Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

Christian Barkald

Model-Based Multiresolution Small Area Estimation

Master's thesis in Applied Physics and Mathematics
Supervisor: Geir-Arne Fuglstad
June 2021

NTNU
Norwegian University of
Science and Technology

Christian Barkald

Model-Based Multiresolution Small Area Estimation

Master's thesis in Applied Physics and Mathematics
Supervisor: Geir-Arne Fuglstad
June 2021

Norwegian University of Science and Technology
Faculty of Information Technology and Electrical Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Abstract

Accurate and reliable estimates for vaccination coverage in low- and middle-income countries is key for planning and enacting public health policies, and when deploying supplementary vaccination programs. Household survey data containing the vaccination status of children is typically analysed with design-based methods, however, these estimates tend to mask geographical inequalities within larger regions. Surveys are usually powered to give reliable estimates for the largest subnational administrative areas, resulting in large uncertainties for estimates in smaller areas. In recent years, effort has been taken to develop model-based approaches for analysing such data to reveal where the need for funding and resources is greatest.

Here, we expand on the well known Besag-York-Mollié (BYM) discrete spatial model, by allowing spatial smoothing between administrative areas on multiple geographical scales. The models we consider are Bayesian hierarchical models with binomial likelihood, which are implemented in the Stan programming language. The new proposed models have a parameter controlling how the geographical variation is split between different spatial scales. We show that multiresolution modeling approaches can improve predictive accuracy compared to the BYM model on a single administrative level, and the interpretability of the model parameters offer valuable insight into the overall variability of vaccination coverage.

A goal of future research is to understand variability in vaccination coverage within countries, and this decomposition of variability can simplify comparison between different regions and countries.

Sammendrag

Nøyaktige og pålitelige estimater for vaksinasjonsdekning i lav- og middelsinntektsland spiller en viktig rolle for planlegging og gjennomføring av folkehelsepolitikk, og for igangsettelse av supplementerende vaksinasjonsprogram. Data fra husstandsundersøkelser som inkluderer informasjon om vaksinasjonsstatus for barn blir typisk analysert med designbaserte metoder, men disse metodene skjuler ofte ulikheter innad i større regioner. Husstandsundersøkelser er designet for å kunne gi pålitelige estimater for de største subnasjonal administrative områdene i et land, som gir store usikkerhetsintervall for mindre områder. Flere modellbaserte metoder har de siste årene blitt brukt for å analysere slike datasett for å finne ut hvor behovet for ressurser er størst.

I denne oppgaven bygger vi på den diskrete romlige Besag-York-Mollié (BYM) modellen, ved å tillate romlig utglatting mellom administrative områder på flere geografiske skalaer. Modellene vi studerer er Bayesianske hierarkiske modeller med binomiske observasjoner, implementert i statistikkprogrammeringsspråket Stan. De nye modellene har en parameter som kontrollerer hvordan den geografiske variasjonen blir delt på forskjellige romlige oppløsninger. Vi viser at modellene med romlig utglatting på flere oppløsninger kan forbedre prediksjonsnøyaktighet sammenlignet med finskalamodeller, og tolkbarheten til modellparametrene gir verdifullt innsyn i geografiske forskjeller i vaksinasjonsdekning.

Målet med videre forskning er å forstå variasjonen av vaksinasjonsdekning innad i ulike land, og dekomponering av variasjonen fleroppløsningsmodellene gir kan forenkle sammenligningen mellom regioner og land.

Preface

This work was conducted during the spring of 2021 at the Department of Mathematical Sciences at the Norwegian University of Science and Technology (NTNU), and completes my Master's degree in Applied Physics and Mathematics.

I would like to thank my supervisor Geir-Arne Fuglstad for invaluable guidance during the work and the preceding specialization project during the fall of 2020. Even under challenging circumstances related to COVID-19 restrictions, the work has been motivating and rewarding.

Finally, I would like to thank my parents who have always supported me during my years at NTNU.

Christian Barkald
Trondheim, June 2021

Contents

1 Introduction

The UN 2030 Agenda for Sustainable Development sets 17 goals with 169 associated targets (United Nations General Assembly, 2015). To inform policymakers on how to best implement programs to improve public health, sustainable development goal (SDG) indicators are regularly monitored in low- and middle-income countries. One program that collects, analyses and disseminates survey data on a wide range of SDG indicators is the Demographic and Health Surveys (DHS) Program (NPC and ICF, 2019). Through over 400 surveys in more than 90 countries, the DHS program has collected representative data on a wide range of demographic and health indicators, such as the vaccination rates. As part of the 2030 agenda, UNICEF has laid out a goal of "leaving no one behind", that all children should be vaccinated (Unicef, 2019). However, in low- and middle-income countries, pockets of children regularly go unvaccinated, mainly due to low availability of health care and lack of human resources.

An efficient and available measles-containing-vaccine first-dose (MCV1) vaccine has been available since 1974, but in 2017 there were still more than 17 million cases of measles globally, and over 80 thousand deaths (Local Burden of Disease Vaccine Coverage Collaborators, 2021). There were big gains in vaccination coverage from 2000 to 2010, but the vaccination coverage has since regressed.

Traditionally, survey data is analysed using design-based methods, which produce estimates of vaccination coverage for larger geographical regions, usually on the largest subnational administrative (admin1) level. However, even in countries with with high vaccination coverages in admin1 regions, design-based estimates tend to mask heterogeneity on finer geographical scales within admin1 regions, such as differences between second-level (admin2) administrative units. Local coldspots in vaccination coverage are often sources of larger outbreaks of disease and can sustain ongoing disease transmission. This is a major obstacle for achieving herd immunity.

Conducting surveys such that design-based estimates are reliable on finer spatial scales, for instance on admin2 level, is unfeasible due to cost and difficult working conditions for the fieldworkers, so in recent years one has sought to

use spatial statistical models for small area estimation with survey data. Some approaches to examine trends in vaccine coverage on fine spatial scales are Utazi et al. (2021), Wang et al. (2018), Utazi et al. (2020), Dong and Wakefield (2020). Typically, these are binomial spatial regression models, with a spatial smoothing random effect.

We will focus on a discrete multiresolution spatial model-based approach for modeling vaccination coverage. Vaccination programs are often administered on admin1 level, where funding and management is allocated. Because of this, it is reasonable to believe that which admin1 region a child lives in plays a key role in their likelihood of being vaccinated. Usually, spatial modeling in small area estimation is done on the finest spatial scale the data allows, before aggregating up to larger regions. One of the goals of the multiresolution approach is to examine whether such fine scale methods are able to detect effects that are determined by admin1 borders. We also want to know if these admin1 effects are large enough for us to be able to detect differences between modeling approaches.

We base our models on the intrinsic conditional autoregressive models, due to Besag (1974). Maps of geographical regions are transformed into graphs, before a joint probability distribution is defined over the nodes, paired with an observation likelihood. The autoregressive models provide spatial smoothing, neighbouring regions are able to borrow information from each other, such that even if the data is sparse, reliable estimates of vaccination coverage can be found in smaller geographical regions. The new approach explored here allows spatial smoothing on multiple administrative scales, which allows a different correlation structure than if only one spatial resolution is utilized.

Nigeria is one of the countries with largest geographical inequality in MCV1 coverage, and has in recent years experienced stagnant or declining vaccination rates due to political instability. To analyse the MCV1 coverage in Nigeria on admin1 and admin2 level, we use the 2018 DHS data from Nigeria, and compare estimates from the multiresolution models, with a discrete fine scale spatial model and with design-based estimates.

We will also score the models by treating design-based estimates as noisy observations of the true vaccination coverages. This allows us to turn predictive distributions of the true coverages from the models, into predictive distributions for the design-based estimates, providing a method for model validation on real data.

Through a simulation study we show that in the presence of a moderate to strong spatial effect on admin1 level, multiresolution modeling approaches outperform a fine scale model in predictive accuracy. However, for the 2018 DHS data from Nigeria the difference between the models' predictive accuracies are minor.

This thesis is organized as follows. In Chapter 2 we present the survey data

used as a motivating example for small area estimation. Then in Chapter 3 we review methods for discrete spatial models and present the models we use to analyze the survey data. An overview of design-based survey statistics, scoring rules for predictive distributions and Bayesian inference based on Hamiltonian Monte Carlo is also given. In Chapter 4 we conduct a simulation study where the predictive accuracy of the different spatial models are compared. The models are applied to analyzing the measles vaccination coverage in Nigeria in Chapter 5, and the results are compared to design-based direct estimates. Finally, we discuss our findings in Chapter 6.

2 | DHS Survey in Nigeria 2018

To motivate the need for small area estimation we consider the vaccination coverage of the MCV1 vaccine, for admin1 and admin2 regions, called states and local government areas, respectively, in Nigeria. The 2018 Nigeria Demographic and Health Surveys (2018 NDHS) survey (NPC and ICF, 2019) is a survey of Nigerian households, providing information about a wide range of demographic and health indicators, including the vaccination status of children. The data collection took place between 14 August and 29 December 2018, and the information is intended to inform policymakers and help design effective programs to improve public health.

The 2018 NDHS used a sampling frame based on the 2006 Population and Housing Census of the Federal Republic of Nigeria. Nigeria is divided into 36 states and one federally controlled area (all 37 are here referred to as states), and further subdivided LGAs, of which there are 774 in total. During the census, each LGA was divided into census enumeration areas (EAs), and it is among the EAs that the primary sampling units (PSUs), referred to as clusters in the survey, were selected for fieldworkers to survey. Figure 2.1 shows a map of Nigeria and its subdivision into administrative areas, together with the location of the survey clusters. Additionally, each enumeration area was classified as either urban or rural.

The sample for the 2018 NDHS was a two-stage stratified sample frame. To achieve higher accuracy for estimates based on survey data, survey units are divided into subgroups known as strata. In the 2018 NDHS, the strata are obtained by separating the EAs in each state by urban/rural status, creating 74 strata in total. Within each stratum, samples are selected following a two stage process. For each stratum a given number of clusters were chosen from the EAs at random with probability equal to their size, then 30 households were selected randomly within each PSU. Both clusters and households were selected without replacement. In total 1400 clusters were selected, but 11 were not surveyed due to dangerous fieldwork conditions.

In the data set, each response corresponds to a child, and it contains informa-

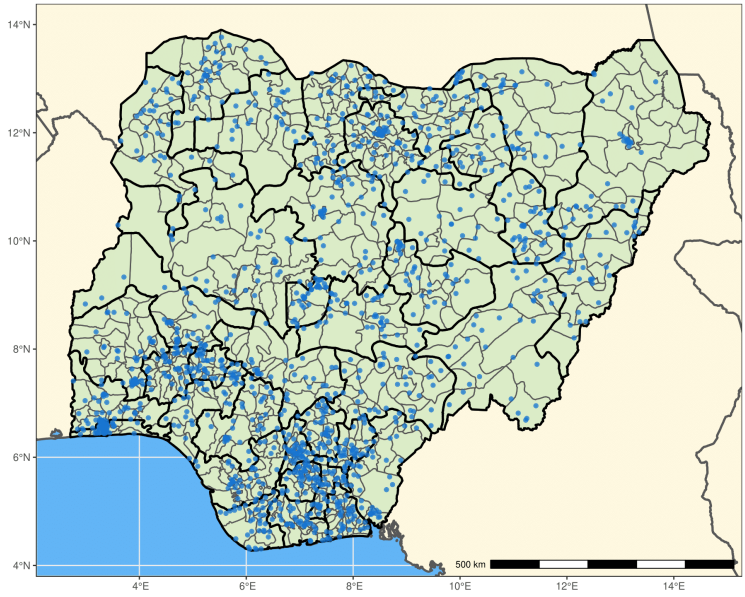


Figure 2.1: Map of Nigeria showing national, state (thick lines) and LGA (thin lines) borders, as well as DHS cluster locations.

tion such as which cluster the response belongs to, the MCV1 status when applicable, and the survey weight (described in Section 3.5). Classical design-based methods, also described in Section 3.5, use the responses directly in weighted estimates. However, for the spatial models we consider, each response is assigned to the LGA its cluster lies in. We then get the aggregated data y_i and n_i, the number of surveyed vaccinated children and total number of surveyed children, respectively, for each LGA i. Note that not all LGAs contain a survey cluster, in which case $y_i = n_i = 0$.

To determine the vaccination status of children in the surveyed households, all eligible women ages 15 to 49 were asked about vaccination status of their children. We consider the MCV1 status for children between ages of 12 and 23 months that were still alive at the time of surveying. The vaccination status of the children is determined either through their vaccination card, or by caregiver recall. In cases with no available vaccination card or caregiver recall, the children are considered unvaccinated. In total the 2018 NDHS contains data for 6036 children, from 637 of the 774 LGAs.

With each survey cluster there is metadata containing the strata to which the cluster belongs, as well as a GPS coordinate of the centroid of the clusters. Due to privacy concerns, the GPS position is scrambled to within 2 km for urban and 10 km for rural clusters, making sure that the scrambled coordinate lies in the correct state and LGA. When processing the data, each cluster is assigned to the LGA its GPS coordinate lies in. However, for some clusters the assigned LGA does not correspond to the reported state. This is likely due to small differences in the maps used. In these cases, the cluster is assigned to the closest LGA that lies in the correct state, as reported in the metadata. Additionally, there are six clusters with missing GPS information. To get a fair comparison between estimates from survey and model based methods, those clusters are omitted from the data.

Finally, Table A.1 in the 2018 NDHS final report (NPC and ICF, 2019) states that in the census frame used in the DHS survey, there are no residents in rural Lagos. However, some clusters are still categorised as rural in the data set. Following a population growth model, these EAs have been reclassified since the last census. We change these observations from rural to urban in our analysis.

In order to aggregate LGA level estimates to state level, we use under five population weighted averages. The under five population counts are found using rasters for 2018 with resolution 100 meters, available from WorldPop (Tatem, 2021). These are estimated population counts extrapolated from the last census, and there is no automatic distinction between urban and rural population. The urban and rural population can be estimated separately using a map of the EAs from the census. However, these maps are not made public, and we do not try to estimate the urban and rural population separately using population density

or other methods.

3 | Theory and Methods

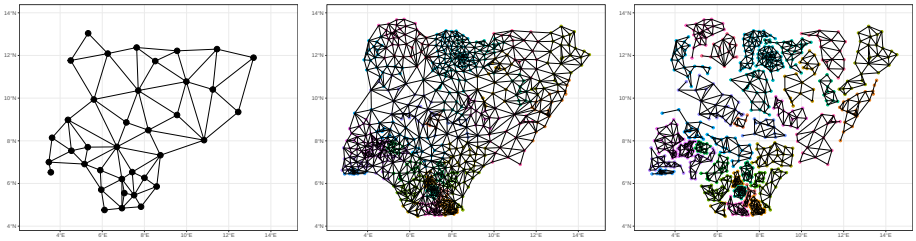
3.1 Gaussian Markov Random Fields

In many applications, such as analysis of spatial data, time series and image processing, a natural modeling approach is to specify the relationship between neighbouring regions. For instance, spatially aggregated data within administrative regions known as areal data, such as the MCV1 data from Nigeria, can be modelled by defining the distribution of the value in each region conditional on the neighbouring regions. Using this approach, we typically view maps as graphs, with regions corresponding to nodes, then model the node values as a multivariate normal random vector.

For a map of non-overlapping regions, let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote the associated graph with nodes $\mathcal{V}$ and edges $\mathcal{E}$, in which all regions are represented by a labelled node. If two regions share a border, the associated nodes are connected by an edge, denoted $i \sim j$ for nodes i and j. Adjacent nodes, two distinct nodes that are both endpoints of the same edge, are called neighbours, and the neighbourhood of a node i is defined as the set of nodes adjacent to i, $N(i) = \{j : i \sim j\}$. Figure 3.1 shows three graph representations of the administrative regions of Nigeria shown in Figure 2.1. The leftmost and middle panel show the graph representation of the states and the LGAs, respectively, while the rightmost panel shows the representation of the LGAs, where only LGAs have to share a border and lie in the same state to be considered neighbours.

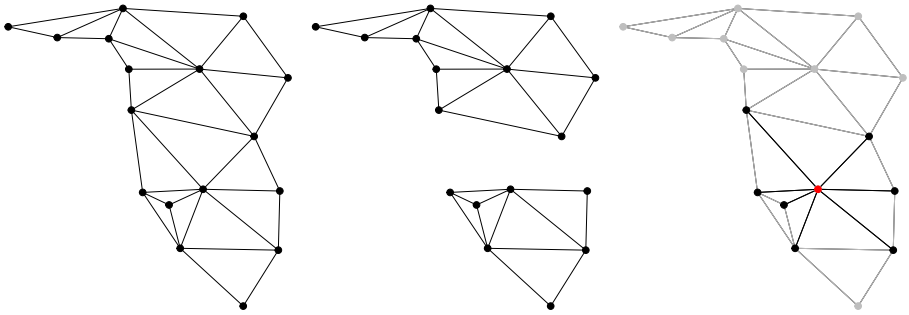
A path from node i_1 to i_m is a sequence of nodes $i_1, i_2, \ldots, i_m$ such that $(i_k, i_{k+1}) \in \mathcal{E}$ for $k = 1, \ldots, m-1$, and sets of nodes such that any two nodes are connected by a path are called connected components. If the graph consists of one connected component, it is said to be connected, and conversely, if there are two or more connected components, it is said to be disconnected. Examples of a connected and a disconnected graph are displayed in Figure 3.2.

We now consider Gaussian Markov random fields (GMRFs), with respect to an undirected graph, such as the graphs of the administrative areas of Nigeria in



(a) States **(b)** LGAs **(c)** LGAs disconnected

Figure 3.1: Graph structure used when analysing DHS data from Nigeria.



(a) Connected graph **(b)** Disconnected graph **(c)** A neighbourhood

Figure 3.2: Examples of graph structures.

