Master's thesis

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

Julie Røste

The Importance of Mesh Resolution When Using the SPDE Approach

Master's thesis in Applied Physics and Mathematics
Supervisor: Geir-Arne Fuglstad
December 2020

NTNU
Norwegian University of Science and Technology

Julie Røste

The Importance of Mesh Resolution When Using the SPDE Approach

Master's thesis in Applied Physics and Mathematics
Supervisor: Geir-Arne Fuglstad
December 2020

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

The objective of this work is to investigate the importance of mesh resolution for the stochastic partial differential equation (SPDE) approach by Lindgren et al. (2011). In this approach, a Gaussian Random Field (GRF) is approximated on a mesh. The resolution of this mesh plays an important role in determining the predictive power, the behaviour of parameter estimates and the fitness of the model. A higher mesh resolution gives better approximations, but at the cost of longer runtime. In addition to increasing the resolution of the mesh, it is possible to extend the mesh beyond the boundary of the domain to reduce possible boundary effects. This, however, adds more nodes to the mesh and gives longer runtime. Therefore, it is interesting to investigate this trade-off between approximation accuracy and runtime. The SPDE approach is widely used for spatial modelling, so many users will benefit from finding guidelines on how to construct the mesh.

We have performed two case studies, one with continuous data - annual precipitation in the conterminous US, and one with count data - prevalence of secondary education for women in Kenya. A Gaussian model and a Binomial model are used. The mesh is created independently of the observation locations inside a given boundary. We have varied two mesh parameters, the maximum edge length between mesh nodes (h) and the outer boundary extension (r). For each configuration, the elements of interest are computed.

Throughout this study, we find that increasing the mesh resolution through h has the strongest impact on the results, both in terms of predictive power, when the parameter estimates stabilize and for model fitness, but only up to a certain mesh resolution. Specifically, a maximum edge length of $1/12$ of the spatial range seems to be sufficient for the Gaussian case. For the Binomial case, a lower resolution is sufficient with an h of $1/4$ of the spatial range. Increasing the mesh resolution more than this will only increase the runtime. These suggestions are only guidelines on where to start when building the mesh, and thus it is important to explore meshes with both lower and higher resolutions to find the optimal mesh for a particular problem.

Sammendrag

Målet med denne oppgaven er å undersøke viktigheten av oppløsningen til triangelnettet (eng: "Mesh") for SPDE-tilnærmingen til Lindgren et al. (2011). I denne tilnærmingen approksimeres et romlig Gaussisk felt (GRF) på et triangelnett. Oppløsningen til dette nettet spiller en viktig rolle i å avgjøre den prediktive styrken, atferden til parameterestimatene, og tilpasningsevnen til dataen for SPDE-modellen. En høyere nettoppløsning gir bedre approksimasjoner, men på bekostning av høyere beregningstid. I tillegg til å øke oppløsningen til nettet så er det mulig å utvide nettet utover grensene til domenet for å redusere eventuelle grenseeffekter. Imidlertid gjør dette antallet noder i nettet øker, og gir derfor en høyere beregningstid. Det er derfor interessant å undersøke dette kompromisset mellom nøyaktigheten til approksimasjonene og beregningstiden. SPDE-modellen er mye brukt til romlig modellering, så mange brukere vil ha nytte av å finne retningslinjer for hvordan triangelnettet bør konstrueres.

Vi har gjort to case-studier, en med kontinuerlig data - logaritmisk transformert årlig nedbørsmengde over det kontinentale USA, og en med diskret data - utbredelse av videregående opplæring for kvinner i Kenya. En Gaussisk modell og en Binomisk modell blir brukt. Triangelnettet lages uavhengig av observasjonslokasjonene innenfor en gitt grense. Vi har variert to nettparametere, nemlig maksimal kantlengde mellom nettnoder (h) og ytre grenseutvidelse av nettet (r). For hver konfigurasjon beregner vi den prediktive styrken, atferden til parameterestimatene, og tilpasningsevnen til dataen.

Gjennom denne studien ser vi at økt nettoppløsningen gjennom h har størst innvirkning på resultatene, både når det gjelder prediktiv styrke og når parameterestimatene stabiliserer seg, men bare inntil en viss oppløsning. Nærmere bestemt ser en maksimal kantlengde på 1/12 av den romlige rekkevidden ρ tilstrekkelig ut for den Gaussiske dataen. For den Binomiske dataen er det tilstrekkelig med en lavere oppløsning på 1/4 av den romlige rekkevidden ρ. Å øke oppløsningen utover disse verdiene vil kun øke beregningstiden. Disse forslagene er bare retningslinjer for hvor man burde begynne når man lager triangelnett i SPDE-modellen, og det er derfor viktig å utforske nett med både finere og grovere oppløsning for å finne det optimale triangelnettet for et spesifikt problem.

Table of Contents

Chapter 1

Introduction

1.1 Introduction and Motivation

In many real-world scenarios, the goal is to predict a value at an unobserved location, based on observations at observed locations. Gaussian Random Fields (GRFs) are one of the most important modeling tools for solving such problems. A GRF assumes that the values at the observed and unobserved locations are multivariate Gaussian distributed and correlated to each other. GRFs are intuitive to work with in the sense that they are specified by a correlation matrix, which gives information on the correlation between locations directly. They are, however, computationally difficult to work with due to the need of factorizing dense matrices. When there are many observations, which often is the case, this becomes almost computationally infeasible. Lindgren et al. (2011) proposed an approach where a Stochastic Partial Differential Equation (SPDE) is used to approximate a GRF. Combining this with the Integrated Nested Laplace Approximations (INLA) approach by Rue et al. (2009) has become a popular choice for modeling spatial data. It is easy to use thanks to the open-source R-package R-INLA[1] by Lindgren et al. (2015), where these approaches are implemented. It also exists another popular R-package named TMB[2], by Kristensen et al. (2016), where the SPDE approach can be used.

