the model:

$$\text{logit}(p_{ic}) = \alpha + \boldsymbol{\beta}^\top \boldsymbol{x}_{ic} + \gamma I\left(\boldsymbol{s}_{ic} \in \text{Urban}\right) + S(\boldsymbol{s}_{ic}). \tag{5.6}$$

Inference for cluster-level coverage p_{ic} can be obtained using posterior samples of contributing parameters in an analogous way to equation (5.4), and the administrative-area-level estimates can be approximated using the aggregation procedure summarized in equation (5.5).

Now we focus on accounting for clustering and consider two mechanisms by which cluster-level non-spatial variation arises.

Within-Cluster Variation

Suppose within a super-population in a cluster there are groups with their own distinct vaccination coverage. If we carried out repeated sampling at this cluster, a different group (or mixtures of groups) would be sampled each time and we would "see" a different coverage beyond sampling variation. This phenomenon has been extensively studied in the statistics literature (McCullagh and Nelder, 1989). In this case, we have **within-cluster variation** that induces **overdispersion (OD)** at the cluster level. What we are really interested in is the overall coverage we would see if we could sample everyone in the cluster. To capture this kind of variation, we can assume that each group's coverage is drawn from a distribution. We describe two possible models that achieve this purpose. First, we can assume that each group's coverage follows a beta distribution. This results in a Beta-Binomial likelihood for

the data:

$$Y_{ic}|\mu_{ic}, d \sim \text{Beta-Binomial}\left(n_{ic}, \mu_{ic}, d\right) \tag{5.7}$$

$$\text{logit}(\mu_{ic}) = \alpha + \boldsymbol{\beta}^\top \boldsymbol{x}_{ic} + \gamma I\left(\boldsymbol{s}_{ic} \in \text{Urban}\right) + S(\boldsymbol{s}_{ic}). \tag{5.8}$$

Specifically, we can think of this model as a result of:

$$\begin{aligned} Y_{ic}|p_{ic} &\sim \text{Binomial}\left(n_{ic}, p_{ic}\right) \\ p_{ic}|\mu_{ic}, d &\sim \text{Beta}\left(\mu_{ic}, d\right), \end{aligned}$$

where the beta distribution is parameterized as

$$\begin{aligned} E\left[p_{ic}|\mu_{ic}, d\right] &= \mu_{ic} \\ var\left[p_{ic}|\mu_{ic}, d\right] &= \frac{\mu_{ic}(1-\mu_{ic})}{d+1}. \end{aligned}$$

Marginally, the data follows a Beta-Binomial distribution with

$$\begin{aligned} E\left[Y_{ic}|\mu_{ic}, d\right] &= n_{ic}\mu_{ic} \\ var\left[Y_{ic}|\mu_{ic}, d\right] &= n_{ic}\mu_{ic}(1-\mu_{ic}) \times \frac{n_{ic}+d}{1+d}. \end{aligned}$$

The parameter $d > 0$ characterizes the degree of overdispersion, with higher values of d corresponding to less overdispersion. The limiting case of a Beta-Binomial distribution at $d = \infty$ is a binomial distribution. The target of inference is $E\left[p_{ic}|\mu_{ic}, d\right] = \mu_{ic}$, which corresponds to the hypothetical coverage we would see if we could sample everyone in the cluster – it can be estimated using posterior samples of contributing parameters in (5.8). We label this model the *Beta-Binomial OD* model.

An alternative is to assume that each group's coverage follows a normal distribution on the logit scale (i.e., a logit-normal distribution). This results in a Lono-Binomial likelihood for the data:

$$\begin{aligned} Y_{ic}|\eta_{ic}, \sigma_\delta^2 &\sim \text{Lono-Binomial}\left(n_{ic}, \eta_{ic}, \sigma_\delta^2\right) \\ \eta_{ic} &= \alpha + \boldsymbol{\beta}^\top \boldsymbol{x}_{ic} + \gamma I\left(\boldsymbol{s}_{ic} \in \text{Urban}\right) + S(\boldsymbol{s}_{ic}). \end{aligned}$$

The Lono-Binomial distribution is the compound distribution where the p parameter in the binomial distribution as being randomly drawn from a logit-normal distribution. We use this name to emphasize the parallel between this distribution and the Beta-Binomial distribution. Specifically, we can think of this model as a result of:

$$\begin{aligned} Y_{ic}|q_{ic} &\sim \text{Binomial}\left(n_{ic}, q_{ic}\right) \\ \text{logit}\left(q_{ic}\right)|\eta_{ic}, \sigma_\delta^2 &= \eta_{ic} + \delta_{ic} \\ \delta_{ic} &\sim_{iid} \text{Normal}\left(0, \sigma_\delta^2\right) \end{aligned}$$

The parameter $\sigma_\delta^2 > 0$ characterizes the degree of overdispersion, with higher σ_δ^2 value corresponding to more overdispersion. The limiting case of a Lono-Binomial distribution at $\sigma_\delta^2 = 0$ is a binomial distribution. The target of inference in this case is $E\left[q_{ic}|\eta_{ic}, \sigma_\delta^2\right]$, which, by the law of the unconscious statistician, equals

$$\begin{aligned} p_{ic} = E\left[q_{ic}|\eta_{ic}, \sigma_\delta^2\right] &= \int_\delta \text{expit}\left(\eta_{ic} + \delta\right) \pi\left(\delta|\sigma_\delta^2\right) d\delta \\ &\approx \text{expit}\left(\frac{\eta_{ic}}{\sqrt{1 + h^2\sigma_\delta^2}}\right) \\ &= \text{expit}\left(\frac{\alpha + \boldsymbol{\beta}^\top \boldsymbol{x}_{ic} + \gamma I\left(\boldsymbol{s}_{ic} \in \text{Urban}\right) + S(\boldsymbol{s}_{ic})}{\sqrt{1 + h^2\sigma_\delta^2}}\right), \end{aligned} \tag{5.9}$$

where $h = \frac{16\sqrt{(3)}}{15\pi}$. For a derivation of this approximation, see Section 9.13.1 of Wakefield (2013). Inference for this target can be obtained using posterior samples of contributing parameters in (5.9). We label this model the *Lono-Binomial OD* model.

Between-Cluster Variation

The cluster-level excess variation may also be **true signal (TS)** that represents **between-cluster variation** in cluster coverage means beyond the spatial field. In this case, if we carried out repeated sampling at this cluster, we would "see" different groups having the same mean coverage. We can model this kind of variation using an independent normal term

in the linear predictor of cluster-level vaccination coverage:

$$
\begin{aligned}
&Y_{ic}|p_{ic} \sim \text{Binomial}\left(n_{ic}, p_{ic}\right) \\
&\text{logit}(p_{ic}) = \alpha + \boldsymbol{\beta}^\top \boldsymbol{x}_{ic} + \gamma I\left(\boldsymbol{s}_{ic} \in \text{Urban}\right) + S(\boldsymbol{s}_{ic}) + \epsilon_{ic} \\
&\epsilon_{ic} \sim \text{Normal}\left(0, \sigma_\epsilon^2\right)
\end{aligned}
$$

We emphasize that ϵ_{ic} is capturing between-cluster differences in coverage, not the within-cluster variability that induces overdispersion in the *Lono-Binomial OD* model. Hence, the target of inference is

$$
p_{ic} = \text{expit}\left(\alpha + \boldsymbol{\beta}^\top \boldsymbol{x}_{ic} + \gamma I\left(\boldsymbol{s}_{ic} \in \text{Urban}\right) + S(\boldsymbol{s}_{ic}) + \epsilon_{ic}\right), \tag{5.10}
$$

which is different from (5.9). Posterior samples of contributing parameters can be used to estimate (5.10), with $\epsilon_{ic}^{(m)}$ randomly sampled from a $\text{N}\left(0, \sigma_\epsilon^{2(m)}\right)$ distribution. We label this model the *Binomial TS* model.

In reality, the cluster-level non-spatial variation is likely to be the result of a mixture of the within-cluster overdispersion and between-cluster true signal, but most household surveys only obtain a single sample of households within each selected cluster, so the survey data does not contain enough information to identify the source of the cluster-level variation. Therefore, map producers need to be extremely careful about their assumptions regarding cluster-level extra variation when choosing the target of inference for vaccination coverage estimation.

5.3.2 Model comparison via 2018 NDHS analysis

We apply the aforementioned four classes of models, namely *Binomial NN*, *Beta-Binomial OD*, *Lono-Binomial OD* and *Binomial TS*, to analyze data from the 2018 NDHS. We fit two models within each class: one with urbanicity as a covariate to account for survey stratification and the other without. In addition, all models include the same set of covariates: poverty, aridity, log-transformed night-time lights, log-transformed travel time and enhanced vegetation index (EVI).

All models were fitted using the INLA approach (Rue et al., 2009) implemented in the `INLA` package in `R`. For the intercept α, a N(0, 0) prior is used, where the format is N(mean, precision); this prior is improper but the intercept is well estimated. For the other fixed effect parameters (i.e., $\boldsymbol{\beta}, \gamma$) We used N(0, 0.001) priors. We used the Penalized Complexity (PC) priors (Simpson et al., 2017) for the hyperparameters of the random effects — these are typically the most sensitive to the prior choice. For the spatial range ρ, we set a PC prior so that the median effective range is at a fifth of the diameter of the spatial domain. For the variance hyperparameters, we set PC priors so that $\Pr(\sigma_\epsilon > 1) = \Pr(\sigma_S > 1) = \Pr(\sigma_\delta > 1) = 0.01$.

For model comparison, we use the widely applicable information criterion (WAIC) (Watanabe, 2010) to evaluate the predictive power of the models. We also conduct a 37-fold cross-validation exercise to assess the performance of out-of-sample predictions. Specifically, we use the hold-out method of cross-validation, setting aside data from an entire state each time. Within a hold-out state, we let y_c and n_c be the number of vaccinated and sampled children at cluster c, and $\hat{p}_c = \frac{y}{n}$ be the observed coverage at cluster c. In addition, we let $\tilde{p}_c$ be the posterior median of the coverage estimate at cluster c obtained from fitting the model on data from other states. Using the hold-out data from m cluster locations, we compute the following model validation criteria:

- Bias $= \frac{1}{m}\sum_c(\tilde{p}_c - \hat{p}_c)$.
- Mean Absolute Error (MAE) $= \frac{1}{m}\sum_c|\tilde{p}_c - \hat{p}_c|$.
- Root Mean Square Error (RMSE) $= \sqrt{\frac{1}{m}\sum_c(\tilde{p}_c - \hat{p}_c)^2}$.

Table 5.1 reports the estimates of selected parameters and the model validation results for each model. Based on the results, the odds of being vaccinated against measles is estimated to be 35% to 42% higher in an urban area than in a rural area. The 95% credible intervals (CIs) of the corresponding coefficients are always strictly greater than 0 (i.e., odds ratio greater than 1), indicating a strong association between urbanicity and MCV1 coverage after

accounting for the other covariates. Across all model classes, adding the urban/rural strata variable to account for survey design always improves predictive performance in terms of WAIC, bias, MAE and RMSE. This result is consistent with the findings of Paige et al. (2020). Table 5.2 reports the posterior medians and 2.5% and 97.5% quantiles of the regression coefficients for covariates. The results are fairly consistent across all models.

We now compare the *Binomial NN* models, which assume there is no cluster-level excess variation beyond the spatial field, to the other three classes of models, which explicitly account for non-spatial excess variation under various assumptions. Figure 5.4 shows the posterior medians of the spatial field in each model. The estimated spatial field from the *Binomial NN* models has shorter range and smaller marginal spatial variance, which characterize a "bumpy" surface with a lot of local residual spatial correlation (the spatial range ≈ 78 km). In contrast, the estimated spatial fields from the other models tend to be "smoother" with spatial correlation over longer distances (spatial range ≈ 322 km). This contrast hints at the presence of considerable small-scale variation that can be better captured by the *Beta-Binomial OD*, *Lono-Binomial OD* and *Binomial TS* models. The estimates for the overdispersion/true signal parameters in these models indicate significant cluster-level variation beyond the spatial field. In addition, their predictive performances measured by WAIC, Bias, MAE and RMSE are noticeably better than that of the *Binomial NN* models (see Table 5.1).

In summary, this analysis illustrates the advantages of accounting for the survey design, especially the survey stratification and clustering, when fitting geostatistical models to estimate vaccination coverage. Once a satisfactory model is selected, it is important to present the model-based estimates with care. This will be the focus of the next section.

5.4 Presenting model-based estimates and their uncertainties

As mentioned in Section 5.1, pixel-level vaccination coverage estimates are routinely produced as output of geostatistical models, and they are often aggregated by population density to form administrative-area-level coverage estimates. Figures 5.5 to 5.8 show the maps of the

Model Class	Strata Included?	Strata	Spatial Field		Overdispersion		True Signal	Model Validation			
		Urban γ	Range ρ	SE σ_S	d	σ_δ	σ_ϵ	WAIC	Bias	MAE	RMSE
Binomial NN	No Strata		0.69 (0.47, 1.0)	1.1 (0.96, 1.3)				3618	-0.016	0.260	0.312
	Strata	0.30 (0.10, 0.49)	0.70 (0.47, 1.1)	1.1 (0.97, 1.3)				3615	-0.016	0.259	0.310
Beta-Binomial OD	No Strata		2.9 (1.7, 5.1)	1.6 (1.1, 2.1)	2.9 (2.6, 8.6)			3581	-0.007	0.254	0.307
	Strata	0.31 (0.10, 0.51)	3.0 (1.8, 5.3)	1.6 (1.1, 2.1)	2.9 (2.6, 8.6)			3573	-0.006	**0.253**	**0.305**
Lono-Binomial OD	No Strata		2.7 (1.6, 4.8)	1.4 (1.1, 1.9)		0.73 (0.63, 1.9)		3456	**-0.004**	0.254	0.307
	Strata	0.35 (0.13, 0.57)	2.8 (1.7, 5.0)	1.4 (1.1, 1.9)		0.73 (0.63, 1.9)		**3451**	-0.005	**0.253**	0.306
Binomial TS	No Strata		2.7 (1.6, 4.8)	1.4 (1.1, 1.9)			0.73 (0.63, 1.9)	3456	-0.006	0.254	0.307
	Strata	0.35 (0.13, 0.57)	2.8 (1.7, 5.0)	1.4 (1.1, 1.9)			0.73 (0.63, 1.9)	**3451**	-0.006	**0.253**	0.306

Table 5.1: Estimates of parameters and model validation results. Reported are the posterior medians and 2.5% and 97.5% quantiles of the regression coefficients and the parameters of the spatial field and cluster-level excess variation. The spatial range parameter ρ is on the longitude-latitude degree scale, which, given the geographical location of Nigeria, equals an average of 111 km per degree. The widely applicable information criterion (WAIC) is calculated based on all the data. The bias, mean absolute error (MAE) and root mean squared error (RMSE) are calculated based on the cross-validation exercise described in the supplementary materials. Bold figures represent the "best" models according to the relevant criteria.

Model Class	Strata Included?	Intercept	Aridity	Poverty	EVI	Travel Time	Night Light
Binomial NN	No Strata	1.7	0.57	-2.2	-0.0085	-0.093	0.31
		(0.21, 3.3)	(0.13, 1.0)	(-4.0, -0.39)	(-0.13, 0.11)	(-0.19, 0.0012)	(0.14, 0.50)
	Strata	1.5	0.53	-2.2	0.0048	-0.046	0.25
		(-0.065, 3.0)	(0.08, 0.97)	(-4.1, -0.45)	(-0.12, 0.13)	(-0.15, 0.053)	(0.071, 0.43)
Beta-Binomial OD	No Strata	2.1	0.26	-2.4	-0.044	-0.067	0.26
		(0.6, 3.7)	(-0.36, 0.83)	(-4.2, -0.75)	(-0.16, 0.074)	(-0.15, 0.021)	(0.09, 0.44)
	Strata	1.9	0.21	-2.5	-0.027	-0.023	0.21
		(0.33, 3.5)	(-0.42, 0.78)	(-4.3, -0.8)	(-0.15, 0.09)	(-0.12, 0.068)	(0.029, 0.39)
Lono-Binomial OD	No Strata	2.1	0.29	-2.8	-0.046	-0.068	0.30
		(0.50, 3.8)	(-0.40, 0.90)	(-4.7, -1.0)	(-0.17, 0.081)	(-0.16, 0.025)	(0.12, 0.49)
	Strata	2.1	0.23	-2.9	-0.028	-0.018	0.23
		(0.50, 3.8)	(-0.47, 0.84)	(-4.8, -1.1)	(-0.16, 0.099)	(-0.12, 0.08)	(0.044, 0.42)
Binomial TS	No Strata	2.1	0.29	-2.8	-0.046	-0.068	0.30
		(0.50, 3.8)	(-0.40, 0.90)	(-4.7, -1.0)	(-0.17, 0.081)	(-0.16, 0.025)	(0.12, 0.49)
	Strata	2.1	0.23	-2.9	-0.028	-0.018	0.23
		(0.50, 3.8)	(-0.47, 0.84)	(-4.8, -1.1)	(-0.16, 0.099)	(-0.12, 0.08)	(0.044, 0.42)

Table 5.2: Estimates of parameters for each model. Reported are the posterior medians and 2.5% and 97.5% quantiles of the regression coefficients for covariates. Note that the estimates for the coefficients of covariates are not directly comparable because the covariates are measured on different scales.

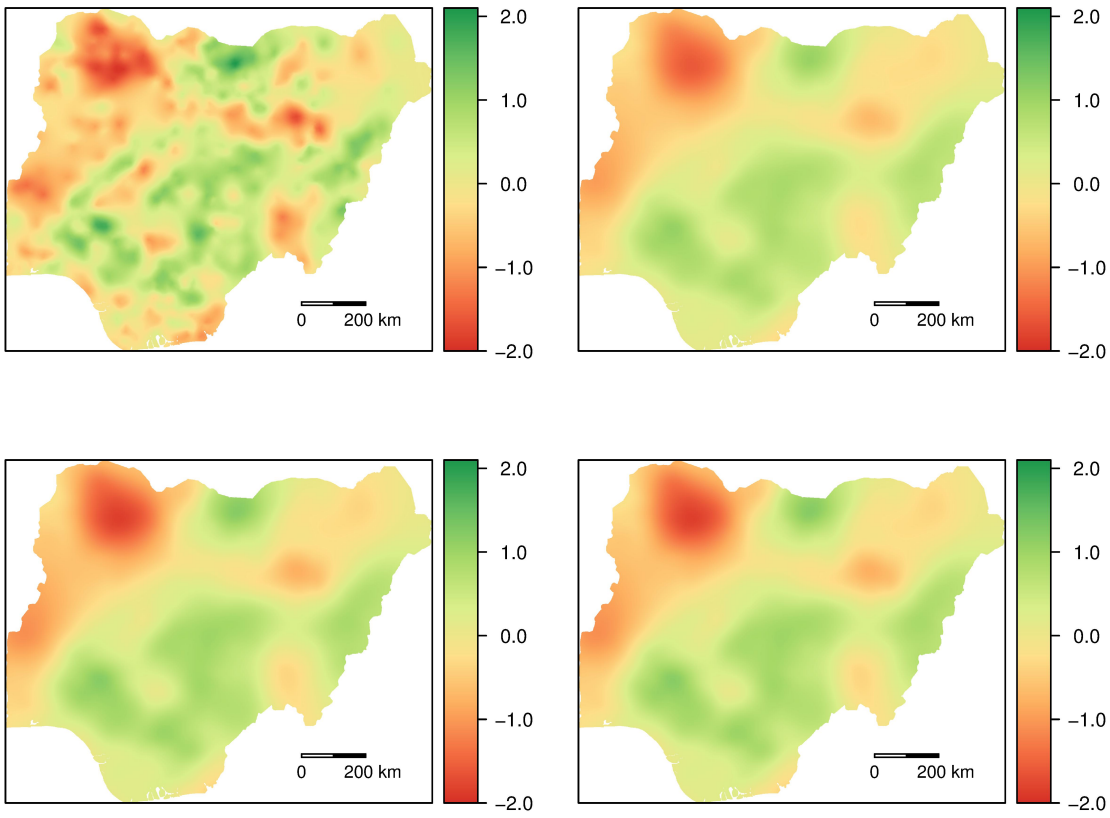

Figure 5.4: Posterior medians of the spatial fields based on the *Binomial NN* (top left), *Beta-Binomial OD* (top right), *Lono-Binomial OD* (bottom left) and *Binomial TS* (bottom right) models that include the urban/rural strata variable.

posterior medians and the widths of 90% credible intervals (CIs) for the estimated MCV1 coverage at the 1×1 km pixel, LGA and state levels based on the *Binomial NN*, *Beta-Binomial OD*, *Lono-Binomial OD* and *Binomial TS* models that include the urban/rural strata variable. These maps show a consistent trend: coverage estimates at a finer spatial resolution tend to have larger associated uncertainty — and hence poorer precision. In particular, the pixel maps are often associated with huge uncertainties — a consequence of the sparsity of the survey data, since the 2018 NDHS was only powered to be representative at the state level (National Population Commission - NPC and ICF, 2019). Although geostatistical models can usually produce smoothed estimates that are slightly biased but more precise than the direct estimates, the 1×1 km pixel is too fine a spatial resolution for reliable coverage estimation using the typical number of clusters in household surveys. Therefore, when presenting vaccination coverage estimates on a map, one needs to carefully consider the trade-off between the geographical scale and precision of estimates, and choose an appropriate spatial resolution such that the resultant map have reasonably high precision for the estimates to be statistically reliable. In addition, the uncertainty associated with the point estimates should always be appropriately presented. Instead of labeling uncertainty maps with ambiguous terms such as "low" and "high" (Utazi et al., 2018, 2019b, 2020), quantitative scales should be used to clearly show how reliable the coverage estimates are.

5.4.1 Coefficient of variation (CV) as an uncertainty measure

Besides the commonly used posterior CI width and standard deviation as quantitative uncertainty measures for vaccination coverage estimates, another sensible choice is the coefficient of variation (CV) of the posterior distribution. Consider the vaccination coverage p for a generic area. Let $\theta = \frac{p}{1-p}$ be the odds, and $\phi = \log(\theta) = \text{logit}(p)$ be the log odds. We assume that the posterior for the log odds ϕ in a generic area approximately follows a Normal$(\hat{\phi}, \hat{\sigma}^2)$ distribution, where $\hat{\phi}$ and $\hat{\sigma}^2$ are posterior estimates of $\text{E}[\phi|\boldsymbol{y}]$ and $\text{var}(\phi|\boldsymbol{y})$. This is equivalent to the posterior for the odds θ following a LogNormal$(\hat{\phi}, \hat{\sigma}^2)$ distribution, with mean,

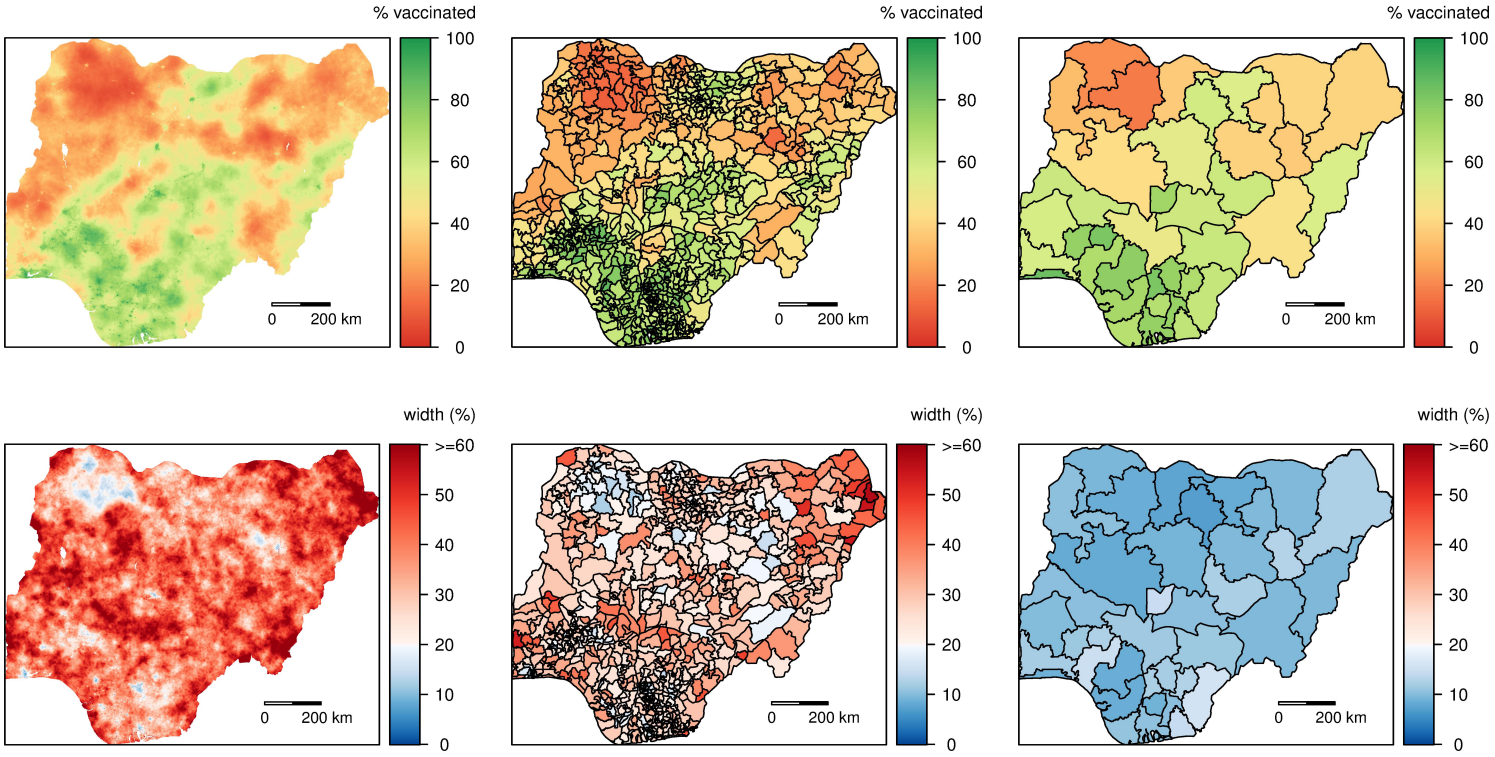

Figure 5.5: Maps of the posterior medians (top row) and the widths of 90% credible intervals (botton row) for the estimated MCV1 coverage at the 1×1 km pixel (left), LGA (middle) and state (right) levels, based on the *Binomial NN* model that includes the urban/rural strata variable.

