

# TMA4250 - Project 3

Ole Riddervold, Ole Kristian Skogly

April 2023

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Simulation and Visualization</b>	<b>3</b>
2.1	Precision matrices of the Besag model . . . . .	3
2.2	Besag model vs. normal distribution on admin1 . . . . .	4
2.3	Besag model vs. normal distribution on admin2 . . . . .	6
2.4	Empirical correlations and Variances in Besag model on admin2 . . . . .	6
<b>3</b>	<b>Small Area Estimation</b>	<b>9</b>
3.1	Display of observations . . . . .	9
3.2	Independent hierarchical model . . . . .	9
3.3	Besag hierarchical model . . . . .	12
3.4	Model update with data from Kaduna . . . . .	12
3.5	Sensitivity of the Besag hyperparameter . . . . .	15
3.6	MLE of Besag hyperparameter . . . . .	16
<b>4</b>	<b>Conclusion</b>	<b>18</b>

# 1 Introduction

This project focuses on Gaussian Markov random fields (GMRFs), specifically the construction and simulation of Besag models on the admin1 and admin2 graphs of Nigeria. admin1 and admin2 refer to the two nested subdivisions in the geography of Nigeria, where admin1 consists of 37 areas and admin2 consists of 775 areas (excluding a lake). The project serves as a demonstration as to how spatial statistics can be applied to do prediction in situations with sparse data. It will investigate frequentist spatial models in comparison to traditional statistical models, a direct- and spatial Bayesian hierarchical approach, updating with additional data, as well as parameter estimation for spatial hierarchical models. Problem 1 involves construction-, simulation- and visualization of results of Besag models. Problem 2 focuses on small area estimation of vaccination coverage for children in the 37 admin1 areas of Nigeria. Coding has been done using R and relevant libraries, among others `ggplot`, `sgdal` and `spdep`.

## 2 Simulation and Visualization

Problem 1 involves constructing and simulating Besag models on the admin1 and admin2 graphs of Nigeria. The precision matrices for these models are constructed using neighborhood matrices, with dimension and rank depending on the number of areas in each graph. The sparsity pattern and proportion of non-zero elements in the precision matrices are computed and discussed, along with the benefits of using GMRFs instead of standard multivariate Gaussian distributions. Realizations from the Besag models and a standard multivariate Gaussian distribution are generated and displayed on a map of Nigeria, with the purpose of the sum-to-zero constraint explained. The process is repeated for the admin2 graph, and the empirical marginal variance and correlation between admin2 areas are computed and analyzed.

### 2.1 Precision matrices of the Besag model

Let  $G_1 = (\mathcal{V}_1, \mathcal{E}_1)$  and  $G_2 = (\mathcal{V}_2, \mathcal{E}_2)$  be a graph representing the admin1 and admin2 spatial structure of Nigeria, respectively. Where  $\mathcal{V}$  denotes the nodes (states in Nigeria) and  $\mathcal{E}$  denotes the edges (neighboring states). To construct the precision matrices, we use the definition of the density function for the Besag model, which is given by:

$$f(\mathbf{x}; \tau) \propto \tau^{-\frac{n-1}{2}} \exp \left( -\frac{\tau}{2} \sum_{i \sim j} (x_i - x_j)^2 \right) \quad (1)$$

Here,  $\mathbf{x} \in \mathbb{R}^n$ , where  $n$  is the number of nodes in  $\mathcal{V}$ . The notation  $i \sim j$  indicates that the sum is taken over all pairs of neighboring nodes  $(i, j) \in \mathcal{E}$ .

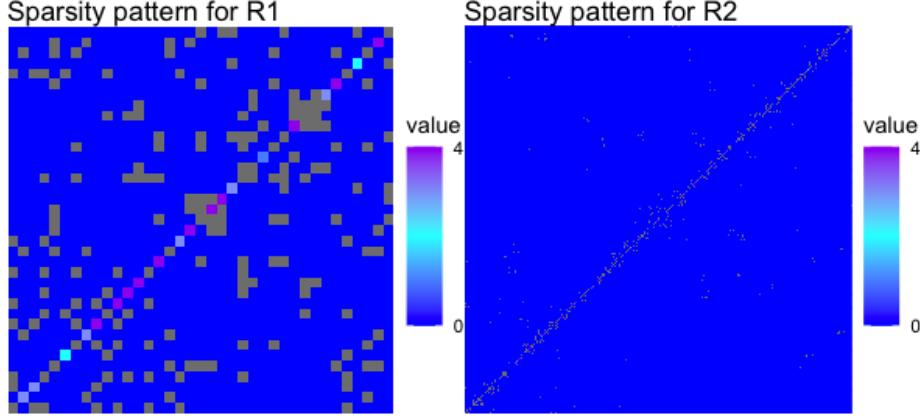
By doing some algebraic manipulation of Equation (1), one can see that, assuming the mean is  $\mathbf{0}$ , the precision matrices for the Besag model on  $G_1$  and  $G_2$  can be defined as  $\mathbf{Q}_1 = \tau_1 \mathbf{R}_1$  and  $\mathbf{Q}_2 = \tau_2 \mathbf{R}_2$  respectively.  $\mathbf{R}_1$  and  $\mathbf{R}_2$  are called structure matrices, and  $\tau_1 > 0$  and  $\tau_2 > 0$  are the precision parameters that one wants to estimate. The structure matrix is defined as

$$R_{ij} = \begin{cases} n_i & i = j \\ -1 & i \sim j \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where  $n_i$  denotes the amount of neighbors for node  $i$ .

The precision matrix  $\mathbf{Q}_1 \in \mathbb{R}^{37 \times 37}$  for the admin1 graph, with a rank of 36; while the precision matrix  $\mathbf{Q}_2 \in \mathbb{R}^{775 \times 775}$  for the admin2 graph has a rank of 774. For the admin1 graph, the proportion of non-zero elements is 0.1526, while for the admin2 graph, it is 0.0088.

Treating Besag models as GMRFs provides several benefits over treating them as standard multivariate Gaussian distributions. For one, many GMRFs, like the Besag model, are improper, meaning that the precision matrix does not have full rank. This makes it difficult to use traditional methods for working with the multivariate Gaussians, since the precision matrix cannot be inverted. GMRFs also often have sparse precision matrices, making them computationally efficient for large datasets. The sparsity patterns for the precision matrices  $R_1$  and  $R_2$  can be seen in Figure 1.