Figure 3.1. A GMRF is a finite dimensional random vector, following a multivariate normal distribution, with additional conditional independence assumptions. We restrict ourselves to cases where the conditional distribution in a node only depends on its neighbours, hence the name Markov. The Markov property is closely tied to the inverse of the covariance matrix, called the precision matrix. Rue and Held (2005) gives the following definition for a GMRF;

Definition 3.1.1 (GMRF)**.** A random vector $\boldsymbol{x} = (x_1, \ldots, x_n)^{\mathrm{T}} \in \mathbb{R}^n$ is a GMRF wrt. to a labelled graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with mean $\boldsymbol{\mu} \in \mathbb{R}^n$ and $n \times n$ precision matrix $\mathbf{Q} > 0$, if its density has the form

$$\pi(\boldsymbol{x}) = (2\pi)^{-n/2} |\mathbf{Q}|^{1/2} \exp\left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^T \mathbf{Q} (\boldsymbol{x} - \boldsymbol{\mu})\right), \quad \text{for } x \in \mathbb{R}^n,$$

and

$$Q_{ij} \neq 0 \iff (i, j) \in \mathcal{E} \quad \text{for all } i \neq j.$$

The relationship between the graph $\mathcal{G}$ and the precision matrix $\mathbf{Q}$ can be expressed as $Q_{ij} = 0 \iff x_i \perp x_j | \boldsymbol{x}_{-ij}$, where $\boldsymbol{x}_{-ij}$ denotes all the components of $\boldsymbol{x}$ apart from the ith and jth. That is, the non-zero pattern of $\mathbf{Q}$ corresponds with the edges of the graph. As an example, the conditional distribution of the red node in Figure 3.2c only depends on the black nodes. Hence, the entries of the precision matrix that would have corresponded to edges between the red and grey nodes are all zero. The sparsity of the precision matrix makes GMRFs attractive from a computational point of view, since Cholesky decomposition and evaluation of the probability density is faster than for a multivariate normal random vector with dense precision matrix. For instance, the Cholesky factorization for an $n \times n$ dense precision matrix requires $\mathcal{O}(n^3)$ flops, while sparse precision matrices in spatial GMRFs only require $\mathcal{O}(n^{2/3})$ flops.

Specifying the conditional distribution of each node can lead to the joint distribution being improper, with rank deficient precision matrices. For instance, the joint distribution might end up being invariant of the addition of a constant in each node. Even though these structures are not by themselves proper distributions, they often play key roles as priors in spatial statistics, and are called intrinsic GMRFs. Rue and Held (2005) states the formal definition.

Definition 3.1.2 (Intrinsic GMRF)**.** A random vector $\boldsymbol{x} = (x_1, \ldots, x_n)^{\mathrm{T}} \in \mathbb{R}^n$ is an intrinsic GMRF (IGMRF) of rank $n - k$ with parameters $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\mathbf{Q}$, if $\mathbf{Q}$ is an $n \times n$ symmetric positive semidefinite matrix and its density is

$$\pi(\boldsymbol{x}) = (2\pi)^{-n/2} (|\mathbf{Q}|^*)^{1/2} \exp\left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^T \mathbf{Q} (\boldsymbol{x} - \boldsymbol{\mu})\right), \quad \text{for } x \in \mathbb{R}^n,$$

where $|\mathbf{Q}|^*$ is defined as the product of all non-zero eigenvalues of $\mathbf{Q}$.

It is an IGMRF wrt. to a labelled graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ if

$$Q_{ij} \neq 0 \iff (i,j) \in \mathcal{E} \quad \text{for all } i \neq j.$$

When simulating from IGMRFs additional constraints have to be imposed. Usually, a weighted sum-to-zero constraint and an intercept are added to each connected component. Such constraints are also strongly recommended when using IGMRFs as components in larger models (Freni-Sterrantino et al., 2018), to make the interpretation of each model component as clear as possible.

3.2 Intrinsic Conditional Autoregressive Models and Scaling of Precision Matrices

A common class of IGMRFs are intrinsic conditional autoregressive (ICAR) models, due to Besag (1974). For applications such as disease mapping and image processing, the goal is to borrow strength between neighbours in a graph structure. This is especially useful when the data is sparse and we believe that neighbouring nodes share similar characteristics.

The density of an ICAR model is

$$\pi(\boldsymbol{x}|\kappa) \propto \exp\left(-\frac{\kappa}{2}\sum_{i\sim j} w_{ij}(x_i - x_j)^2\right),$$

where κ is a overall precision parameter, $i \sim j$ denotes that node i and j are neighbours in the associated graph, and w_{ij} are symmetric weights. The weights can be chosen in a number of ways, for instance the Euclidean distance between nodes or set to all be unitary.

Perhaps the simplest of all ICAR models is the version with unitary weights, often referred to simply as the Besag model. It has the conditional formulation

$$x_i|\boldsymbol{x}_{-i}, \kappa \sim \mathcal{N}\left(\sum_{j\in N(i)} x_j/n_i, (\kappa n_i)^{-1}\right), \quad i = 1, \ldots, n,$$

where n_i denotes the number of neighbours of node i, and n is the total number of nodes.

For a graph with k connected components the joint distribution, expressed as an IGMRF is

$$\pi(\boldsymbol{x}) \propto \kappa^{(n-k)/2} \exp\left(-\frac{\kappa}{2}\sum_{i\sim j}(x_i - x_j)^2\right), \tag{3.1}$$

which can be reformulated in terms of the $n \times n$ structure matrix $\mathbf{R}$, defined by

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

It follows that $\boldsymbol{x}$ is an IGMRF with mean zero and precision $\mathbf{Q} = \kappa\mathbf{R}$. The rank deficiency is resolved through a weighted sum-to-zero constraint.

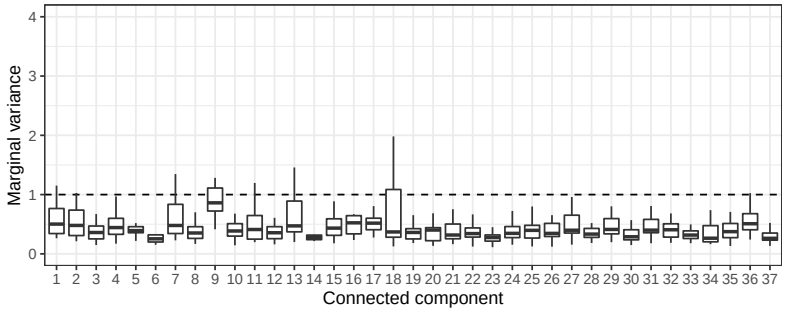
It is a feature of all intrinsic CAR models that the marginal variances in the nodes differ, and will depend on the graph structure itself. In order to have a meaningful interpretation of the precision parameter κ, we need to scale the structure matrix. Following the recommendations of Freni-Sterrantino et al. (2018), for graphs with one connected component we scale the precision matrix using the geometric mean of the marginal variances obtained with $\kappa = 1$. Let $\mathbf{R}^-$ be a generalized inverse of the structure matrix corresponding to a connected graph. Then the scaling factor becomes

$$S = \exp\left(\frac{1}{n}\sum_{i=1}^{n}\log\left(\left[\mathbf{R}^-\right]_{ii}\right)\right),$$

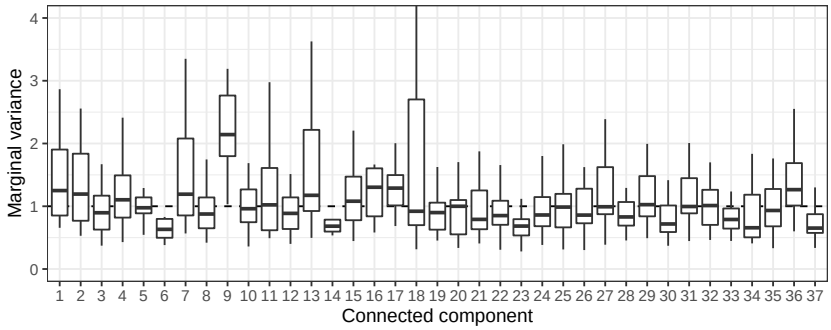
which gives the scaled precision matrix $\mathbf{Q}_{\text{scaled}} = \kappa S\,\mathbf{R}$. This ensures that the marginal variances in the nodes is approximately equal to κ^{-1}.

Similarly, when there are k connected components of sizes greater than one, it is recommended to use a separate scaling factor for each component. Let $\mathbf{R} = \mathbf{R}_1 + \ldots + \mathbf{R}_k$ denote the structure matrices of each of the k connected components (if the nodes are labeled in order of connected component, $\mathbf{R}$ is a block diagonal matrix). Then the scaled precision matrix becomes $\mathbf{Q}_{\text{scaled}} = \kappa\left(S_1\,\mathbf{R}_1 + \ldots + S_k\,\mathbf{R}_k\right)$, where S_i is the scaling factor of the ith connected component.

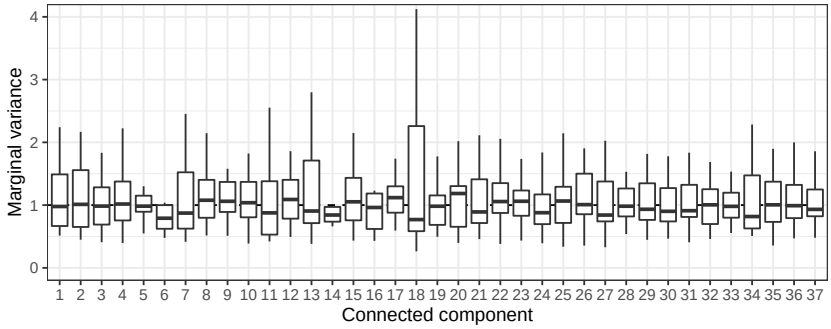
Figure 3.3 shows the effect of different methods of scaling the marginal variances of the ICAR model, defined on the disconnected graph of Nigeria shown in Figure 3.1c. The estimated marginal variances are found empirically using 500 realizations of the Besag model, before being grouped by connected component. The boxplots clearly show that before scaling the marginal variances are about 0.5 on average, and that using a common scaling factor for all connected components just scales the overall marginal variances. The difference between connected components are very large. For instance, connected component 9 has



(a) Unscaled



(b) All connected components scaled by the same factor



(c) Connected components scaled separately

Figure 3.3: Boxplot of the marginal variances in each node of an Besag GMRF, grouped by connected components along the x-axis, with different methods of scaling. The graph used is the disconnected graph of LGAs in Nigeria, shown in Figure 3.1c, with $\kappa = 1$.

marginal variances about twice the average. By using separate scaling factors for each components, shown in the bottom row, the average of the marginal variances in each node grouped by connected component are about equal.

To resolve the issue of improper distributions we want to impose linear constraints on a random vector $\boldsymbol{x}$. There are multiple ways of doing this, with different computational properties. Let $\odot$ denote the element wise product between two vectors, let $\boldsymbol{e}$ be a vector of length k, let $\boldsymbol{w}$ be a weight vector of length n, and let $\mathbf{B}$ denote a $k \times n$ constraint matrix. The two main constraint alternatives are hard constraints

$$\mathbf{B}(\boldsymbol{w} \odot \boldsymbol{u}) = \boldsymbol{e},$$

and soft constraints

$$\mathbf{B}(\boldsymbol{w} \odot \boldsymbol{u}) \sim \mathcal{N}\left(\boldsymbol{e}, \epsilon\, \mathbf{I}_k\right),$$

for some small ϵ. Here we will use soft constraints, as they usually make samplers perform better provided ϵ is not too small.

One alternative when using the Besag model as a part of larger model is to combine it with a n dimensional normal iid. effect. This is called the Besag-York-Mollié (BYM) model (Besag et al., 1991). We will use this model as a part of larger hierarchical models, with the weighted parametrization presented in Simpson et al. (2017)

The BYM model consists of a scaled ICAR component $\boldsymbol{u}$ of the form found in Expression (3.1), and a iid. normal component $\boldsymbol{v}$, both sharing the same variance κ^{-1}. Because of confounding between the spatially correlated effect and the independent effect, it is often much easier to estimate the total variance than the variance in each of the components separately. Therefore, it is useful to parametrize the model using a weight $0 \leq \theta \leq 1$,

$$\psi_i = \sqrt{\theta}\, u_i + \sqrt{1-\theta}\, v_i, \text{ for } 1 \leq i \leq n.$$

3.3 Model Formulation

We consider four different Bayesian hierarchical spatial models to make predictions about vaccination coverage for LGAs. In this section we describe the observation model, the spatial smoothing and the prior distributions, while in Section 3.4 we describe how the estimates for LGA level vaccination coverage are aggregated to estimate state level vaccination coverages, as well as linear constraints put on the GMRFs used.

Let Y_i and n_i denote the surveyed number of vaccinated children and the total number of surveyed children in LGA i, respectively. We assume that the

number of vaccinated children surveyed in each region follow a binomial spatial regression model given by

$$Y_i|n_i, p_i \sim \text{Binomial}\,(n_i, p_i)\,,$$
$$\text{logit}\,(p_i) = \eta_i = \mu + \psi_i,$$

for all LGAs i, where p_i denotes the vaccination coverage in region i, μ is an intercept, and $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_n)$ denotes a spatial random effect, described later in the section.

The spatial random effect $\boldsymbol{\psi}$ is the only component that differs for the four different models. The first model we consider, referred to as the Admin2 model, uses a BYM component with the graph structure of the LGAs, shown in Figure 3.1b. The second model, referred to as the Admin1 model, uses a BYM component with the graph structure of the states, shown in Figure 3.1a.

We will also consider two multiresolution models based on the BYM model. The third model, referred to as the Disconnected model, has one BYM component on state level, and one on LGA level, where the LGAs are disconnected between states, as shown in Figure 3.1c. That is, LGAs $i_1 \sim i_2$ if and only if the two LGAs share a border and they lie in the same state. Finally, the fourth model, referred to as the Connected model, has the same form as the disconnected model, but we use the full graph structure of the LGAs.

Following the idea of the weighted parametrization of the BYM model described in Section 3.1, the total overall variance is distributed between the state effect and LGA effect with a weight w_0 for the multiresolution models. The weight between the Besag and iid. effect on state and LGA level is denoted w_1 and w_2, repectively. Furthermore, the effects on state and LGA level share a common precision parameter κ. This means that for the two multiresolution models the variance of the BYM state component is $w_0\kappa^{-1}$, and $(1 - w_0)\kappa^{-1}$ for the BYM LGA component.

Table 3.1 summarizes the four models. In the expressions for the linear predictors η_i, let $\text{st}[i]$ denote which state the LGA i lies in, N_1 is the number of states, N_2 is the number of LGAs. The Besag and iid. component on state level are denoted by u_i and v_i, respectively, and similarly on LGA level we use ϕ_i and ε_i. Note that ϕ_i uses the disconnected LGA graph, while ϕ_i^* uses the connected version of the graph.

The Admin1 model is in a sense a special case. Even though the model is defined using LGA level data, all LGAs in the same state share the same vaccination coverage. In practice, we aggregate the data over the states when implementing the Admin1 model and it provides immediate estimates for the state level coverage.

Table 3.1: Model specification for the four different models considered. Each of the Besag GMRFs ($\boldsymbol{\phi}$, $\boldsymbol{\phi}^*$ and $\boldsymbol{u}$) and iid components ($\boldsymbol{v}$ and $\boldsymbol{\varepsilon}$) share the precision parameter κ, and we let st$[i]$ denote the state that contains LGA i and N_2 be the number of LGAs.

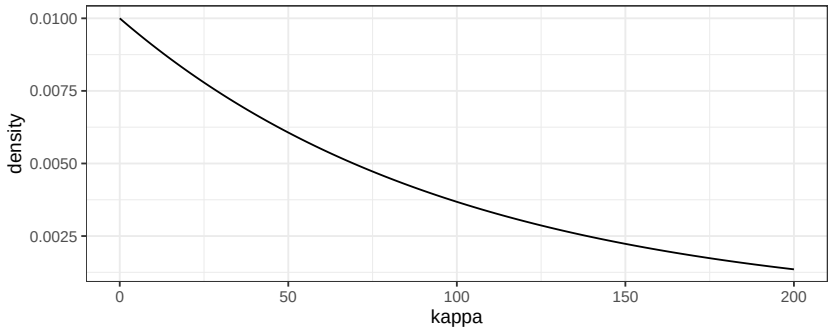
Model	Linear predictor	Regions
Admin2	$\eta_i = \mu + \sqrt{w_2}\,\phi_i^* + \sqrt{1-w_2}\,\varepsilon_i$	$1 \le i \le N_2$
Admin1	$\eta_i = \mu + \sqrt{w_1}\,u_{\text{st}[i]} + \sqrt{1-w_1}\,v_{\text{st}[i]}$	$1 \le i \le N_2$
Disconnected	$\eta_i = \mu + \sqrt{w_0 w_1}\,u_{\text{st}[i]} + \sqrt{w_0(1-w_1)}\,v_{\text{st}[i]} + \sqrt{(1-w_0)w_2}\,\phi_i + \sqrt{(1-w_0)(1-w_2)}\,\varepsilon_i$	$1 \le i \le N_2$
Connected	$\eta_i = \mu + \sqrt{w_0 w_1}\,u_{\text{st}[i]} + \sqrt{w_0(1-w_1)}\,v_{\text{st}[i]} + \sqrt{(1-w_0)w_2}\,\phi_i^* + \sqrt{(1-w_0)(1-w_2)}\,\varepsilon_i$	$1 \le i \le N_2$

For the models in Table 3.1, the intercept μ is equipped with an improper flat prior and the total precision parameter κ with a Gamma$(1, 0.01)$ prior. The prior expected value of κ is 100, resulting in a standard deviations of 0.1. With zero mean, this results in a 95% CI of approximately $[0.45, 0.55]$ on probability scale. The weights w_1 and w_2 have uniform priors on the unit interval. The weight between the variance on state level and LGA level, w_0, has a Beta$(2, 0.86)$ prior. The first shape parameter is set to two such that $\pi(w_0)$ goes linearly to zero as w_0 goes to zero, and the second shape parameter such that $\pi(w_0)$ has a peak at $w_0 = 1$, with a median at 0.75. This is done in order to explain as much variation as possible on state level. The priors for κ, w_0, w_1 and w_2 are displayed in Figure 3.4.