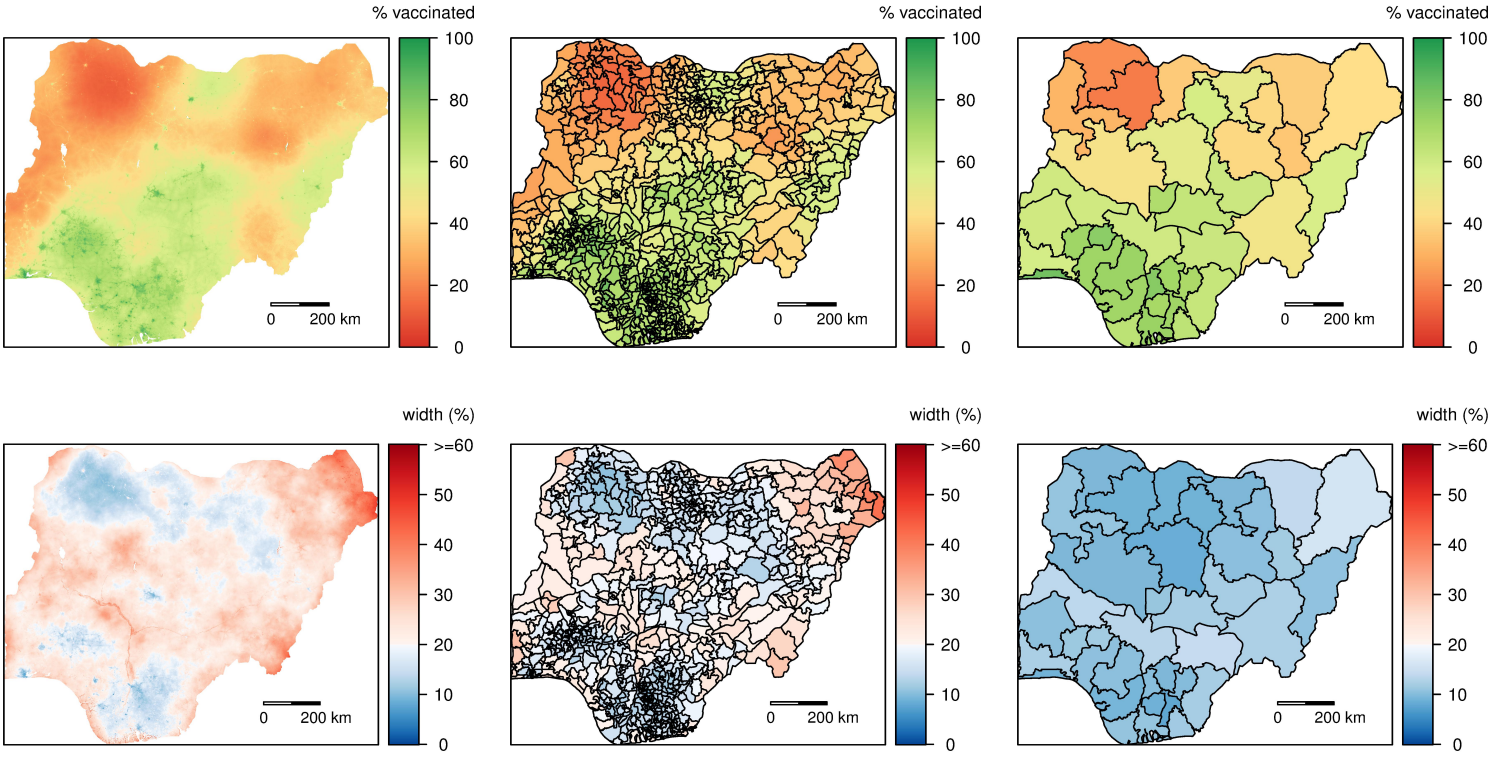

Figure 5.6: Maps of the posterior medians (top row) and the widths of 90% credible intervals (botton row) for the estimated MCV1 coverage at the 1×1 km pixel (left), LGA (middle) and state (right) levels, based on the *Beta-Binomial OD* model that includes the urban/rural strata variable.

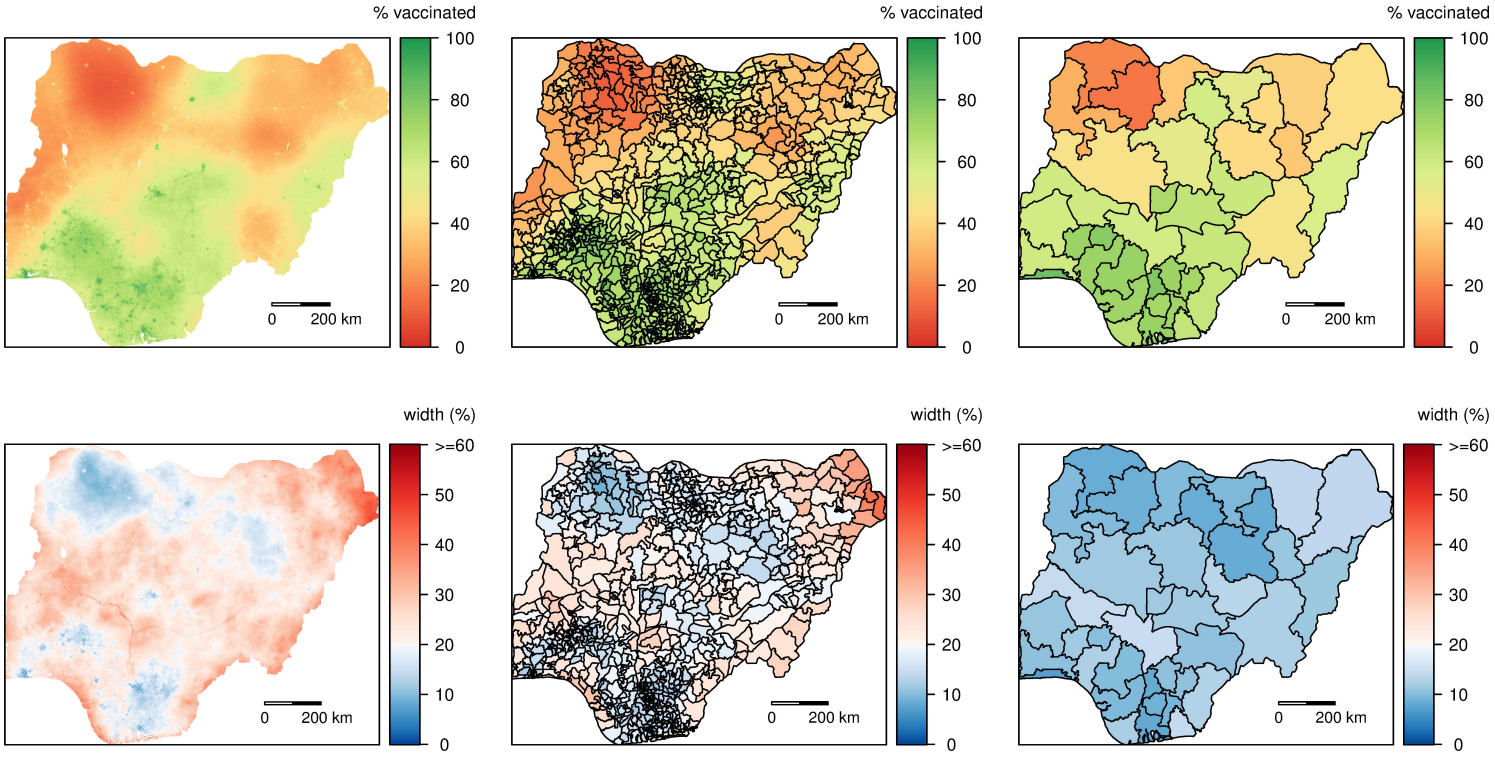

Figure 5.7: Maps of the posterior medians (top row) and the widths of 90% credible intervals (botton row) for the estimated MCV1 coverage at the 1×1 km pixel (left), LGA (middle) and state (right) levels, based on the *Lono-Binomial OD* model that includes the urban/rural strata variable.

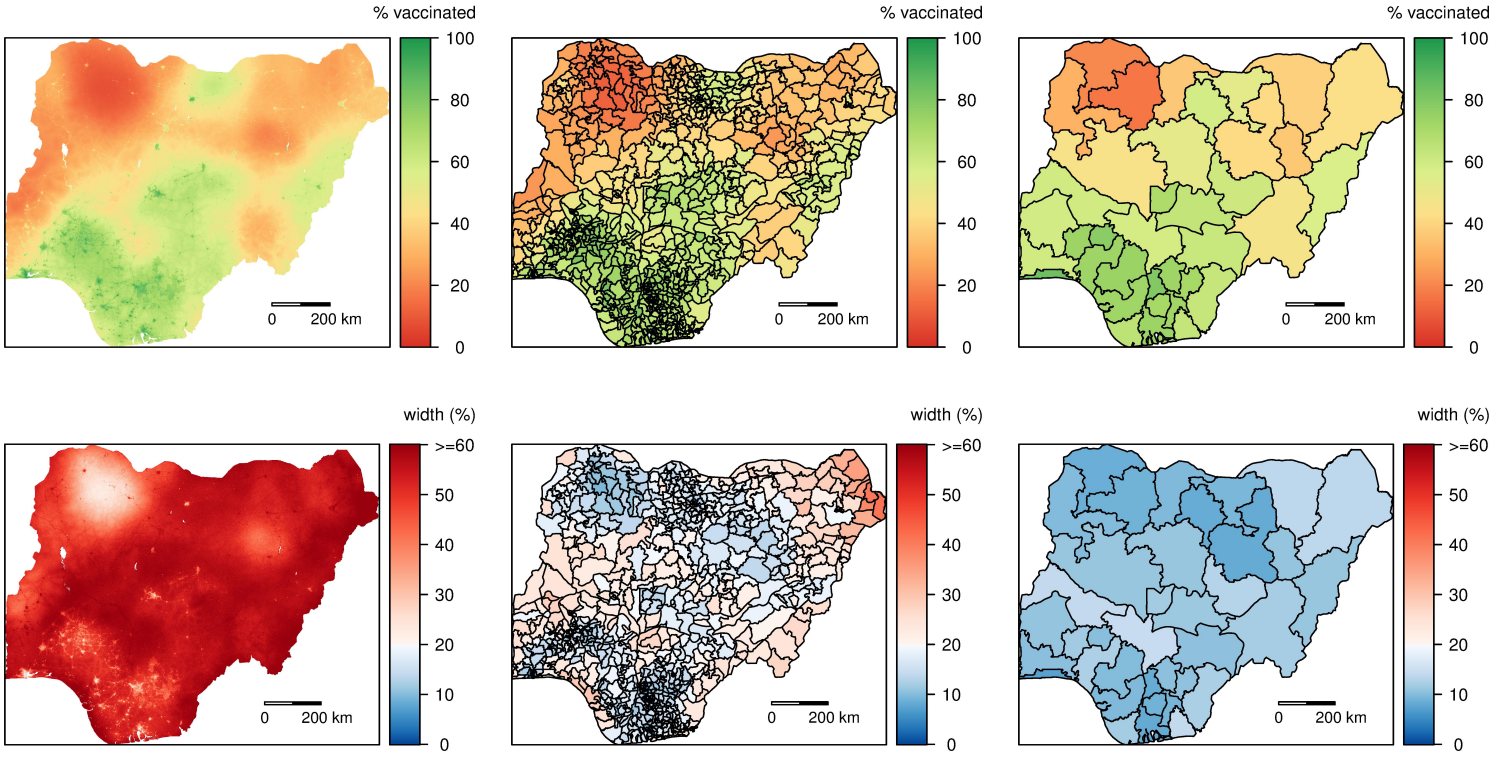

Figure 5.8: Maps of the posterior medians (top row) and the widths of 90% credible intervals (botton row) for the estimated MCV1 coverage at the 1×1 km pixel (left), LGA (middle) and state (right) levels, based on the *Binomial TS* model that includes the urban/rural strata variable.

variance and CV

$$\begin{aligned}
\mathrm{E}\left[\theta|\boldsymbol{y}\right] &= \exp\left(\hat{\phi} + \frac{\hat{\sigma}^2}{2}\right), \\
\mathrm{var}\left(\theta|\boldsymbol{y}\right) &= \mathrm{E}\left[\theta|\boldsymbol{y}\right]^2 \times \left(\exp(\hat{\sigma}^2) - 1\right), \\
\mathrm{CV}\left(\theta|\boldsymbol{y}\right) &= \frac{\sqrt{\mathrm{var}\left(\theta|\boldsymbol{y}\right)}}{\mathrm{E}\left[\theta|\boldsymbol{y}\right]} = \sqrt{\exp(\hat{\sigma}^2) - 1} \approx \sqrt{\hat{\sigma}^2} = \hat{\sigma},
\end{aligned}$$

i.e., the posterior standard deviation $\mathrm{sd}(\phi|\boldsymbol{y})$.

Now we consider the 100(1 - α)% posterior credible interval (CI) for θ. The ratio of the higher bound to the lower bound of the CI is

$$\frac{\mathrm{CI}_{\mathrm{hi}}\left(\theta|\boldsymbol{y}\right)}{\mathrm{CI}_{\mathrm{lo}}\left(\theta|\boldsymbol{y}\right)} = \frac{\exp\left(\hat{\phi} + z_{\alpha/2}\hat{\sigma}\right)}{\exp\left(\hat{\phi} - z_{\alpha/2}\hat{\sigma}\right)} = \exp\left(2z_{\alpha/2}\hat{\sigma}\right) \approx \exp\left(2z_{\alpha/2}\mathrm{CV}\left(\theta|\boldsymbol{y}\right)\right),$$

where $z_{\alpha/2}$ is the $\alpha/2$ quantile of a standard normal random variable. Hence,

$$\mathrm{CV}\left(\theta|\boldsymbol{y}\right) \approx \frac{1}{2z_{\alpha/2}} \log\left(\frac{\mathrm{CI}_{\mathrm{hi}}\left(\theta|\boldsymbol{y}\right)}{\mathrm{CI}_{\mathrm{lo}}\left(\theta|\boldsymbol{y}\right)}\right),$$

so that there is a direct relationship between the posterior CV and the ratio of higher end to lower end of the posterior credible interval.

This justifies using hatching based on the posterior CVs of area-level vaccination coverage odds. For example, Figure 5.9 shows a map of the posterior medians of the MCV1 coverage estimates for Nigeria's 37 states, with hatching density proportional to the posterior CV of the vaccination odds. The results are based on 1000 posterior samples of the MCV1 coverage estimates from the *Lono-Binomial OD* model that includes the urban/rural strata variable.

5.4.2 Visualizing point estimates and uncertainties on the same map

The map with hatching in Figure 5.9 is an example of visualization tools that show point estimates and their associated uncertainties on the same map. A major advantage of presenting coverage estimates in this way is that users do not need to look at a separate map to be informed of the uncertainties, and hence tend to be more aware of the map uncertainty

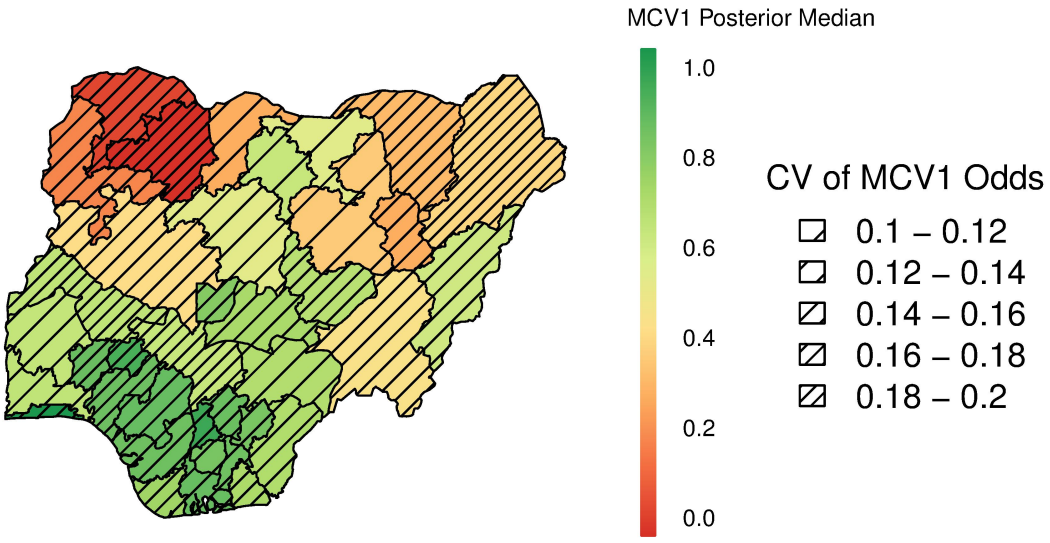

Figure 5.9: Map of the posterior medians of the MCV1 coverage estimates for Nigeria's 37 states, with hatching density proportional to the posterior coefficient of variation (CV) of the vaccination odds. The results are based on 1000 posterior samples of the MCV1 coverage estimates from the *Lono-Binomial OD* model that includes the urban/rural strata variable.

in general. In the example of the hatching map shown in Figure 5.9, the hatching density increases with uncertainty level, resulting in areas with higher uncertainty to be covered by more lines and hence reveal less color that indicates the estimated coverage level. This mechanism of suppressing the color shown in an area based on the uncertainty is also the main idea behind the value-suppressing uncertainty palettes (Correll et al., 2018) used in health indicator mapping (Bhattacharjee et al., 2019). However, depending on specific color choices and spatial scales, these methods can produce rather confusing maps described as "visual puzzles": very sophisticated but not very intuitive (Tufte, 2001). Hence, care must be taken when using these methods for coverage estimates presentation.

When presenting coverage estimates for administrative areas, one can potentially show the posterior distribution for each area while arranging the positions of the plots based on the geographical location of the areas. For example, Figure 5.10 shows the posterior distributions of the MCV1 coverage estimates for Nigeria's 37 states, created using the `geofacet` package (Hafen and Schloerke, 2020) in `R`. The plots for different geographical entities are arranged into a grid that preserves some of the geographical orientation. Although the resultant map is not showing exact locations and sizes of the areas, it still reveals some important spatial patterns in coverage estimates while visualizing their associated uncertainties.

5.4.3 Visualizing rankings with uncertainties

An important aim of vaccination coverage estimation is to rank a set of areas to identify the places that need the most improvement. Often, a single summary statistic, such as the posterior mean or median of the vaccination coverage, is used for ranking areas with little consideration of the associated uncertainties. As Shen and Louis (1998) pointed out, if the ranks of the "unit-specific parameters", in our case, the area-level vaccination coverage, are the target feature, ranking posterior means can perform poorly, and the conditional expected ranks (Laird and Louis, 1989) or a discretized version of them are optimal. To balance the trade-off between optimizing rankings and two other inferential goals, namely good parameter histogram and good unit-specific parameter estimates, Shen and Louis (1998) developed a

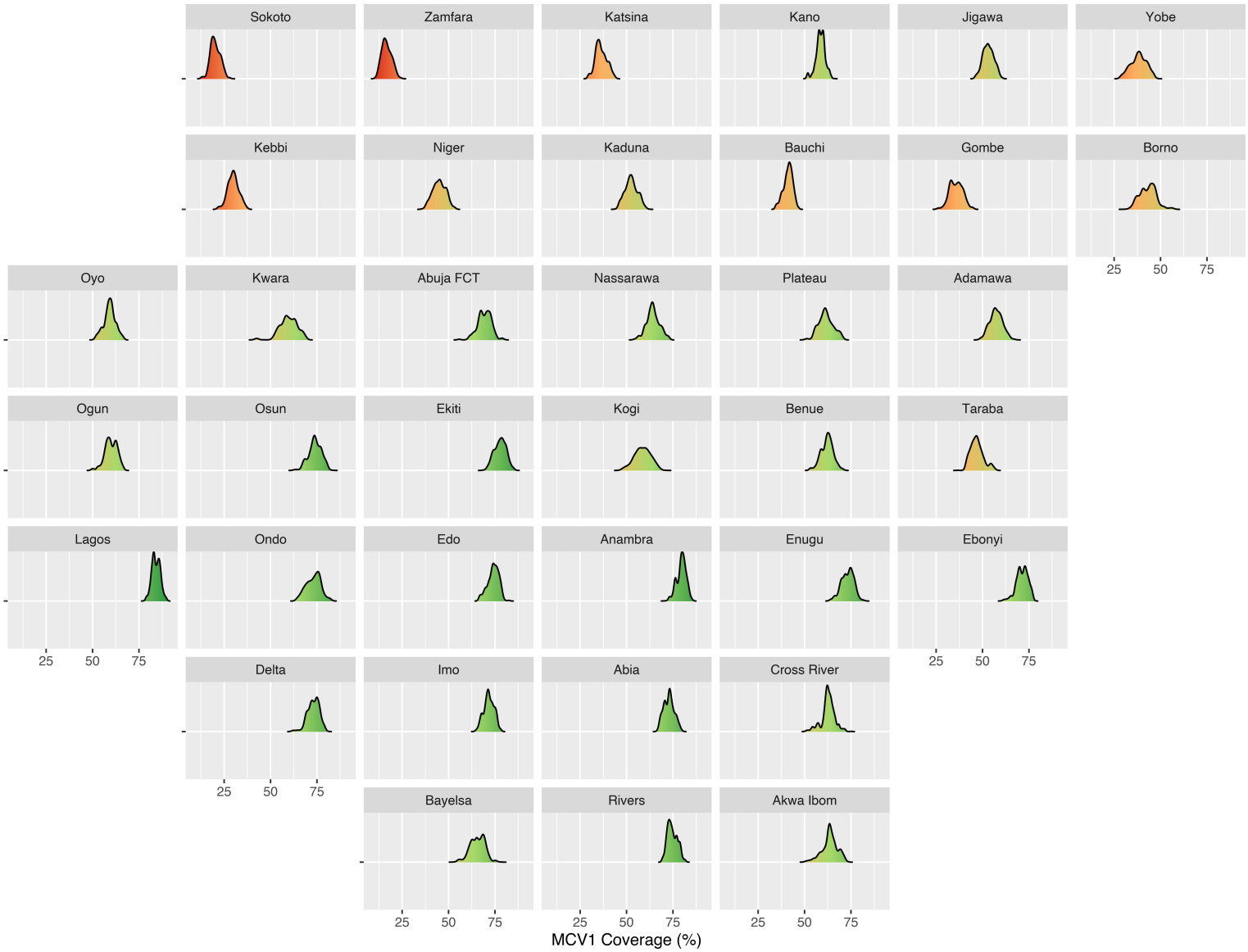

Figure 5.10: The posterior distributions of the MCV1 coverage estimates for Nigeria's 37 states, arranged into a grid that preserved some of the geographical orientation, based on 1000 posterior samples from the *Lono-Binomial OD* model that includes the urban/rural strata variable.

new class of "triple-goal estimates" that optimizes estimation of the histogram and the ranks.