**Figure 1:** Sparsity pattern for the precision matrices  $R_1$  and  $R_2$ .

## 2.2 Besag model vs. normal distribution on admin1

---

**Algorithm 1** Calculating first-order GMRF

---

```

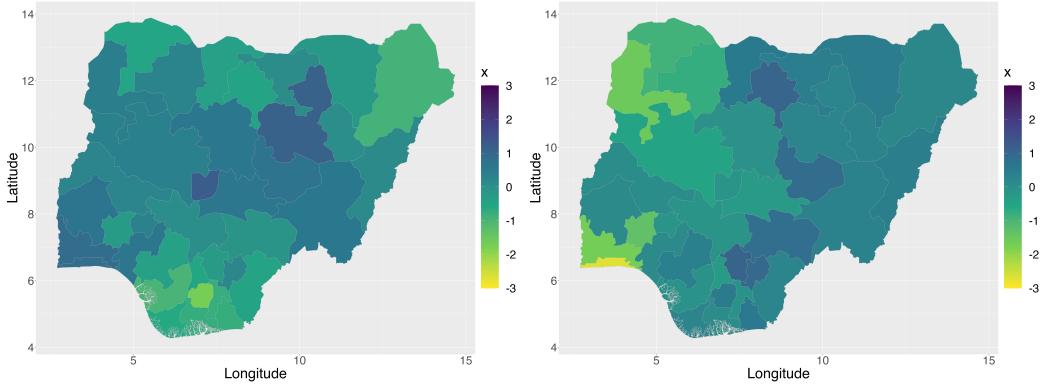
function FIRSTORDERGMRF( $Q$ :  $n \times n$  precision matrix,  $\epsilon$ : small value added to the diagonal of  $Q$  for numerical stability)
     $n \leftarrow$  length of  $Q$ 
     $Q \leftarrow Q + \epsilon \cdot$  Add epsilon to the diagonal of  $Q$ 
     $L \leftarrow$  Cholesky decomposition of  $Q$ 
     $z \leftarrow$  vector of  $n$  random normal variables
     $v \leftarrow$  solution to the linear system  $L \cdot v = z$ 
     $x \leftarrow v - \text{mean}(v) \cdot$  Compute  $x$  by centering  $v$  around its mean
    return  $x$ 
end function

```

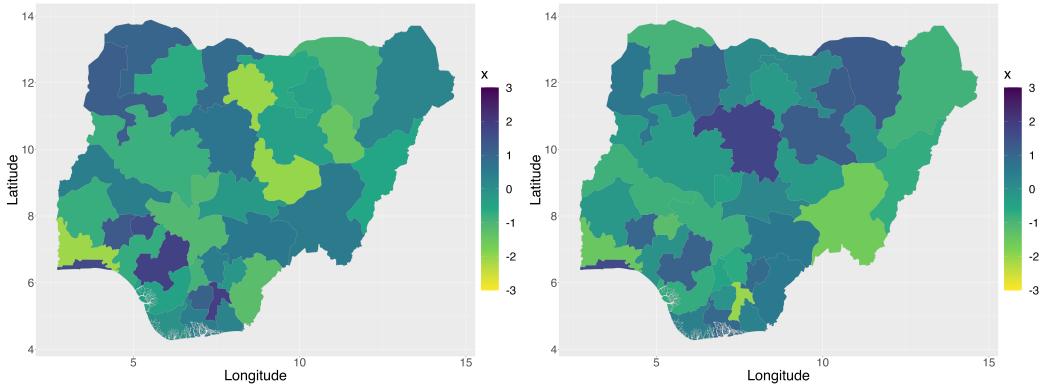
---

To simulate the Besag model on the admin1 graph with  $\tau_1 = 1$  and sum-to-zero constraint, one can apply Algorithm 1. First, construct the precision matrix  $\mathbf{Q}_1$  using the neighborhood matrix  $\mathbf{R}_1$  and precision parameter  $\tau_1$ . Then, add a small value to the diagonal of  $Q_1$  for the purpose of numerical stability, and finally compute its Cholesky decomposition. One can then generate a vector of 37 standard normal random variables, and solve the linear system  $\mathbf{L}_1 \cdot \mathbf{v} = \mathbf{z}$  for  $v$ , where  $\mathbf{L}_1$  is the Cholesky decomposition of  $\mathbf{L}_1$ , and  $\mathbf{z}$  is the vector of random normal variables. Finally, we center  $\mathbf{v}$  around its mean to enforce the sum-to-zero constraint.

The distribution  $N_{37}(\mathbf{0}, \mathbf{I}_{37})$  can be sampled by using the `rnorm` function in R, which returns 37-dimensional vectors of samples from the relevant normal distribution.



**Figure 2:** Realization 1 and 2 from the Besag model on the admin1 graph with  $\tau_1 = 1$ .



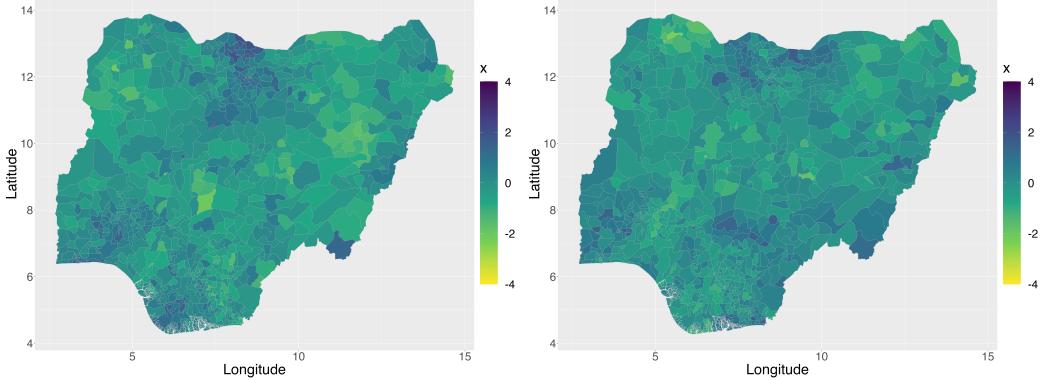
**Figure 3:** Realization 1 and 2 from the normal distribution on the admin1 graph

The sum-to-zero constraint in the Besag model on the admin1 graph enforces that the sum of the values of the random variables over all administrative areas is zero to ensure identifiability of the model parameters and to account for any overall spatial trend in the data.