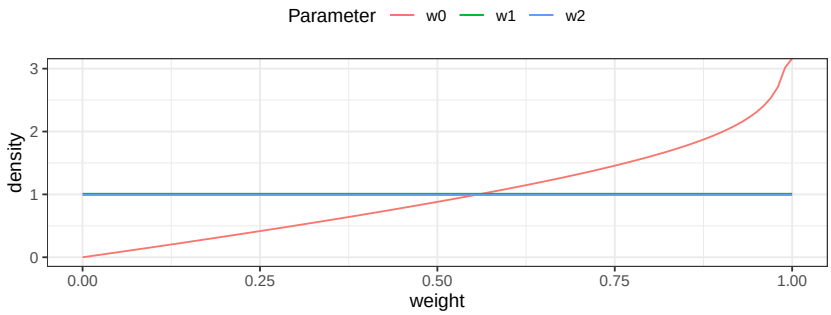
3.4 Aggregation and Linear Constraints

The Disconnected, Connected and Admin2 models in Table 3.1 do not immediately provide estimates of the proportion of vaccinated children in each state. State level estimates using these models are obtained as population weighted averages over all the LGAs within each state. Let q_i be the estimated under five population in LGA i described in Section 2. The state level vaccination coverages are estimated by

$$p_{\text{State } j} = \frac{\sum_{i \in \text{State } j} q_i\, p_i}{\sum_{i \in \text{State } j} q_i}, \quad \text{for } 1 \le j \le N_1,$$
$$\eta_{\text{State } j} = \text{logit}^{-1}\left(p_{\text{State } j}\right).$$



(a) Total precision κ



(b) BYM weights

Figure 3.4: The prior distribution for the parameters κ, w_0, w_1 and w_2.

A diagram giving an overview of the multiresolution models is shown in Figure 3.5. It shows the split of total variance between different geographical scales, and between the structured Besag component and unstructured iid component on each scale. The figure also shows the binomial observations for each LGA, and how LGA level vaccination coverages are aggregated to state level coverages using under five population estimates.

We now consider the constraints put on our Besag components $\boldsymbol{u}$, $\boldsymbol{\phi}$ and $\boldsymbol{\phi}^*$. One of the goals of the Disconnected and Connected models is to model on state level and LGA level separately. That is, we want

$$p_{\text{State } j} = \text{logit}^{-1}\left(\mu + \sqrt{w_0 w_1}\, u_j + \sqrt{w_0(1-w_1)}\, v_j\right).$$

Because of the non-linear logit link function, a sum-to-zero constraint will not achieve this. If we simplify the notation of the linear predictors in Table 3.1, by letting $u_j + v_j$ denote the BYM component for state j, and $\phi_i + \varepsilon_i$ denote the BYM component for LGA i, the desired constraint becomes

$$\text{logit}^{-1}(\mu) = \frac{\sum_{j=1}^{N_1} \text{logit}^{-1}(\mu + u_j + v_j) \times \sum_{i \in \text{State } j} q_i}{\sum_{i=1}^{N_2} q_i}, \tag{3.2}$$

$$\text{logit}^{-1}(\mu + u_j + v_j) = \frac{\sum_{i \in \text{State } j} \text{logit}^{-1}(\mu + u_j + v_j + \phi_i + \varepsilon_i) \times q_i}{\sum_{i \in \text{State } j} q_i}, \tag{3.3}$$

for $1 \leq i \leq N_1$.

This is the ideal case, which allows modeling the national, state and LGA vaccination coverages by separate parameters. However, such a constraint will make the posterior difficult to sample from for large graphs such as that of Nigeria. Instead, we opt to use the first order Taylor expansion with respect to the components u_j and ϕ_i of the Expressions (3.2) and (3.3), respectively, and we get

$$\frac{\sum_{j=1}^{N_1} u_i \times \sum_{i \in \text{State } j} q_i}{\sum_{i=1}^{N_2} q_i} \sim \mathcal{N}(0, \epsilon_1), \tag{3.4}$$

$$\frac{\sum_{i \in \text{State}_j} \phi_i \times q_i}{\sum_{i \in \text{State}_j} q_i} \sim \mathcal{N}(0, \epsilon_2), \text{ for } 1 \leq j \leq N_1. \tag{3.5}$$

The parameters ϵ_1 and ϵ_2 determine how strict these constraints are. A stricter constraint will typically come at the cost of a parameter space that is harder to sample from. As we will see later in Section 3.7, the curvature of the log density in parameter space plays a key role in sampling efficiency. As ϵ_1 and ϵ_2 become very small we are effectively placing the density in some small subspace,

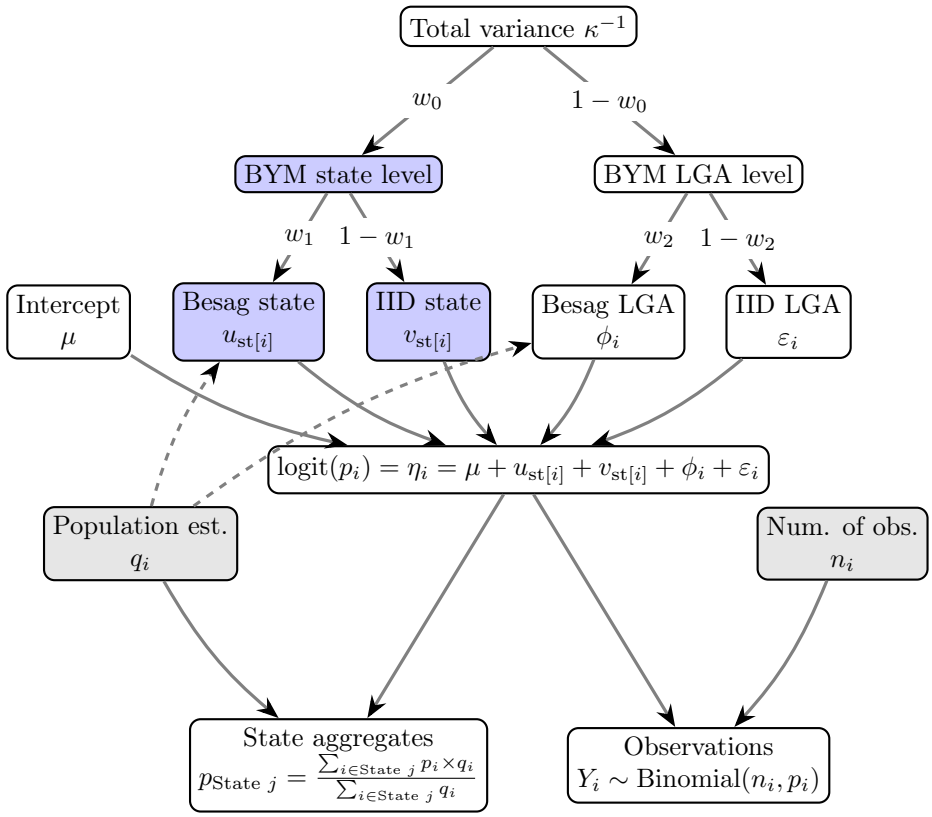


Figure 3.5: Diagram of the multiresolution models for an LGA i. The weight w_0 splits the total variance between different geographical scales, and the weights w_1 and w_2 between a Besag random effect and an unstructured random effect on each scale. The blue boxes indicate the BYM component on state level, which is preferred by the total variance split, while the gray boxes are known quantities. The dashed paths from the population estimates to the Besag components indicate that the population estimates are used in the linear constraints in the Besag model. The state LGA i lies in is denoted $\mathrm{st}[i]$.

making effective sampling difficult. Here $\epsilon_1 = 0.1$ and $\epsilon_2 = 0.01$ are chosen as small as possible while still achieving fast sampling.

The Admin1 model uses the constraint (3.4), and the Admin2 model uses the constraint

$$\frac{\sum_{i=1}^{n_2} \phi_i^* \times q_i}{\sum_{i=1}^{n_2} q_i} \sim \mathcal{N}(0, \epsilon_2).$$

Commonly, the linear predictors η_i will include various covariates, such as urban/rural classification, settlement type, land surface temperature, travel time to health facility, enhanced vegetation index and travel time to cities. See for instance Utazi et al. (2021) and Utazi et al. (2020). When including geospatial covariates in the linear predictors, there is the added complication of how the covariates should be estimated on LGA level. For instance, if the covariates are available in the DHS data, we have to estimate values for LGAs from cluster level data, and if other sources are used we have to account for the scrambled GPS position of the clusters. We choose not to include these extra predictors to avoid these problems.

In the case of urban/rural classification we also have the complication from the survey design. Urban or rural strata for the same state may be under- or oversampled relative to each other, resulting in different survey weights. If there is a small rural population within a state, then this stratum is likely oversampled, leading to significant bias if the urban and rural samples are treated the same. Finally, the effect of urban/rural status may change a lot between states and LGAs. In a country of over 200 million people, living conditions are likely to vary a lot. For this reason, the models are fitted to urban and rural data separately.

One of the reasons we are interested in multiresolution models is that they allow correlation between LGAs in a different way than the Admin2 model. Figure 3.6 shows the correlation between LGAs as a function of distance between their centroids, from 1000 realizations from the Besag model on states, Besag model on LGAs and the a sum of the Besag model on state level and LGA level. In each case the total precision is one. Correlations between LGAs are put into bins of 3.5 km according to the distances between LGAs, and the quantiles in each bin is calculated. The Besag model on states allows much greater correlation for LGAs that are less than 500 km apart than the same model on LGAs. Note that negative correlation for long distances are due to the linear constraints.

3.5 Design-Based Survey Estimates

Analysis of survey data is typically done using design-based methods, which take into account key aspects of the survey design (Lohr, 2010). Most importantly,

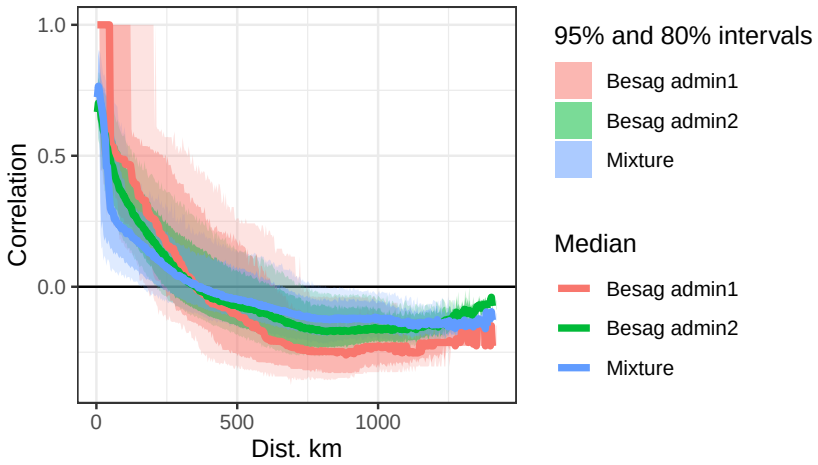


Figure 3.6: Correlation of η_i as a function of distance between centroids of LGAs.

design-based methods differ from model-based methods in assuming that the data is fixed from a finite population. For instance, the responses from a given household will be deterministic in the cases of age and MCV1 status. Hence, all uncertainty arises from the survey procedure itself; which households are surveyed?

Data that arises from complex survey designs differ from simple random samples in three key ways, which has to be accounted for when doing analysis. Firstly, the population is usually divided into subgroups known as strata, and the analyst creating the survey design decides how many samples should be drawn from each stratum. The sample size from each stratum is often about equal, even though their population can be very different. This ensures reliable estimates, even in areas with few people. The second complication is that within each stratum, a number of randomly chosen primary sampling units (PSUs) will be chosen and a fixed number of households will be sampled within each PSU. This will tend to underestimate the variance, but will increase precision per survey cost. Finally, the estimates should reflect that we sample without replacement from a finite population.

Designing a sample frame is usually done with census data, and the population is divided into small enumeration areas. A two-stage cluster sample, consists of sampling in two stages. First, within each stratum, we sample a number of EAs (the number is chosen by the analyst), which become the PSUs. Then, we select survey respondents within each PSU.

Because the units are sampled with unequal probability, each survey response

is given an inverse probability weight, corresponding to how many responses they represent. The weights in the 2018 NDHS are calculated in two stages. Let a_h denote the number of survey PSUs in stratum h, and M_{hi} the number of households in PSU i in stratum h as reported by the sampling frame. Then the probability of choosing PSU i in stratum h is

$$P_{1hi} = \frac{a_h M_{hi}}{\sum_i M_{hi}}.$$

The selection probability on the second stage is calculated as

$$P_{2hi} = \frac{g_{hi}}{L_{hi}},$$

where g_{hi} is the number of households surveyed in PSU i in stratum h, and L_{hi} is the number of households listed in the household listing conducted by the fieldworker at the time of surveying. The overall selection probability is then $P_{hi} = P_{1hi}P_{2hi}$, and the weights are $w_{hi} = 1/P_{hi}$ for the g_{hi} responses.

Let Y_{hij} denote observation j in cluster i from stratum h, with weights w_{hi}. Then the stratified sampling estimator is

$$Y_{h\cdot\cdot} = \frac{\sum_i \sum_j w_{hi} Y_{hij}}{\sum_i \sum_j w_{hi}}.$$

The variance of $Y_{h\cdot\cdot}$ can be estimated in multiple ways, for instance by linearisation or subsampling (Lumley, 2004).

To analyse the 2018 NDHS data, the `survey` package in `R` (Lumley, 2004) is used. We use a quasibinomial model, described in Mercer et al. (2014), producing the design-based direct estimates $\eta_{\mathrm{DE},h}$ and standard error $\sigma_{\mathrm{DE},h}$ on logit scale. The direct estimates $\eta_{\mathrm{DE},h}$ are assumed to be unbiased, and the uncertainties are assumed to be normally distributed on logit scale.

3.6 Scoring Rules

In order to evaluate the predictive distributions obtained in Bayesian inference, there are numerous choices of scoring rules. That is, a numerical score based on the predictive distribution and the observation. A common scoring rule is the mean square error (MSE) for all the predicted vaccination coverages. However, this score does not penalize the uncertainty of the predictive distribution. Ideally, we want a scoring rule that rewards the sharpness of predictive distributions (the concentration of the forecasts) subject to calibration (consistency between the observations and the forecasts) (Gneiting et al., 2007). That is, we want sharp predictive distributions, given that we do not have over- or underdispresion of

the data. For instance, if models are scored using MSE we only use a location measure such as the mean of the posterior distribution, which means that there is no penalty for unreasonably large or small uncertainty in the estimated value. Therefore, we need different methods to score our models.

Numerous such scoring rules exist. Here we will use generalization of the mean absolute error known as the continuous rank probability score (CRPS) (Matheson and Winkler, 1976), and the Dawid-Sebastiani score (DSS) (Dawid and Sebastiani, 1999), which uses only the first and second moment of the predictive distribution. These are two examples of proper scoring rules. Let Z be a random variable and F be a cumulative distribution function (CDF). Then, a proper scoring rule score $S(Z, F)$ has the property $\mathrm{E}\{\mathrm{S}(Z, F)\} \leq \mathrm{E}\{\mathrm{S}(Z, G)\}$ for any CDF G if $Z \sim F(z)$ (Gneiting and Raftery, 2007). This makes proper scoring rules a possible ranking criterion for different models, by scoring their predictive distributions.

In the univariate case, with an observations y and a CDF F, we define the CRPS and DSS as

$$\begin{aligned}\mathrm{CRPS}(y, F) &= \int_{\mathbb{R}} \left(F(z) - \mathbf{1}\{z \geq y\}\right)^2 \mathrm{d}z, \\ \mathrm{DSS}(y, F) &= \log \sigma_F^2 + \frac{(y - \mu_F)^2}{\sigma_F^2},\end{aligned}$$

where μ_F and σ_F^2 are the mean and variance, respectively, of a random variable following the distribution F.

For random vectors we define CRPS and DSS as the average values

$$\begin{aligned}\mathrm{CRPS}(\boldsymbol{y}, \boldsymbol{F}) &= \frac{1}{n} \sum_{i=1}^{n} \mathrm{CRPS}(y_i, F_i), \\ \mathrm{DSS}(\boldsymbol{y}, \boldsymbol{F}) &= \frac{1}{n} \sum_{i=1}^{n} \mathrm{DSS}(y_i, F_i),\end{aligned}$$

where $\boldsymbol{F}$ denotes the vector of marginal distributions.

In a Bayesian framework, we typically do not have closed forms of the posterior predictive distributions. Instead, estimates of the CRPS and DSS are found using the posterior samples with numerical methods (Krüger et al., 2020). These methods are implemented in the `R` package `scoringRules`.

When analysing real world data, the true vaccination coverages are unavailable for LGAs and states. Instead, we use leave-one-state-out cross-validation to assess the prediction accuracy of the models on state level, and K-fold cross-validation for cluster level predictions.

Evaluation of the models' predictive accuracy for state level vaccination coverages is done by comparing the predictive distributions to the design-based direct estimates. Let $\eta_{\text{DE},i}$ be the direct estimate of state i with standard error $\sigma_{\text{DE},i}$ for either urban or rural data. The direct estimates are assumed to be unbiased and normally distributed on logit scale. For each model we have a fitted leave-one-state-out predictive distribution for $\eta_{\text{State } i}$, and by adding a random variable $\epsilon_i \sim \mathcal{N}(0, \sigma_{\text{DE},i})$ to each of the posterior draws, we obtain a predictive distribution for the direct estimates $\eta_{\text{DE},i}$ themselves. The direct estimates and the corresponding predictive distributions are used to estimate MSE, CRPS and DSS for each model.

Additionally, we evaluate the out-of-sample predictive accuracy for clusters through K-fold cross-validation. For each cluster we calculate the MSE, CRPS and DSS for the ratios $\hat{p}_i = y_i/n_i$, i.e. the ratio of vaccinated children to total surveyed children in each cluster, with the predictive distribution of the vaccination coverage in the LGA in which the cluster is located.

3.7 Inference with Hamiltonian Monte Carlo

3.7.1 The Metropolis-Hastings Algorithm

In Bayesian inference we often encounter probability distributions known only up to a normalizing constant. For instance, applying Bayes theorem for some hierarchical model with observation likelihood $\pi_{\text{obs.}}(y|\theta)$, and prior density $\pi_{\text{prior}}(\theta)$ for the latent set of parameters, we get the posterior distribution of the latent parameters given the observations,

$$\pi(\theta|y) = \frac{\pi_{\text{obs.}}(y|\theta)\pi_{\text{prior}}(\theta)}{\pi(y)},$$

where $\pi(y)$ is the marginal distribution of the observations which acts as a normalizing constant. Unfortunately, $\pi(\theta|y)$ is often analytically intractable, and we only know the unnormalized distribution, denoted $\tilde{\pi}(\theta|y)$, often called the kernel of the density function.