In this section, we focus on visualization of the uncertainty associated with ranking and recommend two presentation methods. The first is to use ridgeline plots to show the posterior distribution of coverage estimate for each area. For example, Figure 5.11 shows the ridgeline plots of the posterior distributions of the MCV1 coverage estimates for Nigeria's 37 states, ordered by posterior median, based on 1000 posterior samples from the *Lono-Binomial OD* model that includes the urban/rural strata variable. When comparing MCV1 coverage estimates across different areas, one can clearly see how much the areas' posterior distributions overlap with each other on the ridgeline plots, and hence be informed of the uncertainty associated with the relative ranking of the areas.

Alternatively, one can use histograms of the posterior ranking distributions to identify areas with the highest or lowest vaccination coverage. For a given set of areas, each Monte Carlo draw of posterior coverage estimates gives a ranking of the areas. One can therefore obtain a posterior ranking distribution for each area by repeatedly drawing and ranking posterior samples, and the expected rank (ER) for each area can be estimated by taking the average of its ranking samples. For example, Figure 5.12 shows the histograms of the posterior ranking distributions of the 5 states with the lowest and 5 with the highest ERs based on 1000 posterior Monte Carlo samples of the state-level MCV1 coverage estimates. Zamfara, Sokoto and Kebbi seem to have the highest posterior probabilities of being ranked as the lowest three states. As for Gombe and Katsina, the similarity in their ranking distributions reveals high uncertainty in their relative rankings. On the other hand, the state of Lagos is likely to have the highest MCV1 coverage, followed by the states of Anambra and Ekiti. However, there is considerable uncertainty in the ranking of the next two states: Rivers and Osun, as is evident in the wide spread of their ranking distributions.

The posterior ranking distributions as well as the complete range of the distributions for all 37 states of Nigeria are shown in Figure 5.13 and Figure 5.14. We see that the states with the lowest and highest ERs tend to have narrower posterior ranking distributions, whereas the states ranked in the middle tend to have higher uncertainties associated with their relative

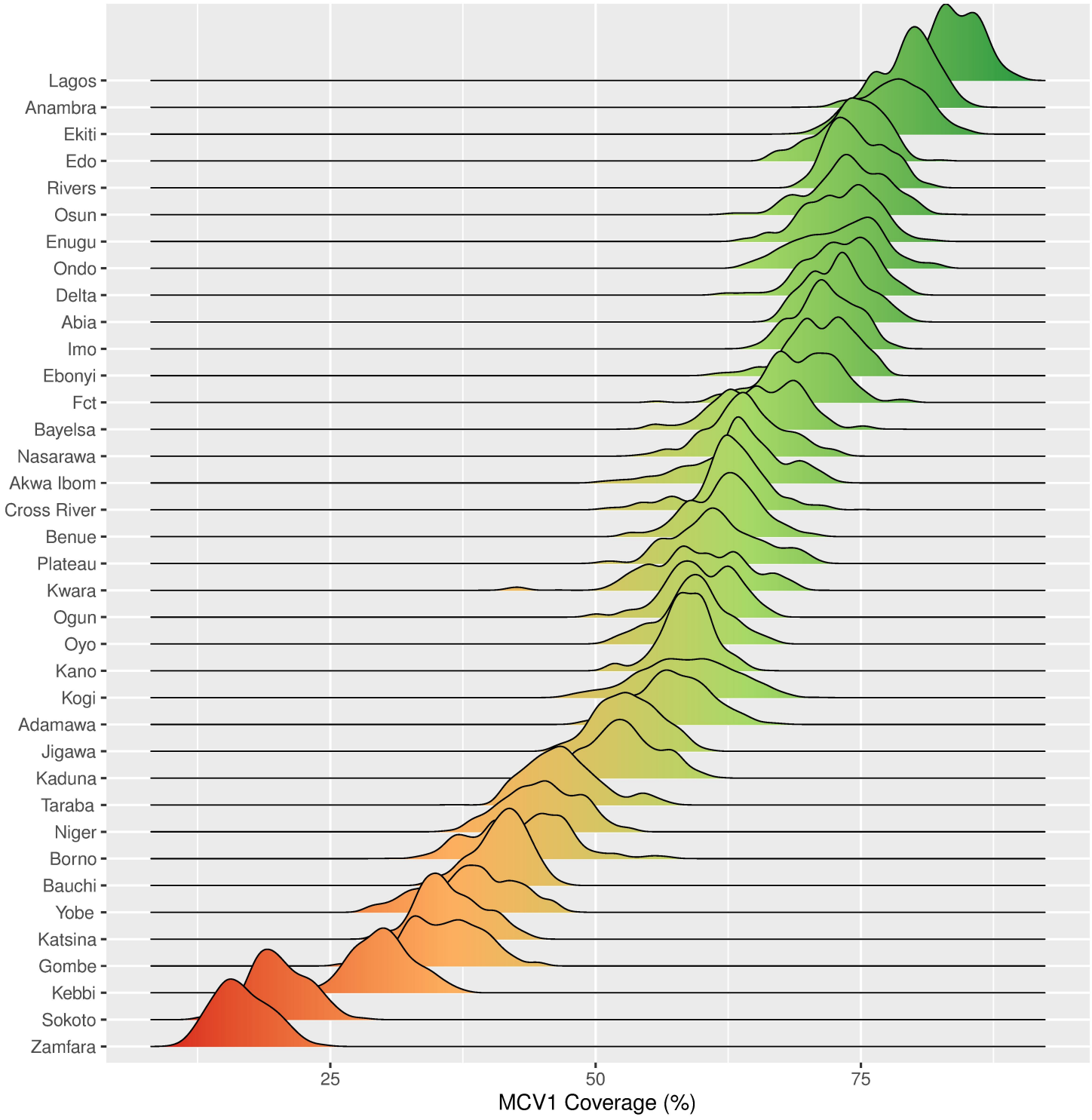

Figure 5.11: Ridgeline plots of the posterior distributions of the MCV1 coverage estimates for Nigeria's 37 states, ordered by posterior median, based on 1000 posterior samples from the *Lono-Binomial OD* model that includes the urban/rural strata variable.

rankings.

For the ranking of the LGAs, we show the posterior ranking distributions of the 5 LGAs with the lowest and 5 with the highest ERs in Figure 5.15, and the posterior distribution range for all 774 LGAs in Nigeria in Figure 5.16, based on 1000 posterior samples from the *Lono-Binomial OD* model that includes the urban/rural strata variable. Since the LGA-level coverage estimates generally have lower precision than the state-level estimates, the LGA ranking distributions have much wider spread than those of the states.

In summary, both the ridgeline plots of posterior distributions and histograms of posterior ranking distributions are useful tools for identifying areas with the highest or lowest vaccination coverage. They are not only straightforward to calculate, but also effective in visualising the uncertainty associated with the ranking of the areas.

5.4.4 Visualizing posterior exeedance probabilities

Another common objective of vaccination coverage estimation is to identify areas with vaccination coverage higher or lower than a certain threshold. In this case, it is useful to examine maps of posterior exceedance probabilities. For example, Figure 5.17 shows the maps of the marginal posterior probabilities of LGA-level MCV1 coverage exceeding the 80%, 50% and 20% thresholds respectively. In particular, the exceedance probability map corresponding to the 80% threshold is especially relevant in evaluating to what extent Nigeria has attained the Global Vaccine Action Plan (GVAP) target of reaching 80% coverage with all vaccines in all districts by 2020 (World Health Organization, 2012). The map highlights that substantial efforts are needed in most places to meet the target. Fewer than 7 percent of the LGAs (i.e., 53 out of 774 LGAs) have 50% or greater posterior probability of being at the target of 80% MCV1 coverage. In addition, the exceedance probability map corresponding to the 20% threshold identifies 28 LGAs in northwest Nigeria that have 50% or lower posterior probability of reaching an MCV1 coverage level of 20%. This demonstrates how exceedance probability maps can be used to identify areas with low coverage estimates while keeping users informed of the uncertainties associated with the results.

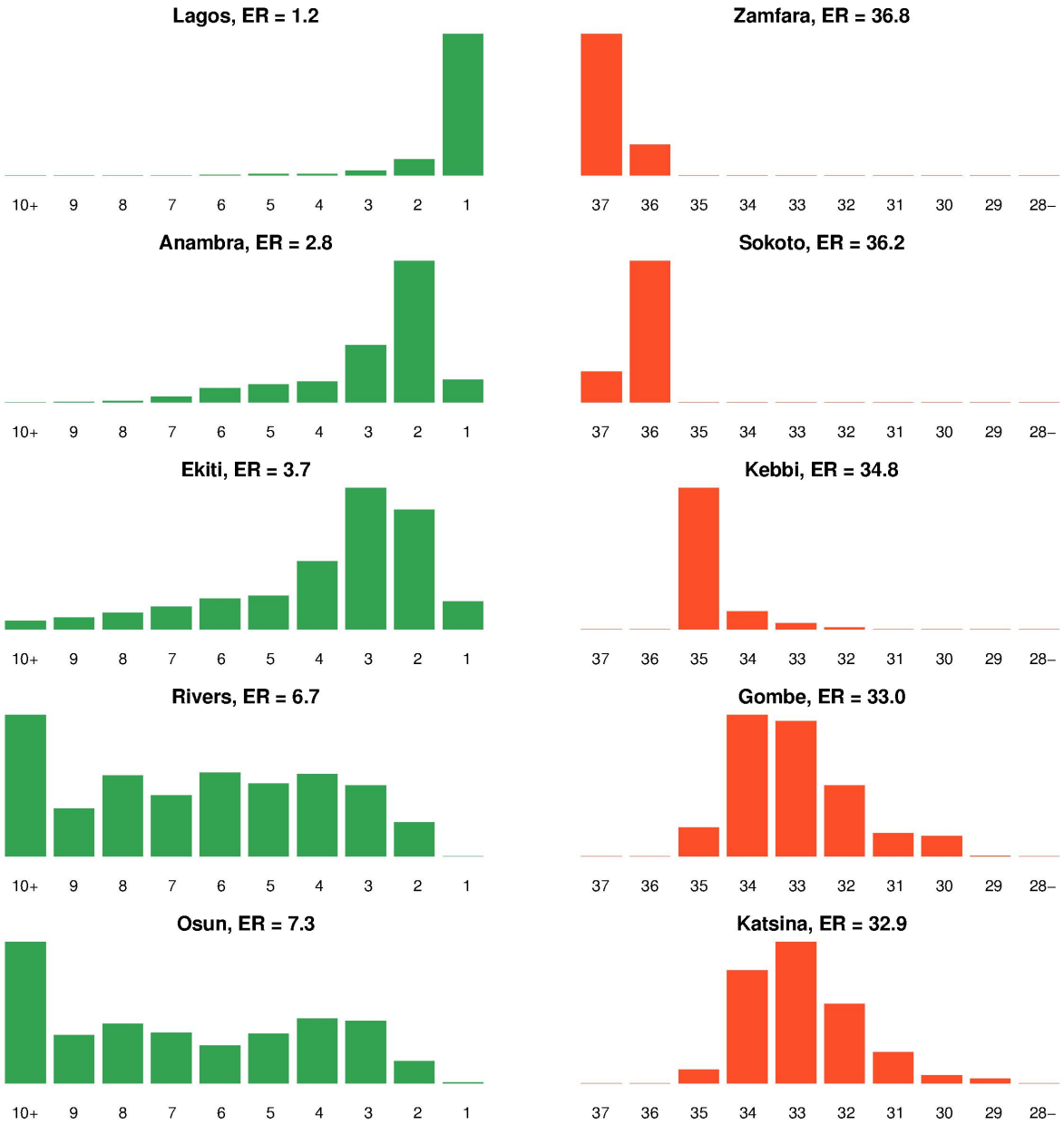

Figure 5.12: Histograms of the posterior ranking distributions of the 5 states with the lowest (left column) and 5 with the highest (right column) expected ranks (ERs), based on 1000 posterior samples of the MCV1 coverage estimates of Nigeria's 37 states from the *Lono-Binomial OD* model that includes the urban/rural strata variable.

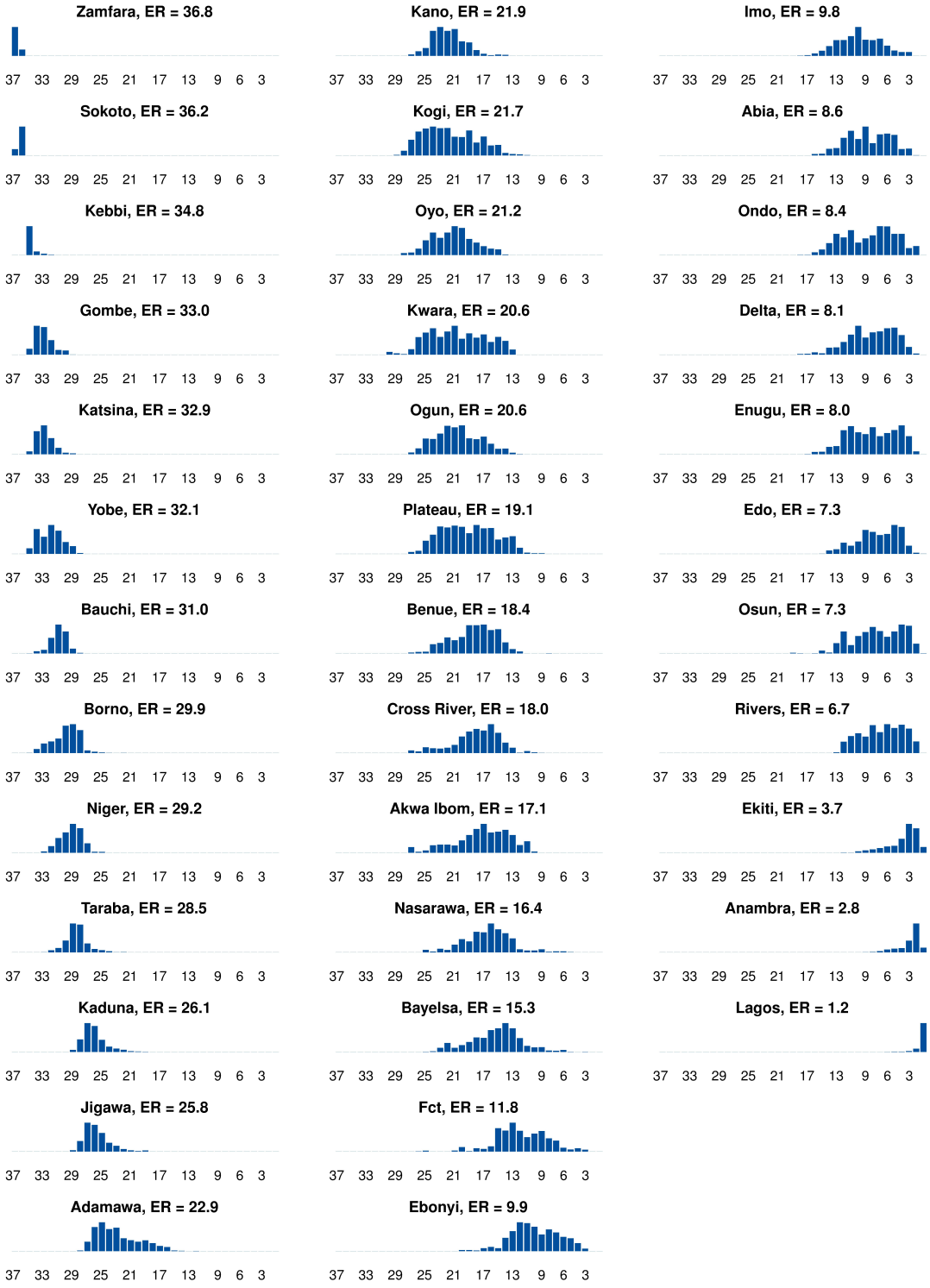

Figure 5.13: Posterior ranking distributions of all 37 states in Nigeria with their expected ranks (ER) based on 1000 posterior samples of the MCV1 coverage estimates from the *Lono-Binomial OD* model that includes the urban/rural strata variable.

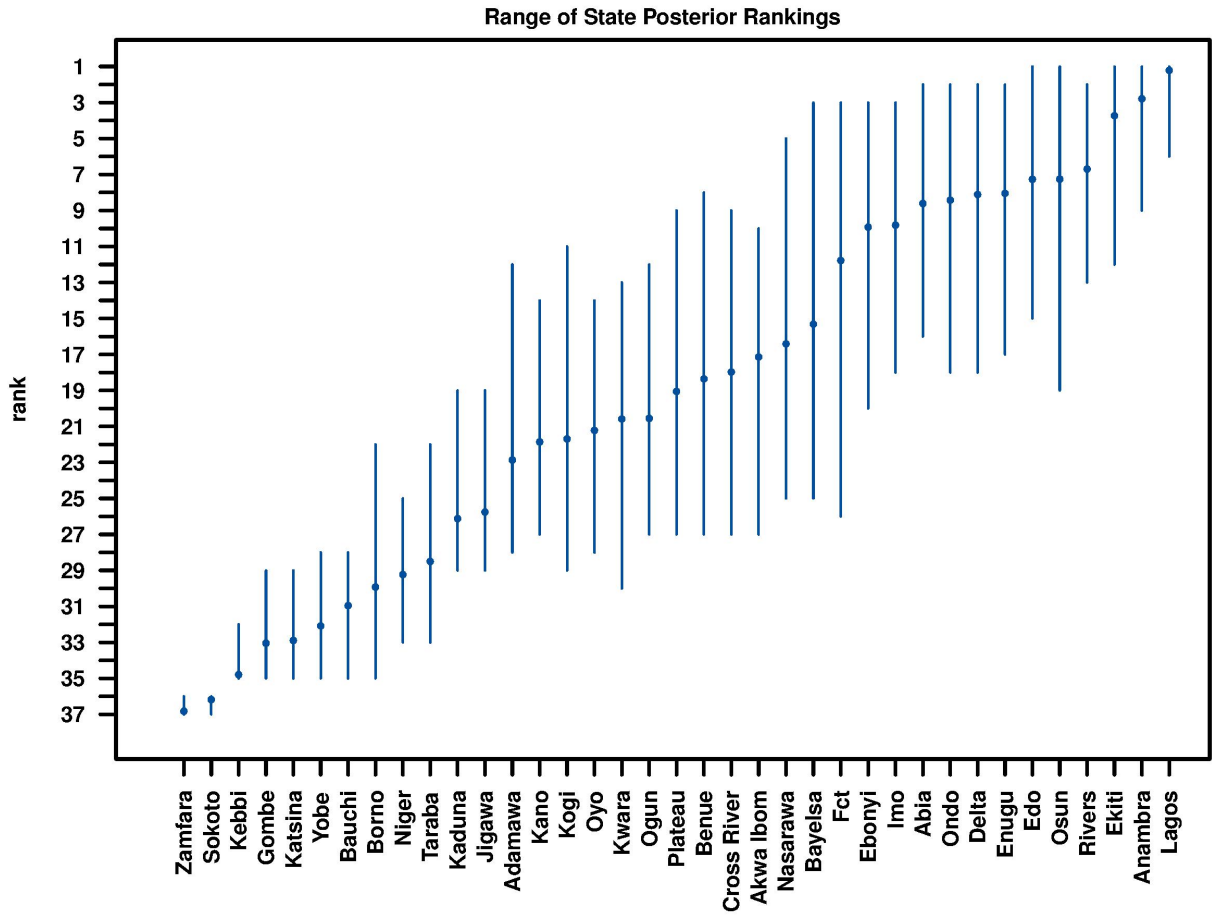

Figure 5.14: The range of the posterior rankings of all 37 states in Nigeria with their expected ranks (ERs) based on 1000 posterior samples of the MCV1 coverage estimates from the *Lono-Binomial OD* model that includes the urban/rural strata variable. The points indicate the ERs of the states, and the vertical line segments mark the complete range of the posterior rankings of the states.

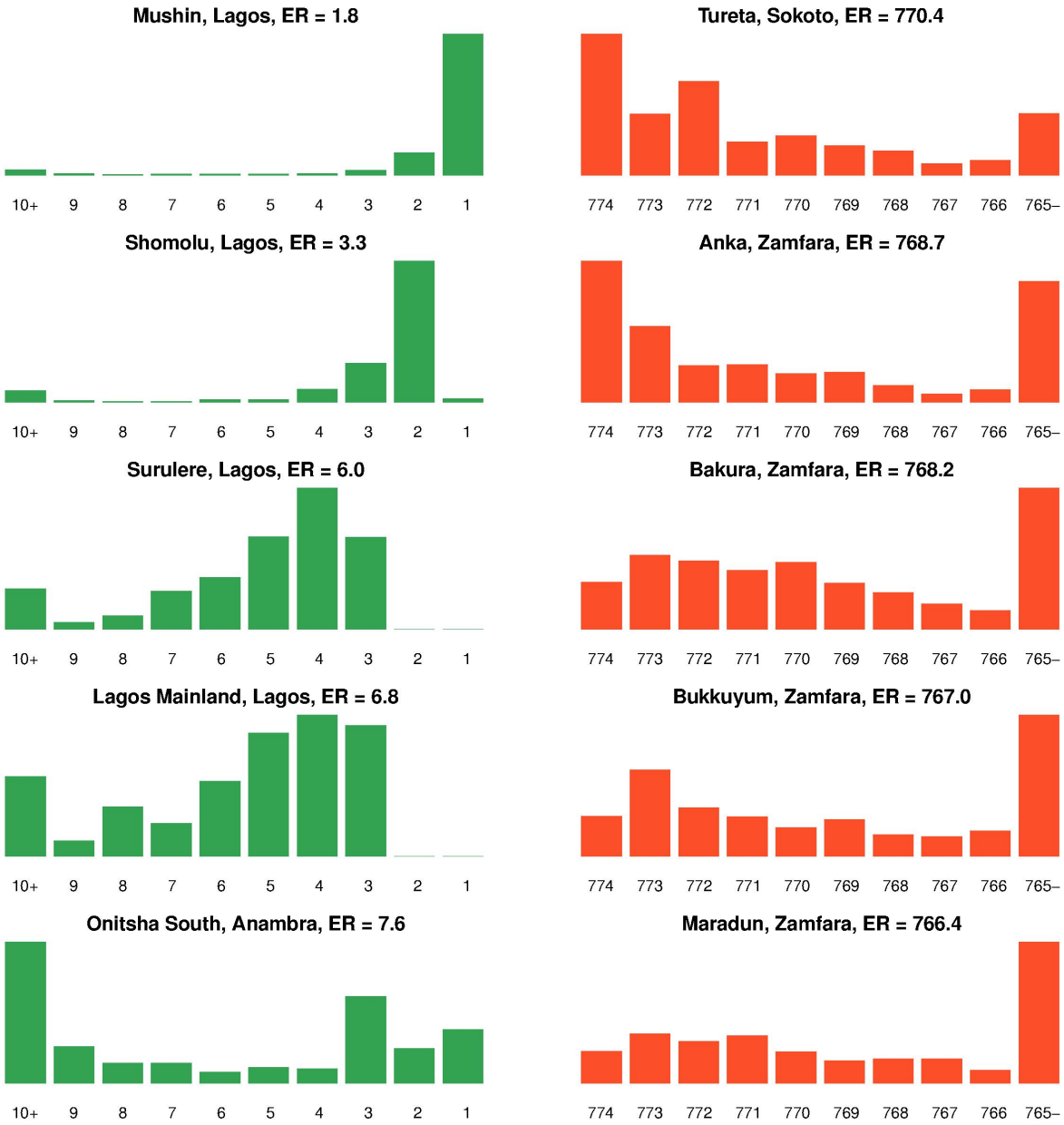

Figure 5.15: Posterior ranking distributions of the 5 LGAs with the lowest (left column) and 5 with the highest (right column) expected ranks (ERs) based on 1000 posterior samples of the MCV1 coverage estimates of Nigeria's 774 LGAs from the *Lono-Binomial OD* model that includes the urban/rural strata variable.