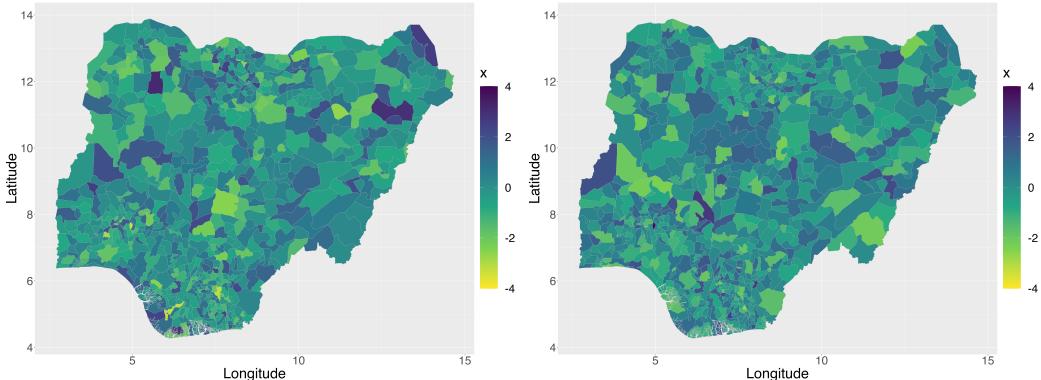
The Besag model and standard-normal distribution assume different dependence spatial structures between random variables representing each region in admin1. The Besag model enforces stronger dependence between neighboring areas by weighting the pdf of each variable by the differences to their neighbors, while the normal distribution assumes no specific dependence structure, and uniform variance across all regions. Besag model simulations show smoother spatial patterns due to the enforced dependence as seen in Figure 2, while normal distribution simulations are more variable and less spatially coherent, seen in Figure 3.

### 2.3 Besag model vs. normal distribution on admin2

This section will cover the exact same analysis as in Section 2.2 on the higher-resolution admin2 graph.



**Figure 4:** Realization 1 and 2 from the Besag model on the admin2 graph with  $\tau_2 = 1$

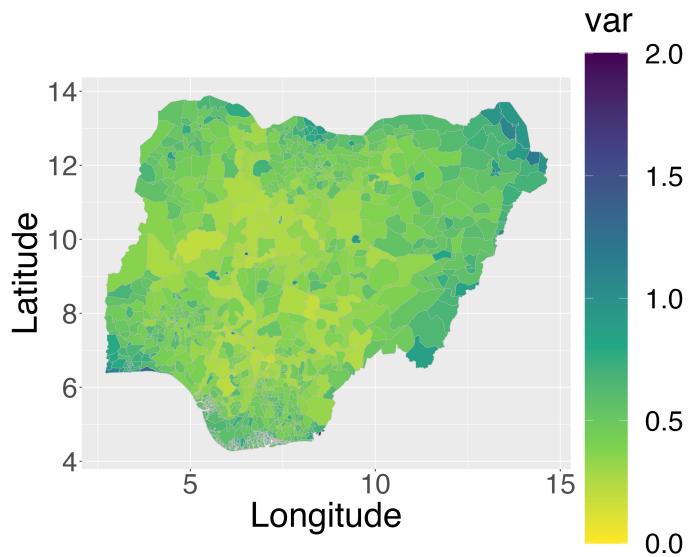


**Figure 5:** Realization 1 and 2 from the normal distribution on the admin2 graph

The realizations from the Besag model on  $G_2$  and the normal distribution  $N_{775}(\mathbf{0}, \mathbf{I}_{775})$  yielded large differences in spatial patterns as seen in Figures 4 and 5. The Besag model's realizations show smoother and more spatially coherent patterns, while the realizations based on the normal distribution are more variable and less coherent. Compared to the admin1 graph discussed in the previous problem, the dissimilarities between the two distributions appear to be more apparent in the admin2 graph. The sum-to-zero constraint is once again imposed to guarantee that the random variables' values across all administrative areas add up to zero.

### 2.4 Empirical correlations and Variances in Besag model on admin2

One can observe from the 100 simulations using the Besag model on the admin2 graph with  $\tau_2 = 1$ , the variance is non-constant along the edges of the map, indicating that the model is non-stationary



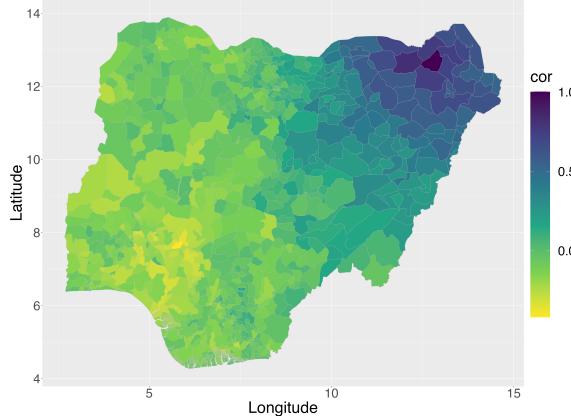
**Figure 6:** 100 realizations from the Besag model on the admin2 graph with  $\tau_2 = 1$  where the marginal variance is plotted.

(but only slightly so), with reference to Figure 6.

We examine Gubio, designated as state 150, and determine the empirical correlations between Gubio and all other admin2 areas based on 100 realizations. From Nigeria's map, Gubio appears at the upper right corner, with dependencies on states all over the country, including those situated far away. Although the pairwise Markov property dictates that  $X_i$  and  $X_j$  are conditionally independent, given  $X_{i,j}$ , if  $i$  and  $j$  are not neighbors, the Besag model may still indicate a dependence between non-neighboring states. This is because the pairwise Markov property only guarantees conditional independence between non-neighboring states, and not unconditional independence, meaning that intermediate states can cause correlation between states that are far apart. Negative correlations can also arise when the spatial effects between adjacent areas have opposing directions. Furthermore, the sum-to-zero constraint implies that negative correlations must exist, as the sum of all  $x_i$  values is constrained to zero:

$$\sum_{i \in \mathcal{V}} x_i = 0$$

The Figure 7 clearly shows that there are negative correlations as well.



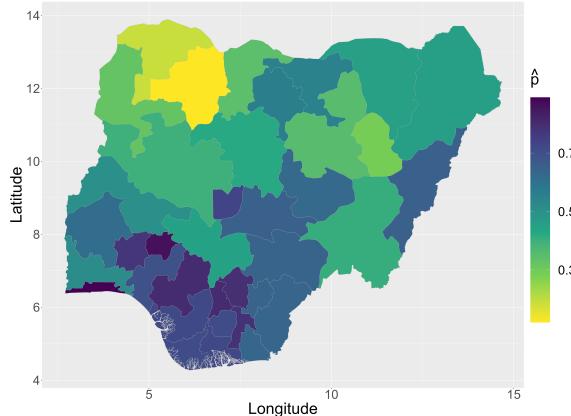
**Figure 7:** 100 realizations for the Gubio (admin2 area number 150), where the correlations is plotted.