Our primary goal is to be able to estimate the expectation of functions of $\theta \sim \pi(\theta)$. For Bayesian models, such as the ones presented in Section 3.3, Markov chain Monte Carlo (MCMC) methods are flexible and widespread methods of obtaining draws from the desired distribution $\pi(\theta)$ (Givens and Hoeting, 2012), that do not require the normalizing constant to be known. To estimate expectations we can then use Monte Carlo integration using the MCMC draws.

One of the most general MCMC methods is the Metropolis-Hastings (MH) algorithm. The MH algorithm proposes new candidate draws θ^* from a proposal

distribution $q(\theta^*|\theta)$, where θ is the previous value, before accepting the new candidate draw with probability

$$\alpha = \min\left\{1, \tilde{\pi}(\theta^*)q(\theta|\theta^*)/\tilde{\pi}(\theta)q(\theta^*|\theta)\right\}. \tag{3.6}$$

Otherwise, the new draw is set equal to the previous draw θ. The accept/reject step ensures detailed balance and reversibility with respect to the stationary target distribution $\pi(\theta)$.

Designing a fast converging MH algorithm can be very difficult, and usually involves finding a good parametrization, suitable proposal distributions and tuning of the proposal distribution's parameters, all specific to the modeling problem at hand. Typically, the proposal distribution for new draws takes the form of a random walk, for instance a multivariate Gaussian proposals centered at the current iteration. This raises the question of how the step length, i.e. the covariance matrix, should be chosen. If the step length is too short, the Markov chain will explore the distribution too slowly, and if the step length is too long, the acceptance probability of the proposal transitions might be very low. Finding a good proposal density is even harder if there are regions of the parameter space with high curvature, e.g. all the density is in one direction.

Figure 3.7 shows 200 draws from the bivariate normal distribution with zero mean, unit variances, and correlation 0.98 , using a MH normal random walk, with standard deviation 0.2. The chain starts at the origin, and because almost all the density is concentrated along the diagonal $x = y$, 42% of the new proposed states are rejected. We also see that the Markov chain fails to explore the tails of the distribution. An alternative alternative MCMC approach, illustrated in the right panel, is the No-U-Turns (NUTS) sampler, described in detail later in this section, which utilises the geometry of the posterior density when obtaining proposals.

3.7.2 Hamiltonian Monte Carlo

To overcome the problems of poor mixing and exploration of probability densities, Hamiltonian Monte Carlo (HMC) (Girolami and Calderhead, 2011) and the No-U-Turns (NUTS) sampler (Homan and Gelman, 2014) are increasingly favored. At a high level, HMC is a method of proposing new candidate samples informed by the gradient of the target density, and the NUTS sampler is an extension of HMC that greatly reduces the need for parameter tuning. To fit the models in our analysis we use `Stan` (Stan Development Team, 2020), which is an implementation of the NUTS sampler.

Let $\boldsymbol{\theta}$ be the random vector with density $\pi(\boldsymbol{\theta})$, and let $\mathcal{L}(\boldsymbol{\theta}) = \log \pi(\boldsymbol{\theta})$. In HMC, we first augment the parameter space with draws from a multivariate normal random momentum vector $\boldsymbol{r}$ with equal length as $\boldsymbol{\theta}$, with the component

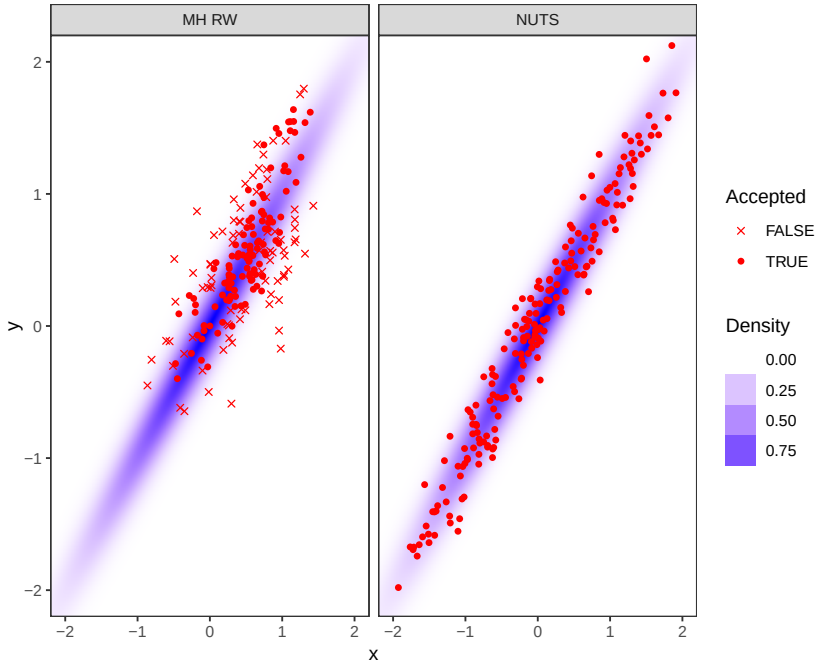


Figure 3.7: Two hundred samples of a MH random walk with multivariate normal proposal, compared to two hundred NUTS samples, superimposed on the true density. The proposals rejected during the MH accept/reject stage are marked with a cross.

r_i corresponding to the parameter θ_i. In the simplest case the components r_i are independent, with unit variance. The unnormalized joint density of $\boldsymbol{\theta}$ and $\boldsymbol{r}$ then becomes $\exp\left(\mathcal{L}(\boldsymbol{\theta}) - \boldsymbol{r}\cdot\boldsymbol{r}/2\right)$.

This augmented parameter space can be be interpreted physically as a Hamiltonian system, where some imagined particle with position $\boldsymbol{\theta}$ and momentum $\boldsymbol{r}$ moves around in the augmented parameter space. The Hamiltonian is defined as $H(\boldsymbol{\theta}, \boldsymbol{r}) = V(\boldsymbol{\theta}) + K(\boldsymbol{r})$, where $V(\boldsymbol{\theta}) = -\mathcal{L}(\boldsymbol{\theta})$ and $K(\boldsymbol{r}) = \boldsymbol{r}\cdot\boldsymbol{r}/2$ is potential and kinetic energy, respectively. We can then simulate the system's evolution over time following Hamiltonian dynamics,

$$\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t} = \frac{\partial H}{\partial \boldsymbol{r}}; \quad \frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}t} = -\frac{\partial H}{\partial \boldsymbol{\theta}},$$

which preserves the total energy. Since energy is preserved, the joint density is constant along trajectories. By proposing new draws from the joint density of $\boldsymbol{\theta}$ and $\boldsymbol{r}$ along the trajectories the Metropolis acceptance probability is one. We then discard the momentum parameter, and obtain new samples from the target distribution $\pi(\boldsymbol{\theta})$.

In practice, the Hamiltonian systems arising from interesting models do not have analytical solutions, so we have to use numerical integration to compute paths in parameter space. To ensure that detailed balance is preserved, the leapfrog estimator is used, which proceeds with the updates

$$\begin{aligned}
\boldsymbol{r}_{t+\epsilon/2} &= \boldsymbol{r}_t + \frac{\epsilon}{2}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}_t), \\
\boldsymbol{\theta}_{t+\epsilon} &= \boldsymbol{\theta}_t + \epsilon\, \boldsymbol{r}_{t+\epsilon/2}, \\
\boldsymbol{r}_{t+\epsilon} &= \boldsymbol{r}_{t+\epsilon/2} + \frac{\epsilon}{2}\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}_{t+\epsilon}),
\end{aligned}$$

for some small step size ϵ, for a total of L steps.

Because we sample $\boldsymbol{r} \sim \mathcal{N}(0, \mathbf{I})$, then transform this proposal to a proposal for the joint position-momentum state, in general we have to consider the Jacobian of the transformation when computing the MH ratio. This is necessary to keep detailed balance intact. However, the leapfrog scheme is volume preserving, so the absolute value of its Jacobian is one. The leapfrog scheme is also time reversible, since you can do the steps in reverse. Together, these properties ensure that detailed balance is kept intact.

A naive HMC approach would be to sample $\boldsymbol{r}_t \sim \mathcal{N}(0, \mathbf{I})$, then do L leapfrog steps, starting at $(\boldsymbol{\theta}_t, \boldsymbol{r}_t)$, giving us the proposal $(\boldsymbol{\theta}^*, \boldsymbol{r}^*) = (\boldsymbol{\theta}_{t+L\epsilon}, -\boldsymbol{r}_{t+L\epsilon})$. Due to discretization error we accept this proposal with probability

$$\alpha = \min\left\{1, \exp\left(H(\boldsymbol{\theta}_t, \boldsymbol{r}_t) - H(\boldsymbol{\theta}^*, \boldsymbol{r}^*)\right)\right\}. \tag{3.7}$$

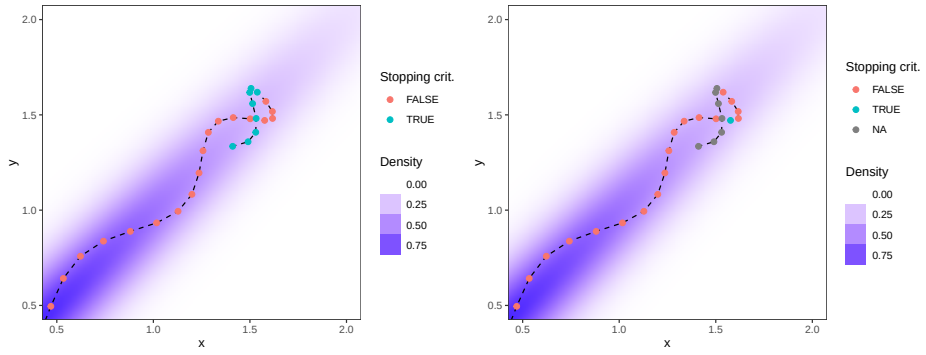
3.7.3 The NUTS Sampler

This naive HMC approach raises the question of how to tune L and ϵ. An intuitive method of determining the length of the trajectory, is to add iterations until the trajectory doubles back on itself. That is, for the earliest and latest states $(\boldsymbol{\theta}^-, \boldsymbol{r}^-)$ and $(\boldsymbol{\theta}^+, \boldsymbol{r}^+)$ in the trajectory, terminate whenever

$$\boldsymbol{r}^+ \cdot (\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-) < 0, \tag{3.8}$$

which effectively means that the trajectory will not do a U-turn.

Unfortunately, when we do the leapfrog steps in reverse, there is no guarantee that the trajectory reaches the original state $(\boldsymbol{\theta}^-, \boldsymbol{r}^-)$ when starting from $(\boldsymbol{\theta}^+, \boldsymbol{r}^+)$. The sampler might satisfy its stopping condition before the original position momentum state is reached. One such example is displayed in Figure 3.8, which shows a leapfrog trajectory for a bivariate zero mean normal distribution, with unit variance and correlation 0.95. Here, the states along the trajectory satisfying (3.8) are marked as blue. However, integrating backwards in time from the first blue state satisfying the stopping condition, results in a much shorter trajectory, as shown in the rightmost panel. The stopping condition is met already after four steps. This means that for a naive trajectory the proposal densities are not symmetric, $q\left((\boldsymbol{\theta}^+, \boldsymbol{r}^+)|(\boldsymbol{\theta}^-, \boldsymbol{r}^-)\right) \neq q\left((\boldsymbol{\theta}^-, \boldsymbol{r}^-)|(\boldsymbol{\theta}^+, \boldsymbol{r}^+)\right)$, and the MH acceptance probability in Equation (3.6) is not determined by the Hamiltonians in the final and first states.



(a) Stopping criterion **(b)** Stopping criterion in reverse

Figure 3.8: Non-reversible leapfrog trajectory using a naive u-turn stopping condition.

Instead, we build a set of candidate states $\mathcal{C}$ starting from $(\boldsymbol{\theta}_t, \boldsymbol{r}_t)$. By repeatedly expanding $\mathcal{C}$ by doubling its size using leapfrog steps forward or backward, we create a binary tree with leaves corresponding to position-momentum states.

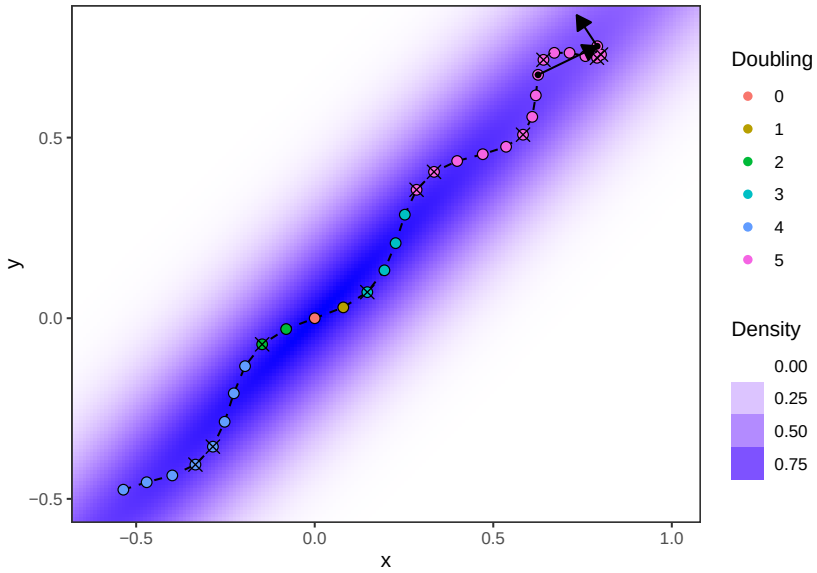


Figure 3.9: Example of a trajectory from NUTS with slice and stopping criterion, superimposed on the true density. The arrows mark the momentum and difference in position for a balanced sub-tree that has reached the stopping condition (3.9). The crossed out states are outside the slice, and the color of the states marks in which doubling they were added. All the states added in the fifth doubling must be omitted from the candidate set to preserve detailed balance, and the crossed out states because they are not in the slice.

An example of such a set $\mathcal{C}$ is displayed in Figure 3.9. The colours of each position-momentum state show in which doubling the states are added. The figure also shows a generalized form of the stopping condition (3.8) with arrows, and states that have to be excluded to preserve detailed balance crossed out, both explained further in this section.

Starting from a position $\boldsymbol{\theta}_t$, the sampling procedure of NUTS consists of four steps.

1. Generate a random momentum vector $\boldsymbol{r}_t \sim \mathcal{N}(0, \mathbf{I})$.
2. Generate a slice variable $u \sim \text{Uniform}[0, \exp\left(-H(\boldsymbol{\theta}_t, \boldsymbol{r}_t)\right)]$.
3. Generate a set of states $\mathcal{B}$ starting at $(\boldsymbol{\theta}_t, \boldsymbol{r}_t)$, and choose a subset $\mathcal{C} \subset \mathcal{B}$ of candidate states deterministically, in such a way that any position-momentum state in $\mathcal{C}$ has the same probability of recreating $\mathcal{C}$ if used as

the the starting state.

4. Choose a proposal position-momentum state from $\mathcal{C}$, and accept it with probability α given by Equation (3.7).

The central idea of NUTS is how the sets $\mathcal{B}$ and $\mathcal{C}$ are generated. We first generate $\mathcal{B}$, starting with $(\boldsymbol{\theta}_t, \boldsymbol{r}_t)$, by repeatedly doubling the size of the set using leapfrog steps either forward or backward in time. This means that $\mathcal{B}$ is really a trajectory in position-momentum space, forming a binary tree illustrated in Figure 3.10. For the jth doubling, we either make 2^j leapfrog steps backwards from the earliest state in the trajectory, or 2^j leapfrog steps forwards in time from the latest state in the trajectory, each with probability $1/2$. As a stopping condition, we check whether

$$\boldsymbol{r}^+ \cdot (\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-) < 0, \text{ or } \boldsymbol{r}^- \cdot (\boldsymbol{\theta}^+ - \boldsymbol{\theta}^-) < 0, \tag{3.9}$$

holds true for the earliest and latest states $(\boldsymbol{\theta}^-, \boldsymbol{r}^-)$ and $(\boldsymbol{\theta}^+, \boldsymbol{r}^+)$ of any balanced subtree, or if $\log(u) - H(\boldsymbol{\theta}', \boldsymbol{r}') > \Delta_{\max}$, for any $(\boldsymbol{\theta}', \boldsymbol{r}')$ in $\mathcal{B}$ for some large $\Delta_{\max}$. When either of the stopping conditions are met, the set of candidate states $\mathcal{C}$ are chosen from the entire leapfrog trajectory $\mathcal{B}$, such that $(\boldsymbol{\theta}', \boldsymbol{r}') \in \mathcal{B}$ satisfies $u \leq \exp\left(-H(\boldsymbol{\theta}', \boldsymbol{r}')\right)$ and omitting the states from the last doubling.

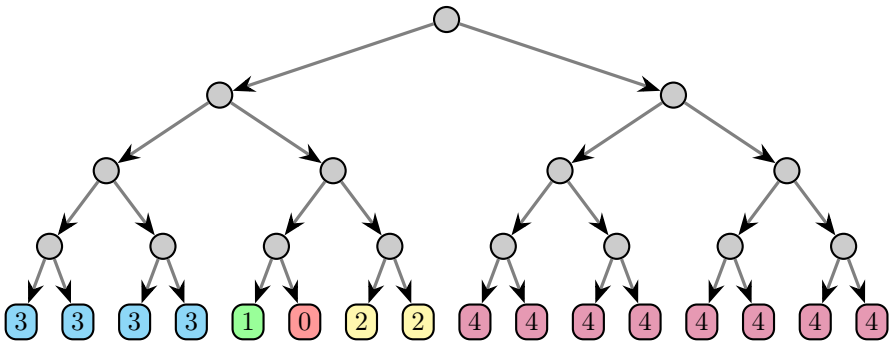


Figure 3.10: Example of a trajectory $\mathcal{B}$ as a binary tree. The labels and colours show in which doubling the leaf nodes are added. Balanced subtrees consists of all leaf nodes that share a common ancestor (gray nodes). For instance, all the leaf nodes added in the third doubling, or the four leftmost leaf nodes added in the fourth doubling. Note that to recreate a specific binary tree from any leaf node, there is only one choice of adding nodes either to the left or to the right in each doubling.