In the SPDE approach, the GRF is approximated on a mesh, by a finite element (FEM) representation with Gaussian distributed weights and piece-wise linear basis functions. The mesh is a triangulation with a set of nodes and edges that form non-intersecting triangles. The weights are located on the mesh nodes and are the values of the GRF approximation. The piece-wise linear basis functions interpolate the weights so that we can obtain values at locations of interest. The mesh resolution is given by the number of triangles in the mesh, so a mesh with more triangles gives an approximation with more weights. The mesh plays a central part in the SPDE approach, and specifically, it determines both the quality of our results and the computational cost. *The quality of results* in this context means both the predictive power of the model, how well the model fits the data and how parameter estimates behave with models with varying mesh resolutions. Pre-

[1] Available at `www.r-inla.org`

[2] Available at `http://tmb-project.org/`

dictions made by the SPDE model can be compared with the observed values, by holding out some observations in the estimation process, and then predict on the locations of the held out observations. In this work, we use a 10-fold cross-validation as well as hold-out regions for different mesh resolution. To quantify the predictive power, we use the scoring rules root-mean-square-error(RMSE) and the continuous ranked probability score (CRPS). RMSE measures point-wise errors between a prediction and the observed value at a location, while CRPS measures how well a predicted distribution fits with the observed value. Better predictions mean lower values of the scoring rules.

Our interest in the mesh has both a practical and a theoretical aspect. The practical aspect comes from the fact that there is no automacy in creating this mesh. Therefore, users of the SPDE approach have to consider carefully how to construct the mesh. Most importantly, which resolution and how large buffer region should the mesh have? There are many users of the SPDE approach, as highlighted by Bakka et al. (2018), but there are few guidelines for how to construct the mesh. Thus coming up with general recommendations for mesh resolution will be of great use for many. Generally, high resolutions with large buffer regions yield good predictions that are almost unaffected by boundary effects. On the other hand, there is a computational complexity involved when working with spatial models. The SPDE approach makes modeling more computationally efficient than using only a GRF, but it still has a computational complexity worth attention, which is mainly determined by the mesh resolution. Therefore, users of the SPDE approach have to balance between accurate estimates with long running time and less accurate estimates with shorter running time. The purpose of this work is to gain insight in how to determine this balance.

A more theoretical aspect revolves around the fact that we want to approximate an exact GRF. In this work, we use Bayesian hierarchical models formulated at three levels, (1) an observation model conditioned on a latent field and parameters, (2) a latent field that includes fixed covariates and the GRF approximation, and (3) a prior for the parameters range ρ, marginal variance σ_{s}^2 and the nugget effect σ_{N}^2. When a mesh has too low resolution relative to the range, the GRF approximation might deviate strongly from the desired Matérn structure, as shown by Fuglstad and Beguin (2018). In particular, the marginal variance will vary over the domain instead of being constant. This effect will be explained more in Section 2.2.2. When we work with a finite element method that has piece-wise linear basis functions, there is a limitation on how precisely the GRF can be approximated, and we get a discretization error which will be captured by the nugget effect. This discretization error is variability that cannot be resolved on a mesh with piece-wise linear basis functions, i.e., a subgrid variation. The nugget effect contains a combination of measurement error and a small scale variability as well, where the latter is variation on a smaller scale than we observe. For Gaussian observation models, the nugget effect can be seen as the variance in the Gaussian likelihood. For non-Gaussian observation models on count data, for example, when the model has a Binomial likelihood, the nugget effect will contain an observation noise as well. This observation noise comes from the fact that the true underlying field is continuous, but we observe discrete outcomes that can only take a few values on the continuous scale. Since the non-Gaussian observation model on top makes us unable to observe the latent field directly, we have less information about the latent field, and it is less obvious how the mesh resolution affects the results.

From the SPDE approach, we will obtain both predictions based on some observed

data, as well as parameter estimates. The SPDE model is an approximation of the GRF, so the parameter estimates, i.e., the range $\hat{\rho}$ and marginal variance $\hat{\sigma}_s^2$, cannot be interpreted exactly as the parameters of the Matérn covariance function in the GRF. However, the purpose of spatial modeling is in most cases to predict, in which case we try to find a well-suited model to make as good predictions as possible. It should therefore be noted that even though parameter estimates cannot be interpreted exactly as the true GRF parameters, this is fine, since the goal is to obtain the best possible predictions, and not to approximate the GRF itself as accurately as possible.

We believe that the mesh in the SPDE approach has a different impact for Gaussian and non-Gaussian models. We will therefore use two datasets, one with continuous responses on which we will use a Gaussian likelihood, and one with count data using a Binomial likelihood. Note that our interest lies not specifically in the results of these datasets, but in which conclusions and recommendations we generally can draw from modeling on these types of datasets.

We will use R-INLA as data processing and modeling tool, since the SPDE approach is implemented there. The git-book "Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA" by Krainski et al. (2018) has been a major inspiration for theory development and the code implementation. In this book, there is a lot of information about the mesh, especially on how it is constructed both theoretically and practically in R-INLA. There are however no clear guidelines on which resolution to use.