### 3 Small Area Estimation

Small area estimation is a statistical technique used to estimate parameters of a small geographic area using survey data. This problem will concern the estimation of vaccination coverage (proportion vaccinated) for children in the 37 admin1 areas in Nigeria. This is a challenging task due to the absence of a complete registration system for vaccines in Nigeria, requiring large efforts to manually collect data in each state. The goal of this analysis is to learn about the true proportions of vaccinated children based on observed proportions using hierarchical Bayesian spatial models, based only on observations from surveys and known variances.

#### 3.1 Display of observations

Spatial correlation in vaccination coverage is apparent in Figure 8, with generally decreasing vaccination coverage from north to south. To determine whether using a spatial model is appropriate for reducing uncertainty in small area estimation, it is crucial to consider the validity of certain key assumptions. Specifically, it is essential to assume that the true proportions of vaccination coverage vary spatially and that the model can adequately capture the correlation between neighboring areas. Despite the potential bias introduced by the spatial model, the reduction in variance achieved may justify its use for borrowing strength in space. Therefore, the decision to use a spatial model to reduce uncertainty should be based on the validity of these assumptions.



**Figure 8:** Observed proportions on the Nigeria map.

#### 3.2 Independent hierarchical model

This subsection will define a simple hierarchical model of vaccination probabilities, based on the (unrealistic) assumption that the vaccination rate in each state is spatially independent.

Let  $p_a$ ,  $a \in \{1, \dots, n\}$  ( $n = 37$ ) be the true proportion of kids vaccinated in state  $a$  in Nigeria,

and let  $\hat{P}_a$ ,  $a \in \{1, \dots, n\}$  be an estimator of  $p_a$ . Furthermore, let

$$\begin{aligned}\mathbf{Y} &= (\text{logit}(\hat{P}_1), \dots, \text{logit}(\hat{P}_n))^T \\ \mathbf{X} &= (\text{logit}(P_1), \dots, \text{logit}(P_n))^T\end{aligned}$$

where  $\mathbf{Y}$  is observed and  $\mathbf{X}$  is a latent quantity that we wish to estimate. It is assumed that  $Y \sim \mathcal{N}_n(\text{logit}(p_a), V_a)$ , and thus, the likelihood function for the data  $\mathbf{Y}$  is given by

$$\mathbf{Y} | \mathbf{X} = \mathbf{x} \sim \mathcal{N}_n(\mathbf{x}, \mathbf{D}) \quad (3)$$

where  $\mathbf{D} = \text{diag}(V_1, \dots, V_n)$  are the observed variances.  $\mathbf{X}$  gets the prior

$$\mathbf{X} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

What we have here is a conjugate prior distribution, meaning that the posterior will also be normal. We will start by introducing a theorem from [Fuglstad, 2023] which will be applied many times throughout this problem.

**Theorem 3.1.** *Let  $\mathbf{X}$  be some parameter of interest, and let the data  $\mathbf{Y}$  have the likelihood function defined by  $\mathbf{Y} | \mathbf{X} = \mathbf{x} \sim \mathcal{N}_m(\mathbf{Ax}, \mathbf{Q}^{-1})$ , where  $\mathbf{A} \in \mathbb{R}^{m \times n}$ . Using the prior on  $\mathbf{X}$  defined by  $\mathbf{X} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{Q}_0^{-1})$ , the posterior distribution is given by*

$$\mathbf{X} | \mathbf{Y} = \mathbf{y} \sim \mathcal{N}_n \left( (\mathbf{Q}_0 + \mathbf{A} \mathbf{Q}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{Q} \mathbf{y}, (\mathbf{Q}_0 + \mathbf{A}^T \mathbf{Q} \mathbf{A})^{-1} \right)$$

*Proof.* The posterior pdf can be written as

$$\begin{aligned}f(\mathbf{x}, \mathbf{y}) &= f(\mathbf{y})f(\mathbf{x}|\mathbf{y}) \\ &\propto \exp \left( -\frac{1}{2} \mathbf{x}^T \mathbf{Q}_0 \mathbf{x} \right) \exp \left( -\frac{1}{2} (\mathbf{y} - \mathbf{Ax})^T \mathbf{Q} (\mathbf{y} - \mathbf{Ax}) \right) \\ &= \exp \left( -\frac{1}{2} \mathbf{x}^T \mathbf{Q}_0 \mathbf{x} - \frac{1}{2} \mathbf{y}^T \mathbf{Q} \mathbf{y} + \frac{1}{2} \mathbf{y}^T \mathbf{Q} \mathbf{A} \mathbf{x} + \frac{1}{2} \mathbf{x}^T \mathbf{A}^T \mathbf{Q} \mathbf{y} - \frac{1}{2} \mathbf{x}^T \mathbf{A} \mathbf{Q} \mathbf{A} \mathbf{x} \right) \\ &= \exp \left( -\frac{1}{2} \mathbf{x}^T (\mathbf{Q}_0 + \mathbf{A}^T \mathbf{Q} \mathbf{A}) \mathbf{x} + \mathbf{x}^T \mathbf{A}^T \mathbf{Q} \mathbf{y} - \frac{1}{2} \mathbf{y}^T \mathbf{Q} \mathbf{y} \right) \quad (4)\end{aligned}$$

The strategy of this proof is to write out the pdf of a multivariate normal distribution, compare the terms with the ones in Equation (4), and finally write the pdf out as a conditional normal distribution. The multivariate normal pdf has the form

$$\begin{aligned}f(\mathbf{x}, \mathbf{y}) &\propto \exp \left( -\frac{1}{2} (\mathbf{x}^T \mathbf{Q}_{xx} \mathbf{x} + \mathbf{y}^T \mathbf{Q}_{yy} \mathbf{y} - 2 \mathbf{x}^T \mathbf{Q}_{xy} \mathbf{y}) \right) \\ &= \exp \left( -\frac{1}{2} \mathbf{x}^T \mathbf{Q}_{xx} \mathbf{x} - \mathbf{x}^T \mathbf{Q}_{xy} \mathbf{y} - \frac{1}{2} \mathbf{y}^T \mathbf{Q}_{yy} \mathbf{y} \right) \quad (5)\end{aligned}$$

Comparing the terms in Equation (5) to those in Equation (4), we get

$$\begin{aligned}\mathbf{Q}_{xx} &= \mathbf{Q}_0 + \mathbf{A}^T \mathbf{Q} \mathbf{A} \\ \mathbf{Q}_{xy} &= \mathbf{Q}_{yx}^T - \mathbf{A}^T \mathbf{Q} \\ \mathbf{Q}_{yy} &= \mathbf{Q}\end{aligned}\tag{6}$$