By omitting the the states from the last doubling we make sure that for any state in $\mathcal{C}$ the probability of recreating the same trajectory $\mathcal{B}$ is 2^{-j}, since at each doubling we have to expand $\mathcal{B}$ in the correct direction. The condition that

$u \leq \exp\left(-H(\boldsymbol{\theta}', \boldsymbol{r}')\right)$ is a sampling method called slice sampling (Neal, 2003), which ensures that conditional on u all the states in $\mathcal{C}$ have equal probability density.

Returning to the example of a trajectory $\mathcal{B}$ in Figure 3.9, in the fifth doubling, there is a balanced subtree that satisfies the u-turn criterion (3.9) (the angle between the arrows is less than 90°), so the pink states are omitted from $\mathcal{C}$. Additionally, the states that are omitted from $\mathcal{C}$ due to the slice u are crossed out.

We can now choose proposals states from $\mathcal{C}$ with any kernel that leaves the uniform distribution invariant, while still preserving detailed balance. The transition kernel used in NUTS favors the candidate states most recently added to the trajectory $\mathcal{B}$, which makes the average number of leapfrog steps between the original state and the proposed state larger on average.

Stan is largely based on the original NUTS sampler, with some changes to improve performance (Betancourt, 2016). Firstly, slice sampling is no longer utilized when constructing $\mathcal{C}$. Instead, proposals are drawn from $\mathcal{C}$ following a multinomial distribution with probabilities

$$P(\boldsymbol{\theta}', \boldsymbol{r}') = \frac{\exp(-H(\boldsymbol{\theta}', \boldsymbol{r}'))}{\sum_{(\boldsymbol{\theta}'', \boldsymbol{r}'') \in \mathcal{C}} \exp(-H(\boldsymbol{\theta}'', \boldsymbol{r}''))},$$

for $(\boldsymbol{\theta}', \boldsymbol{r}') \in \mathcal{C}$. Additionally, a generalized version of the No-U-Turns criterion (3.9) is used.

3.7.4 Adaptively Tuning the Step Length

We now have a way of choosing the number of steps L during sampling, and we turn to choosing a suitable step length ϵ. During the warmup phase of Stan's sampling procedure, the step length is adaptively tuned to achieve a target MH acceptance probability. For a tunable parameter x of any MCMC algorithm, and some statistic $H_t(x)$ of the Markov chain, there are multiple schemes to dynamically tune x so that the average expectation $T^{-1} \sum_{t=1}^{T} \mathrm{E}(H_t|x)$ goes to zero. Our goal is to tune ϵ such that a target MH acceptance rate is achieved. NUTS uses the dual averaging algorithm of Nesterov (Homan and Gelman, 2014),

$$x_{t+1} \leftarrow \mu - \frac{\sqrt{t}}{\gamma} \frac{1}{t + t_0} \sum_{i=1}^{t} H_i; \quad \bar{x}_{t+1} \leftarrow \eta_t \, x_{t+1} + (1 - \eta_t)\bar{x}_t, \tag{3.10}$$

with $\bar{x}_1 = x_1$ and $\eta_t = t^{-\kappa}$, and free variables $\mu, t_0 \geq 0$.

Let $H_t = \delta - H_t^{\mathrm{NUTS}}$ for some target acceptance probability δ, and

$$H_t^{\text{NUTS}} = \frac{1}{|\mathcal{B}_t|} \sum_{(\boldsymbol{\theta},\boldsymbol{r})\in\mathcal{B}_t} \min\left\{1, \exp\left(H(\boldsymbol{\theta}_t, \boldsymbol{r}_t) - H(\boldsymbol{\theta}, \boldsymbol{r})\right)\right\},$$

where $\mathcal{B}_t$ is the trajectory from state $(\boldsymbol{\theta}_t, \boldsymbol{r}_t)$. Then H_t^{NUTS} is a function of the step size ϵ, and we use Equation (3.10), setting $x = \log \epsilon$, to update the step size after each iteration in the warmup phase. This ensures that ϵ will converge to a value such that the average acceptance rate is approximately δ.

The NUTS sampler therefore both chooses sensible number of integration steps L and leapfrog step length ϵ, given some target acceptance rate δ specified by the user. In practice a high target acceptance rate will force the sampler to use shorter step lengths, which might help in regions of parameter space with high curvature, but increase computational cost.

3.7.5 Stan

The Stan probabilistic programming language is efficient, flexible, and requires minimal tuning. However, model parametrization can greatly affect the sampling efficiency. For the models introduced in Section 3.3, three weights, $0 \leq w_0, w_1, w_2 \leq 1$, are used to control the proportion of total variance placed on different components of the linear predictor. If the prior of these parameters are declared directly, there will be boundaries in parameter space where the log density suddenly drops to minus infinity. If the leapfrog trajectory crosses one of these boundaries, Stan reports a divergent transition, and if it happens during warmup the integration step length is forced to be very small. To avoid this problem, the parameters are implemented on logit scale, before being transformed and used as usual.

The default target acceptance ratio is 0.8, but this can increased if the step length is too long and the stopping condition $\log(u) - H(\boldsymbol{\theta}', \boldsymbol{r}') > \Delta_{\max}$ is met, resulting in a divergent transition. This occurs if the curvature varies greatly between different parts of the parameter space, and often require a reparametrization of the model. We also specify the number of chains and iterations during sampling and warmup.

4 | Simulation Study with Multiresolution Data

In this section we conduct simulation experiments in order to compare the predictive performance of the models presented in Section 3.3 in the presence of multiresolution variation.

4.1 Purpose

There are three main questions we want to answer.

(i) Do the multiresolution models offer improved predictive performance of the state level vaccination coverages over the Admin2 model if there is a state random effect?

(ii) If the true state vaccination coverages are not known exactly, only with added noise, are we able to detect potential differences between the models' predictive accuracies?

(iii) How uncertain are the parameter estimates, and are they interpretable?

In Section 3.6 we present a method of scoring models by viewing the direct estimates as noisy observations of the true values. The second point examines how noisy these observations can be, before potential differences in predictive accuracy between the multiresolution approaches and the Admin2 model disappears.

4.2 Simulation Setup

All code is run using `R` version 4.03 (R Core Team, 2020), and the models are fitted using `rstan` version 2.21.2 (Stan Development Team, 2020). For all simulations

we use the graph structure the LGAs and states of Nigeria, as well as the total number of survey respondents in the 2018 NDHS data for each LGA. The under five population counts in Nigeria are used for aggregation from LGA to state level and Besag linear constraints. To generate data that is representative of a wide range of plausible real world scenarios, we choose suitable values of μ, κ, w_0, w_1 and w_2, found in Table 4.1, and generate data with the Connected model. State vaccination coverages are calculated using a population weighted average over the LGAs. Most importantly, we consider cases with different variation in vaccination coverage over the country as a whole (total precision κ low and high), and with different proportions of the state and LGA variances (weight w_0 low, medium and high). Four typical realizations of the vaccination coverages on logit scale is shown in Figure 4.1.

Table 4.1: Combinations of parameters used to simulate data.

Parameter	Levels	Value	Interpretation
μ	Low / high	$\{-0.5, 0.5\}$	Vaccination level
κ	Low / high	$\{3, 0.4\}$	Variability
w_0	Low / medium / high	$\{0.2, 0.5, 0.8\}$	adm1 vs. adm2
w_1	Low / medium / high	$\{0.2, 0.5, 0.8\}$	Struct. adm1
w_2	Low / medium / high	$\{0.2, 0.5, 0.8\}$	Struct. adm2

To select parameter values we first consider how small the total precision parameter κ can be. Because Nigeria has large geographical inequality in vaccination coverage (Local Burden of Disease Vaccine Coverage Collaborators, 2021), we want LGA coverages in simulated realizations to lie in the range 5% to 95%, corresponding to roughly -3 to 3 on logit scale. Since κ is approximately the marginal precision in each node, the smallest suitable κ is chosen as 0.4 so that two standard deviations is 3.16. We also want to see if the multiresolution models perform better than the Admin2 model if the spatial variation is much smaller, so the high value for κ is 3, resulting in two standard deviations of roughly 1.

Using the design-based methods with the 2018 NDHS data, the vaccination coverage of Nigeria as a whole is approximately 54%, so the intercept μ is set to ± 0.5. The weights w_1 and w_2 both have the same levels, 0.2, 0.5 and 0.8, which corresponds to mostly noise driven vaccination coverage, noise and spatial smoothing, and mostly spatial smoothing. Most importantly, the split between variation between states and variation within states, w_0, has the levels 0.2, 0.5 and 0.8. For $w_0 = 0$ there is no state effect, and for $w_0 = 1$ all spatial variation happens on state level, while the vaccination coverage is constant within each state. What we are most interested in is what happens to the predictive score of the Admin2 model compared to the Connected and Disconnected models for

$w_0 \in \{0.5, 0.8\}$.

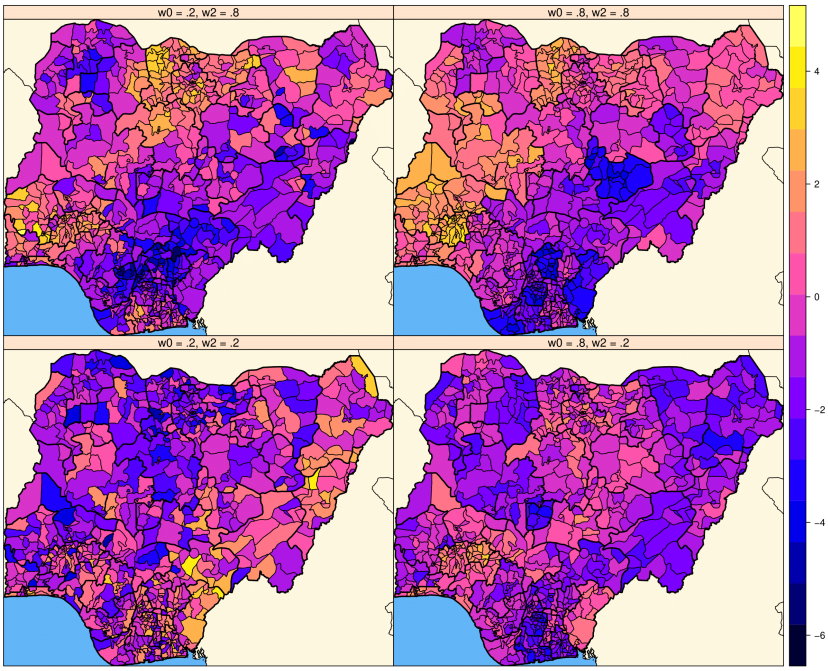


Figure 4.1: Four realizations of LGA probabilities, on logit scale, from the Connected multiresolution model. For all realizations $\mu = -0.5$, $\kappa = 0.4$ and $w_1 = 0.8$.

The fitted models are scored using CRPS and DSS on the known state and LGA vaccination coverages on logit scale. In order to reduce the computational requirement needed to compare the three models, we first generate five realizations for each of the 108 combinations of parameters found in Table 4.1. We then determine which parameters have an effect on how well the models perform, as measured by MSE, CRPS and DSS, before holding these parameters fixed and generating 50 realizations for each remaining parameter combination. For instance, if the value of the intercept makes no difference for the relative model performance, μ is held fixed for the 50 subsequent realizations. Using 50 realizations ensures reasonably accurate estimates of the CRPS and DSS scores of the different models.

All models are fitted using `Stan` with four chains of 6000 iterations (2000 of which are warmup iterations), resulting in effective sample sizes of at least 2000

for each parameter. The maximum tree depth is set to 15, otherwise the default values of the `Stan` parameters are used.

4.3 Importance of Allowing Multiresolution Variation

The initial five realizations reveal that the values of the parameters μ and w_1 only have a minor effect on the CRPS for the four models, compared to the parameters κ, w_0 and w_2. Therefore, for the subsequent realizations $\mu = 0$ and $w_1 = 0.8$ are held fixed, and 50 data sets are generated using each of the 18 remaining combinations of κ, w_0 and w_2 in Table 4.1.

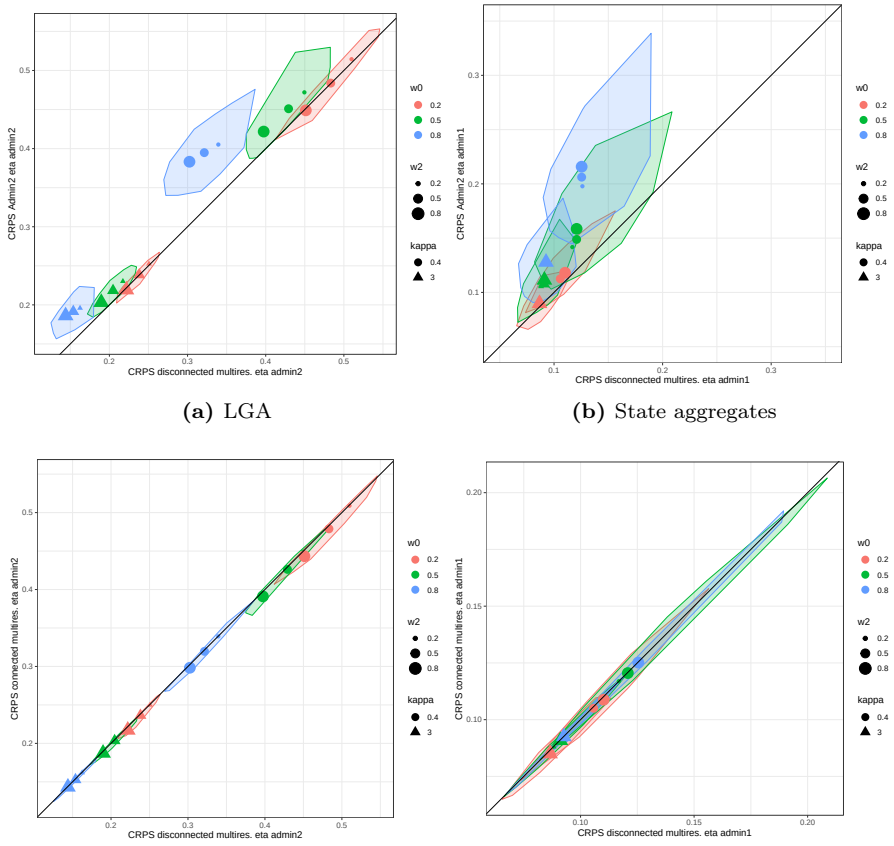
To compare the Admin2 and Connected models with the Disconnected model, we draw a crossplot of the CRPS estimates, both for LGA and state vaccination coverage predictions, displayed in Figure 4.2. In each comparison the average CRPS of the 50 realizations with the same simulation parameters are shown as points. To show the variation, we also draw a convex hull, containing all the crossplot points for each of the 150 combinations of w_0 and κ. That is, if the polygon is above the diagonal, the Disconnected model scores better for each of the simulated 150 data sets.

The simulation results demonstrate that when a significant portion of the total variance happen on state level, i.e. $w_0 \in \{0.5, 0.8\}$, the Admin2 model performs worse than both the multiresolution models. Since the data is generated using the Connected model, we expect the Connected model to perform best on average, as shown in the bottom two panes of Figure 4.2. However, the penalty for using the Disconnected model is minor.

4.4 Model Comparison Using Noisy Observations

As described in Section 3.6, we wish to score models on real world data by predicting design-based direct estimates. Can we detect potential differences in predictive performance when only noisy observations are known? We now use the data sets with $w_2 = 0.5$, $\kappa = 0.4$ and $w_0 \in \{0.5, 0.8\}$ and add normally distributed noise with standard deviation in $\{0, 0.2, \ldots, 1\}$ to both the true known state vaccination coverages, and the predictive distributions of the models (both on logit scale). This is done 50 times for each realization. Finally, the noisy predictive distributions with the noisy true values are scored using CRPS.

The results of this check for sensitivity is displayed in Figure 4.3. The Disconnected model scores better than the Admin2 model for realizations with $\kappa = 0.4$ and $w_0 \in \{0.5, 0.8\}$, if there is no noise. However, we see that as noise is added to



(a) LGA **(b)** State aggregates

(c) LGA **(d)** State aggregates

Figure 4.2: Crossplot of average CRPS in simstudy using Disconnected multiresolution, Admin2 and connected multiresolution model. Using fixed $\mu = 0$ and $w_1 = 0.8$, and 50 samples for each of the 18 combinations of the other parameters found in Table 4.1. For each combination of κ and w_0, the convex hull of the CRPS from all the realizations is shown (150 in total), as well as the mean CRPS for each combination of κ, w_0 and w_2.

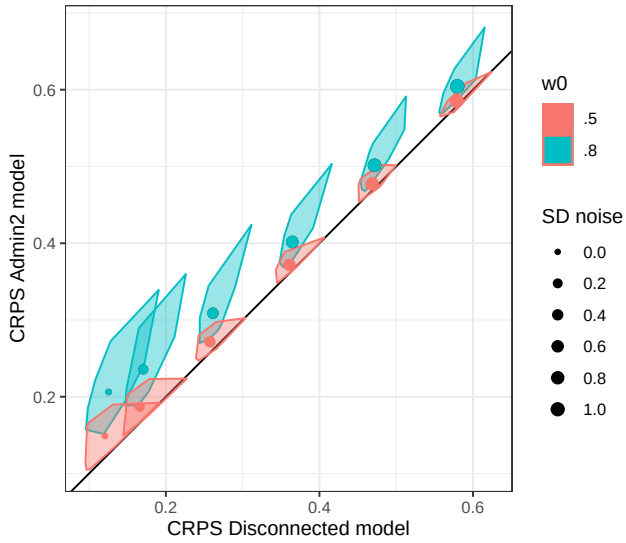


Figure 4.3: Crossplot of the CRPS score comparing the Disconnected and Admin2 models, when the true values are scrambled. The means are shown as points, while the polygons are the convex hulls, each containing 50 points.

the true values, corresponding with greater uncertainty of the direct estimates, the overall CRPS increases, and the difference between the models' scores decreases. In practice, this means that scoring models using the direct estimates, requires the models' differences in predictive accuracy to be high enough compared to direct estimates' uncertainties.