In the next subsection, we briefly present some relevant previous work. Then, in the next sections, we present the datasets to be used in the case studies. Chapter 2 consists of theory behind the SPDE approach, spatial modeling with Bayesian hierarchical models and finally a section with model assessment and prediction scores. In the third chapter, the case studies are presented. This includes both the aim and design of the studies, specific details for the two models and the results. A discussion of the obtained results follows in Chapter 4, with some recommendations based on the results, and a conclusion is given in Chapter 5.

1.2 Previous Work on the Importance of Mesh Resolution

In the article by Righetto et al. (2020), "On the choice of mesh for the analysis of geo-statistical data using R-inla", they study a similar problem. In particular, they construct different meshes with the SPDE approach and measure the impact of the mesh on inference and prediction. They have used the locations of the observations to create the mesh, i.e., the locations are some or all nodes in the mesh. They then investigate different configurations, specified by a parameter c (`cutoff` in R-INLA) which denotes the shortest allowed distance between mesh nodes, and h, the maximum allowed distance edge length. They find that the optimal value of c depends on the number of observations, and must therefore be determined in accordance with the observation size. A change in this parameter has a greater effect than a change in the maximum allowed edge length between nodes.

We, on the other hand, are interested in creating a mesh without knowing the spatial distribution of the locations. The mesh will only be constructed based on the domain of the locations, so that the mesh does not adapt to the spatial design. We then do not need the

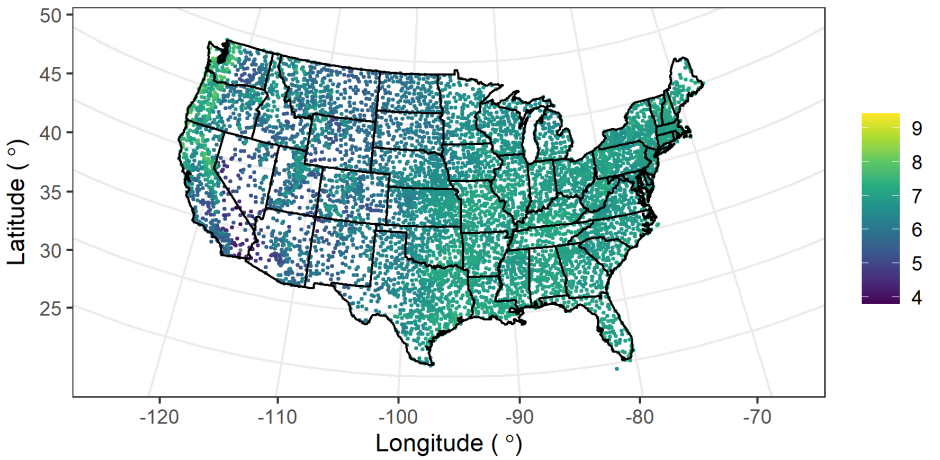


Figure 1.1: Map over conterminous US with annual log precipitation (mm / year) at 7040 stations.

cutoff parameter c, since this parameter only applies if one uses the observation locations as mesh nodes, and observations are so close that we want a mesh node at only one of the locations. We are mainly concerned with the mesh resolution, and when we make the mesh independent of the observation locations, the resolution is controlled by h. This approach makes it possible to use the same mesh when doing for instance k-fold cross-validation or hold-out region validation.

1.3 Case Study 1 - Precipitation in the conterminous US

In this work, we use a dataset of the conterminous US (the USA except Alaska and Hawaii), with monthly total precipitation at different measurement locations. The dataset is a binary file called `RData.USmonthlyMet.bin` and can be found at `https://www.image.ucar.edu/public/Data/`. A description of the dataset can be found at `https://www.image.ucar.edu/Data/US.monthly.met/USmonthlyMet.shtml`. Note that the total monthly precipitation unit is in centimeters and not in millimeters as in the description.

Specifically, we use annual log precipitation for the year 1981 by summing up the monthly precipitation at each location and taking the logarithm, inspired by Fuglstad et al. (2015). This dataset was chosen due to its many observations, the simplicity of preprocessing and that a Gaussian likelihood can be used in the fitting process.

Only measurement stations without missing data are included in the dataset, which gives a total of 7040 stations. Figure 1.1 shows the log-precipitation at each of the locations. The Albers projection (Snyder, 1987) with $lat_0 = 39$ and $lat_1 = 45$ is used in the visualization. The dataset also contains the elevation (in meters) at the stations, which will be an explanatory variable in this analysis. We will use the unit kilometers for elevation.

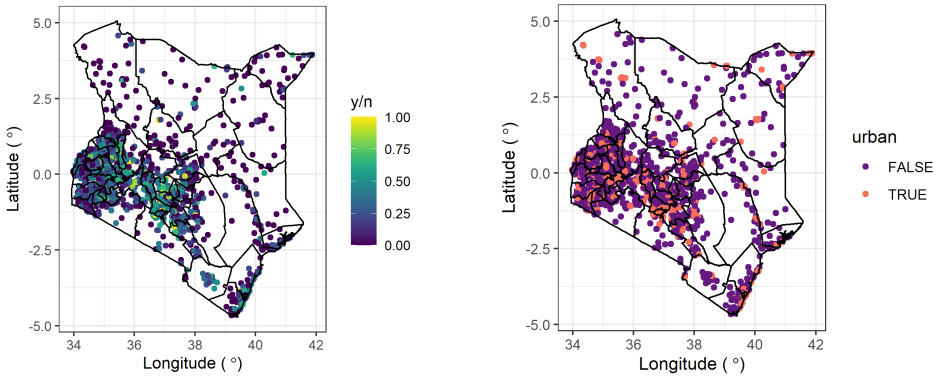
1.4 Case Study 2 - Prevalence of Secondary Education for Women in Kenya