Finally, writing out  $\boldsymbol{\mu}_{x|y}$  and  $\mathbf{Q}_{x|y}$  based on the results from Equation (6) yields

$$\begin{aligned}\boldsymbol{\mu}_{x|y} &= (\mathbf{Q}_0 + \mathbf{A}^T \mathbf{Q} \mathbf{A})^{-1} \mathbf{A} \mathbf{Q} \mathbf{y} \\ \mathbf{Q}_{x|y} &= \mathbf{Q}_0 + \mathbf{A}^T \mathbf{Q} \mathbf{A}\end{aligned}$$

giving to the posterior distribution

$$\mathbf{X} | \mathbf{Y} = \mathbf{y} \sim \mathcal{N}_n \left( (\mathbf{Q}_0 + \mathbf{A} \mathbf{Q}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{Q} \mathbf{y}, (\mathbf{Q}_0 + \mathbf{A}^T \mathbf{Q} \mathbf{A})^{-1} \right)$$

as we wanted to show.  $\square$

Applying Theorem 3.1 to our model where  $\mathbf{A} = \mathbf{I}_{37}$ ,  $\mathbf{Q} = \mathbf{D}^{-1}$  and  $\mathbf{Q}_0 = \frac{1}{\sigma^2} \mathbf{I}_n$ , we get a posterior distribution on the form

$$\mathbf{X} | \mathbf{Y} = \mathbf{y} \sim \mathcal{N}_n \left( \left( \mathbf{D}^{-1} + \frac{1}{\sigma^2} \mathbf{I}_n \right)^{-1} \mathbf{D}^{-1} \mathbf{y}, \left( \mathbf{D}^{-1} + \frac{1}{\sigma^2} \mathbf{I}_n \right)^{-1} \right)\tag{7}$$

Since both  $\mathbf{D}$  and  $\mathbf{I}_n$  are diagonal matrices, the components of  $\mathbf{X} | \mathbf{Y}$  are independent. Furthermore, it is relatively easy to see that

$$X_a | Y_a = y_a \sim \mathcal{N} \left( \frac{1}{\frac{1}{V_a} + \frac{1}{\sigma^2}} \frac{y_a}{V_a}, \frac{1}{\frac{1}{V_a} + \frac{1}{\sigma^2}} \right) = \mathcal{N} \left( \frac{\sigma^2}{V_a + \sigma^2} y_a, \frac{\sigma^2}{V_a + \sigma^2} V_a \right), a \in \{1, \dots, n\}$$

If  $\mathbf{X}$  is an improper prior (i.e. the variance approaches infinity), then Equation (7) becomes:

$$\mathbf{X} | \mathbf{Y} = \mathbf{y} \sim \mathcal{N}_n (\mathbf{y}, \mathbf{D})\tag{8}$$

We can also see that the distribution of  $P_a | \mathbf{Y} = \mathbf{y}$  is simply an expit-transformation of the posterior, i.e.

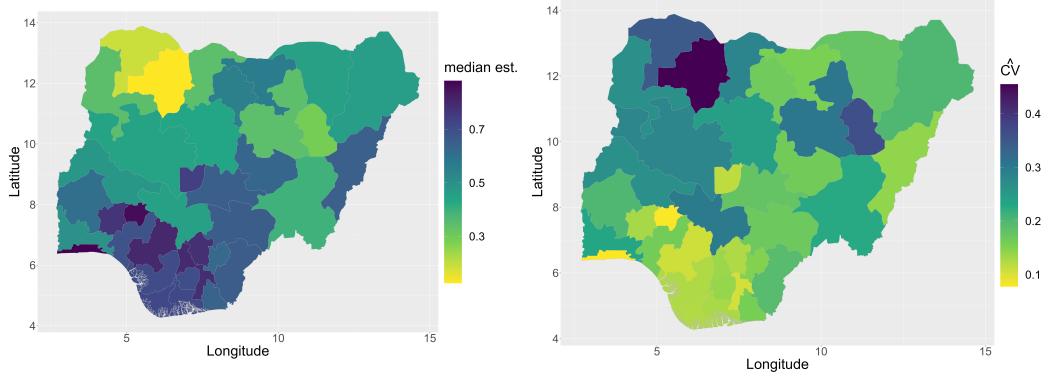
$$P_a | \mathbf{Y} = \mathbf{y} = \text{expit}(X_a | \mathbf{Y} = \mathbf{y}) = \text{expit}(X_a | Y_a = y_a)$$

Its pdf can be derived by using the formula for transformation of random variables, and is given by

$$f_{P_a | Y_a}(p_a) = \frac{1}{\sqrt{2\pi V_a}} \frac{1}{p_a(1-p_a)} \exp \left( -\frac{1}{2} \frac{\left( \ln \left( \frac{1}{p_a} - 1 \right) - y_a \right)^2}{V_a} \right)$$

Empirical estimates of the posterior median, as well as an estimate of the coefficient of variation, defined as  $CV = \sigma/\mu$ , can be seen in Figure 9.

Comparing the estimated medians from Figure 9 with the observations in Figure 8, one can see a striking resemblance between the plots. This is to be expected, as the prior assumption assumed no spatial correlation and a very high variance, making it behave like an improper prior, which doesn't really contain any information. Consequently, the resulting bias is low, and the simulated values are predominantly influenced by the observed proportions.



**Figure 9:** Estimated median and the coefficient of variation for  $P_a|Y = y$  following a spatially independent and approximately improper (very large variance) normal prior, empirically based on 100 samples.

### 3.3 Besag hierarchical model

In this section, we will go one step further from Section 3.2 and add spatial modeling by assuming that  $\mathbf{X}$  a priori follows a Besag model. Let  $G = (\mathcal{V}, \mathcal{E})$  be a graph representing the admin1 spatial structure of Nigeria corresponding to  $G_1$  in Section 2.1. Recall also the pdf of the Besag model in Equation (1) and the definition of the structure matrix in Equation (2) from Section 2.1.