4.5 Parameter Estimation

We now want to interpret how the Disconnected, Connected and Admin2 models uses the model parameters κ, w_0, w_1 and w_2, when fitted to data from the Connected model. Importantly, we want to know if fitting the Connected model recovers the true parameter values, if the Disconnected and Connected models use the weights in the same way, and how the Admin2 model estimates total precision parameter κ.

For each simulated data set, with a given set of parameters, the posterior means and 2.5% and 97.5% quantiles of κ, w_0, w_1 and w_2 are calculated for the relevant fitted models. We then take the average of these statistics for each of the 18 sets of simulation parameters and compare the average to the true known

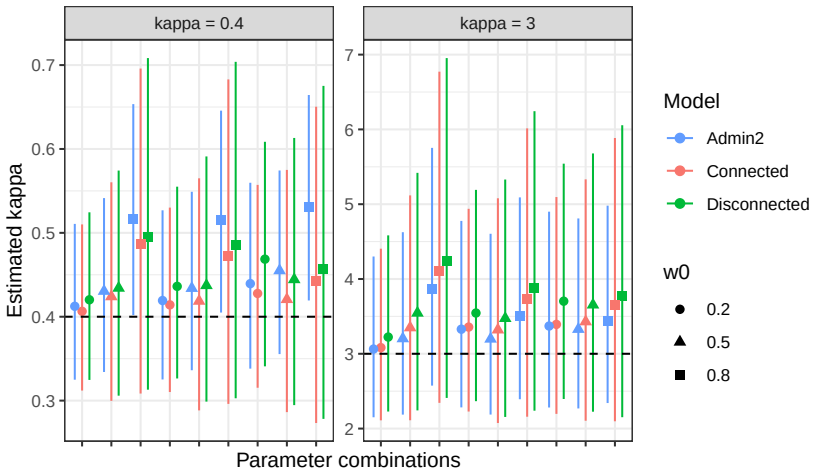


Figure 4.4: The average of the means and 95% quantiles of the total precision parameter κ, for each parameter combination of κ, w_0 and w_2 found in Table 4.1, with $\mu = 0$ and $w_1 = 0.8$ held fixed.

value. The average of the means are marked by points, and the segments go from the average 2.5% quantiles to the average 97.5% quantiles. The parameter combinations are grouped by the true parameter value, and then by model.

Figure 4.4 shows that the posterior means of the total precision from all three models is very similar, and usually larger than the true value. This means that the models tend to underestimate the uncertainty on logit scale. Additionally, the CIs from the Admin2 model are smaller than for the Disconnected and Connected models.

Even though the predictive scores of the Disconnected and Connected models were very similar, the posterior distributions of w_0 displayed in Figure 4.5 show that the models use their components differently. By disconnecting the graph of LGAs between states, the Disconnected model is forced to put more of the spatial variation on state level. When the true w_2 goes to one, meaning that there is only ICAR on LGA level in the simulated data, the Disconnected model places much more of the variation on state level. For all parameter combinations, the Connected model recovers the true value fairly well, however as w_2 grows the Connected model also places more variation on state level, indicating that the model is unable to distinguish between state and LGA level variation.

Figure 4.6 shows that the credible intervals for the weight w_1, which splits variance between Besag and iid. on state level, is typically $[0.2, 0.95]$ for the Connected and Disconnected model. Such wide credible intervals indicate that the

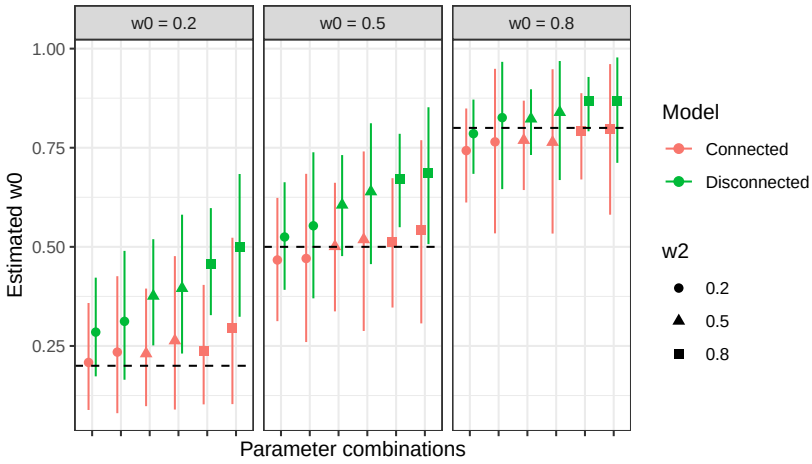


Figure 4.5: The average of the means and 95% quantiles of the weight parameter w_0, for each parameter combination of κ, w_0 and w_2 found in Table 4.1, with $\mu = 0$ and $w_1 = 0.8$ held fixed.

Disconnected and Connected models are unable to properly distinguish between the iid. and ICAR component on state level, likely because the graph of states in Nigeria only has 37 nodes.

In Figure 4.7 we compare the posterior values of w_2 for the different models. The Disconnected model places much less weight on the ICAR component on LGA level than the Connected model, but we also see that if there is little variation on LGA level (w_0 is large), the CIs are become very large. The Admin2 model is forced to use the LGA level ICAR component for spatial variation on state level, meaning that it estimates a w_2 close to one as the true w_0 goes to one. It is clear that there is no shared interpretation of the weight parameters for all the models, rather the models use the components in different ways.

We calculate the empirical coverages for the 95% intervals with the fitted Connected models. For the 50 simulated data set with each parameter of the 18 parameter combinations of κ, w_0 and w_2, we check whether the true simulation parameter lies in the 95% posterior CI. The proportion of the data sets for which this is true is reported in Table 4.2. The table reports the empirical coverages for each value of each parameter. For κ each empirical coverage is based on 450 data sets, 300 for w_0, 900 for w_1 and 300 for w_2. The empirical coverages for the weights are all slightly larger than the nominal coverage, and for the total precision slightly smaller.

If we assume that there is a 95% probability of a parameter being in the 95%

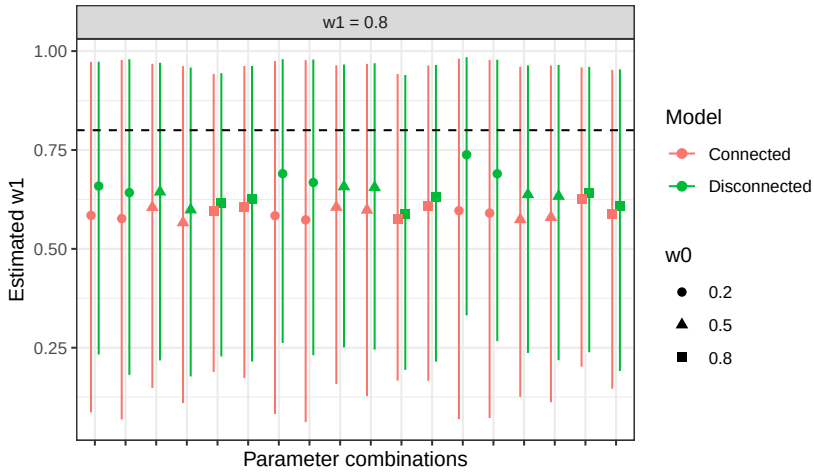


Figure 4.6: The average of the means and 95% quantiles of the state weight parameter w_1, for each parameter combination of κ, w_0 and w_2 found in Table 4.1, with $\mu = 0$ and $w_1 = 0.8$ held fixed.

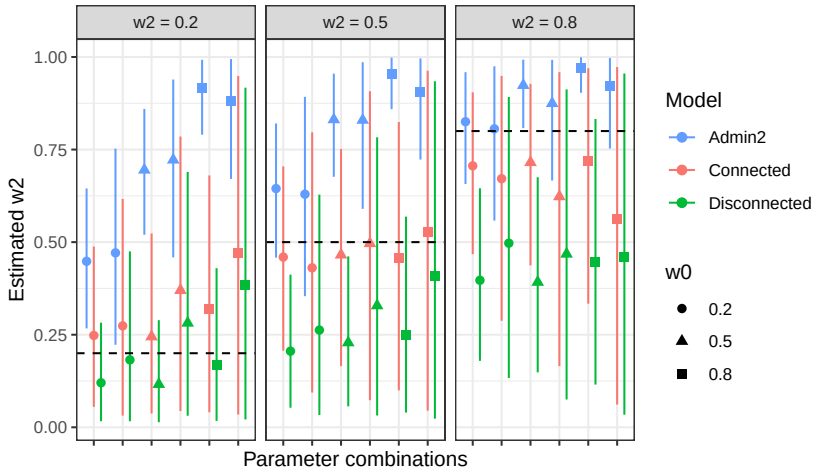


Figure 4.7: The average of the means and 95% quantiles of the LGA weight parameter w_2, for each parameter combination of κ, w_0 and w_2 found in Table 4.1, with $\mu = 0$ and $w_1 = 0.8$ held fixed.

Table 4.2: Empirical 95% coverages for the fitted Connected models.

Parameter	Simulation value	Empirical 95% coverage
κ	0.4	0.94
	3	0.93
w_0	0.2	0.96
	0.5	0.97
	0.8	0.96
w_1	0.8	0.99
w_2	0.2	0.97
	0.5	0.98
	0.8	0.96

CIs from the fitted models, the 95% empirical coverages is the mean of Bernoulli trials, with a standard deviation of approximately 0.007. Values of 0.93 to 0.97 is therefore reasonable, but 0.98 and 0.99 indicate that the model does not gain much information about the parameters from the data.

5 | Case Study: Vaccination Coverage in Nigeria

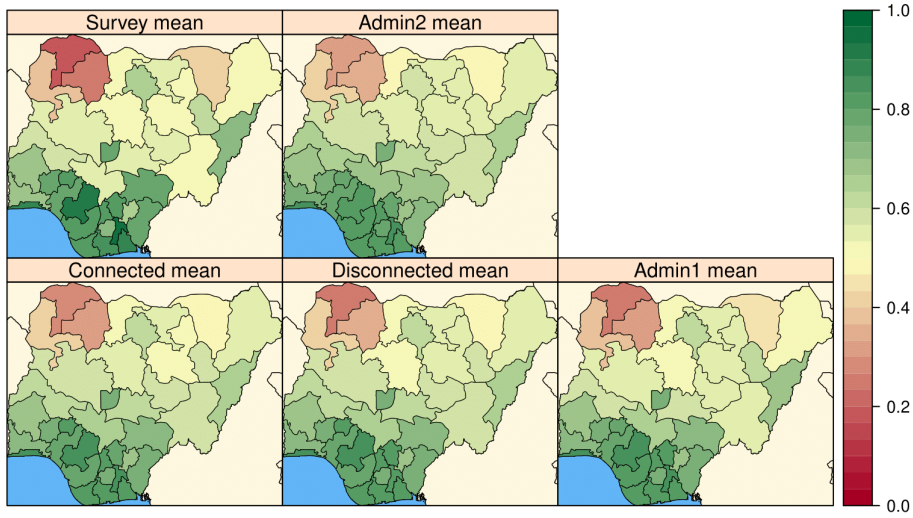
5.1 Estimated Vaccination Coverage Maps

We apply the the four models described in Section 3.3 to the 2018 DHS data of Nigeria, to produce estimates of LGA and state level vaccination coverage. The model-based estimates are compared to design-based direct estimates. The models are fitted to data from rural and urban strata separately to avoid complications in how to jointly model the observations. All the models are fitted using `Stan`, with 4 chains of 6000 iterations (of which 2000 are warmup iterations), and with default values for `Stan` parameters, leading to a minimum effective sample size of 1500 for all model parameters.

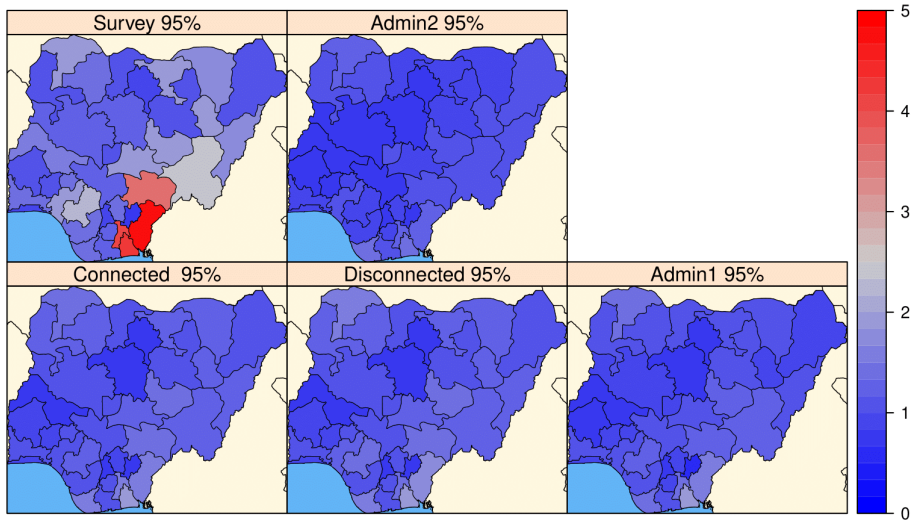
The posterior state means for each fitted model are displayed as maps in Figure 5.1 and 5.2 for urban and rural strata, respectively. The figures also show the size of the centered 95% credible interval for the state vaccination coverages on logit scale for each model. We also display the direct estimates, and the size of the associated 95% confidence intervals. All models result in largely the same spatial pattern for the vaccination coverage, with low coverage in the north-west, and high coverage in the south along the coast. Notably, the vaccination coverage is much lower in rural than urban areas.

Overall the model-based approaches result in smaller credible intervals than the design-based method. For the estimates of the urban areas, states in the south-east have very high uncertainty in the survey estimates. The model-based approaches results in credible intervals for these states that are roughly the same size as the other states. Among the four models, the Admin2 model offers the greatest amount of spatial smoothing, and on average the smallest credible intervals, while the three other models have very similar uncertainties.

The clearest example of how the different models borrow strength across state borders can be seen in the rural estimates for Lagos along the coast, in Figure

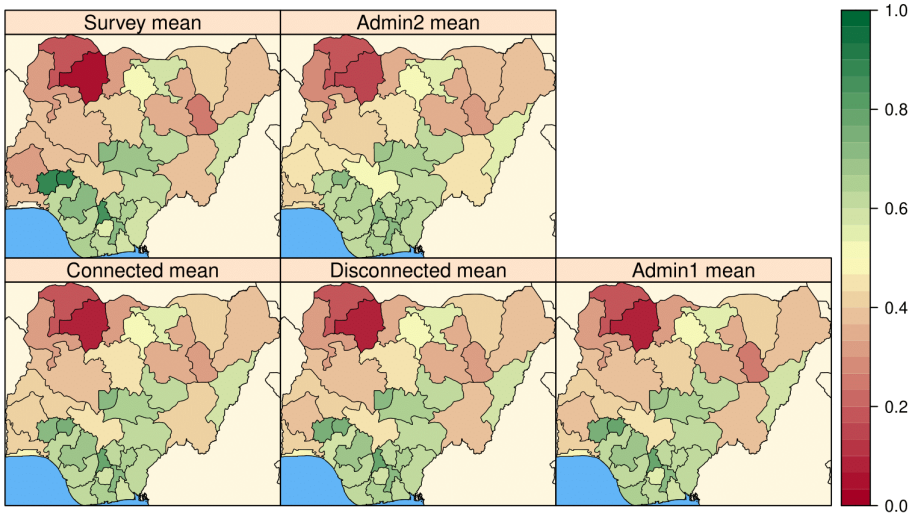


(a) Means

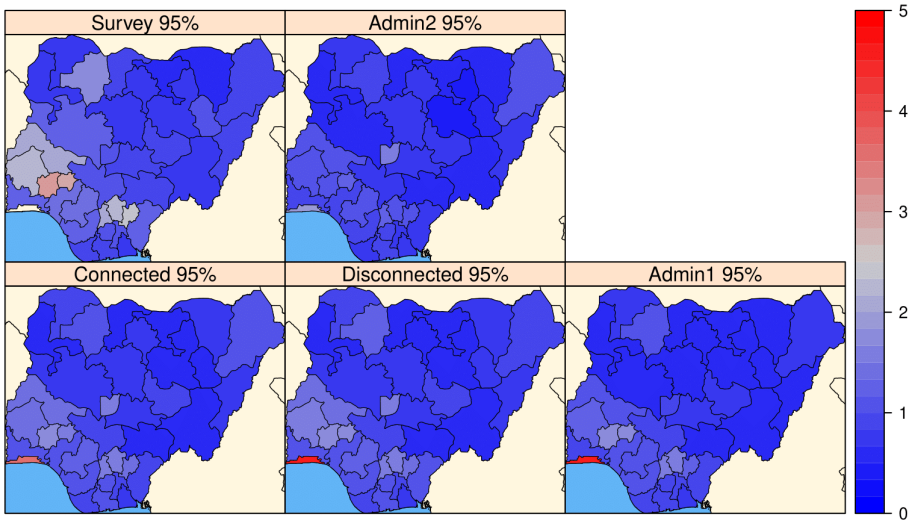


(b) 95% interval size logit scale

Figure 5.1: Maps of posterior means and size of 95% confidence interval and credible intervals on logit scale for state level predictions of urban vaccination coverage.