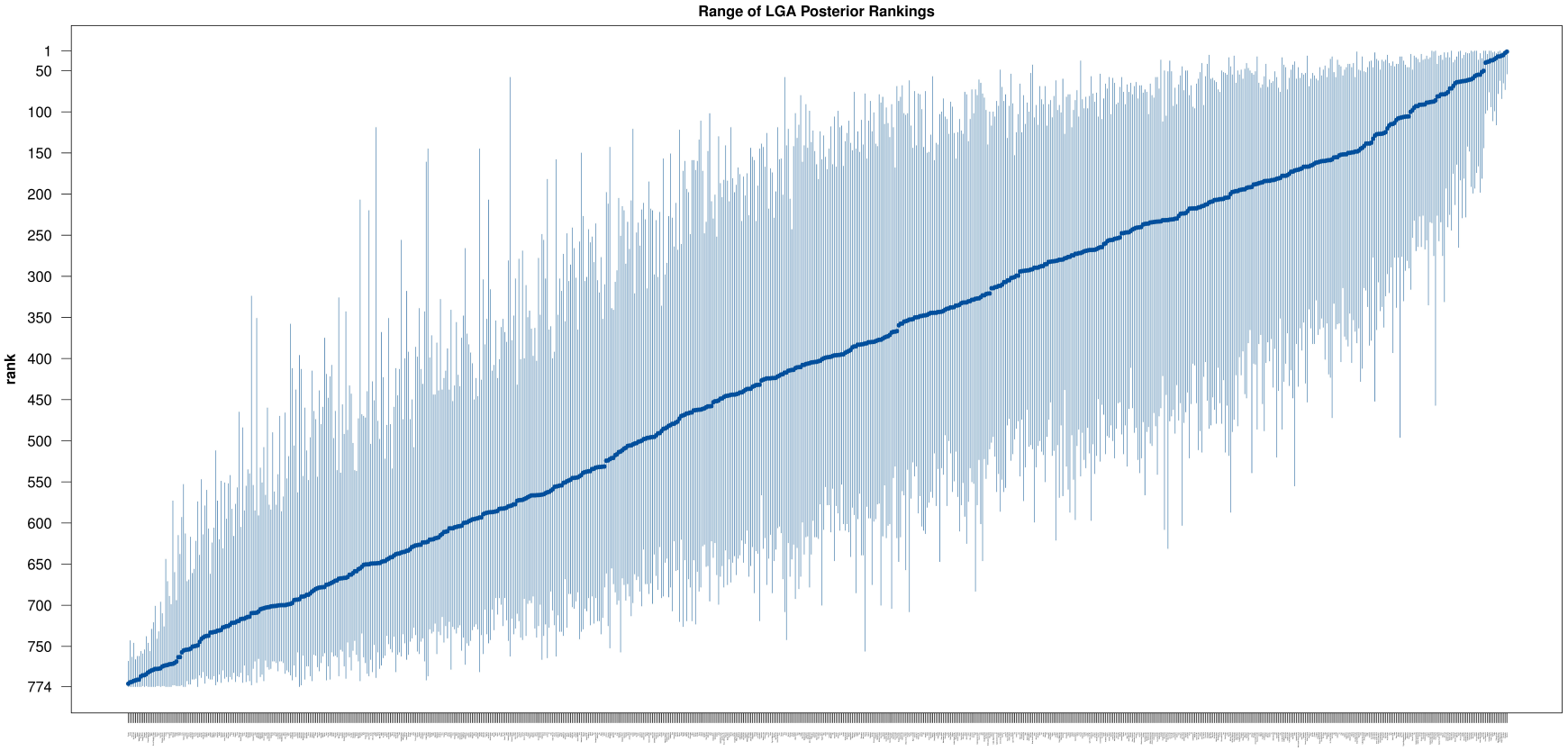

Figure 5.16: The range of the posterior rankings of all 774 LGAs in Nigeria with their expected ranks (ER) based on 1000 posterior samples of the MCV1 coverage estimates from the *Lono-Binomial OD* model that includes the urban/rural strata variable. The points indicate the expected ranks (ER) of the LGAs, and the vertical line segments mark the complete range of the posterior rankings of the LGAs.

It should be noted that the examples shown in Figure 5.17 are based on the marginal posterior probabilities that an area/pixel having vaccination coverage higher or lower than a certain threshold. The results should only be interpreted in a marginal sense, and they do not provide inference for multiple areas' joint exceedance of a certain level. When joint probabilities are of interest, the estimation of regions so that the probability for exceeding a certain level in the entire set is equal to some predefined value is a difficult task. Bolin and Lindgren (2015) provided a formulation of the excursion and contour uncertainty problems for random fields using a set theory framework. They developed methods for solving these problems, as well as the related problem of finding credible regions for contour curves for latent Gaussian models. Their method is based on using a parametric family for the excursion sets in combination with a sequential importance sampling method for estimating joint probabilities. In follow-up work (Bolin and Lindgren, 2017), measures of the statistical uncertainty of contour maps were proposed, and computational methods are constructed to be used in combination with INLA for latent Gaussian models. In our context of vaccination coverage for various spatial scales, their methods can be used to understand the pixel-level coverage estimates and uncertainties, but have limited application for the area-level results.

5.4.5 Novel discrete-color maps for precision control

In addition to visualizing the uncertainties associated with coverage estimates, one should also avoid showing maps with low overall precision. We propose a novel approach to coverage map presentation that allows comparison and control of the overall map uncertainty level. Our method is inspired by Bayesian decision theory and is applicable to presenting coverage estimates at any spatial scale.

The basic idea is that we move away from the usual continuous color scales; instead, we use a discrete set of colors to represent a partition of $[0\%, 100\%]$, the range of vaccination coverage. The partition can be defined based on some pre-specified threshold(s) of interest. For example, we can use the thresholds 20%, 50% and 80% to create four intervals: $[0\%, 20\%)$, $[20\%, 50\%)$, $[50\%, 80\%)$ and $[80\%, 100\%]$, that represent the *extremely low*, *low*,

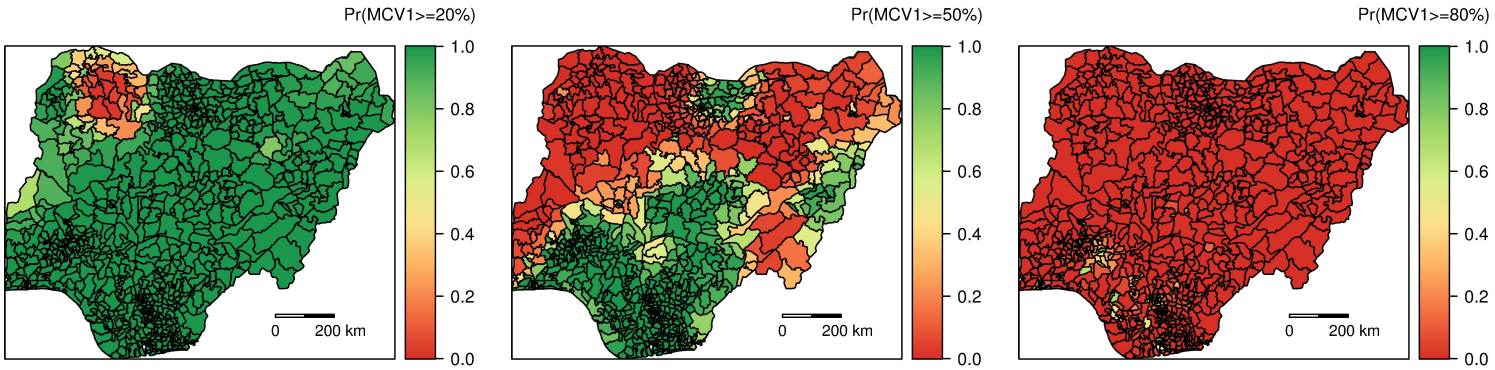

Figure 5.17: Maps of the posterior probabilities of LGA-level MCV1 coverage estimates exceeding the 20%, 50% and 80% thresholds respectively, based on the *Lono-Binomial OD* model that includes the urban/rural strata variable.

medium and *high* coverage intervals respectively. Next, we examine the posterior distribution of the coverage estimate for each area and assign each area to the interval that contains the greatest posterior probability. We call this maximum the *true classification probability* (TCP), so one minus the TCP is the probability of misclassification. The key idea here is that, as clearly shown in the ridgeline plots, for each color assigned to a pixel there is a posterior probability that this color is correctly classified but also a corresponding probability of incorrect classification. This step is inspired by Bayesian decision theory for univariate probabilities: the classification of each area follows a Bayes decision rule that minimizes a 0-1 loss. Finally, we calculate the average of the TCPs across all areas and call it the *average true classification probability* (ATCP) of the map. The ATCP serves as a measure of the overall precision of a vaccination coverage map. A higher ATCP indicates a higher overall map precision, since on average we are more certain about the classification of the areas under the assumed model.

Using this approach, we present the MCV1 coverage estimates at the 1×1 km pixel, LGA and state levels based on the *Lono-Binomial OD* model in Figure 5.18. The top row shows

the maps with the pre-specified discrete color scale, and below each map is a histogram of the TCPs with the ATCP highlighted by a blue vertical line. We see that the state map has the highest ATCP (0.94), followed by the LGA map (ATCP = 0.87) and the pixel map (ATCP = 0.83). This reflects the decreasing precision associated with the state-, LGA- and pixel-level coverage estimates.

If one does not have any specific threshold in mind to form a discrete color scale, one can consider using the quantiles of the posterior samples of the coverage estimates. We take the pixel-level MCV1 coverage estimates as an illustrate example: for a fixed number of levels K, one can form intervals $[L_0, L_1), \ldots, [L_{K-1}, L_K]$, where $L_0 = 0\%$, $L_K = 100\%$, and L_k equals the $100 \times k/K$ quantile of the posterior samples pooled across all pixels. For example, for $K = 2$, we have $L_1 = 44\%$, the median of the pooled posterior samples, which forms two intervals: $[0\%, 44\%)$ and $[44\%, 100\%]$. Figure 5.19 shows three pixel maps with discrete color scales formed by setting $K = 2$, 3 and 4 respectively. Below each map is a histogram of the TCPs with the ATCP highlighted by a blue vertical line. We see that as the number of intervals K increases, the ATCP becomes lower. This is expected: as we partition the range into more quantile intervals, the width of each interval would generally become narrower, and the TCPs would generally become smaller by definition. In fact, two extreme cases are possible when presenting model-based vaccination coverage estimates using this approach. At one extreme, a perfect ATCP can be achieved for showing a flat map where every pixel is assigned to the interval $[0\%, 100\%]$. At the other extreme, we can assign each pixel to a unique color using a finely partitioned, almost-continuous color scale, but the ATCP would be essentially zero. The latter is the standard practice.

To achieve a balance between the precision of a map and the granularity of its color scale, we can set a minimally acceptable level of ATCP and use the most granular color scale that satisfies this requirement. For example, if we set the minimally acceptable ATCP to be 0.70, then among the three pixel maps in Figure 5.19, the one corresponding to $K = 3$ would be the "best" choice, since its ATCP (0.76) is above the 0.70 threshold and its color scale has more quantile intervals than the other eligible pixel map ($K = 2$, ATCP = 0.87). The

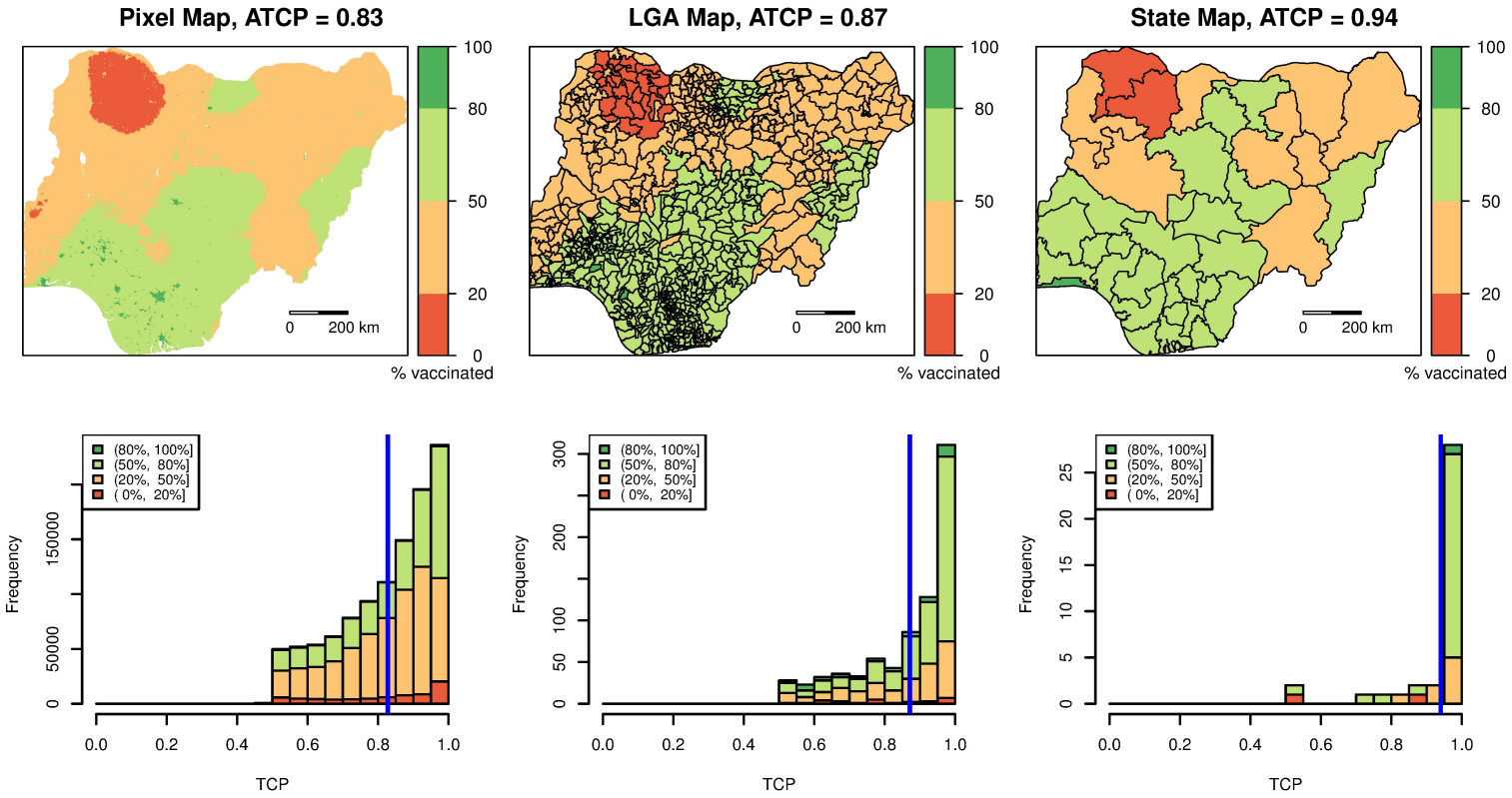

Figure 5.18: Top: Maps of MCV1 coverage estimates at the 1×1 km pixel, LGA and state levels using the discrete color scale composed of the *extremely low* $[0\%, 20\%)$, *low* $[20\%, 50\%)$, *medium* $[50\%, 80\%)$ and *high* $[80\%, 100\%]$ coverage intervals. Bottom: The corresponding histograms of the *true classification probabilities* (TCPs) with the *average true classification probability* (ATCP) highlighted by the blue vertical line.

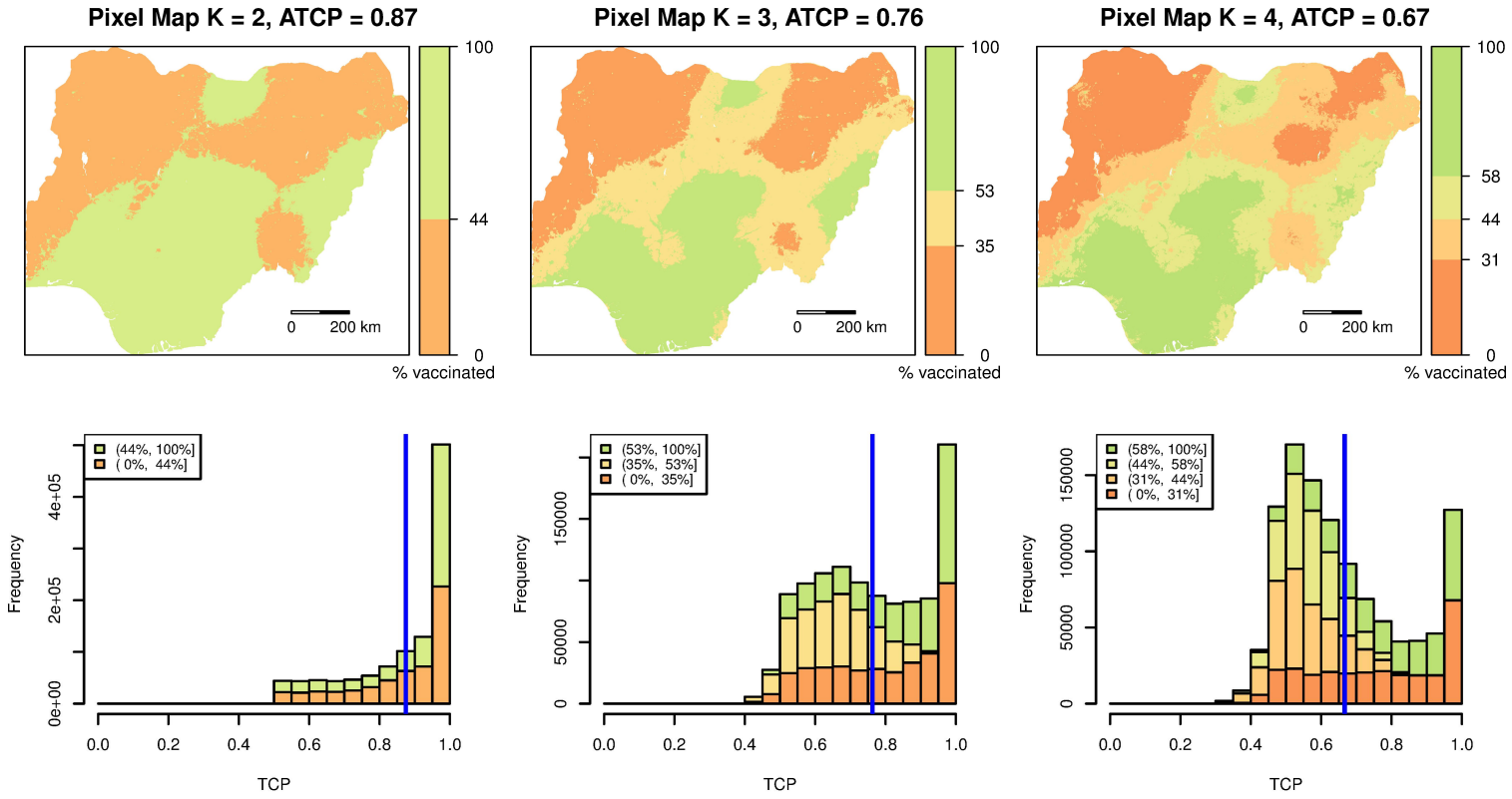

Figure 5.19: Top: Maps of MCV1 coverage estimates at the 1×1 km pixel level using discrete color scales with $K = 2, 3$ and 4 quantile intervals. For each K, the color scale is formed by creating the intervals $[L_0, L_1), \ldots, [L_{K-1}, L_K]$, where $L_0 = 0\%$, $L_K = 100\%$, and L_k equals the $100 \times k/K$ quantile of the pooled posterior samples of the pixel-level coverage estimates based on the *Lono-Binomial OD* model that includes the urban/rural strata variable. Bottom: The corresponding histograms of the *true classification probabilities* (TCPs) of the pixels with the *average true classification probability* (ATCP) highlighted by the blue vertical line.

map corresponding to $K = 4$ has an ATCP of 0.67, hence does not satisfy the minimally acceptable ATCP level of 0.70.

The same procedure can be applied to produce LGA- and state-level MCV1 coverage maps: if we set the same minimally acceptable ATCP at 0.70, we would have $K = 4$ and 5 for the "best" LGA and state maps respectively (Figure 5.20). In Figure 5.21, we show the ATCP values corresponding to the state, LGA and 1×1 km pixel maps with discrete color scales defined by various number of quantile intervals K. For a fixed K, the ATCP is higher for maps with lower spatial resolution. For a fixed spatial resolution, the ATCP decreases with increasing K. This further illustrates the trade-off between the geographical scale and precision of estimates: given the same ATCP threshold, the state-level coverage estimates can be presented using a more granular color scale than the LGA- and pixel-level estimates, because the state-level estimates generally have higher precision. By setting a minimum threshold for ATCP, we can control the overall uncertainty level of a map, and hence avoid presenting maps with low overall precision.

5.5 Discussion

In this chapter, we have discussed a number of crucial choices with respect to the modeling and presentation of vaccination coverage estimates derived from household survey data. Using the MCV1 coverage among children aged 12–23 months in Nigeria as a motivating example, we fitted several Bayesian geostatistical models to the 2018 NDHS survey data and illustrated the importance of properly accounting for survey stratification and cluster-level non-spatial variation in survey outcome. In addition, we demonstrated the trade-off between the geographical scale and precision of estimates by showing the higher uncertainties associated with estimates at finer spatial resolutions. We also demonstrated several visualization methods for mapping and ranking that emphasize the probabilistic interpretation of results, including maps of posterior medians with CI widths, ridgeline plots of posterior distributions, histograms of posterior ranking distributions, and maps of posterior exceedance probabilities, Finally, a novel approach inspired by Bayesian decision theory was introduced to present

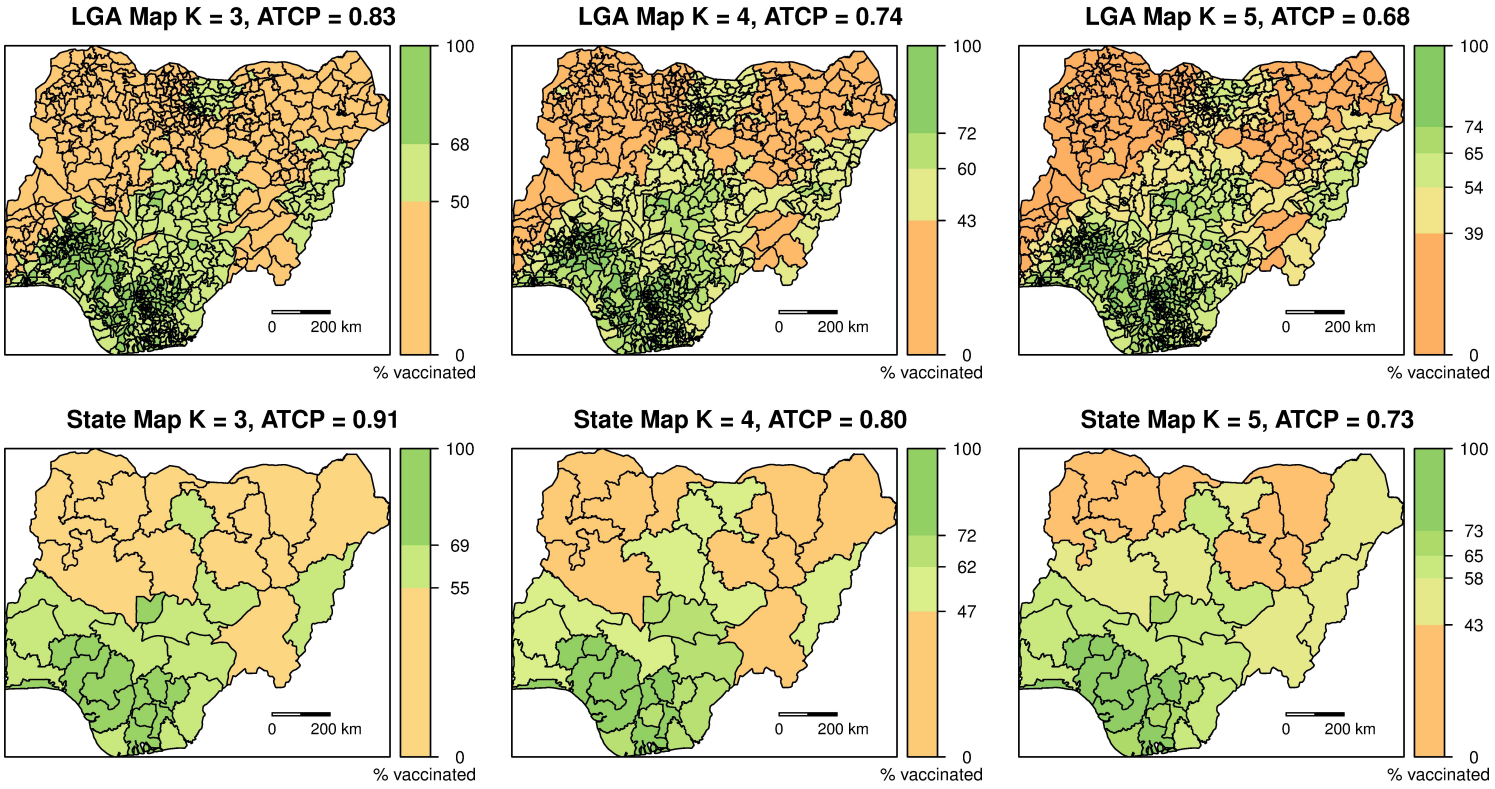

Figure 5.20: Maps of MCV1 coverage estimates at the LGA level (top row) and the state level (bottom row) using a discrete color scale with $K = 3, 4$ and 5 quantile intervals. The color scales are formed by creating the intervals $[L_0, L_1), \ldots, [L_{K-1}, L_K]$, where $L_0 = 0\%$, $L_K = 100\%$, and L_k equals the $100 \times k/K$ quantile of the pooled posterior samples of the coverage estimates at the LGA and state level respectively, based on the *Lono-Binomial OD* model that includes the urban/rural strata variable.