The Besag model is an intrinsic GMRF, meaning it is an improper GMRF of rank  $n - 1$ . Consequently, the precision matrix (which also has rank  $n - 1$ ) is not invertible, and the prior therefore does not have a well-defined normal distribution. It turns out, however, that Theorem 3.1 can be applied regardless of this fact, and even give a proper posterior. Using Equation (3) (i.e. the same likelihood as in Section 3.2), we can apply Theorem 3.1 to get the posterior

$$\mathbf{X}|\mathbf{Y} = \mathbf{y} \sim \mathcal{N}_n \left( (\mathbf{Q} + \mathbf{D}^{-1})^{-1} \mathbf{D}^{-1} \mathbf{y}, (\mathbf{Q} + \mathbf{D}^{-1})^{-1} \right)$$

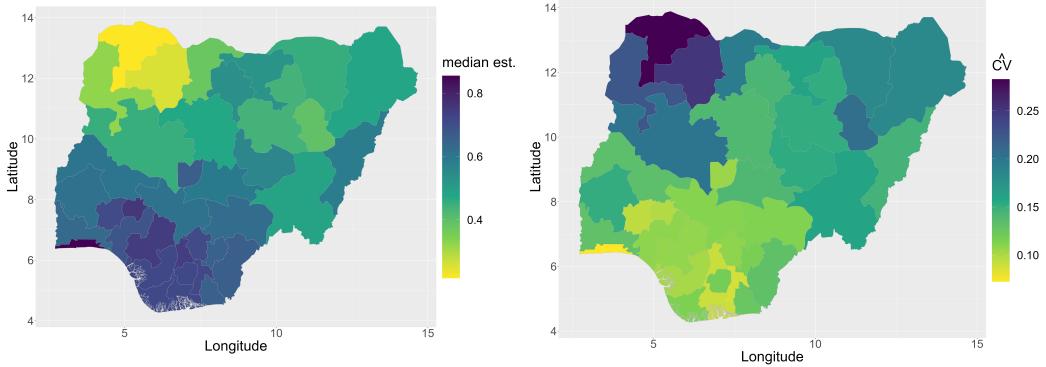
This is indeed a proper GMRF, since the precision matrix  $\mathbf{Q}_C$  given by Equation (9) has full rank.

$$\mathbf{Q}_C = \mathbf{Q} + \mathbf{D}^{-1} \quad (9)$$

Figure 10 illustrates that the estimated median of the Besag hierarchical model covers a slightly smaller range compared to the independent hierarchical model from Section 3.2, and is in general more spatially smooth. Moreover, Figure 10 shows that the estimated coefficient of variation (CV) is generally lower with the new spatial Besag model. This implies that our approach of introducing spatial bias in the model helped in reducing the variance, as anticipated.

### 3.4 Model update with data from Kaduna

This section will study opportunities for updating the Besag posterior distribution from Section 3.3 using additional data for the state of Kaduna, based on data which is independent on the data which is already used.



**Figure 10:** Estimated median and the coefficient of variation for  $P_a|Y = y$  following a hierarchical Besag model, based on 100 samples with  $\tau = 1$ .

We are working with a new independent survey in Kaduna which is much more accurate than the one giving our original data. The new survey as the proposed prior

$$Y_{38}|P_{\text{Kaduna}} \sim \mathcal{N}(\text{logit}(P_{\text{Kaduna}}), 0.1^2)$$

Note that with respect to the graph  $G$ ,  $P_{\text{Kaduna}} = P_{19}$ . Let  $\tilde{\mathbf{Y}} = (Y_1, \dots, Y_{37}, Y_{38})$ . Since the new survey and its data are independent of the original, the updated likelihood function can be written

$$f(\tilde{\mathbf{y}}|\mathbf{x}) = f(\mathbf{y}|\mathbf{x})f(y_{38}|x_{19})$$

The joint likelihood can therefore be described by defining  $\mathbf{A} \in \mathbb{R}^{38 \times 37}$

$$A = \begin{cases} 1, & i = j \vee (i = 38 \wedge j = 19) \\ 0, & \text{otherwise} \end{cases}$$

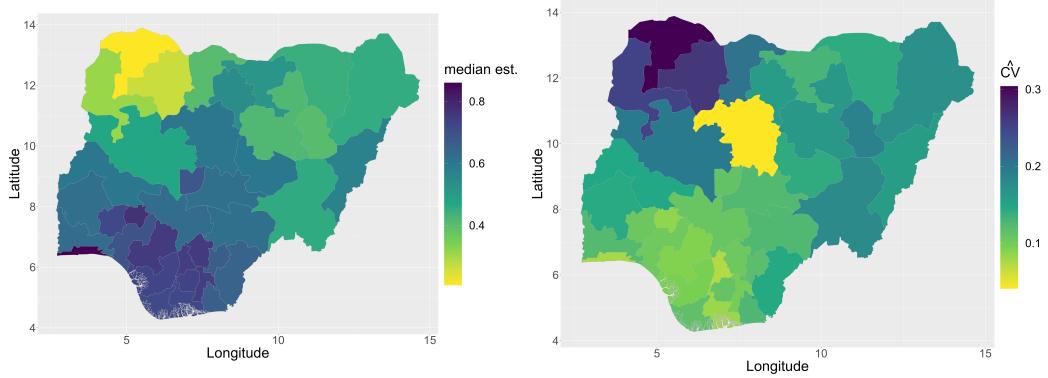
i.e. an identity matrix with an added row where all instances except the 19th are 0. Furthermore, let  $V_{38} = 0.1^2$ , and let  $\tilde{\mathbf{D}} = \text{diag}(V_1, \dots, V_{37}, V_{38})$ . The joint likelihood can then be written as

$$\tilde{\mathbf{Y}}|\mathbf{X} = \mathbf{x} \sim \mathcal{N}_{38}(\mathbf{Ax}, \tilde{\mathbf{D}}) \quad (10)$$

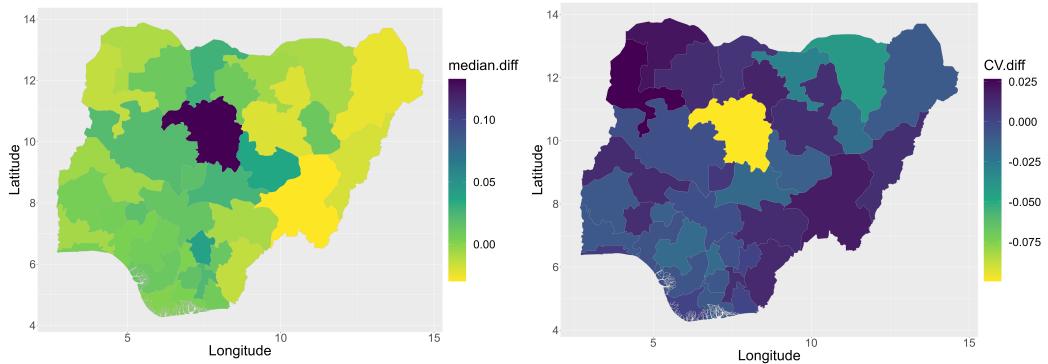
Our situation is exactly the same as for the hierarchical Besag model from Section 3.3, except that we have now introduced a linear transformation in mean through  $\mathbf{A}$ . Applying Theorem 3.1, using the intrinsic Besag prior defined in Section 3.3, and the likelihood from Equation 10, we get a posterior on the form

$$\mathbf{X}|\mathbf{Y} = \mathbf{y} \sim \mathcal{N}_{38}\left((\mathbf{Q} + \mathbf{A}^T \mathbf{D}^{-1} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{D}^{-1} \mathbf{y}, (\mathbf{Q} + \mathbf{A}^T \mathbf{D}^{-1} \mathbf{A})^{-1}\right)$$