In the second case study, we want to use a dataset with count data that can be modelled using a Binomial likelihood. Specifically, we use a dataset from the Demographic and Health Services (DHS) Program, "Kenya: Standard DHS, 2014 Dataset". This survey covers many responses, like education, mortality, diseases and general health, for children and adults in Kenya. The structure of the survey is built upon interviewing people that belong to a certain cluster. These clusters contain a number of households in for example a village or a city, and are selected based on a survey design. More information about the dataset and the DHS Program can be found in Kenya National Bureau of Statistics et al. (2015) and at their website `https://dhsprogram.com/`.

We choose to look at the proportion of women in Kenya between the ages 20-29 that have completed their secondary education, with inspiration from Paige et al. (2019). In this dataset, there are 1580 clusters, with GPS locations for the clusters given in longitude-latitude format. In a cluster at location $\mathbf{s}_i$, n_i women are interviewed and y_i women have completed their secondary education. In total, 11290 women are interviewed, where 3268 answered that they have completed their secondary education, which is a proportion of about 30%. In Figure 1.2, the proportion of women with secondary education is shown for the cluster locations. A cluster is also assigned a categorical label, rural or urban, which is a fixed covariate that we use in the model. A likelihood needs to be determined for the model. For this, we make the following assumptions:

- There are two outcomes for each woman in the cluster, i.e., has completed or has not completed secondary education
- Each woman in the cluster has the same probability p of completing secondary education
- The completion of secondary education for one woman is independent of the other women's outcomes.

In practice, for a survey like this, there might be violations of these statements. We assume, however, that they are adequate, and that we therefore can use a Binomial likelihood.



(a) **(b)**

Figure 1.2: (a) Proportion of women that have completed secondary education, y_i/n_i, on cluster locations $\mathbf{s}_i$ for all $i = 1, \ldots, 1580$ clusters. (b) Urbanicity in clusters.

Chapter 2

Background

In this chapter, we begin with a short introduction to Gaussian Random Fields (GRFs) and Gaussian Markov Random Fields (GMRFs). We then explain the theory behind the SPDE approach and its elements, among other the mesh and prior specifications. We further introduce Latent Gaussian Models (LGMs) and how to perform inference computationally efficient using INLA. Finally, the model assessment criteria WAIC and scoring methods RMSE and CRPS will be given.

2.1 GRFs and GMRFs

Data in a spatial domain $\mathcal{D}$ often have the property that the closer two observations are in space, the higher is the similarity between them. When making a suitable model to fit the data, this spatial dependency should be taken into account, hence a spatial field is required. For point-referenced data, i.e., data with observations that are observed at some given locations on our map, we commonly use a continuously indexed random field.

A Gaussian Random Field (GRF) $\{u(\mathbf{s}) : \mathbf{s} \in \mathcal{D} \subset \mathbb{R}^d\}$ is a spatial process with random variables occurring in a fixed, continuous domain $\mathcal{D}$, where every collection with a finite number of these variables follows a multivariate Gaussian distribution. Let $\{\mathbf{s}_i\}_{i=1,\dots,n}$ be the locations of the observed data. The spatial effects $u(\mathbf{s}_1), \dots, u(\mathbf{s}_n)$ are then realizations of the spatial process such that $u(\mathbf{s}_1), \dots, u(\mathbf{s}_n)$ can be modelled by a multivariate Gaussian distribution with a mean $\boldsymbol{\mu} = [\mu_1, \dots, \mu_n] \in \mathbb{R}^n$ and a covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ with entries $\boldsymbol{\Sigma}_{i,j} = c\big(u(\mathbf{s}_i), u(\mathbf{s}_j)\big)$ for a given positive definite covariance function, $c(\cdot, \cdot)$.

We use spatial covariance functions from known families to ensure that they are positive definite. In this work we use the Matérn covariance function, as defined in Definition 2.1.

Definition 2.1 (Matérn covariance function)**.** *The Matérn covariance function is given by*

$$C_\nu(d) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \Big(\sqrt{8\nu}\frac{d}{\rho}\Big)^\nu K_\nu\Big(\sqrt{8\nu}\frac{d}{\rho}\Big), \quad d \in \mathbb{R}_\oplus, \tag{2.1}$$

where d is the distance between two locations, $\nu > 0$ is a smoothness parameter, ρ is a range parameter, σ^2 is the marginal variance and $K_\nu(\cdot)$ is the modified Bessel function of the second kind with order ν.

Hence, the covariance matrix for the GRF at locations $s_1, \ldots, s_n$ has entries $\mathbf{\Sigma}_{i,j} = C_\nu(|\mathbf{s}_i - \mathbf{s}_j|)$. With the Matérn covariance, the spatial process $u(\mathbf{s})$ is often stationary and isotropic. The spatial process is second-order stationary if it has a constant mean μ for every location $\mathbf{s}_i$ and the spatial covariance function only depends on the distance vector between two locations, $(\mathbf{s}_i - \mathbf{s}_j)$. If, in addition, the covariance function only depends on the distance (Euclidean) and not the direction between the locations, the process is isotropic.