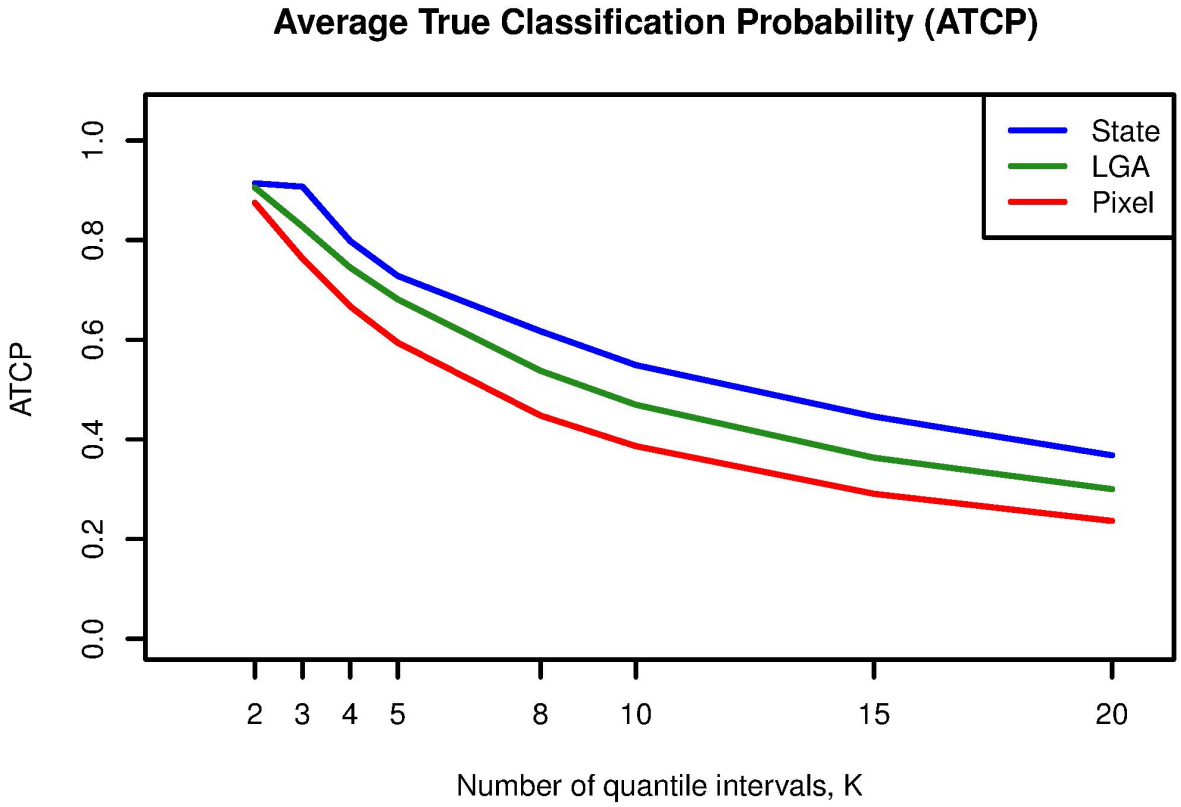

Figure 5.21: The average ture classification probabilities (ATCP) corresponding to the state, LGA and 1×1 km pixel maps with discrete color scales defined by various number of quantile intervals K. For each K within each spatial resolution (state, LGA or pixel), the quantile intervals are formed by creating the intervals $[L_0, L_1), \ldots, [L_{K-1}, L_K]$, where $L_0 = 0\%$, $L_K = 100\%$, and L_k equals the $100 \times k/K$ quantile of the pooled posterior samples of the coverage estimates at that spatial scale based on the *Lono-Binomial OD* model that includes the urban/rural strata variable.

vaccination coverage estimates using discrete color scales, allowing comparison and control of the overall map precision level.

Based on what we have discussed, we recommend the following guidelines for modeling and presenting vaccination coverage estimates using data from household surveys:

- Survey stratification must be acknowledged in the model by including urbanicity as a covariate to avoid bias and improve predictive power.

- Sources of cluster-level non-spatial variation in survey outcomes must be carefully considered, and appropriate model and prediction scheme must be used to properly account for within-cluster variation that induces overdispersion or between-cluster variation that represents true signal.

- With most household surveys being designed to provide statistically reliable estimates at administrative-1 levels, it is critical to acknowledge the potentially high uncertainties associated with coverage estimates at fine spatial resolutions. Quantitative measures of uncertainty should always be properly labelled and presented along with point estimates for user discretion. Map producers should choose appropriate spatial scales to present model results and refrain from showing high-resolution estimates with extremely low precision.

- It is beneficial to use visualization methods that emphasize the probabilistic interpretation of results to identify areas with relatively low or high vaccination coverage.

- Instead of using continuous color scales to show exact point estimates of vaccination coverage, map producers are advised to use our proposed approach to present maps with discrete color scales, which allows comparison and control of the overall precision level of the maps.

It should be noted that the quantile-based procedure we proposed for choosing coverage thresholds is one of the many ways that map producers may use to determine the discrete

color scale — other possible procedures include equal division of the 0% – 100% coverage range given an integer K, or a mixture of posterior quantiles and pre-determined threshold(s) of interest. For any given set of coverage estimates to be presented, the "best" choice for discrete color scale would depend on the specific purpose of the map. The definition and interpretation of ATCP, however, is not influenced by which discrete color scale is used. Therefore, ATCP provides a consistent measure of map quality in terms of the overall map uncertainty and can be used to compare and evaluate any discrete color maps.

An issue that we briefly mentioned in the introduction but did not discuss at length is the challenge of accounting for the jittering of survey cluster locations in vaccination coverage estimation. Most regression models rely heavily on using geospatial covariates to capture outcome variability and make predictions at locations without survey samples. When survey GPS coordinates are displaced to preserve respondents privacy, the traditional method of overlaying covariate surfaces on point locations is going to match the survey clusters with incorrect covariate values. Unfortunately, this issue has been routinely ignored (Takahashi et al., 2017; Mosser et al., 2019) or countered with ad-hoc methods (Gething et al., 2015; Utazi et al., 2018, 2019a,b). Diligent investigation through simulation studies is needed to understand how much an impact the location displacement could make on model prediction accuracy, especially in the context of covariate modeling. The development of statistically sound methods to overcome this challenge is still an active area of research (Wilson, 2019). This issue again points to not displaying pixel-level maps, at least not with pixels that are smaller than the level of jittering.

The accurate mapping of vaccination coverage is an important endeavor in the era of the Sustainable Development Goals (SDGs) with the central focus of "leaving no one behind" (United Nations, 2015). It paves the way towards achieving equity in vaccination by helping us understand the current spatial disparities in attainment of coverage targets. Advances in statistical modeling tools in combination with increased availability of high-resolution geospatial data have enhanced our ability to utilize household surveys to produce vaccination coverage estimates at fine spatial scales. However, it is crucial for us to acknowledge the

limitations in the data and methods and use procedures that are appropriate and statistically vetted. Paying attention to details such as survey design and map uncertainty takes care, patience and diligence; but it is definitely worth the effort considering the ultimate goals it helps achieve for our world.

Chapter 6

DISCUSSION AND FUTURE WORK

Comprehensive evaluation of immunization program performance is essential for monitoring progress and planning future interventions for preventing life-threatening infectious diseases. In resource-limited settings where high quality administrative data is rare, the efficient use of data from disease surveillance systems and household surveys is crucial for obtaining reliable estimates for policy guidance. In this dissertation, we developed several statistical methods for analyzing infectious disease incidence and complex survey data to facilitate assessing and planning immunization programs in LMICs. In particular, we considered estimation of vaccination coverage and immunization campaign effectiveness across space and time.

In Chapter 3, we developed a space-time smoothing model for estimating sub-national RI-specific MCV1 coverage using complex survey data. The proposed method is able to account for the impact of SIAs on the overall MCV1 coverage measured by cross-sectional surveys by incorporating information from the WHO SIA calendar. It can also be used to analyze data from multiple surveys with different data collection schemes and construct coverage estimates with uncertainty that reflects the various sampling designs. We conducted a simulation study that considered a number of scenarios with surveys of various sizes to demonstrate how our method improves estimation through more efficient use of data. We also used the method to analyze data from nine household surveys conducted in Nigeria and constructed the RI-specific MCV1 coverage in each of Nigeria's 37 states from 2003 to 2018.

In Chapter 4, we developed a discrete-time hidden Markov model under the TSIR framework for estimating SIA efficacy with measles incidence data. We proposed a two-stage estimation procedure that separately estimates reporting rate and the other model param-

eters via an OLS regression model with robust standard errors and an MCMC algorithm with uncertainty propagation. Our approach accounts for under-reporting and seasonality of measles transmission, and accommodates monthly reported incidence data that are publicly available from the WHO Measles and Rubella Surveillance Database for most member countries. The model can be used to estimate the underlying susceptible population dynamics, assess how many susceptible people were immunized by past SIAs, and forecast incidence trends in the future under various hypothetical SIA scenarios. Through a simulation study, we investigated the effect of uncertainty propagation under a range of reporting rate scenarios. The model was then used to analyze the monthly reported measles incidence in Benin between 2012 and 2018 as an illustrative example.

In Chapter 5, we discussed a number of crucial choices with respect to the modeling and presentation of vaccination coverage estimates derived from household survey data. Specifically, we focused on the acknowledgement of the survey design in modeling, and the appropriate presentation of estimates and their uncertainties. Using the MCV1 coverage among children aged 12–23 months in Nigeria as a motivating example, we fitted several Bayesian geostatistical models to the 2018 NDHS survey data and illustrated the importance of properly accounting for survey stratification and cluster-level non-spatial variation in the survey outcome. We also discussed the trade-off between the geographical scale and precision of model-based estimates, and demonstrated visualization methods for mapping and ranking that emphasize the probabilistic interpretation of results, including maps of posterior medians with CI widths, ridgeline plots of posterior distributions, histograms of posterior ranking distributions, and maps of posterior exceedance probabilities. Finally, a novel approach inspired by Bayesian decision theory is introduced to present vaccination coverage estimates using discrete color scales, allowing comparison and control of the overall map precision level.

The methods described in this dissertation were developed to provide reliable and useful information for guiding immunization program planning and resource allocation. Future work in analysis of survey and incidence data in space and time includes understanding and

addressing sources of bias in modeling, identifying and accounting for uncertainties associated with different data sources and model components, and developing model assessment tools for validating spatial models fitted to data from surveys with stratified cluster designs. Additionally, software that implements all visualization methods discussed in this dissertation would be useful for quantitative researchers and program officers to analyze data and judge the impact of interventions. An `R` package or an `RShiny` app would be a good starting point to improve the accessibility and utility of these methods.

BIBLIOGRAPHY

Anderson, R. M., Anderson, B., and May, R. M. (1992). *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press.

Angela, N. (2019). Underreporting and reporting delays. In Held, L., Hens, N., O'Neill, P. D., and Wallinga, J., editors, *Handbook of Infectious Disease Data Analysis*, pages 437–454. CRC Press.

Bakka, H. (2018). How to solve the stochastic partial differential equation that gives a Matérn random field using the finite element method. *arXiv preprint arXiv:1803.03765*.

Begon, M., Bennett, M., Bowers, R. G., French, N. P., Hazel, S., and Turner, J. (2002). A clarification of transmission terms in host-microparasite models: numbers, densities and areas. *Epidemiology & Infection*, 129(1):147–153.

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.

Bhattacharjee, N. V., Schaeffer, L. E., Marczak, L. B., Ross, J. M., Swartz, S. J., Albright, J., Gardner, W. M., Shields, C., Sligar, A., Schipp, M. F., et al. (2019). Mapping exclusive breastfeeding in Africa between 2000 and 2017. *Nature medicine*, 25(8):1205–1212.

BID Initiative (2016). Empowering countries to enhance immunization and overall health service delivery through improved data collection, quality, and use. Available at `http://bidinitiative.org/wp-content/uploads/BID_GlobalFactSheet.pdf`. [Accessed Apr-02-2020].

Biellik, R. J. and Orenstein, W. A. (2018). Strengthening routine immunization through measles-rubella elimination. *Vaccine*, 36(37):5645–5650.

Binyaruka, P. and Borghi, J. (2018). Validity of parental recalls to estimate vaccination coverage: evidence from Tanzania. *BMC health services research*, 18(1):440.

Bjørnstad, O. N., Finkenstädt, B. F., and Grenfell, B. T. (2002). Dynamics of measles epidemics: estimating scaling of transmission rates using a time series SIR model. *Ecological Monographs*, 72(2):169–184.

Bolin, D. and Lindgren, F. (2015). Excursion and contour uncertainty regions for latent Gaussian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(1):85–106.

Bolin, D. and Lindgren, F. (2017). Quantifying the uncertainty of contour maps. *Journal of Computational and Graphical Statistics*, 26(3):513–524.

Bosco, C., Alegana, V., Bird, T., Pezzulo, C., Bengtsson, L., Sorichetta, A., Steele, J., Hornby, G., Ruktanonchai, C., Ruktanonchai, N., et al. (2017). Exploring the high-resolution mapping of gender-disaggregated development indicators. *Journal of The Royal Society Interface*, 14(129):20160825.

Chesher, A. (1991). The effect of measurement error. *Biometrika*, 78(3):451–462.

Chiles, J.-P. and Delfiner, P. (2009). *Geostatistics: Modeling Spatial Uncertainty*, volume 497. John Wiley & Sons.

Correll, M., Moritz, D., and Heer, J. (2018). Value-suppressing uncertainty palettes. In *ACM Human Factors in Computing Systems (CHI)*.

Cox, D. R. and Miller, H. D. (1977). *The Theory of Stochastic Processes*, volume 134. CRC Press.

Cutts, F. T., Claquin, P., Danovaro-Holliday, M. C., and Rhoda, D. A. (2016). Monitoring vaccination coverage: defining the role of surveys. *Vaccine*, 34(35):4103–4109.

Cutts, F. T., Izurieta, H. S., and Rhoda, D. A. (2013). Measuring coverage in MNCH: design, implementation, and interpretation challenges associated with tracking vaccination coverage using household surveys. *PLoS medicine*, 10(5).

Diggle, P. J. and Giorgi, E. (2019). *Model-based Geostatistics for Global Public Health: Methods and Applications, Section 5.3*. Chapman and Hall/CRC.

Dolan, S. B. and MacNeil, A. (2017). Comparison of inflation of third dose diphtheria tetanus pertussis (DTP3) administrative coverage to other vaccine antigens. *Vaccine*, 35(27):3441–3445.

Dong, T. Q., Rhoda, D. A., and Mercer, L. D. (2020a). Impact of state weights on national vaccination coverage estimates from household surveys in Nigeria. *Vaccine*, 38(32):5060–5070.

Dong, T. Q. and Wakefield, J. (2020). Space-time smoothing models for estimating sub-national measles routine immunization coverage using complex survey data. *Unpublished manuscript*.

Dong, T. Q., Wakefield, J., McCarthy, K., Thakkar, N., and Frey, K. (2020b). Estimating efficacy of measles supplementary immunization activities via discrete-time modeling of disease incidence time series. *Unpublished manuscript*.

Dunkle, S. E., Wallace, A. S., MacNeil, A., Mustafa, M., Gasasira, A., Ali, D., Elmousaad, H., Mahoney, F., and Sandhu, H. S. (2014). Limitations of using administratively reported immunization data for monitoring routine immunization system performance in Nigeria. *The Journal of Infectious Diseases*, 210(suppl_1):S523–S530.

Feller, W. (1950). *An Introduction to Probability Theory and Its Applications*. New York: John Wiley & Sons Inc.

Ferrari, M. J., Grais, R. F., Bharti, N., Conlan, A. J., Bjørnstad, O. N., Wolfson, L. J., Guerin, P. J., Djibo, A., and Grenfell, B. T. (2008). The dynamics of measles in sub-Saharan Africa. *Nature*, 451(7179):679.

Finkenstädt, B. F. and Grenfell, B. T. (2000). Time series modelling of childhood diseases: a dynamical systems approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49(2):187–205.

Frost, C. and Thompson, S. G. (2000). Correcting for regression dilution bias: comparison of methods for a single predictor variable. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 163(2):173–189.

Fuller, W. A. (2009). *Measurement Error Models*, volume 305. John Wiley & Sons.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016.

Gething, P., Tatem, A., Bird, T., and Burgert-Brucker, C. (2015). Creating spatial interpolation surfaces with DHS data. DHS Spatial Analysis Reports No. 11. Rockville, Maryland, USA: ICF International. Available at `http://dhsprogram.com/pubs/pdf/SAR11/SAR11.pdf`.

Gething, P. W., Casey, D. C., Weiss, D. J., Bisanzio, D., Bhatt, S., Cameron, E., Battle, K. E., Dalrymple, U., Rozier, J., Rao, P. C., et al. (2016). Mapping plasmodium falciparum mortality in Africa between 1990 and 2015. *New England Journal of Medicine*, 375(25):2435–2445.

Glass, K., Xia, Y., and Grenfell, B. (2003). Interpreting time-series analyses for continuous-time biological modelsmeasles as a case study. *Journal of Theoretical Biology*, 223(1):19–25.

Golding, N., Burstein, R., Longbottom, J., Browne, A. J., Fullman, N., Osgood-Zimmerman, A., Earl, L., Bhatt, S., Cameron, E., Casey, D. C., et al. (2017). Mapping under-5 and

neonatal mortality in Africa, 2000–15: a baseline analysis for the Sustainable Development Goals. *The Lancet*, 390(10108):2171–2182.

Graetz, N., Friedman, J., Osgood-Zimmerman, A., Burstein, R., Biehl, M. H., Shields, C., Mosser, J. F., Casey, D. C., Deshpande, A., Earl, L., et al. (2018). Mapping local variation in educational attainment across Africa. *Nature*, 555(7694):48.

Grenfell, B. T., Bjørnstad, O. N., and Finkenstädt, B. F. (2002). Dynamics of measles epidemics: scaling noise, determinism, and predictability with the TSIR model. *Ecological Monographs*, 72(2):185–202.

Griliches, Z. and Ringstad, V. (1970). Error-in-the-variables bias in nonlinear contexts. *Econometrica: Journal of the Econometric Society*, pages 368–370.

Hafen, R. and Schloerke, B. (2020). geofacet: 'ggplot2' Faceting Utilities for Geographical Data. `https://hafen.github.io/geofacet/index.html`. [Accessed Jun-02-2020].

Hájek, J. (1971). Comment on An essay on the logical foundations of survey sampling, part one. *The Foundations of Survey Sampling*, 236.

Hancioglu, A. and Arnold, F. (2013). Measuring coverage in MNCH: tracking progress in health for women and children using DHS and MICS household surveys. *PLoS medicine*, 10(5):e1001391.

Held, L., Höhle, M., and Hofmann, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling*, 5(3):187–199.

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.

Hu, Y., Chen, Y., Wang, Y., and Liang, H. (2019). Validity of maternal recall to assess

vaccination coverage: evidence from six districts in Zhejiang province, China. *International Journal of Environmental Research and Public Health*, 16(6):957.

ICF (2020a). The Demographic and Health Surveys. The DHS Program website. Funded by USAID. https://dhsprogram.com/. [Accessed Apr-02-2020].

ICF (2020b). The DHS Program - GPS Data Collection. https://dhsprogram.com/What-We-Do/GPS-Data-Collection.cfm. [Accessed Apr-02-2020].

IHME (2020a). Institute for Health Metrics and Evaluation. http://www.healthdata.org/. [Accessed Apr-02-2020].

IHME (2020b). Institute for Health Metrics and Evaluation - Acting on Data. http://www.healthdata.org/acting-on-data. [Accessed Apr-02-2020].

Insitutue for Disease Modeling (2020). Epidemiological MODeling software (EMOD) documentation. http://www.idmod.org/documentation. [Accessed Apr-02-2020].

Institute for Health Metrics and Evaluation (2019). Low- and Middle-Income Country Neonatal, Infant, and Under-5 Mortality Geospatial Estimates 2000-2017. https://doi.org/10.6069/9ABZ-XG84. [Accessed Jun-10-2020].

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

Laird, N. M. and Louis, T. A. (1989). Empirical Bayes ranking methods. *Journal of Educational Statistics*, 14(1):29–46.

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

Lohr, S. L. (2009). *Sampling: Design and Analysis*. Nelson Education.

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

Mahmud, A. S., Metcalf, C. J. E., and Grenfell, B. T. (2017). Comparative dynamics, seasonality in transmission, and predictability of childhood infections in Mexico. *Epidemiology & Infection*, 145(3):607–625.

Mbabazi, W. B., Nanyunja, M., Makumbi, I., Braka, F., Baliraine, F. N., Kisakye, A., Bwogi, J., Mugyenyi, P., Kabwongera, E., and Lewis, R. F. (2009). Achieving measles control: lessons from the 2002–06 measles control strategy for Uganda. *Health Policy and Planning*, 24(4):261–269.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall.

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.

Metcalf, C., Bjørnstad, O., Ferrari, M., Klepac, P., Bharti, N., Lopez-Gatell, H., and Grenfell, B. (2011). The epidemiology of rubella in Mexico: seasonality, stochasticity and regional variation. *Epidemiology & Infection*, 139(7):1029–1038.

Metcalf, C., Cohen, C., Lessler, J., McAnerney, J., Ntshoe, G., Puren, A., Klepac, P., Tatem, A., Grenfell, B., and Bjørnstad, O. (2013). Implications of spatially heterogeneous vaccination coverage for the risk of congenital rubella syndrome in South Africa. *Journal of the Royal Society Interface*, 10(78):20120756.

Morton, A. and Finkenstädt, B. F. (2005). Discrete time modelling of disease incidence time series by using Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):575–594.

Mosser, J. F., Gagne-Maynard, W., Rao, P. C., Osgood-Zimmerman, A., Fullman, N., Graetz, N., Burstein, R., Updike, R. L., Liu, P. Y., Ray, S. E., et al. (2019). Mapping diphtheria-pertussis-tetanus vaccine coverage in Africa, 2000–2016: a spatial and temporal modelling study. *The Lancet*, 393(10183):1843–1855.

National Bureau of Statistics (2016). Nigeria National Nutrition and Health Surveys. `http://www.nigerianstat.gov.ng/nada/index.php/catalog/54`, `http://www.nigerianstat.gov.ng/nada/index.php/catalog/53`. [Accessed Apr-02-2020].

National Population Commission - NPC and ICF (2014). Nigeria Demographic and Health Survey 2013 - Final Report. Abuja, Nigeria: NPC and ICF. Available at `http://dhsprogram.com/pubs/pdf/FR293/FR293.pdf`.

National Population Commission - NPC and ICF (2019). Nigeria Demographic and Health Survey 2018 - Final Report. Abuja, Nigeria: NPC and ICF. Available at `http://dhsprogram.com/pubs/pdf/FR359/FR359.pdf`.

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.

Paige, J., Fuglstad, G.-A., Riebler, A., and Wakefield, J. (2020). 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*. To appear.

Patel, M. K., Dumolard, L., Nedelec, Y., Sodha, S. V., Steulet, C., Gacic-Dobo, M., Kretsinger, K., McFarland, J., Rota, P. A., and Goodson, J. L. (2019). Progress toward regional measles eliminationworldwide, 2000–2018. *Morbidity and Mortality Weekly Report*, 68(48):1105.

Porth, J. M., Wagner, A. L., Tefera, Y. A., and Boulton, M. L. (2019). Childhood immu-

nization in Ethiopia: accuracy of maternal recall compared to vaccination cards. *Vaccines*, 7(2):48.