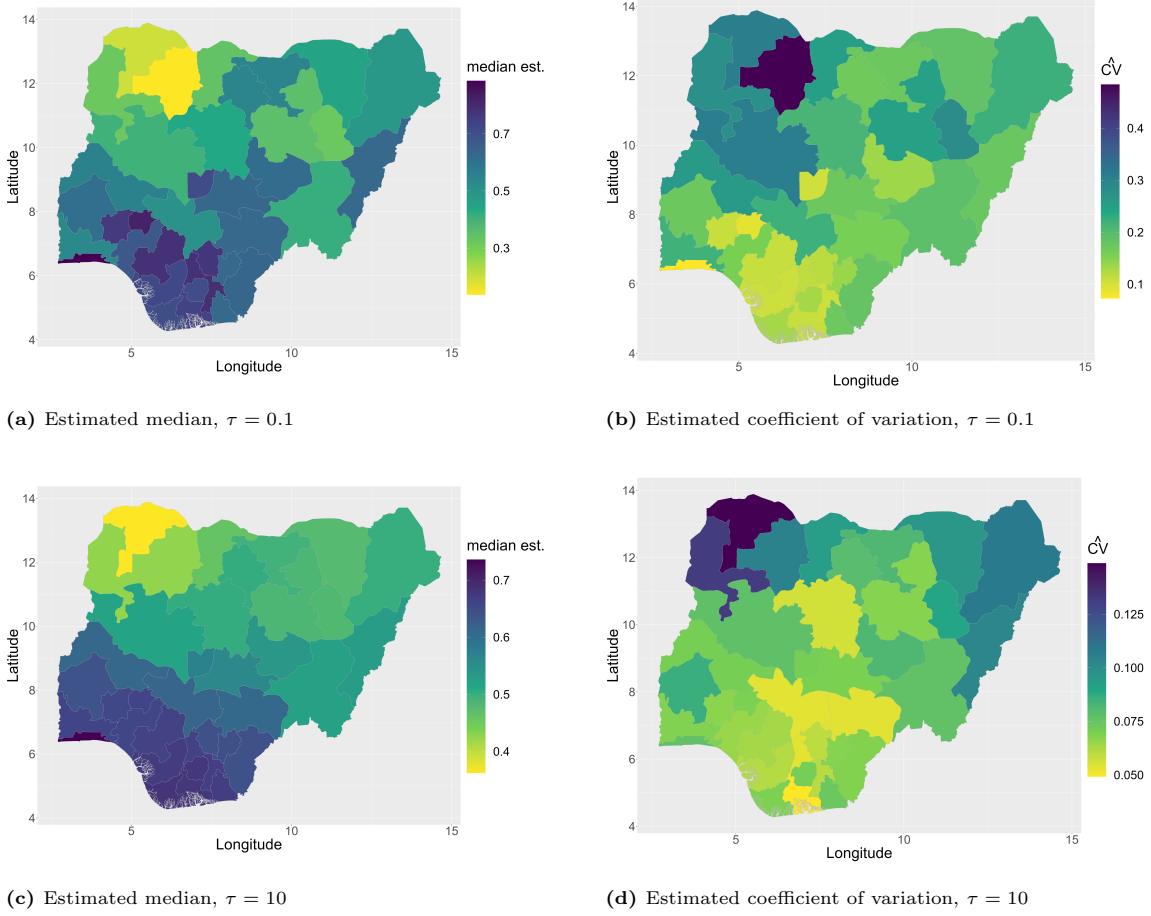
The model has been realized in Figure 11. The updated model looks very similar to the one displayed in Figure 10, except that Kaduna has a higher vaccination proportion estimate and a marginally lower variance. The updated realization shares more similarities with Figure 10 than with Figure 9. To make the differences more clear, they are plotted in Figure 12. This goes to show that model updating not only factors in desirable information about a specific variable in  $\mathbf{X}$ , but also has an updating effect on the others.



**Figure 11:** Estimated median and the coefficient of variation for  $P_a|Y = y$  following a hierarchical Besag model with additional data from a more accurate independent survey, based on 100 samples with  $\tau = 1$ .



**Figure 12:** Estimated difference between median and coefficient of variation between the updated and the original Besag hierarchical model in Section 3.3, based on 100 samples with  $\tau = 1$ .



**Figure 13:** Estimated median and coefficient of variation for  $P_a|Y = \mathbf{y}$  following a hierarchical Besag model, based on 100 samples with different values of  $\tau$ .

### 3.5 Sensitivity of the Besag hyperparameter

Figure 13 shows the Besag hierarchical spatial model for two new values of  $\tau$  ( $\tau = 0.1$  and  $\tau = 10$ ). Comparing the plots of Figure 13 with Figure 10 from Section 3.3, one can observe the following effects of changing  $\tau$ :

- Small  $\tau$  decreases the interspatial effect of each node in the graph (each state), and makes the model less spatially dependent, as can be observed in Figure 13a. Looking at the posterior distribution, this is quite clear, as  $\lim_{\tau \rightarrow 0} \mathbf{X}|Y \sim \mathcal{N}_n(\mathbf{y}, \mathbf{D})$ , which is equivalent to the case from Section 3.2 when  $\sigma^2 \rightarrow \infty$  in the posterior distribution, as seen in Equation (8). Smaller  $\tau$  also increases the variance of all estimates slightly, since it leads to more bias.
- Large  $\tau$  provides a large amount of spatial smoothing for all estimates, and makes the model more spatially dependent as can be observed in Figure 13c. Looking at the effect of a large  $\tau$  in the  $\boldsymbol{\mu}_C = \mathbf{Q}_C^{-1} \mathbf{D}^{-1} \mathbf{y}$  and  $\mathbf{Q}_C$  (defined in Equation (9)), it is clear that the spatial term

becomes dominant, neglecting the "independent" term of  $\mathbf{Q}_C$  provided by  $\mathbf{D}$ . In this case, the variance is also reduced, implying that spatial predictions gives a less variant model.

It is clear that different  $\tau$  can lead to vastly different estimates, and it is therefore very important to estimate  $\tau$  correctly.