GRFs are intuitive to work with to capture spatial dependency as they have a defined expectation and correlation, and a simple covariance structure. They have good analytical properties as only linear algebra is needed for computing the conditional distributions at unobserved locations, based on a set of observations. On the other hand, GRFs have computational challenges. In spatial statistics, it is normal to have many observations, such as on the order of 10 000. To calculate the likelihood of a GRF, we need to factorize the $n \times n$-matrix $\mathbf{\Sigma}$. This has a computational cost of $\mathcal{O}(n^3)$ because $\mathbf{\Sigma}$ is generally a dense matrix. With this amount of observations, it becomes practically impossible to work with the GRF. This is often referred to as the "big n problem". Gaussian Markov Random Fields (GMRFs) can become a solution to this challenge (Rue and Held, 2005).

A GMRF is a GRF where observations are assumed to be conditionally dependent on for example first and second-order neighbours. Conditional independence is assumed for higher-order neighbours, which gives Markov properties to the GMRF. The conditional independence information is "hidden" in the covariance matrix, but stated explicitly in the precision (inverse covariance) matrix $\mathbf{Q}$, as in the following theorem.

Theorem 2.1 ((Rue and Held, 2005)). *Let $\mathbf{x} = (x_1, \ldots, x_n)^\top$ follow a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and precision matrix $\mathbf{Q} > 0$. Then for $i \neq j$,*

$$x_i \perp x_j \mid \mathbf{x}_{-\{i,j\}} \iff Q_{ij} = 0. \tag{2.2}$$

The notation $\mathbf{x}_{-\{i,j\}}$ means all elements in $\mathbf{x}$ that are not x_i and x_j, i.e. $(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n)$. Thus, if two variables x_i and x_j are conditionally independent given all other elements, the precision matrix for $\mathbf{x}$ has value 0 for $\mathbf{Q}_{ij} = \mathbf{Q}_{ji}$. This yields a sparse precision matrix if we assume many variables being conditionally independent. Now let $\mathcal{G}$ be a labelled graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with a set of nodes, $\mathcal{V} = \{1, \ldots, n\}$, and a set of edges $\mathcal{E}$, where it only exists edges between nodes that are conditionally dependent. Thus the graph $\mathcal{G}$ and the precision matrix $\mathbf{Q}$ both can give information on the conditional independence.

The formal definition of a GMRF is given in Definition 2.2.

Definition 2.2 (Gaussian Markov Random Field (Rue and Held, 2005)). *A random vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)^\top \in \mathbb{R}^n$ is called a Gaussian Markov Random Field (GMRF) with respect to a labelled graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with mean $\boldsymbol{\mu}$ and precision matrix $\mathbf{Q} > 0$ if and only if its density has the form*

$$\pi(\mathbf{x}) = (2\pi)^{-n/2} |\mathbf{Q}|^{1/2} \exp\Big(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{Q} (\mathbf{x} - \boldsymbol{\mu}) \Big), \quad \mathbf{x} \in \mathbb{R}^n, \tag{2.3}$$

and

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

Now, when working with a precision matrix that is sparse, the factorization of $\mathbf{Q}$ can be performed using sparse matrix algorithms, which results in a lower computational cost than for a GRF. In particular, a temporal sparsity structure with dimension $\mathbb{R}$ in gives a computational cost of about $\mathcal{O}(n)$, a spatial structure with dimension $\mathbb{R}^2$, gives about $\mathcal{O}(n^{3/2})$ and a spatiotemporal structure gives about $\mathcal{O}(n^2)$. (Rue and Held, 2005)

2.2 The Stochastic Partial Differential Equations Approach

Using GMRFs alone has practical limitations for irregular observation locations, so Lindgren et al. (2011) proposed an approach where a Stochastic Partial Differential Equation (SPDE) is used to create a link between GRFs and GMRFs. In this way, the continuously indexed GRF properties are kept and the computational efficiency of GMRFs is gained.

2.2.1 Discretizing the Random Field

In the SPDE approach, a GMRF represents a GRF with Matérn covariance structure in such a way that an SPDE has the GRF as a stationary solution. This SPDE is defined in Definition 2.3.

Definition 2.3. *(Stochastic partial differential equation) A GRF with the Matérn covariance function is a stationary solution to the SPDE*

$$\left[\kappa^2 - \Delta\right]^{\alpha/2} \tau\, u(\mathbf{s}) = \mathcal{W}(\mathbf{s}), \quad \mathbf{s} = [s_1, s_2, \ldots, s_d] \subset \mathbb{R}^d, \tag{2.4}$$

with a smoothness parameter $\nu = \alpha - d/2$ and scaling parameters $\kappa > 0$ and $\tau > 0$. These parameters are related to the parameters of the Matérn covariance function by a range $\rho = \sqrt{8\nu}/\kappa$ and marginal variance $\sigma_s^2 = \frac{1}{\tau^2\kappa^{2\nu}} \frac{\Gamma(\nu)}{\Gamma(\nu+1/2)(4\pi)^{d/2}}$. $\mathcal{W}(\mathbf{s})$ denotes a spatial Gaussian white noise process with unit variance. $\Delta = \sum_{i=1}^{d} \frac{\partial^2}{\partial s_i^2}$ is the Laplacian.