Portnoy, A., Jit, M., Helleringer, S., and Verguet, S. (2018). Impact of measles supplementary immunization activities on reaching children missed by routine programs. *Vaccine*, 36(1):170–178.

Ramakrishnan, R., Venkata Rao, T., Sundaramoorthy, L., and Joshua, V. (1999). Magnitude of recall bias in the estimation of immunization coverage and its determinants. *Indian Pediatrics*, 36:881–886.

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

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.

Rue, H. and Tjelmeland, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields. *Scandinavian Journal of Statistics*, 29(1):31–49.

Shen, W. and Louis, T. A. (1998). Triple-goal estimates in two-stage hierarchical models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):455–471.

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.

SMART Technical Advisory Group (2006). Measuring Mortality, Nutritional Status, and Food Security in Crisis Situations: SMART Methodology Manual v1.0. SMART, Action Against Hunger Canada, `https://smartmethodology.org/survey-planning-tools/smart-methodology/`.

SMART Technical Advisory Group (2017). Measuring Mortality, Nutritional Status, and Food Security in Crisis Situations: SMART Methodology Manual v2.0. SMART, Action Against Hunger Canada, `https://smartmethodology.org/survey-planning-tools/smart-methodology/`.

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

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):485–493.

Steele, J. E., Sundsøy, P. R., Pezzulo, C., Alegana, V. A., Bird, T. J., Blumenstock, J., Bjelland, J., Engø-Monsen, K., de Montjoye, Y.-A., Iqbal, A. M., et al. (2017). Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface*, 14(127):20160690.

Strategic Advisory Group of Experts on Immunization (2019). The Global Vaccine Action Plan 2011-2020. Review and lessons learned. Geneva: World Health Organization; Licence: CC BY-NC-SA 3.0 IGO.

Takahashi, S., Metcalf, C. J. E., Ferrari, M. J., Tatem, A. J., and Lessler, J. (2017). The geography of measles vaccination in the African Great Lakes region. *Nature Communications*, 8(1):1–9.

Team, R Core (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Thakkar, N., Gilani, S. S. A., Hasan, Q., and McCarthy, K. A. (2019). Decreasing measles burden by optimizing campaign timing. *Proceedings of the National Academy of Sciences*, 116(22):11069–11073.

The Measles and Rubella Initiative (2012). Global Measles and Rubella Strategic Plan 2012-2020. https://measlesrubellainitiative.org/learn/the-solution/the-strategy/. [Accessed Apr-02-2020].

The Measles and Rubella Initiative (2019). Routine Immunization. https://measlesrubellainitiative.org/learn/the-impact/routine-immunization/. [Accessed Apr-02-2020].

The World Bank (2019). The World Bank DataBank. https://databank.worldbank.org/data/home.aspx.

Tufte, E. R. (2001). *The Visual Display of Quantitative Information*, volume 2. Graphics Press Cheshire, CT.

UNICEF (2020). The UNICEF Multiple Indicator Cluster Surveys. https://mics.unicef.org/. [Accessed Apr-02-2020].

United Nations (2015). Transforming our world: the 2030 Agenda for Sustainable Development. https://sustainabledevelopment.un.org/post2015/transformingourworld. [Accessed Apr-02-2020].

United Nations (2019). World Population Prospects - Population Division. https://population.un.org/wpp/Graphs/. [Accessed Apr-07-2020].

Utazi, C., Thorley, J., Alegana, V., Ferrari, M., Nilsen, K., Takahashi, S., Metcalf, C., Lessler, J., and Tatem, A. (2019a). A spatial regression model for the disaggregation of areal unit based data to high-resolution grids with application to vaccination coverage mapping. *Statistical Methods in Medical Research*, 28(10-11):3226–3241.

Utazi, C. E., Thorley, J., Alegana, V. A., Ferrari, M. J., Takahashi, S., Metcalf, C. J. E., Lessler, J., Cutts, F. T., and Tatem, A. J. (2019b). Mapping vaccination coverage to explore the effects of delivery mechanisms and inform vaccination strategies. *Nature Communications*, 10(1):1633.

Utazi, C. E., Thorley, J., Alegana, V. A., Ferrari, M. J., Takahashi, S., Metcalf, C. J. E., Lessler, J., and Tatem, A. J. (2018). High resolution age-structured mapping of childhood vaccination coverage in low and middle income countries. *Vaccine*, 36(12):1583–1591.

Utazi, C. E., Wagai, J., Pannell, O., Cutts, F. T., Rhoda, D. A., Ferrari, M. J., Dieng, B., Oteri, J., Danovaro-Holliday, M. C., Adeniran, A., et al. (2020). Geospatial variation in measles vaccine coverage through routine and campaign strategies in Nigeria: Analysis of recent household surveys. *Vaccine*.

Valadez, J. J. and Weld, L. H. (1992). Maternal recall error of child vaccination status in a developing nation. *American Journal of Public Health*, 82(1):120–122.

Verguet, S., Johri, M., Morris, S. K., Gauvreau, C. L., Jha, P., and Jit, M. (2015). Controlling measles using supplemental immunization activities: a mathematical model to inform optimal policy. *Vaccine*, 33(10):1291–1296.

Wakefield, J. (2013). *Bayesian and Frequentist Regression Methods*. Springer Science & Business Media.

Wakefield, J., Dong, T. Q., and Minin, V. N. (2019). Spatio-temporal analysis of surveillance data. In Held, L., Hens, N., O'Neill, P. D., and Wallinga, J., editors, *Handbook of Infectious Disease Data Analysis*, pages 455–476. CRC Press.

Wakefield, J., Haneuse, S., Dobra, A., and Teeple, E. (2011). Bayes computation for ecological inference. *Statistics in Medicine*, 30(12):1381–1396.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(Dec):3571–3594.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, pages 817–838.

Wilson, K. (2019). *Combining Survey and Census Data in Time and Space in a Developing World Context*. PhD thesis, University of Washington.

Wilson, K. and Wakefield, J. (2018). Pointless spatial modeling. *Biostatistics*, 21(2):e17–e32.

World Health Organization (2012). Global Vaccine Action Plan 2011-2020. https://www.who.int/immunization/global_vaccine_action_plan/GVAP_doc_2011_2020/en/. [Accessed Apr-02-2020].

World Health Organization (2017a). Measles position paper. https://www.who.int/immunization/policy/position_papers/measles/en/. [Accessed Apr-02-2020].

World Health Organization (2017b). Substantial decline in global measles deaths, but disease still kills 90000 per year. https://www.who.int/news-room/detail/26-10-2017-substantial-decline-in-global-measles-deaths-but-disease-still-kills-90-00 [Accessed Apr-07-2020].

World Health Organization (2019). Measles fact sheets. https://www.who.int/news-room/fact-sheets/detail/measles. [Accessed Apr-02-2020].

World Health Organization (2020a). Health topics - Immunization. https://www.who.int/topics/immunization/en/. [Accessed Apr-07-2020].

World Health Organization (2020b). Immunization Agenda 2030: A Global Strategy to Leave No One Behind. https://www.who.int/immunization/immunization_agenda_2030/en/. [Accessed Apr-07-2020].

World Health Organization (2020c). Immunization, vaccines and biologicals - Data, statistics and graphics. https://www.who.int/immunization/monitoring_surveillance/data/en/. [Accessed Apr-07-2020].

World Health Organization (2020d). Immunization, vaccines and biologicals - measles. https://www.who.int/immunization/diseases/measles/en/. [Accessed Apr-02-2020].

World Health Organization (2020e). Measles and Rubella Surveillance Data. https://www.who.int/immunization/monitoring_surveillance/burden/vpd/surveillance_type/active/measles_monthlydata/en/. [Accessed Apr-07-2020].

WorldPop (2017). Benin 1km births. Version 2.0 2015 estimates of numbers of live births per grid square, with national totals adjusted to match UN national estimates on numbers of live births. www.worldpop.org.

WorldPop (2019). Nigeria total number of people per 100m grid-cell. https://www.worldpop.org/geodata/summary?id=28029. [Accessed Nov-19-2019].

WorldPop (2020a). The WorldPop Project. https://www.worldpop.org/. [Accessed Apr-02-2020].

WorldPop (2020b). The WorldPop Project - About. https://www.worldpop.org/about. [Accessed Apr-02-2020].

Appendix A

APPENDIX FOR CHAPTER 3

A.1 Autoregressive model of order 1 (AR1) for Gaussian random vectors

The autoregressive model of order 1 (AR1) for the Gaussian vector $\boldsymbol{x} = [x_1, \ldots, x_n]^\intercal$ is defined as

$$
\begin{aligned}
x_1 &\sim \mathrm{N}\left(0, \left(\tau(1-\rho^2)\right)^{-1}\right) \\
x_i &= \rho x_{i-1} + \epsilon_i; \quad \epsilon_i \sim \mathrm{N}\left(0, \tau^{-1}\right) \qquad \text{for } i = 2, \ldots, n
\end{aligned}
$$

where $|\rho| < 1$. The marginal precision κ is equal to

$$\kappa = \tau(1-\rho^2)$$

Hence, the marginal standard deviation σ is equal to

$$\sigma = \kappa^{-1/2} = \left(\tau(1-\rho^2)\right)^{-1/2}$$

A.2 Additional results from the sensitivity analysis using 12-month birth cohorts

Model	ICAR σ^2_α	Space IID σ^2_γ	RW2 σ^2_δ	Time IID σ^2_τ	Space×Time σ^2_ϕ	Survey σ^2_ϵ
IID-IID	81.5	1.4	3.1	0.7	10.0	3.3
ICAR-IID	80.2	1.7	3.1	0.9	11.0	3.2
IID-RW2	23.9	9.8	3.3	4.6	46.6	11.7
ICAR-RW2	17.0	4.0	1.6	0.6	75.1	1.8
IID-AR1	74.4	1.4	3.2	0.9	16.6	3.5
ICAR-AR1	43.0	1.9	3.8	1.1	45.8	4.4

Table A.1: Proportions of variance explained by each random effect component (percent) in the sensitivity analysis with 12-month birth cohort data.

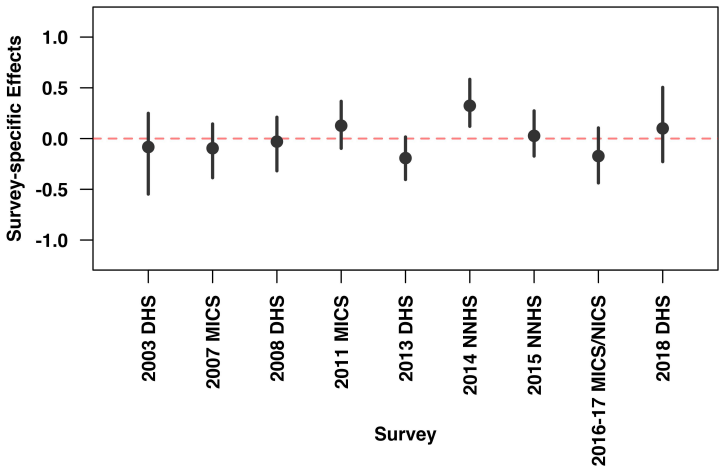

Figure A.1: The posterior medians and 95% CIs of the independent survey-specific effects ϵ_s in the sensitivity analysis with 12-month birth cohort data..

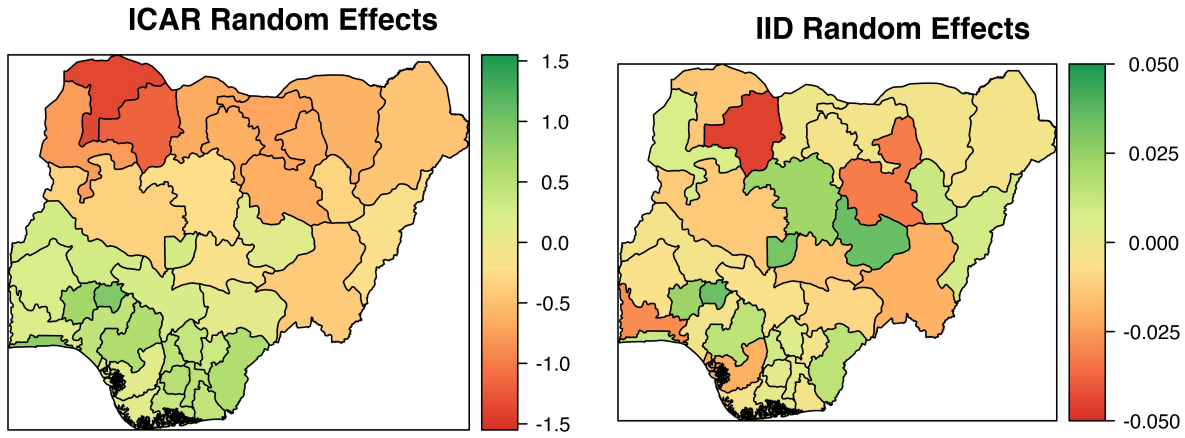

Figure A.2: The posterior medians of the ICAR spatial random effects α_i (left) and the IID random effects γ_i (right) from the fitted *ICAR-AR1 Model* in the sensitivity analysis with 12-month birth cohort data.

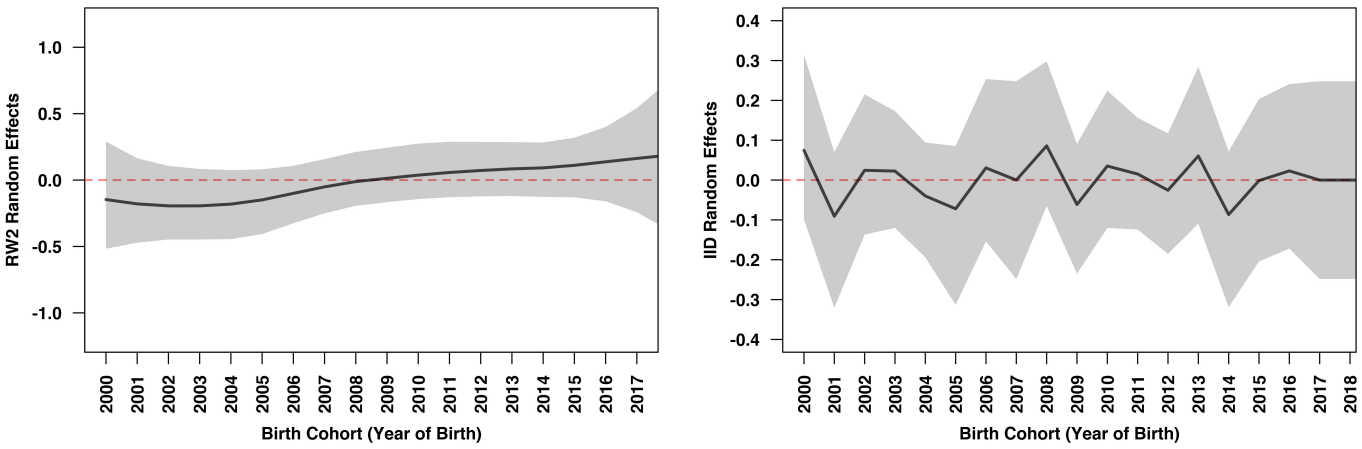

Figure A.3: The posterior medians and 95% CIs of the RW2 temporal random effects δ_b (left) and the IID random effects τ_b (right) from the fitted *ICAR-AR1 Model* in the sensitivity analysis with 12-month birth cohort data.

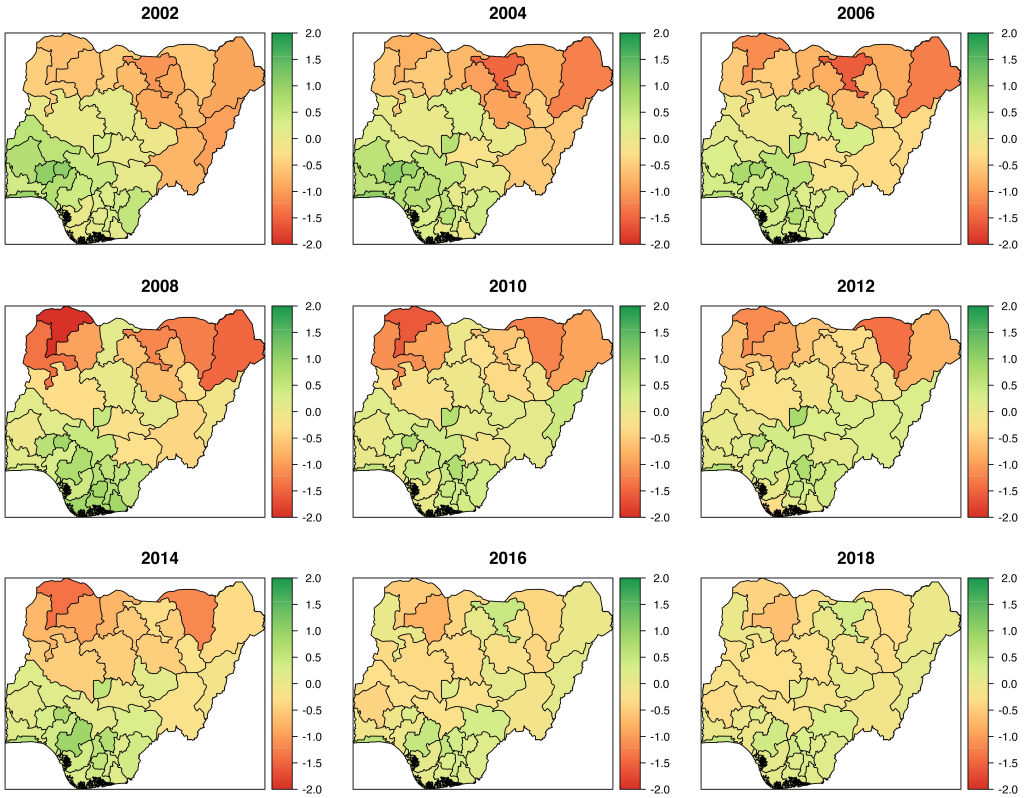

Figure A.4: The posterior medians of the space-time interactions ϕ_{ib} across space over selected 12-month birth cohorts from the fitted *ICAR-AR1 Model* in the sensitivity analysis. Specifically, each map shows the estimated interaction for the cohorts born in the corresponding year.

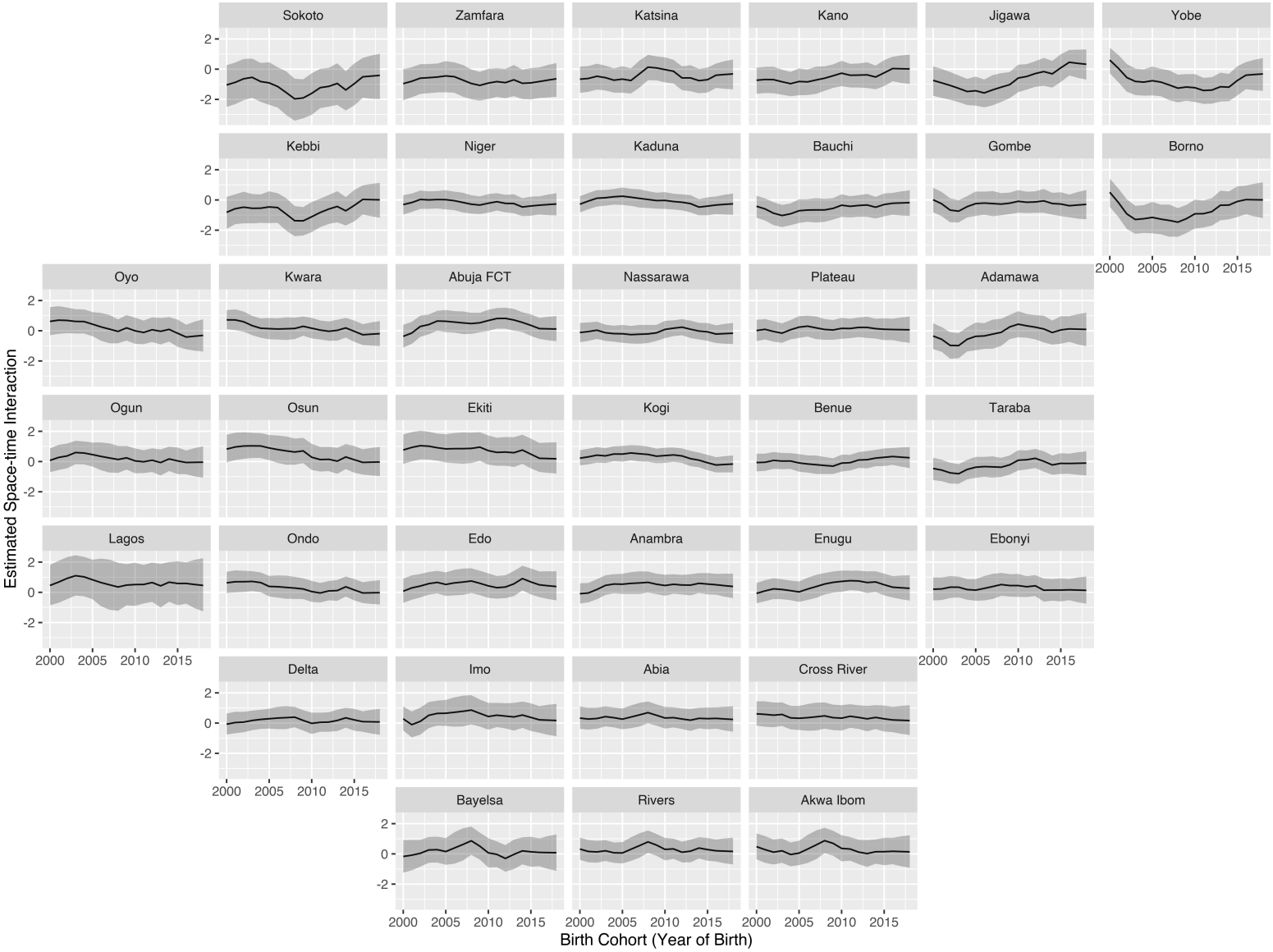

Figure A.5: The posterior medians of the space-time interactions ϕ_{ib} over time in all states from the fitted *ICAR-AR1 Model* in the sensitivity analysis with 12-month birth cohort data.

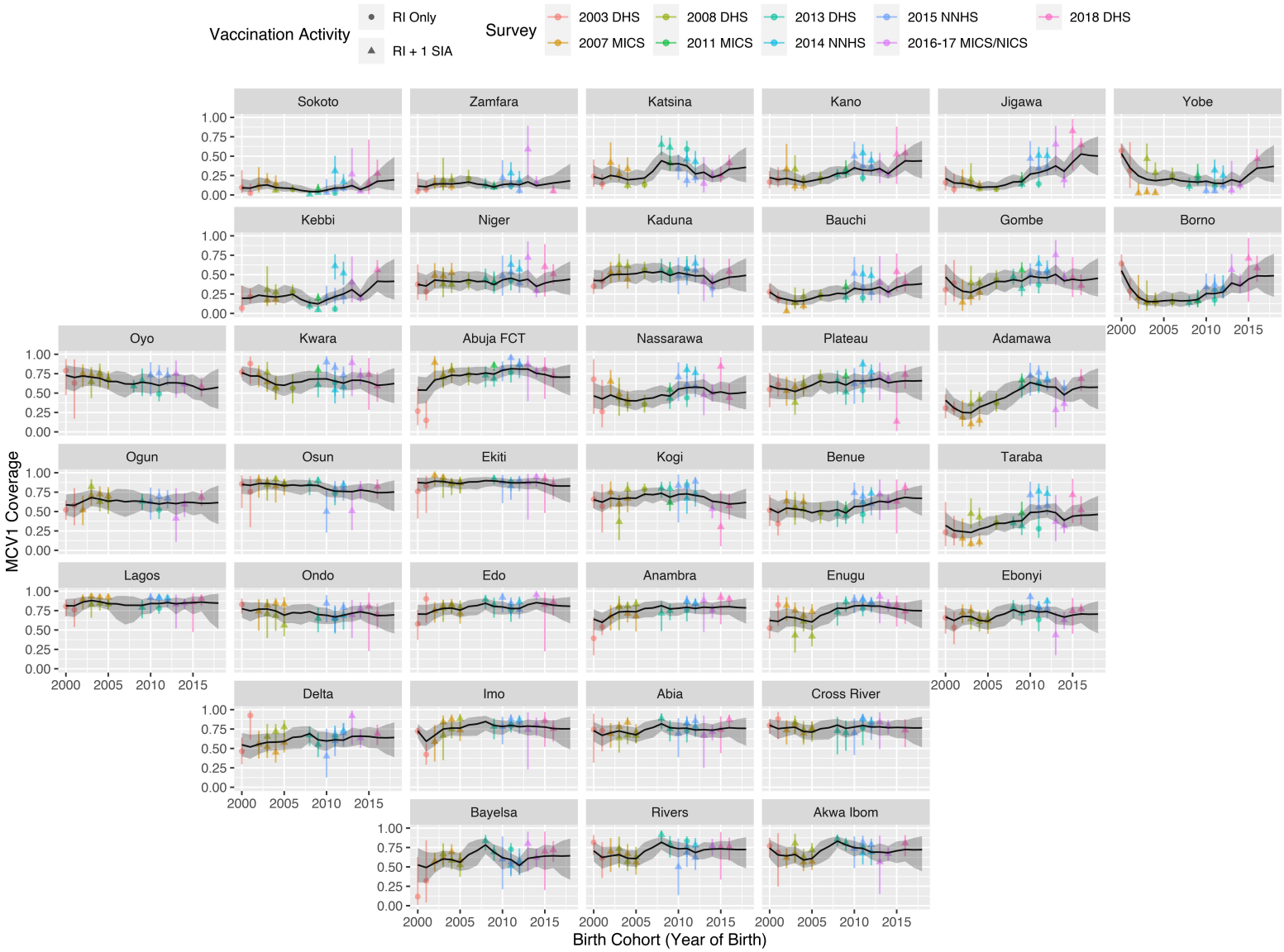

Figure A.6: The posterior medians of the state-level RI-specific MCV1 coverage (dark grey lines) and the associated 95% CIs (light grey ribbons) for the 12-month birth cohorts born between 2000 and 2018 based on the fitted *ICAR-AR1 Model* in the sensitivity analysis. The design-based overall MCV1 coverage estimates were also shown for references.