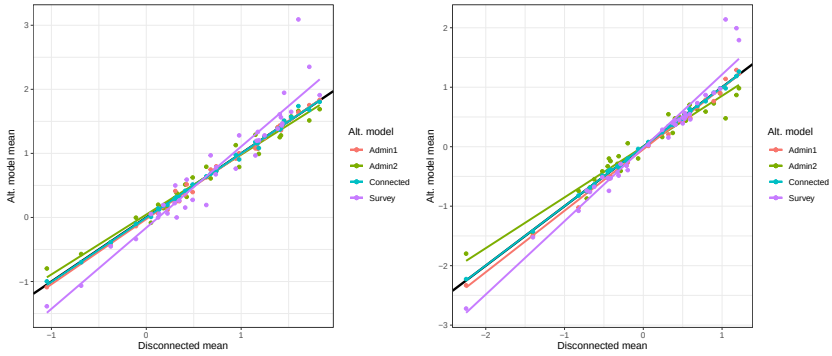


(a) Means

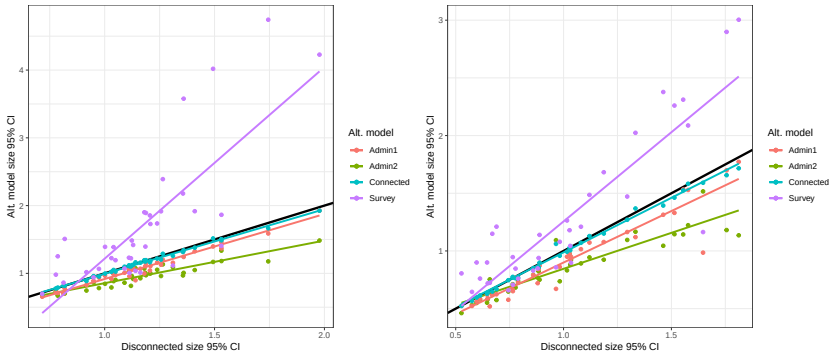


(b) 95% interval size logit scale

Figure 5.2: Maps of posterior means and size of 95% confidence interval and credible intervals on logit scale for state level predictions of rural vaccination coverage.



(a) Urban strata **(b)** Rural strata



(c) Urban strata **(d)** Rural strata

Figure 5.3: Crossplot of state estimates and the sizes of a 95% CI (both on logit scale), using the Disconnected model along the x-axis, and the three other models and survey estimates along the y-axis. The black line is the diagonal $x = y$, and the colored lines are linear regressions for each of the alternative models.

5.2. There are no rural clusters in Lagos, hence there is no direct estimate. The Admin2 model has a credible interval of width 1.99 (on logit scale), which is much narrower than for the Admin1, Disconnected and Connected models, with credible intervals of 4.51, 4.57 and 3.57, respectively.

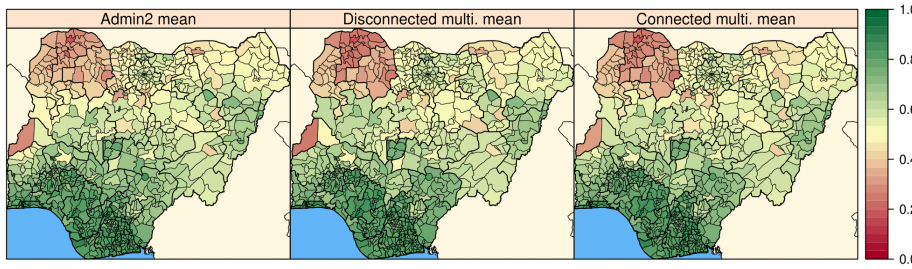
The estimates and size of the CIs for the four different models are compared directly in Figure 5.3. Note that the rural Lagos data point is omitted from the data, since there are no people living there and hence no survey estimate to compare with. The upper two panels show that compared with the survey estimates, all models have much greater smoothing between states. The vaccination coverage estimated by the models tend toward the country mean compared to the direct estimates, and the Admin2 model offers the greatest amount of spatial smoothing. The lower two panels show that the estimated uncertainties of the predictions are greatly reduced using a spatial models instead of direct estimates. In addition to a greater amount of smoothing, the Admin2 model gives sharper predictive distributions than the three other models.

We also display the vaccination coverage estimates for LGAs in Figures 5.4 and 5.5, for urban and rural estimates, respectively. For LGAs, there are too few survey respondents per area to generate reliable direct estimates, and the Admin1 model produces the same estimates for all LGAs within each state, so neither are included in this comparison. For the Disconnected model, with no direct sharing of information between neighbouring LGAs across state borders, we expect to see sharp differences in estimated vaccination coverages along state borders. We also see some of these sharp differences between LGAs across state borders for the Connected model. Overall the Admin2 model appears to have the largest amount of spatial smoothing. A feature of the ICAR model is that the marginal variances in each node is proportional to its number of neighbours. This can be seen in the size of the CIs of the Admin2 model along the border of Nigeria, where the Admin2 model estimates greatest uncertainty.

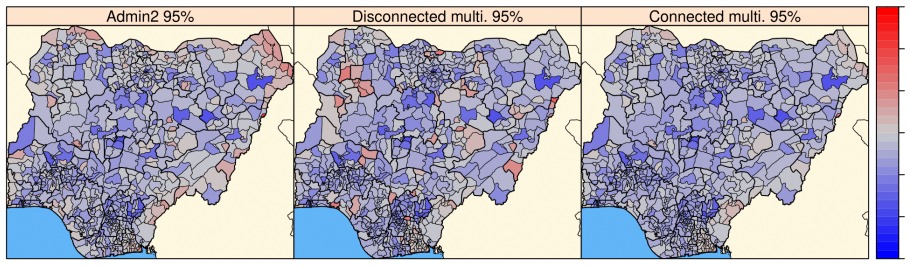
In Figure 5.6 the posterior means and sizes of 95% CIs for LGAs from the fitted Disconnected, Connected and Admin2 models are compared in a crossplot. Notably, the 95% CIs from the fitted Admin2 model are typically slightly larger than the CIs obtained using either of the multiresolution models, which is the opposite of what happens when aggregating to state level. We also see that the posterior means are very similar for all three models.

Table 5.1 compares the estimates of the total precision for the four fitted models. The Admin2, Disconnected and Connected model all share similar total precision values, both for the urban and rural observations. However, the Admin1 model estimates a far greater total precision.

For the Disconnected and Connected model, Table 5.1 also shows the posterior mean and the 95% CI for the weight w_0. Since the Disconnected model is unable to share strength of information across state borders on LGA level, it is forced

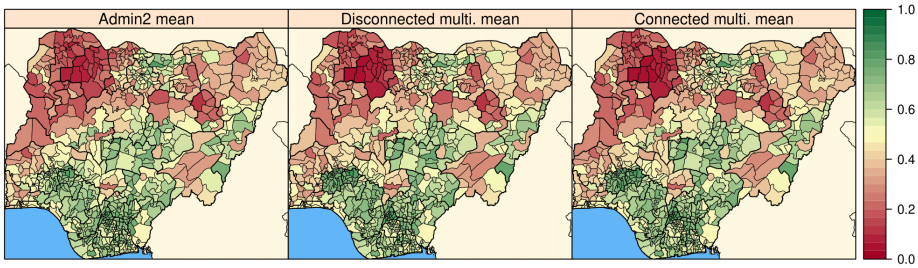


(a) Means

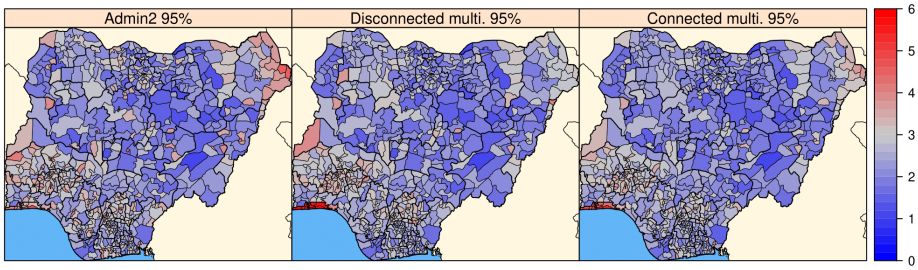


(b) 95% interval size logit scale

Figure 5.4: Maps of posterior means and size of 95% confidence interval and credible intervals on logit scale for LGA level predictions of urban vaccination coverage.

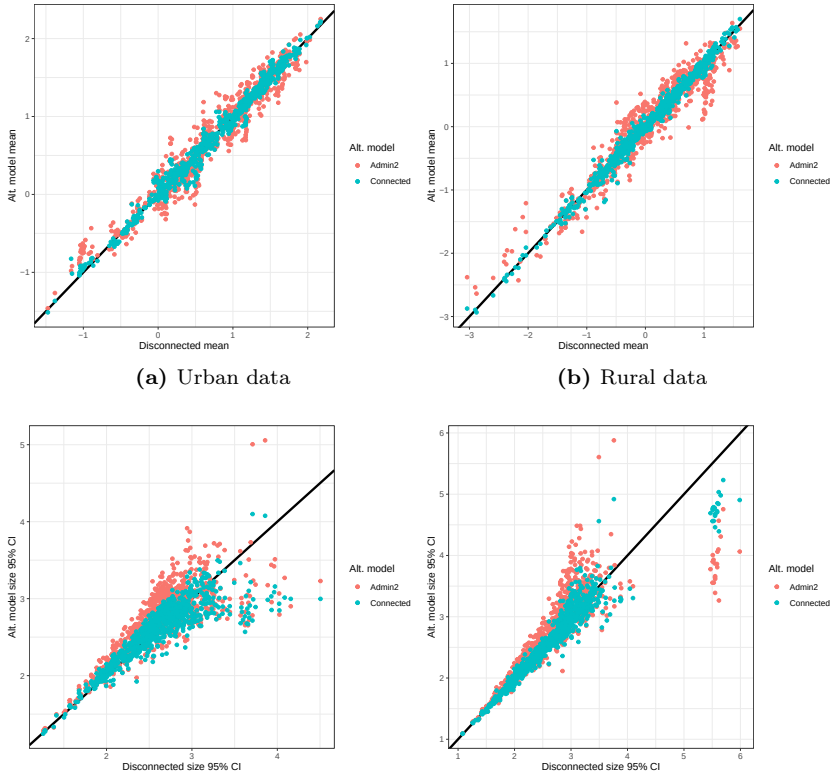


(a) Means



(b) 95% interval size logit scale

Figure 5.5: Maps of posterior means and size of 95% confidence interval and credible intervals on logit scale for LGA level predictions of rural vaccination coverage.



(a) Urban data **(b)** Rural data

(c) Urban data **(d)** Rural data

Figure 5.6: Crossplot of LGA level estimates of the vaccination coverage and the size of the 95% CIs (both on logit scale), using the Disconnected model along the x-axis, and the Connected and Admin2 models along the y-axis.

Table 5.1: Posterior means and 95% CIs of the total precision κ and weight w_0 for the fitted models. The weight w_0 is only a parameter in the Disconnected and Connected model.

Residence	Model	Total precision κ			Weight w_0		
		Mean	2.5%	97.5%	Mean	2.5%	97.5%
Urban	Admin2	1.41	0.88	2.21			
	Admin1	3.59	1.76	6.58			
	Disconnected	1.44	0.86	2.31	0.48	0.28	0.70
	Connected	1.38	0.82	2.24	0.31	0.08	0.58
Rural	Admin2	0.98	0.70	1.33			
	Admin1	1.81	0.94	3.06			
	Disconnected	0.88	0.57	1.26	0.56	0.38	0.72
	Connected	0.92	0.61	1.30	0.35	0.16	0.57

to explain more of the variation state level. This is consistent with the results in the simulation study.

In Figure 5.7 we compare the aggregate value of the state vaccination coverages $\eta_{\text{State } j}$ with the component shared by all LGAs in a given state, i.e. $\eta_j = \mu + \sqrt{w_0 w_1}\, u_j + \sqrt{w_0(1-w_1)}\, v_j$, for the Disconnected and Connected models. If model interpretation is desirable, not just predictive accuracy, we want the two values to agree. We see that for the Disconnected model, there is good agreement between the values, but for the Connected model the state effect does not really control the overall state vaccination coverages.

5.2 Prediction of Direct Estimates

As described in Section 3.6, we evaluate the different models using leave-one-state-out cross-validation for each state. The predictive distributions from the four models are scored using MSE, CRPS and DSS, and the average scores over all the states and urban and rural areas, for each of the models, are shown in Table 5.2. The CRPS and DSS for the different models is also compared in Figure 5.8.

For states where all models manage to predict well, with a low CRPS and DSS, the Admin2 model performs slightly better than the three other models, but the Admin2 model is worse for states that are harder to predict. Table 5.2 shows that on average, the Admin2 and Connected model performs slightly better than the Admin1 and Disconnected model. Overall, the Connected model scores best.

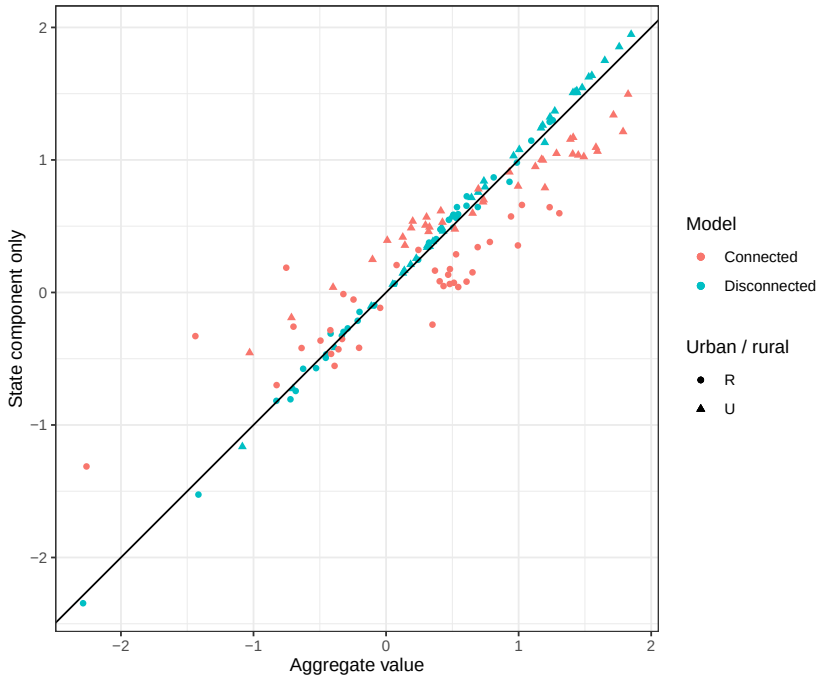


Figure 5.7: Comparison between the aggregated probability estimates on logit scale, and the state component alone, for the Connected model and the Disconnected model.

Table 5.2: Average MSE, CRPS and DSS for the predictive distributions of the survey estimates on logit scale, using leave-one-state-out cross-validation. Lower score is better.

Model	MSE	CRPS	DSS
Admin2	0.480	0.386	0.263
Admin1	0.526	0.397	0.263
Disconnected	0.519	0.392	0.246
Connected	0.479	0.379	0.169

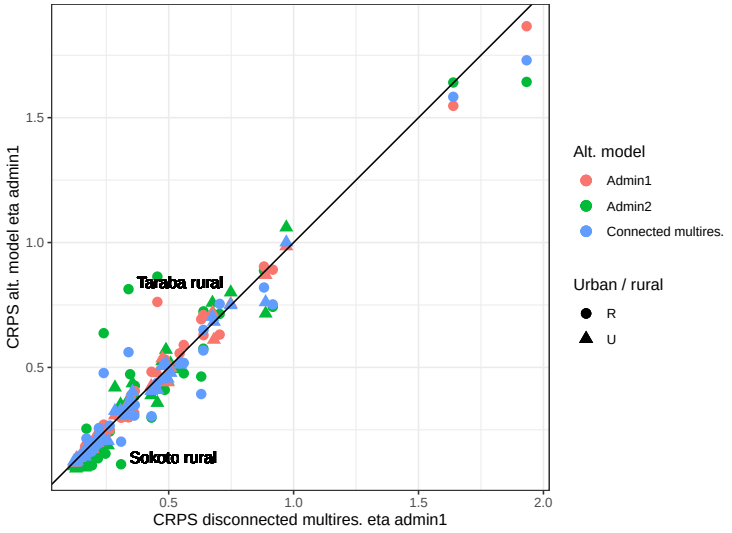
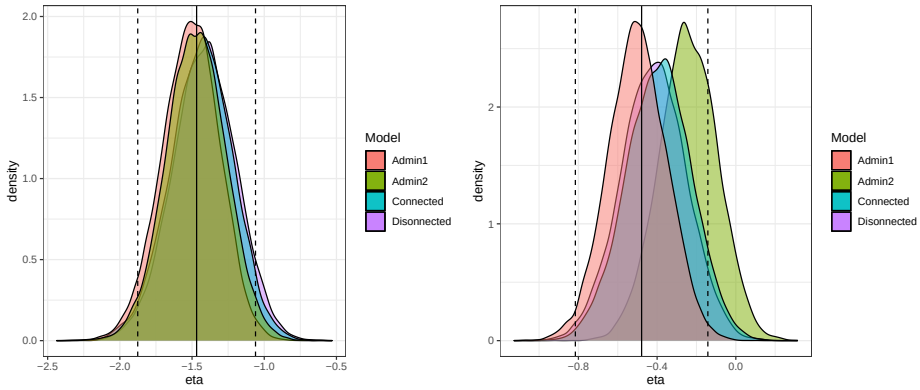


Figure 5.8: Crossplot of the CRPS for the four models using leave-one-state-out cross-validation. All models are compare to the Disconnected model, and two strata where there is a large difference between the Disconnected model and the Admin2 model are marked. Lower score is better.



(a) Sokoto rural **(b)** Taraba rural

Figure 5.9: The leave-one-state-out posterior densities for each of the four models, for the indicated strata. The full black line denotes the direct estimate mean, and the stippled line to the corresponding 95% confidence interval.

In Figure 5.8 we have named two strata, Taraba rural and Sokoto rural, with a large score discrepancy between the Admin2 model and Disconnected model. The posterior distributions on logit scale for the four models are displayed in Figure 5.9, which shows that the large differences occur simply when the location of one of the posterior densities is off. In particular, it is not the case that the models have vastly different variances.

As seen in the simulation study, in particular Figure 4.3, the multiresolution models are only slightly better at predicting on state level when there is moderate to strong state effects ($w_0 = .5$ and $w_0 = .8$), when the data is generated by the Connected model. Furthermore, as more noise is added to the true state vaccination coverages, the difference in score between the Disconnected and the Admin2 model decreases. For the real world data the estimate of w_0 is between 0.31 and 0.35 and the total precision is between 0.92 and 1.38 with the Connected model. The typical standard error of the direct estimates is approximately 0.4. In Figure 4.3 this corresponds to $w_0 = 0.5$ and standard deviation of the noise of 0.4, where we see that the score difference between the Disconnected and Admin2 model is marginal. Since the estimated w_0 for the Connected model fitted to the MCV1 data is roughly 0.3, we expect only minor differences in score between the Connected, Disconnected and Admin2 model that we observe.