This SPDE is solved by a stochastic weak approach on a limited domain $\mathcal{D}$, as derived in the author's project thesis, Røste (2020). Here, the dimension was 1D, but the procedure is the same for 2D. We use the Neumann boundary conditions, i.e., the normal derivative at the boundary is zero. We use $\alpha = 2$ so that the smoothness $\nu = 1$. We divide the domain $\mathcal{D}$ into a mesh consisting of non-intersecting triangles with m nodes and edges between these nodes. The solution to the SPDE, $u(\mathbf{s})$, can be approximated by a finite element representation (FEM),

$$u(\mathbf{s}) = \sum_{k=1}^{m} \psi_k(\mathbf{s}) w_k, \quad \mathbf{s} \in \mathcal{D}, \tag{2.5}$$

where w_k are Gaussian-distributed weights located at the mesh nodes and $\{\psi_k(\mathbf{s})\}_{k=1,\ldots,m}$ are piece-wise linear basis functions that are used interpolate the weights for any location $\mathbf{s}$.

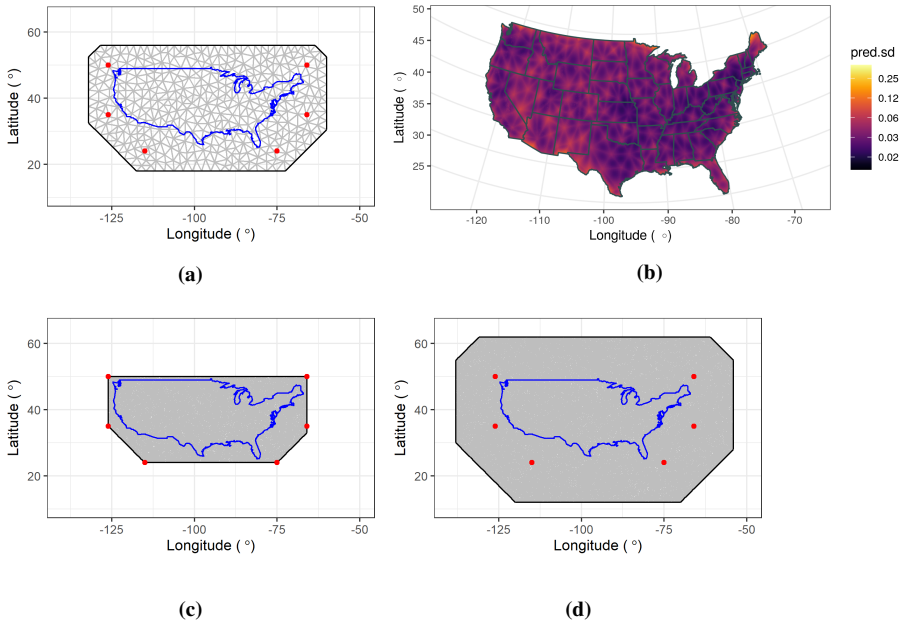


Figure 2.1: (a) Low-resolution mesh over conterminous US with the boundary with parameters $h = 4$, $r = 10\%$. (b) Predicted standard deviation on 400×200 grid using an SPDE model with mesh from (a). (c) Mesh with parameters $h = 1$, $r = 0\%$. (d) Mesh with parameters $h = 1$, $r = 20\%$.

The resulting derivation when solving the SPDE, yields the precision matrix

$$\mathbf{Q} = \tau^2\left(\kappa^4\mathbf{C} + 2\kappa^2\mathbf{G} + \mathbf{G}\mathbf{C}^{-1}\mathbf{G}\right), \tag{2.6}$$

where $\mathbf{C}_{i,j} = \langle \psi_i, 1 \rangle$ and $\mathbf{G}_{i,j} = \langle \nabla\psi_i, \nabla\psi_j \rangle$. The parameters τ and κ are as in Definition 2.3, and can be interpreted as the parameters of the Matérn covariance function, range ρ and marginal variance σ_s^2, for sufficiently fine meshes and small boundary effects. Note that we use these parameter symbols for all mesh resolutions, even though they cannot be interpreted as the Matérn parameters for low resolutions. The Gaussian-distributed weights, $\mathbf{w} = [w_1, \ldots, w_m]^\top$ will then follow the joint distribution $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}^{-1}(\tau, \kappa))$ with $\mathbf{Q}$ from (2.6).

The choice of mesh plays an important role in the resulting SPDE model, and will therefore be discussed more thoroughly in Section 2.2.2.

2.2.2 Mesh

The mesh in the SPDE approach is a collection of nodes inside a boundary with edges between the nodes, creating triangles. It must cover the study region, i.e., where we are going to observe data. For example, inside the boundary of a country. A mesh is described by two elements in addition to the boundary region. The first element is the mesh resolution, that is, how close the mesh nodes are. This can be controlled by introducing a maximum

allowed edge length between the nodes, which will be denoted h. The second element is the region that the mesh covers. To avoid boundary effects, it is common to add an outer extension to the domain boundary. This can be set by a relative factor, r, that determines how large the extension should be compared to the domain size. There are other ways to define the mesh as well, but we will in this work focus on h and r.

When the spatial domain has two dimensions, every location $\mathbf{s} \in \mathcal{D} \subset \mathbb{R}^2$ will be either inside a triangle, at an edge between two mesh nodes or at one mesh node. The piece-wise linear basis functions $\boldsymbol{\psi}(\mathbf{s}) = \psi_1(\mathbf{s}), \ldots, \psi_m(\mathbf{s})$ are centered at the mesh nodes, and are constructed such that for a given location $\mathbf{s}$, only the three mesh nodes defining the triangle can contribute to the field value $u(\mathbf{s})$. They have value 1 at node k and 0 at the other nodes.