Appendix B

APPENDIX FOR CHAPTER 4

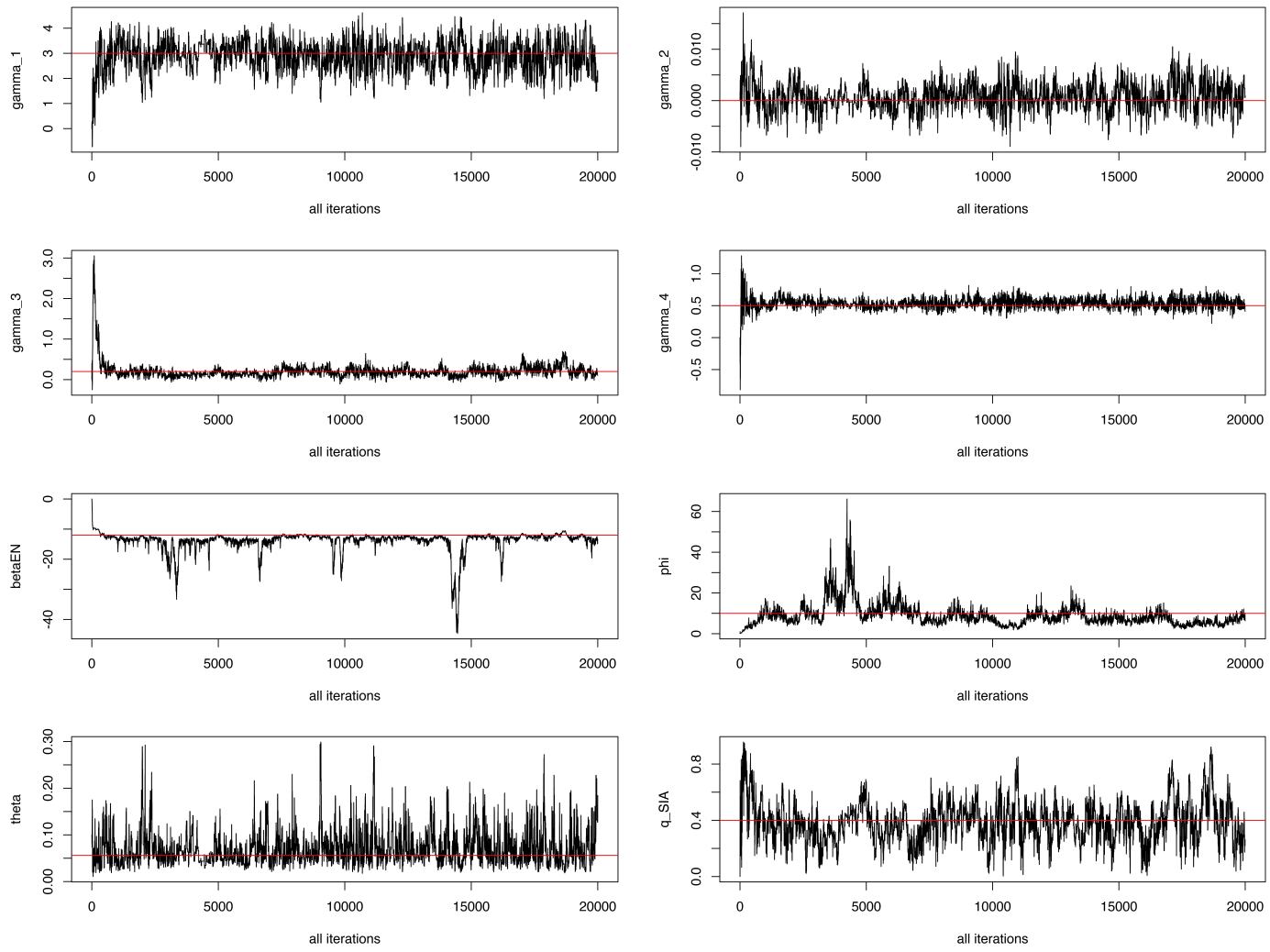

Figure B.1: Traceplots of posterior samples of model parameters in the simulation study computed with uncertainty propagation for when the reporting rate $\rho = 0.01$.

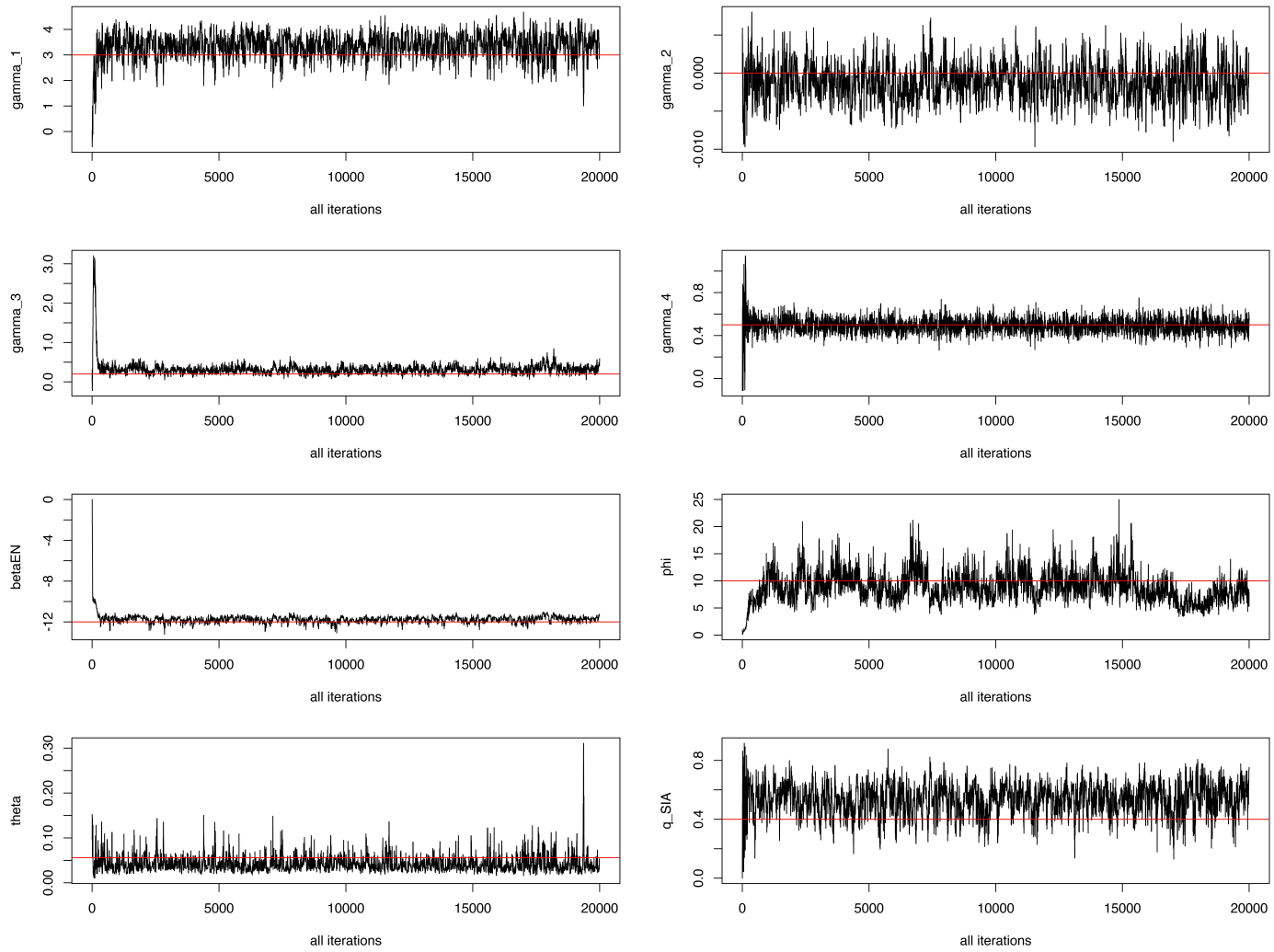

Figure B.2: Traceplots of posterior samples of model parameters in the simulation study computed with uncertainty propagation for when the reporting rate $\rho = 0.1$.

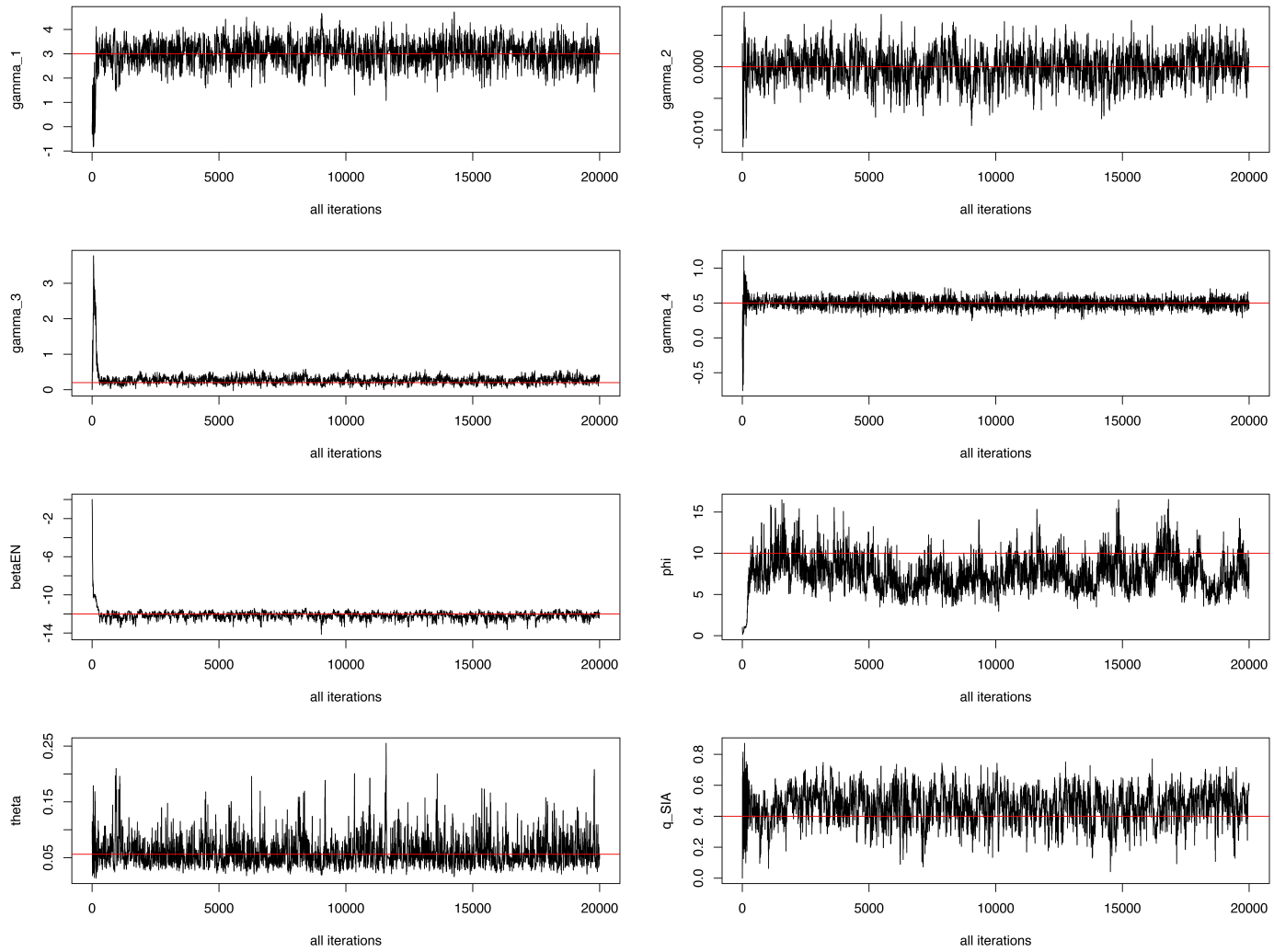

Figure B.3: Traceplots of posterior samples of model parameters in the simulation study computed with uncertainty propagation for when the reporting rate $\rho = 0.3$.

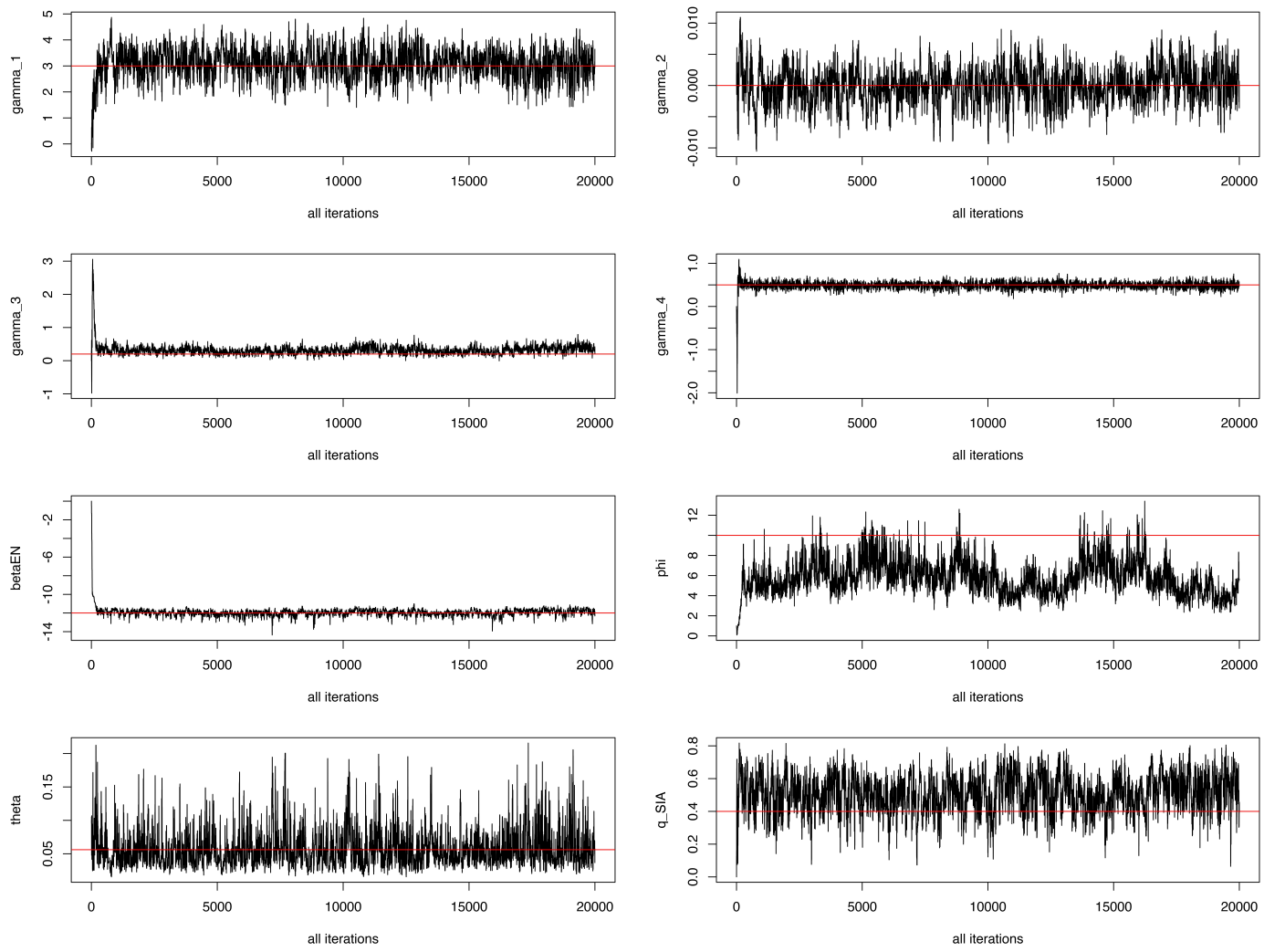

Figure B.4: Traceplots of posterior samples of model parameters in the simulation study computed with uncertainty propagation for when the reporting rate $\rho = 0.5$.

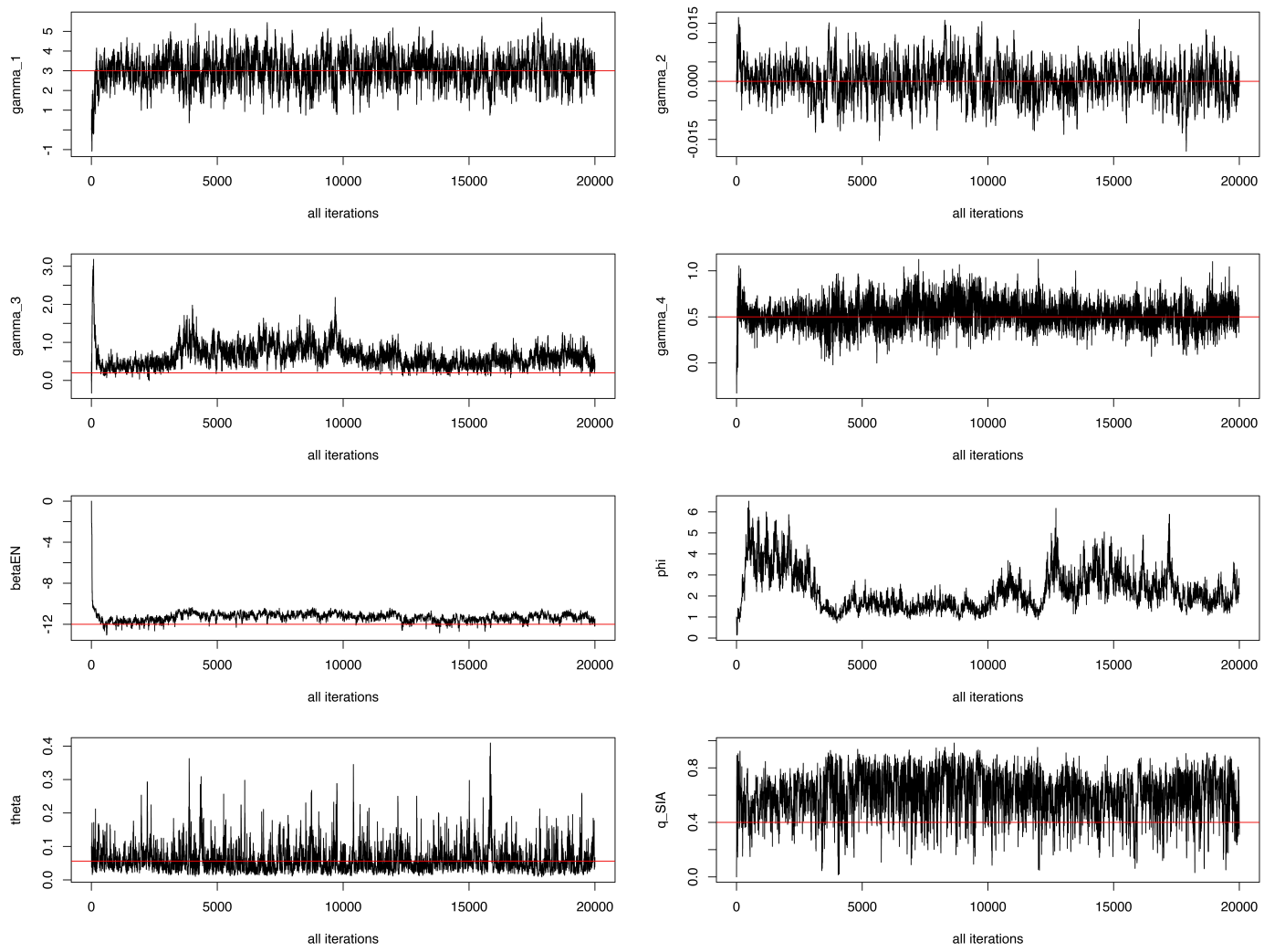

Figure B.5: Traceplots of posterior samples of model parameters in the simulation study computed with uncertainty propagation for when the reporting rate $\rho = 0.7$.

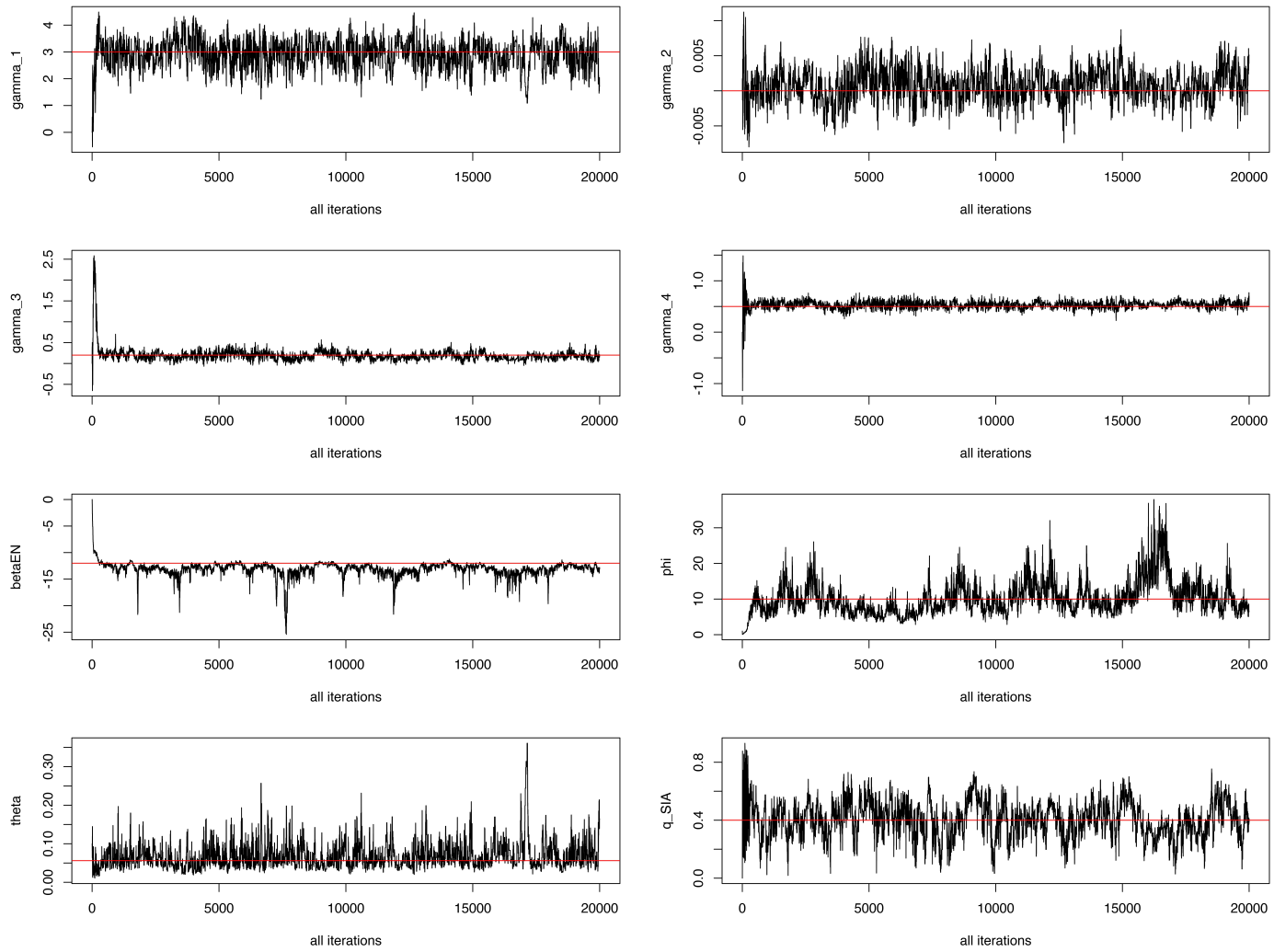

Figure B.6: Traceplots of posterior samples of model parameters in the simulation study computed without uncertainty propagation (i.e., true ρ plugged-in) where reporting rate $\rho = 0.01$.