### 3.6 MLE of Besag hyperparameter

The marginal pdf of  $y$  given  $\tau$  can be expressed as

$$f(\mathbf{y}; \tau) = \frac{f(\mathbf{y}|\mathbf{x})f(\mathbf{x}; \tau)}{f(\mathbf{x}|\mathbf{y}; \tau)} \quad (11)$$

Since there is only one observation, the log likelihood function is defined as  $l(\tau; \mathbf{y}) = \log f(\mathbf{y}; \tau)$ . Applying this to Equation (11) yields

$$l(\tau; \mathbf{y}) = \log(f(\mathbf{y}|\mathbf{x})) + \log(f(\mathbf{x}; \tau)) - \log(f(\mathbf{x}|\mathbf{y}; \tau)) \quad (12)$$

In order to evaluate  $l(\tau; \mathbf{y})$ , one can evaluate each term of Equation (12):

$$\log(f(\mathbf{x}; \tau)) = \text{Const.} + \frac{n-1}{2} \log \tau - \frac{\tau}{2} \mathbf{x}^T \mathbf{R} \mathbf{x} \quad (13)$$

$$\log(f(\mathbf{y}|\mathbf{x})) = \text{Const.} - \frac{1}{2} \log |\mathbf{D}| - \frac{1}{2} (\mathbf{y} - \mathbf{x})^T \mathbf{D}^{-1} (\mathbf{y} - \mathbf{x}) \quad (14)$$

$$\log(f(\mathbf{x}|\mathbf{y}; \tau)) = \text{Const.} + \frac{1}{2} \log |\mathbf{Q}_C| - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_C)^T \mathbf{Q}_C (\mathbf{x} - \boldsymbol{\mu}_C) \quad (15)$$

where

$$\begin{aligned} \mathbf{Q}_C &= \tau \mathbf{R} + \boldsymbol{\Sigma}^{-1} \\ \boldsymbol{\mu}_C &= \mathbf{Q}_C^{-1} \boldsymbol{\Sigma}^{-1} \mathbf{y} \end{aligned}$$

Note that in Equation (13), we used the fact that the besag pdf is proportional to  $\tau^{\frac{n-1}{2}}$ , as was shown in Equation (1). There are more parts of the pdf, which is an improper GMRF, but that can be pushed into the constant.

Substituting Equations (13), (14) and (15) into Equation (12), we get the result

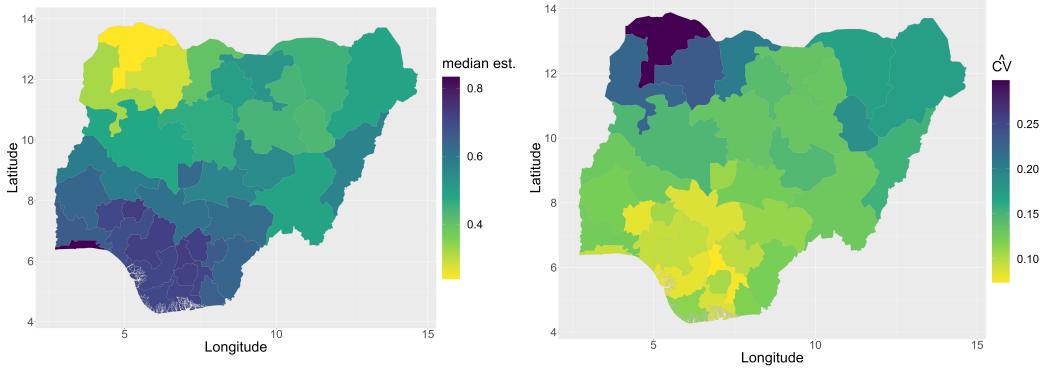
$$\begin{aligned} l(\tau; \mathbf{y}) &= \text{Const.} + \frac{37-1}{2} \log \tau - \frac{\tau}{2} \mathbf{x}^T \mathbf{R} \mathbf{x} - \frac{1}{2} (\mathbf{y} - \mathbf{x})^T \mathbf{D}^{-1} (\mathbf{y} - \mathbf{x}) \\ &\quad - \frac{1}{2} \log |\mathbf{Q}_C| + \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_C)^T \mathbf{Q}_C (\mathbf{x} - \boldsymbol{\mu}_C) \end{aligned} \quad (16)$$

The  $-\frac{1}{2} \log |\mathbf{D}|$  was in Equation (14) was also moved into the constant. In this case, it is clear from the left hand side of Equation (16) that the log-likelihood is dependent of  $\mathbf{x}$ , despite the fact that it appears in the right hand side. This does not make a difference for optimization as long as one chooses  $\mathbf{x}$  for each loop. In our case, we chose  $\mathbf{x} = \boldsymbol{\mu}_C$  for every iteration of the optimization, giving the equation

$$l(\tau; \mathbf{y}) = \text{Const.} + 18 \log \tau - \frac{\tau}{2} \boldsymbol{\mu}_C^T \mathbf{R} \boldsymbol{\mu}_C - \frac{1}{2} \log |\mathbf{Q}_C| - \frac{1}{2} (\mathbf{y} - \boldsymbol{\mu}_C)^T \mathbf{D}^{-1} (\mathbf{y} - \boldsymbol{\mu}_C) \quad (17)$$

Using the observed coverages, optimizing Equation (17) gives  $\hat{\tau} \approx 1.473$ .

The estimated median and coefficient of variation are shown in Figure 14, and live up to the discussion of Section 3.5. It is clear that the spatial effect is slightly smoother than that displayed in Figure 10, but not as extreme as the display in Figure 13c. This estimate of  $\tau$  is reasonable, as one would certainly expect some spatial effect in vaccination coverage. Two very rich states are, for example, very likely to be close to each other. Nevertheless, Many other state-specific factors are likely to affect the vaccination-coverage as well. The point of spatial statistics is, after all, to capture the spatial effect which acts as a covariate for some response, which is difficult to do with more traditional statistical methods.



**Figure 14:** Estimated median and the coefficient of variation for  $P_a|Y = y$  following a hierarchical Besag model, based on 100 samples with  $\tau = \hat{\tau} \approx 1.473$ .

## 4 Conclusion

The purpose of this project was to study regular and Bayesian Besag models and their properties, as well as compare them to non-spatial multivariate Gaussian models through application on vaccine coverage in Nigeria with sparse data. The frequentist Besag model in Section 2 showed considerably more smoothing than the frequentist Gaussian on both low- and high-resolution graphs of regions in Nigeria. Furthermore, in Section 3, use of a hierarchical Bayesian model provided not only the desired spatial smoothing, but also a generally lower variance as a consequence of increased bias through spatial correlation. In addition, model updating can be done in a straightforward and effective way using Theorem 3.1. The maximum likelihood estimate for the Besag parameter based on sparse data seemed to yield a very reasonable degree of spatial correlation. Overall, one can conclude that hierarchical Bayesian models with a Besag prior are very efficient for doing inference on variables in a spatial context.

## References

[Fuglstad, 2023] Fuglstad, G. A. (2023). Lecture notes in spatial statistics.