When a location $\mathbf{s}$ is inside a triangle, three inner triangles are formed when drawing an edge from $\mathbf{s}$ to each of the three mesh nodes. We denote the area of these three triangles to be T_1, T_2 and T_3, and the total area of all three triangles is T. The basis function values between the mesh nodes are computed as proportions of the total area of the triangle T, such that $u(\mathbf{s}) = w_{k-1}\frac{T_1}{T} + w_k\frac{T_2}{T} + w_{k+1}\frac{T_3}{T}$. Note that the value connected to w_k is made by the inner triangle on the opposite side, such that if $\mathbf{s}$ is located near w_k, the opposite triangle is large, and the contribution from this mesh node to $u(\mathbf{s})$ is large.

When we have a finite number of observations, n, we can write the approximation in (2.5) on matrix form

$$\mathbf{u} = \mathbf{A}\mathbf{w}, \tag{2.7}$$

where $\mathbf{u}$ is an $n \times 1$ vector with the approximated GRF values and $\mathbf{w}$ is the $m \times 1$ vector with Gaussian distributed weights. $\mathbf{A}$ is an $n \times m$-matrix consisting of the values of the m the basis functions for all n observations. Each row in $\mathbf{A}$ consists of maximum three non-zero elements, which together sum up to 1. Thus observing $u(\mathbf{s})$ is to observe linear combinations of the discrete underlying representation.

The number of triangles in the mesh will then be determined by how high the mesh resolution is, and how large the extension is relative to the domain. The number of triangles m affects the computational cost of the SPDE approach. In particular, the number of triangles determines the number of basis functions in (2.5), and thus, the size of the precision matrix $\mathbf{Q}$. It has a sparse structure, which yields a computational cost of $\mathcal{O}(m^{3/2})$ for the SPDE approach in 2D. (Rue and Held, 2005). Increasing the mesh resolution by halving the maximum edge length h will increase the number of triangles by a factor of four, which again will result in an increased computational cost of a factor of 8.

To get a better understanding of the properties and impact of the mesh, we will now discuss the impact of mesh resolution for the variance of the GRF, with a numerical example. In Figure 2.1, three meshes with different configurations are shown in (a), (b) and (d). Note that the same configurations will be used in the analysis in Chapter 3. The predicted standard deviation on a regular 400×200 grid is shown in (b). The SPDE model with the mesh shown in (a) is used. The colorbar to the right is shown on $\log_2$-scale, and we clearly see that there is an underlying structure here which we do not expect to come from the spatial model. At the mesh node locations and on the edges between these, the predicted standard deviation is higher (typically 0.05-0.1), while in the middle of the triangles, the value is lower (typically around 0.025).

This phenomenon can be explained by the following. The prediction $\hat{u}_i = a_{i,1}\hat{w}_{i,1} + a_{i,2}\hat{w}_{i,2} + a_{i,3}\hat{w}_{i,3}$ on a location $\mathbf{s}_i$ is an interpolation of the values at the three closest node

Table 2.1: Matérn correlation $\mathcal{C}$ and variances for three locations of $\mathbf{s}_i$ are shown for different maximum allowed edge lengths h.

h	$\mathcal{C}(h)$	$\mathrm{V}_{a_1=a_2=a_3=1/3}$	$\mathrm{V}_{a_1=a_2=0.5,a_3=0}$	$\mathrm{V}_{a_1=1,a_2=a_3=0}$
2	0.01	0.34	0.51	1.00
1	0.14	0.43	0.57	1.00
0.5	0.44	0.63	0.72	1.00
0.25	0.73	0.82	0.87	1.00
0.125	0.89	0.93	0.95	1.00
0.0625	0.96	0.98	0.98	1.00

locations in the mesh, $\hat{w}_{i,1}$, $\hat{w}_{i,2}$ and $\hat{w}_{i,3}$, with factors $a_{i,1}$, $a_{i,2}$ and $a_{i,3}$, . These factors sum up to 1 and are the elements of the piece-wise linear basis functions. They define how much the three node weights contribute to the predicted value $\hat{u}_i$. For a location $\mathbf{s}_i$ inside a triangle, all three factors are greater than 0. At an edge between two mesh nodes, one factor is zero, and at a node location, one factor equals 1 and the two others are zero. For simplicity we now let a_1, a_2, a_3 denote the factors for a location $\mathbf{s}_i$, w_1, w_2, w_3 denote the mesh node weights, and we let the these weights have the same variance so that $\mathrm{Var}(w_1) = \mathrm{Var}(w_2) = \mathrm{Var}(w_3) = \sigma^2$. The variance of $\hat{u}_i$ is given by

$$\begin{aligned}
\mathrm{Var}(\hat{u}_i) &= \mathrm{Var}(a_1 w_1 + a_2 w_2 + a_3 w_3), \quad 0 \leq a_1, a_2, a_3 \leq 1 \\
&= \sigma^2 \big(a_1^2 + a_2^2 + a_3^2 + \\
&\quad 2a_1 a_2 \mathrm{Corr}(w_1, w_2) + 2a_1 a_3 \mathrm{Corr}(w_1, w_3) + \\
&\quad 2a_2 a_3 \mathrm{Corr}(w_2, w_3) \big) \\
&\leq \sigma^2 (a_1 + a_2 + a_3)^2, \quad \text{since } -1 \leq \mathrm{Corr}(\cdot, \cdot) \leq 1 \\
&= \sigma^2, \quad \text{since } a_1 + a_2 + a_3 = 1.
\end{aligned} \tag{2.8}$$

From this we can see that the variance of $\hat{u}_i$ will be small when the prediction location $\mathbf{s}_i$ is in the middle of the mesh triangle, i.e., $a_1 \approx a_2 \approx a_3$, because $a_j^2 \leq a_j$ on the interval [0,1]. The variance is larger towards an edge and largest when a prediction location is on a mesh node. Therefore it is natural that we observe a lower standard deviation inside triangles and a higher on the edge and on the node locations. If the correlations between mesh nodes are low, i.e., a low mesh resolution, this difference in variance is even larger.