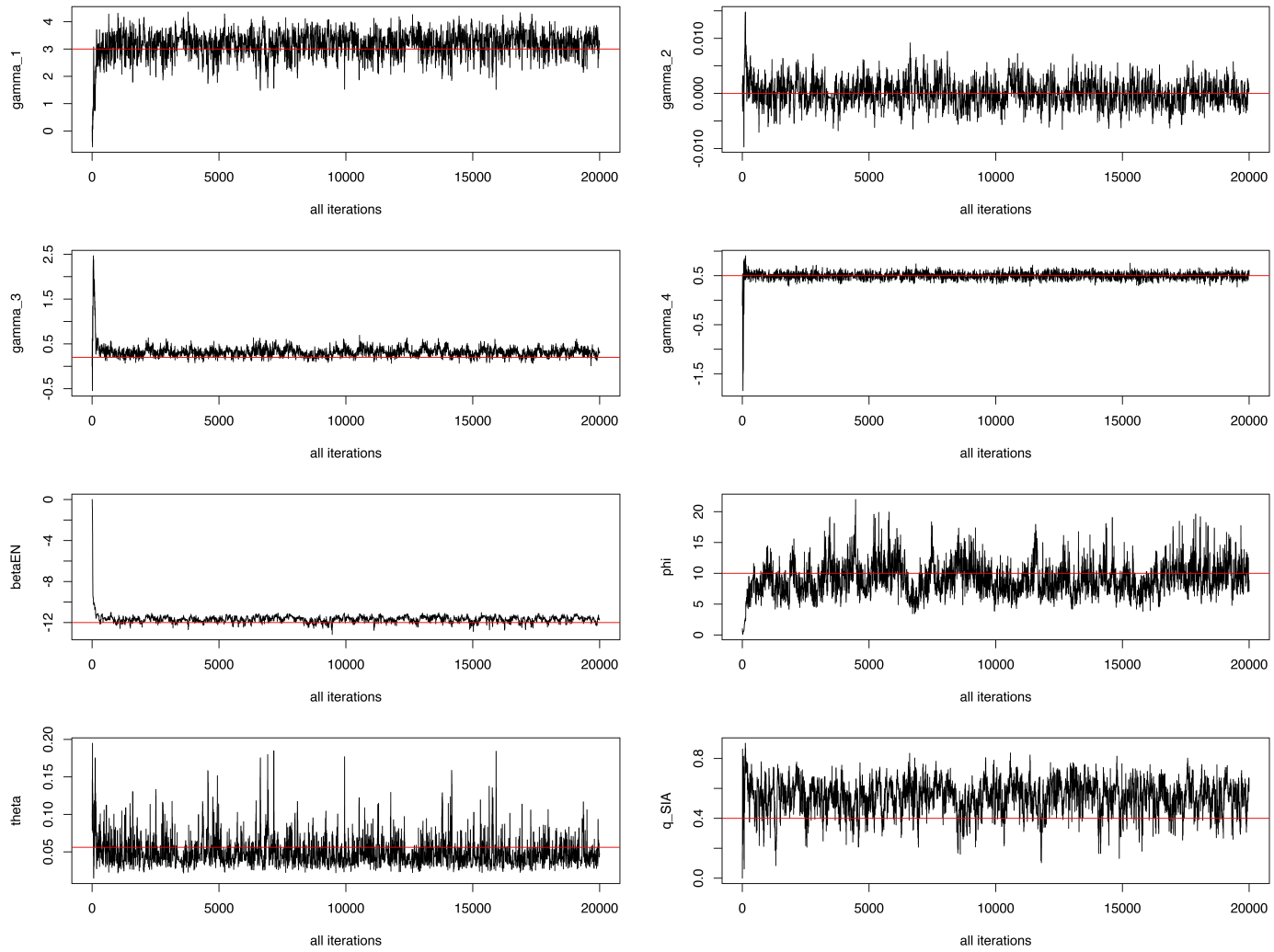

Figure B.7: Traceplots of posterior samples of model parameters in the simulation study computed without uncertainty propagation (i.e., true ρ plugged-in) where reporting rate $\rho = 0.1$.

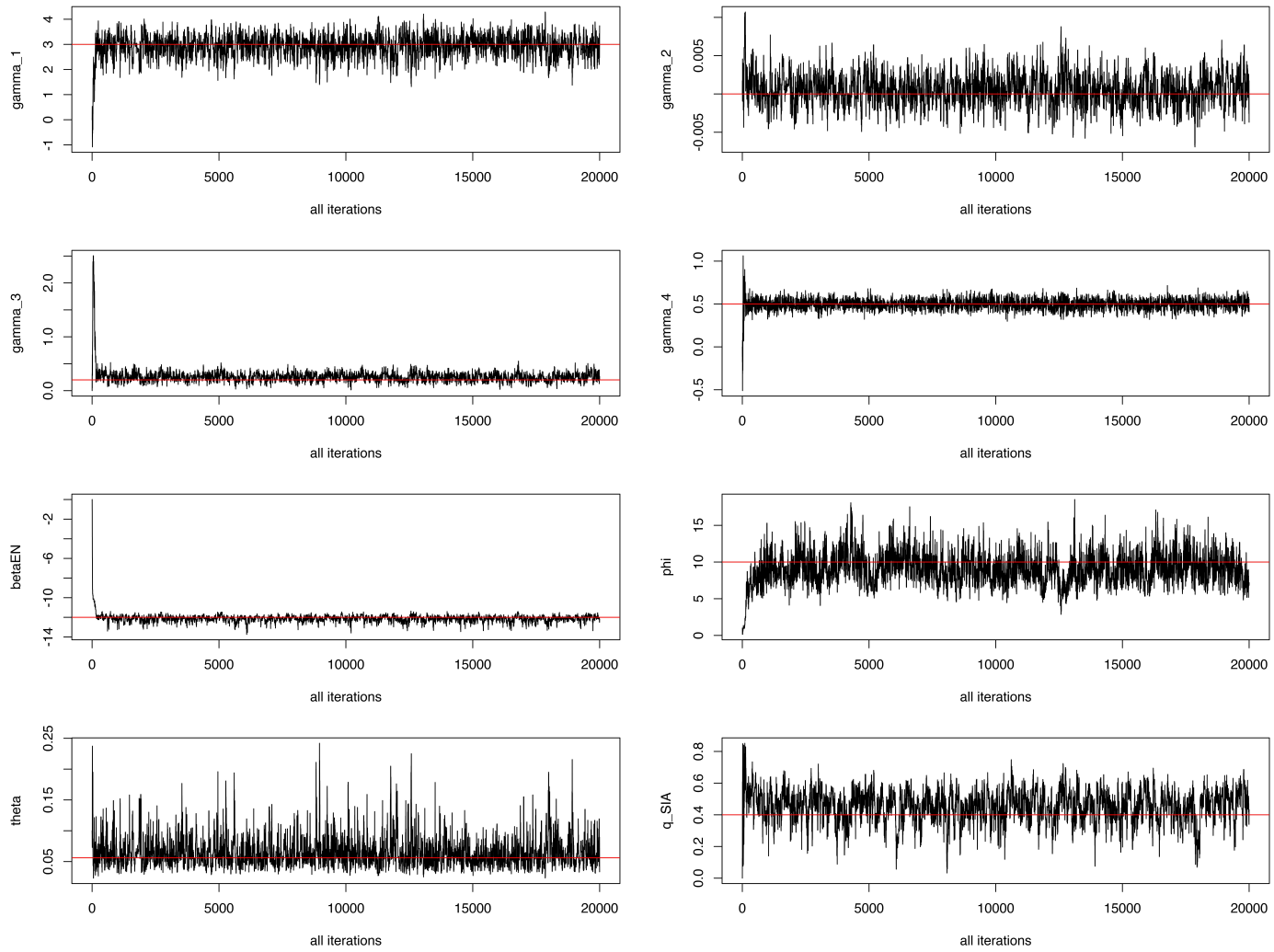

Figure B.8: Traceplots of posterior samples of model parameters in the simulation study computed without uncertainty propagation (i.e., true ρ plugged-in) where reporting rate $\rho = 0.3$.

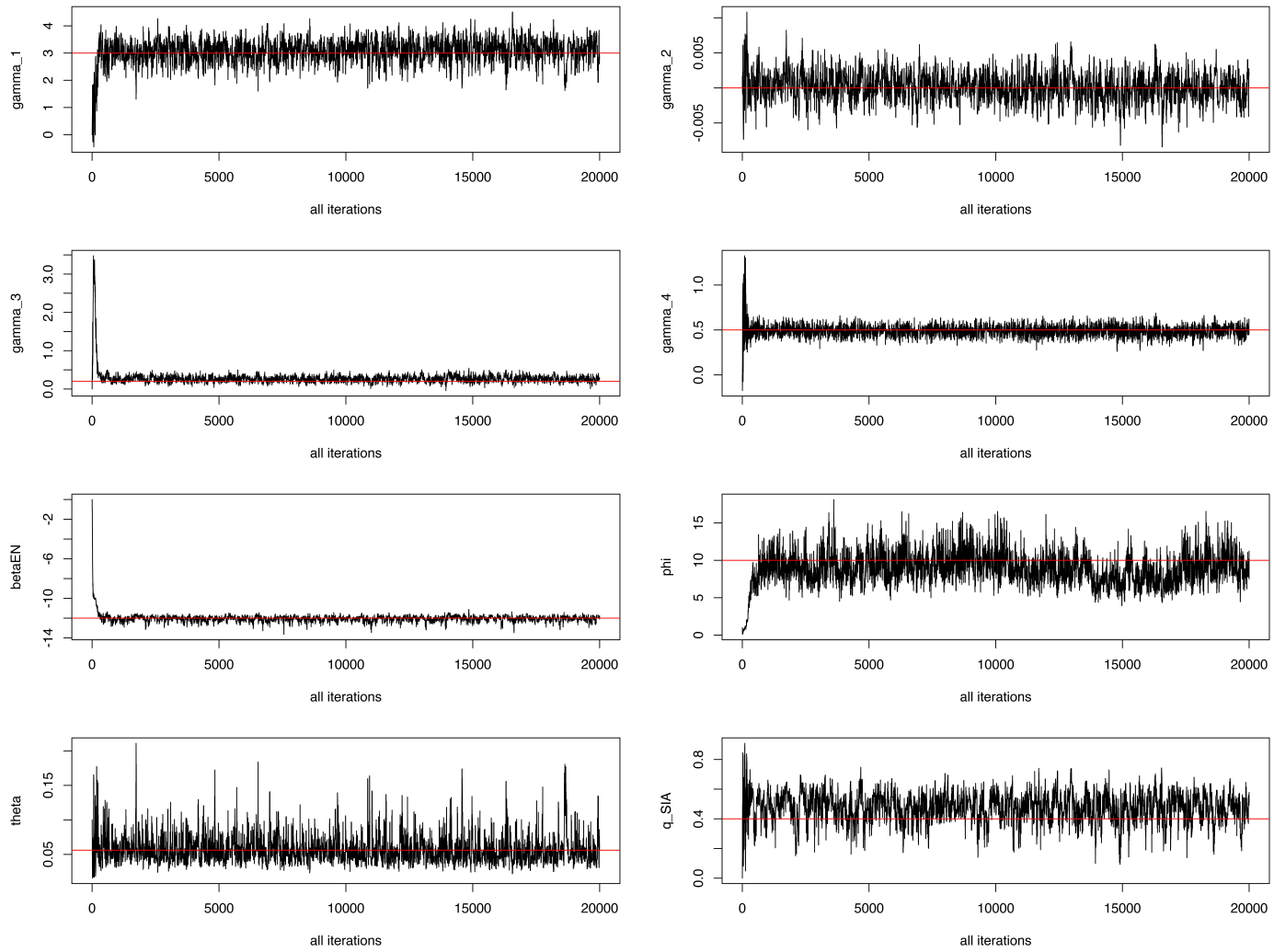

Figure B.9: Traceplots of posterior samples of model parameters in the simulation study computed without uncertainty propagation (i.e., true ρ plugged-in) where reporting rate $\rho = 0.5$.

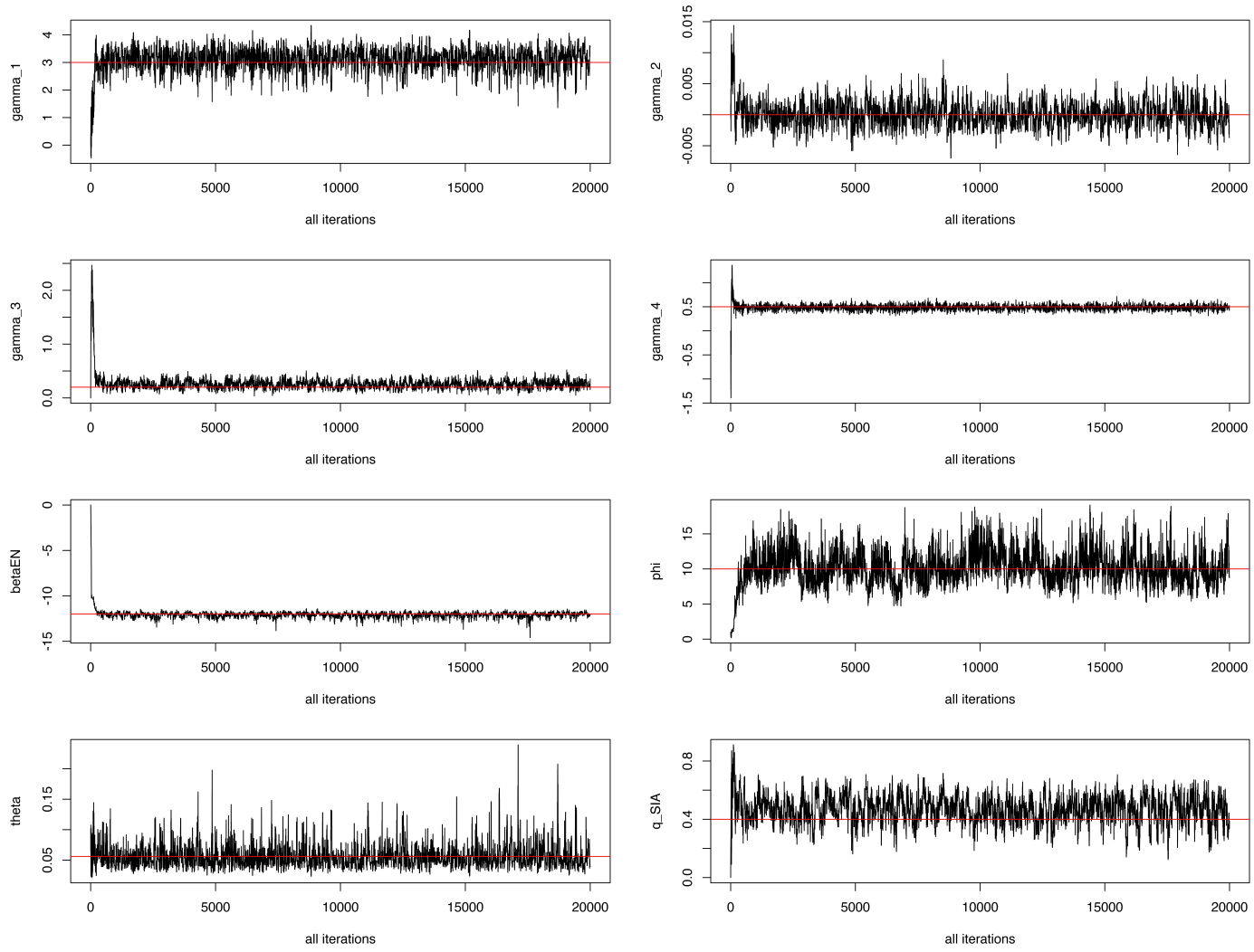

Figure B.10: Traceplots of posterior samples of model parameters in the simulation study computed without uncertainty propagation (i.e., true ρ plugged-in) where reporting rate $\rho = 0.7$.

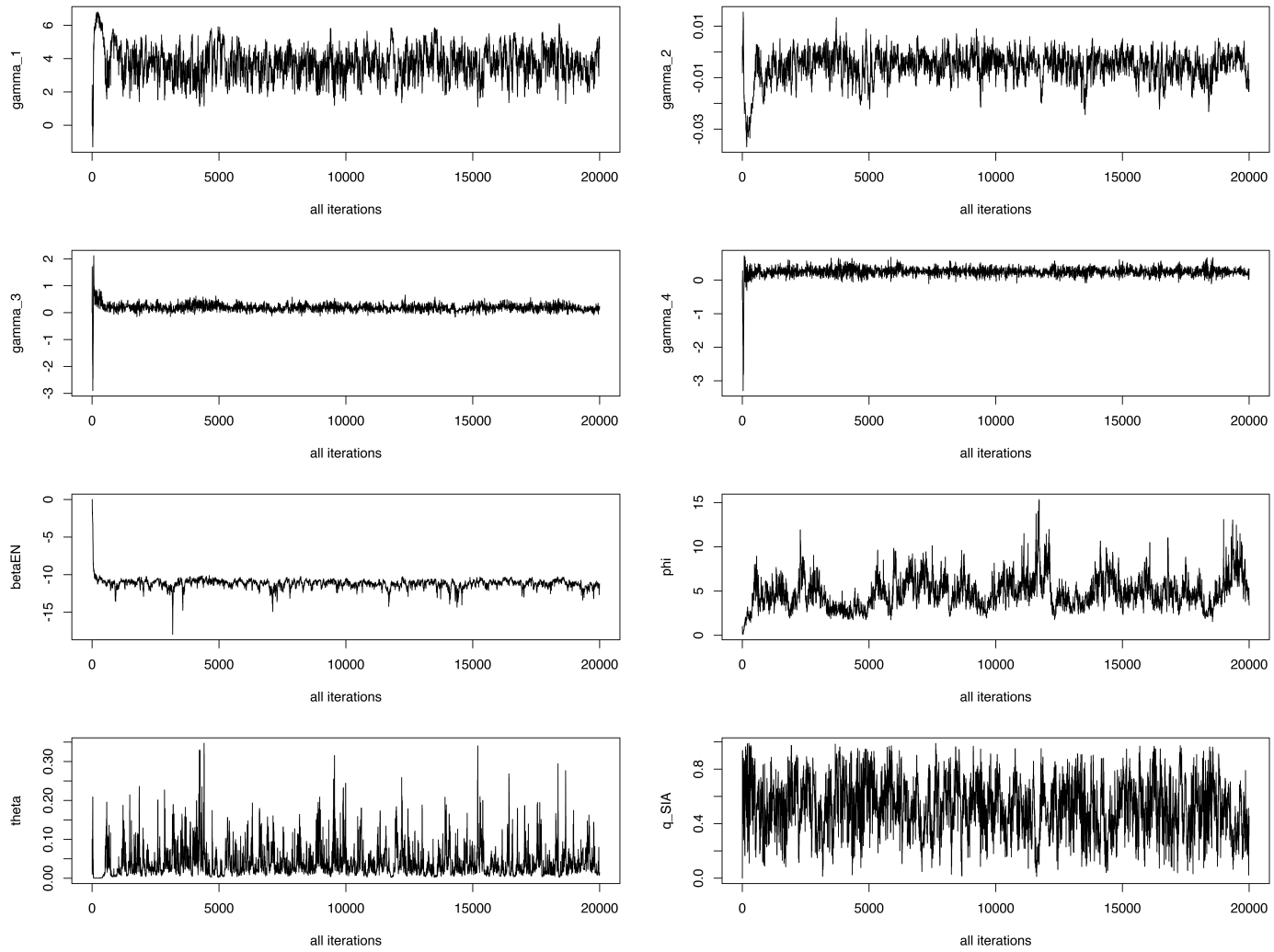

Figure B.11: Traceplots of posterior samples of model parameters in the Benin analysis.

Appendix C

APPENDIX FOR CHAPTER 5

C.1 Covariate data sources

Five geospatial covariates have been used in our analyses: poverty, aridity, log-transformed night-time lights, log-transformed travel time and enhanced vegetation index (EVI). Below are the sources of the covariate data:

- Poverty: Worldpop 2010 estimates of proportion of people per grid square living in poverty, as defined by $2 a day threshold. URL: `https://www.worldpop.org/geodata/summary?id=1267`
- Aridity: Consortium for Spatial Information (CGIAR-CSI) 2018 Global Aridity Index map. URL: `http://www.cgiar-csi.org/data/global-aridity-and-pet-database`)
- Nighttime lights: NOAA VIIRS Nighttime Lights 2018 October data. (`https://www.ngdc.noaa.gov/eog/viirs/download_dnb_composites.html`).
- Travel time: Malaria Mapping Project 2015 Map of Travel Time to Cities. URL: `https://malariaatlas.org/explorer/`
- EVI: MOD13A3 MODIS/Terra vegetation Indices Monthly L3 Global 1km SIN Grid V006, 2018 October data. URL: `https://doi.org/10.5067/MODIS/MOD13A3.006`

C.2 The approximated enumeration area (EA) map

We standardize the Worldpop UN adjusted population layer for 2018 (WorldPop, 2019) into a gridded population surface at 1×1 km resolution. To classify each pixel into urban/rural, we take the following steps:

1. Using Table A.1 in the 2018 DNHS report (National Population Commission - NPC and ICF, 2019), we obtain the urban population proportion within each state.

2. For a given state, we rank the pixels on the gridded population surface **within the state** by descending population size. We start from the top of the ranked list and classify each pixel as urban until the total urban population proportion reach the DHS reported proportion for that state (from step 1). The rest of the pixels in the state will be classified as rural.

3. We repeat step 2 for each state in Nigeria.

This procedure essentially classify the more densely populated pixels within each state to be urban such that the urban population proportion within each state matches what was stated in the DHS report. We emphasize that this is an approximate procedure, since the result of our classification procedure will not match the urban/rural areas in Nigeria exactly — our approach is carried out at the pixel level, whereas the urban/rural areas in Nigeria have been classified at the census enumeration area (EA) level.

Next, we take the following steps to create an approximated EA map:

1. Using Table A.2 in the 2018 DNHS report (National Population Commission - NPC and ICF, 2019), we obtain the number of urban and rural EAs within each state.

2. We overlay the gridded urban/rural population surface with the state/LGA boundaries and assign each pixel to a state/LGA based on its centroid location. We can then calculate the total urban and rural population within each state and within each LGA.

3. Within each state, we can calculate the state-level **average urban/rural EA size** by dividing the urban/rural population by the number of urban/rural EAs in the state.

4. Assuming all the LGAs within a state have the same average urban/rural EA population sizes as the state-level average urban/rural EA population sizes, we can calculate the number of urban/rural EAs within each LGA, denoted $C_j^{(u)}$ and $C_j^{(r)}$ for LGA j, by dividing the urban/rural population by the average urban/rural EA size.

5. Within the urban pixels in LGA j, we can create an approximated urban EA map by sampling $C_j^{(u)}$ pixels with probability proportional to population size with replacement. Pixels with large population counts might be sampled multiple times, meaning that there might be more than one EAs located in a pixel. We can do the same for the rural pixels and create an approximate rural EA map.

6. We repeat steps 3-5 for each state and LGA in Nigeria.

This procedure creates an approximate EA map that matches the summary EA information provided in the DHS report.

VITA

Qi Dong was born to Hongchang Dong and Suben Chen in Qingdao, Shandong, China. She attended boarding school at River Valley High School in Singapore between 2007 and 2010. In 2014, she graduated from Duke University in Durham, North Carolina with a B.S. in Statistical Science and Mathematics. She worked as a statistical analyst at comScore in the Washington D.C. metro area between 2014 and 2015. In 2020, she earned her Ph.D. in Biostatistics from the University of Washington in Seattle, Washington under the supervision of Dr. Jon Wakefield.