5.3 Prediction on Cluster Level

We evaluate the cluster level prediction accuracy of the four models presented in Section 3.3, through a 10-fold cross-validation. First, urban and rural data is separated, then for each of the two groups the clusters are divided into 10 equal sized folds. This results in each fold containing approximately 10% of the urban or the rural survey responses. For the clusters in each fold, the ratios of vaccinated to total number of children, $\hat{p}_i = y_i/n_i$ are calculated, and the data in the nine remaining folds is used to fit the models. Finally, the ratios $\hat{p}_i$ are used with the predictive distributions for the LGAs in which the clusters lie, to estimate the average MSE, CRPS and DSS for the omitted clusters. The results are presented in Table 5.3. Additionally, we compare the cluster CRPS for the different models using a crossplot in Figure 5.10.

Ideally, we would compute the scores on logit scale, but there are many clusters with all vaccinated or all unvaccinated children, making this impossible. This means that the cluster prediction scores (done on probability scale) are not directly comparable to those for state predictions.

From the average scores shown in Table 5.3, we see that the Admin2 model performs marginally better than the Connected and Disconnected model, although the MSE is higher. As expected, the Admin1 model, which predicts the same vaccination coverage for all LGAs in the same state, performs much worse

Table 5.3: Average MSE, CRPS and DSS for cluster level predictions using 10-fold cross-validation. Lower score is better.

Model	MSE	CRPS	DSS
Admin2	0.0908	0.193	2.36
Admin1	0.0909	0.224	38.3
Disconnected	0.0907	0.195	2.86
Connected	0.0901	0.194	2.86

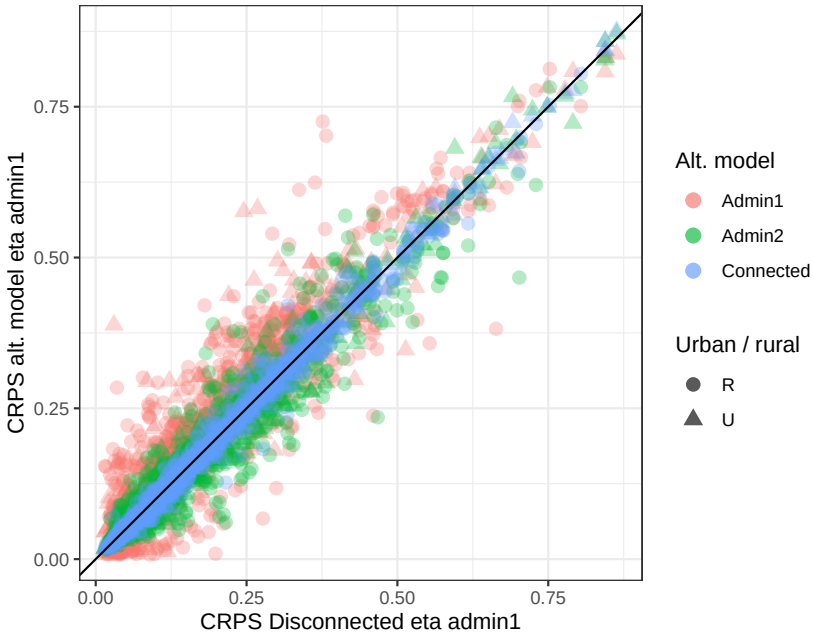


Figure 5.10: Crossplot of the CRPS scores of the four models, using 10-fold cross-validation on cluster level prediction. Lower score is better.

than the three other models when predicting on cluster level. Clearly, there are within state spatial patterns that the three other models are able to capture.

6 | Discussion

We have proposed two new multiresolution discrete spatial models for small area estimation using household survey data, and shown through a simulation study that they outperform the fine scale discrete spatial Admin2 model in the presence of a large or moderate coarse scale variation. The new models also perform equally well or slightly better than the Admin2 model for producing state and LGA level estimates of MCV1 coverage, using data from the 2018 NDHS, with urban and rural data treated separately. Multiresolution models therefore constitute credible alternatives to the fine scale model.

Through the simulation study we showed that for some data sets, the multiresolution models outperform the fine scale model using CRPS as a scoring criterion, when predicting vaccination coverage on coarse and fine geographical scale. However, when fitting the multiresolution models, we are unable to gain much information from the data about some of the model parameters. In particular, it is difficult to distinguish between the structured Besag random effect and the unstructured iid random effect on coarse geographical scale, which is likely because the graph representation of Nigerian states only has 37 nodes. We also show that the score advantage of the multiresolution Disconnected model over the fine scale Admin2 model decreases as noise is added to the true known state vaccination coverages on logit scale. In practice this makes validation through prediction of design-based direct estimates difficult if the standard error of the direct estimates are too large.

The fitted Connected model was able to recover the true parameters used to simulate the data, but the parameter interpretations are not shared by the different models we consider. One of the goals of the multiresolution models is to model on state and LGA level separately. We show that of the two mutliresolution models, the Disconnected and Connected model, only the Disconnected model offers this interpretability. For practitioners, who might want to compare results between countries or set informative priors, this is an attractive feature, and may make the Disconnected model preferable over the Connected model, even though the Connected model tend to perform better in predictive accuracy.

The Disconnected, Connected and Admin2 models all have the ability to highlight geographical heterogeneity that is masked by design-based direct estimates, because they produce estimates of vaccination rates at LGA level, which is the finest geographical scale we consider here. Detecting coldspots in vaccination coverage is of great importance, as they act as sources of outbreaks for diseases such as measles. Analysing survey data with design-based methods is therefore insufficient when combating transmittable diseases. However, when using LGA level estimates produced by model-based methods to set public policies such as vaccination campaigns, one should carefully consider the associated uncertainties. When conducting cross-validation on cluster level data, the MSE on probability scale of the models were roughly 0.09, corresponding to a RMSE of 0.3. On probability scale this is quite large, meaning that even with model-based approaches the sparsity of the data causes considerable uncertainty.

The estimates of state level vaccination coverages from the different models largely agreed with direct estimates, but with smaller uncertainties. Using direct estimates to score the models reveal that the Connected model performs marginally better than the three other models. The validation method also does not reveal if the predictive distributions are well calibrated. For instance, we could examine whether the direct estimates, which are treated as noisy observations of the true vaccination coverages, are over- or underdispersed.

In recent years, continuously indexed models spanning multiple countries have been used to assess the success of vaccination efforts through fine scale pixel maps. For instance, in Local Burden of Disease Vaccine Coverage Collaborators (2021) they map the MCV1 coverage at a $5 \times 5 - \text{km}^2$ resolution in 101 low- and middle-income countries. Here, it appears that there exists sharp differences across national borders. Analysing such data from many countries jointly may therefore be a possible application of multiresolution models. The geographical scale, spanning most of the globe, introduces complications due to the Earth's curvature for continuous models, which are not present for the models presented in Section 3.3.

To properly examine the viability of the models presented here, it is necessary to compare them with best practices modeling approaches. Commonly, continuous spatial models are employed to analyse survey data, with the benefits of avoiding having to define a meaningful neighborhood structure, being able to easily compare results between countries, and using the exact cluster locations. There are also other discrete spatial models, such as the Leroux model, that can outperform the BYM model for disease mapping.

We have also not included a cluster effect or covariates in the model. With no cluster effect one might expect overdispersion in binomial regression models with clustered sample, because potential differences in vaccination rates between clusters in the same LGAs are ignored. By using cluster random effects we

introduce additional complications of how well we are able to distinguish between random effects on state, LGA and cluster level. Most clusters have less than five responses, and for the relevant data there are 1324 clusters distributed over 637 LGAs.

Using covariates in the models would likely increase the predictive accuracy for surveyed households, but raises the question of how one should aggregate over LGAs and states to obtain estimates of MCV1 coverages, when covariate data is unknown for most households. There are covariates with continuous estimates that can be used. For instance, urban and rural stratified data can be analysed jointly. The survey weights are also omitted from the data when doing model-based analysis. Ideally, we want to this information to improve predictive accuracy.

We encounter a similar problem when aggregating from LGA level estimates up to state level. The population counts make no distinction between urban or rural population. Table A.1 in NPC and ICF (2019) clearly shows that the ratio between urban and rural population varies a lot between states, and some LGAs do not contain both urban and rural enumeration areas. It is reasonable to believe the ratios between urban and rural under five population in LGAs varies even more. Thus, there might be significant uncertainty in the aggregation to state level MCV1 coverage, shared by all the models. The population counts are also model-based estimates, which increases the uncertainty of the state level estimates.

7 | Conclusion

The aim of this thesis was to develop multiresolution spatial models as a credible alternative to popular fine scale discrete spatial models. We have shown through a simulation study that for data arising from hierarchical data generating processes, for instance if the vaccination coverage in small local regions are determined by sums of random effects on different geographical resolutions, multiresolution models may outperform simple fine scale models. However, the results indicate that random effects on coarse geographical scale, such as between states or between countries, have to be large compared to fine scale random effects for the multiresolution models to perform better.

Two multiresolution models were applied the 2018 NDHS data for MCV1 coverage, and the predictive accuracies were compared to a fine scale discrete spatial model based on the BYM model. One of the multiresolution models performed best when predicting state level vaccination coverage, but the differences in predictive score between the models considered are small. For prediction on cluster level, the multiresolution models performed equally well as the fine scale model. Since the models were scored by comparing the fitted models to design-based direct estimates, it is difficult to get a large difference between the scores even if there is a strong random effect on coarse geographical scale.

In future work the multiresolution models should be expanded upon by properly acknowledging the survey design, and how to use covariates when aggregating predictions from fine to coarse spatial scales. Our analysis is limited by only comparing the proposed multiresolution models to one fine scale spatial model, and the new multiresolution models should be compared to other spatial model, such as continuously indexed spatial models.

Bibliography

Julian Besag. Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225, 1974.

Julian Besag, Jeremy York, and Annie Mollie. Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43:1–59, 1991.

Michael Betancourt. Identifying the Optimal Integration Time in Hamiltonian Monte Carlo. 2016. arXiv:1601.00225.

A. Philip Dawid and Paola Sebastiani. Coherent Dispersion Criteria for Optimal Experimental Design. *The Annals of Statistics*, 27(1):65–81, 1999. ISSN 00905364.

Tracy Qi Dong and Jon Wakefield. Modeling and presentation of vaccination coverage estimates using data from household surveys, 2020. arXiv:2004.03127.

Anna Freni-Sterrantino, Massimo Ventrucci, and Håvard Rue. A note on intrinsic conditional autoregressive models for disconnected graphs. *Spatial and Spatio-temporal Epidemiology*, 26:25–34, 2018. ISSN 1877-5845.

Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

Geof H. Givens and Jennifer A. Hoeting. *Computational statistics*. John Wiley & Sons, Hoboken, NJ, USA, 2 edition, 2012.

Tilmann Gneiting and Adrian E Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102 (477):359–378, 2007.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

Matthew D. Homan and Andrew Gelman. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15 (1):1593–1623, January 2014. ISSN 1532-4435.

Fabian Krüger, Sebastian Lerch, Thordis Thorarinsdottir, and Tilmann Gneiting. Predictive Inference Based on Markov Chain Monte Carlo Output. *International Statistical Review*, 2020. doi: https://doi.org/10.1111/insr.12405.

Local Burden of Disease Vaccine Coverage Collaborators. Mapping routine measles vaccination in low- and middle-income countries. *Nature*, 589(7842): 415–419, Jan 2021. ISSN 1476-4687.

Sharon L. Lohr. *SAMPLING Design and Analysis, second edition.* CHAPMAN and HALL CRC, 2 edition, 2010.

Thomas Lumley. Analysis of Complex Survey Samples. *Journal of Statistical Software, Articles*, 9(8):1–19, 2004. ISSN 1548-7660.

James E. Matheson and Robert L. Winkler. Scoring Rules for Continuous Probability Distributions. *Management Science*, 22(10):1087–1096, 1976.

Laina Mercer, Jon Wakefield, Cici Chen, and Thomas Lumley. A comparison of spatial smoothing methods for small area estimation with sampling weights. *Spatial Statistics*, 8:69–85, 2014. ISSN 2211-6753. Spatial Statistics Miami.

Radford M. Neal. Slice sampling. *The Annals of Statistics*, 31(3):705 – 767, 2003.

National Population Commission NPC and ICF. Nigeria Demographic and Health Survey 2018 - Final Report. Technical report, Abuja, Nigeria, 2019. URL `http://dhsprogram.com/pubs/pdf/FR359/FR359.pdf`.

R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

Håvard Rue and Leonhard Held. *Gaussian Markov Random Fields, Theory and Applications.* Chapman and Hall/CRC, New York, 1 edition, 2005.

Daniel Simpson, Håvard Rue, Andrea Riebler, Thiago G. Martins, and Sigrunn H. Sørbye. Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors. *Statistical Science*, 32(1):1 – 28, 2017.

Stan Development Team. RStan: the R interface to Stan, 2020. URL `http://mc-stan.org/`. R package version 2.21.2.

Andrew J. Tatem. WorldPop, open data for spatial demography, 2021. URL `https://www.worldpop.org/doi/10.5258/SOTON/WP00645`. [Online; accessed 08-February-2021].

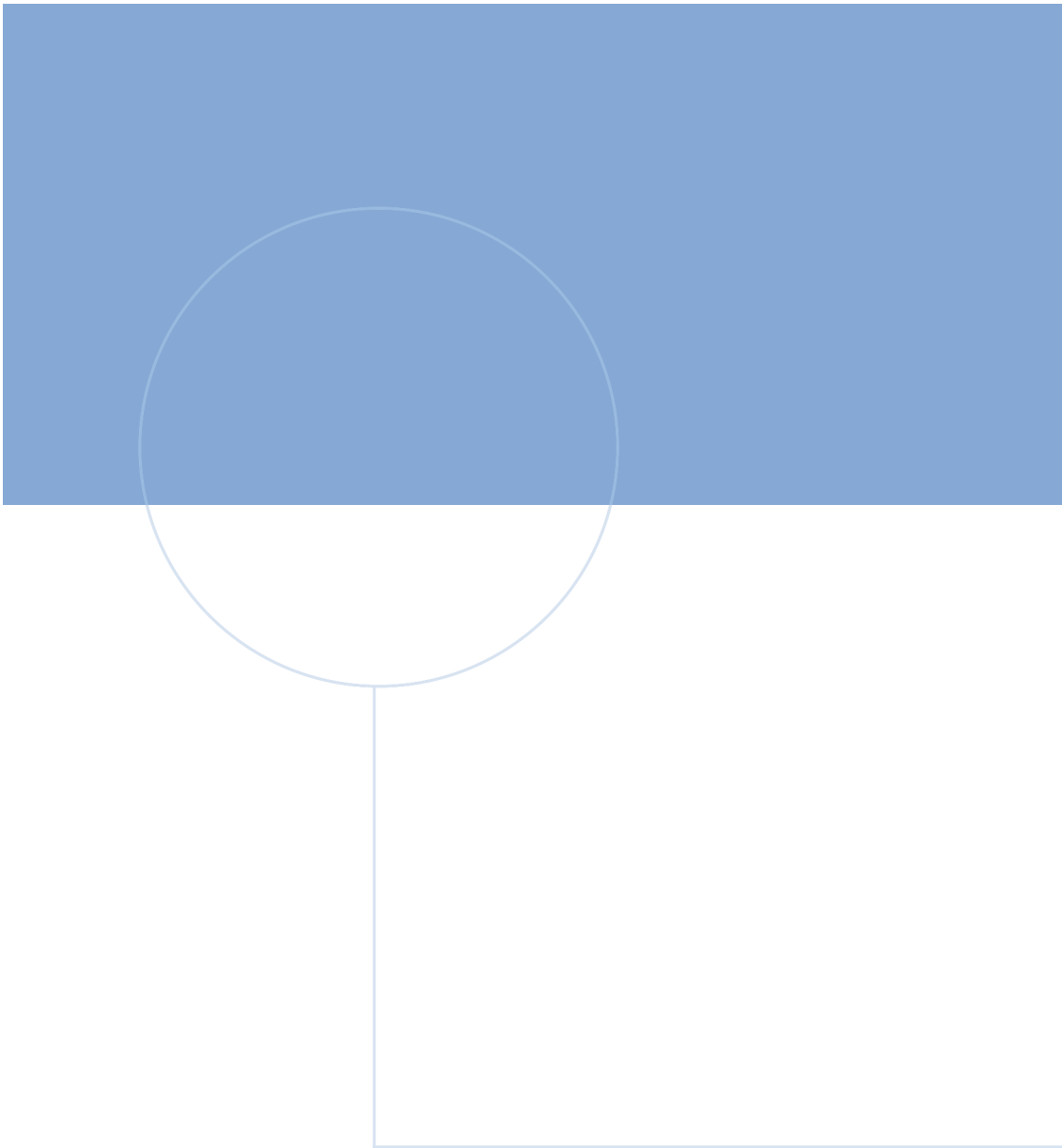
Unicef. Leaving No One Behind: All children immunized and healthy, 2019. URL `https://data.unicef.org/resources/all-children-immunized-and-healthy/`. [Online; accessed 23-May-2021].

United Nations General Assembly. Transforming our World: the 2030 Agenda for Sustainable Development. Technical report, 2015. URL `http://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1`. [Online; accessed 01-May-2021].

C. Edson Utazi, John Wagai, Oliver Pannell, Felicity T. Cutts, Dale A. Rhoda, Matthew J. Ferrari, Boubacar Dieng, Joseph Oteri, M. Carolina Danovaro-Holliday, Adeyemi Adeniran, and Andrew J. Tatem. Geospatial variation in measles vaccine coverage through routine and campaign strategies in Nigeria: Analysis of recent household surveys. *Vaccine*, 38(14):3062–3071, 2020. ISSN 0264-410X.

C. Edson Utazi, Kristine Nilsen, Oliver Pannell, Winfred Dotse-Gborgbortsi, and Andrew J. Tatem. District-level estimation of vaccination coverage: Discrete vs continuous spatial models. *Statistics in Medicine*, 40(9):2197–2211, 2021.

Craig Wang, Milo A Puhan, and Reinhard Furrer. Generalized spatial fusion model framework for joint analysis of point and areal data. *Spatial Statistics*, 23:72–90, 2018. ISSN 2211-6753.



NTNU

Norwegian University of
Science and Technology