To see how this changes when varying the mesh resolution, we construct the following example. Let w_1, w_2, w_3 be nodes in an equilateral triangle, with side length h. Reducing this side length h corresponds to increasing the mesh resolution. Let the variance at the nodes be $\sigma^2 = 1$. The correlation between these mesh nodes are given by a Matérn correlation function with parameters range $\rho = 1$ and smoothness $\nu = 1$. In Table 2.1, the correlation $\mathcal{C}(h)$ between two mesh nodes with internal distance h is given for h = [2, 1, 0.5, 0.25, 0.125, 0.0625]. For coarse meshes, this correlation is low. Let $\mathbf{s}_i$ be inside the triangle formed by w_1, w_2, w_3. Then $\hat{u}_i = a_1 w_1 + a_2 w_2 + a_3 w_3$. We define three cases. In the first case, the location $\mathbf{s}_i$ is in the middle of the triangle, yielding $a_1 = a_2 = a_3 = 1/3$, i.e., all three mesh nodes contributes equally to the value of $\hat{u}_i$. The variance for $\hat{u}_i$ is computed as in Equation (2.8) for different side lengths h. In the second case, the location $\mathbf{s}_i$ is at an edge so that $a_1 = a_2 = 0.5$ and $a_3 = 0$, and in the third case, the location is at

a mesh node.

The resulting variances for different locations $\mathbf{s}_i$ and increasing mesh resolution are shown in Table 2.1. Here, we clearly see that for a high value of h, i.e., low mesh resolution, the variance is substantially lower in the center of the triangles compared to on the edges and nodes. This is in accordance with Figure 2.1, which also has a coarse mesh given by $h = 4$. In addition, the difference in variances between locations in the middle of the triangle and locations close to the mesh node, shrinks with higher mesh resolution, with a change from [0.34, 0.51, 1] for $h = 2$ and up to [0.98, 0.98, 1] for $h = 0.0625$. Thus the observed variance will be less influenced by the mesh itself when using a higher mesh resolution in the SPDE approach.

2.3 Spatial Modeling with Bayesian Hierarchical Models

To model spatial data, Bayesian Hierarchical Models are often used. These are defined the following way: For observed data $\mathbf{y} = (y_1, \dots, y_n)$ given unknown parameters $\boldsymbol{\theta}$, we have a probability distribution $\pi(\mathbf{y} \mid \boldsymbol{\theta})$, called the likelihood. The unknown parameters also have an associated probability specified as the prior distribution $\pi(\boldsymbol{\theta} \mid \boldsymbol{\tau})$ given some hyperparameters τ, which again can have a hyperprior distribution, or be fixed. For the following explanation, we let it be fixed, and therefore disregard τ. The prior distribution $\pi(\boldsymbol{\theta})$ is unrelated to the observed data, and says something about our prior beliefs for $\boldsymbol{\theta}$.

In this work, we will use a class of Bayesian models called Latent Gaussian Models, which now will be introduced.

2.3.1 Latent Gaussian Models

Latent Gaussian Models (LGMs) are a class of Bayesian models that consist of three elements: A likelihood $\pi(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta})$, a latent Gaussian field $\mathbf{x}$ and a vector of parameters $\boldsymbol{\theta}$. The data are assumed to be conditionally independent given the latent field, and (often) the mean of y_i is linked to a Gaussian linear predictor η_i through a link function $g(\cdot)$. Generally, the linear predictor can be written as

$$\eta_i = \mu + \sum_{j=1}^{p} \beta_j z_{i,j} + \sum_{k=1}^{K} f_k(v_{i,k}) + \epsilon_i, \tag{2.9}$$

where μ is the intercept, $\mathbf{z}$ are covariates with linear coefficients $\boldsymbol{\beta}$, and $\{\mathbf{f}_k(\cdot) = [f_k(v_{1,k}), \dots, f_k(v_{i,k}), \dots, f_k(v_{n,k})]^\top\}_{k=1,\dots,K}$ are functions on covariates $\mathbf{v}$, which, among others, can be random iid, temporal, spatial or spatio-temporal effects. The latent field consists of the elements in the linear predictor and the linear predictor itself, $\mathbf{x} = \{\boldsymbol{\eta}, \mu, \boldsymbol{\beta}, \mathbf{f}_1(\cdot), \mathbf{f}_2(\cdot), \dots\}$. This latent field is assumed to be a GMRF with zero mean and precision (inverse covariance) matrix $\mathbf{Q}$.

We can then write the latent Gaussian model as

$$\begin{aligned} \mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}_1 &\sim \prod_i \pi(y_i \mid \eta_i, \boldsymbol{\theta}_1), \\ \mathbf{x} \mid \boldsymbol{\theta}_2 &\sim \mathcal{N}(\mathbf{0}, \mathbf{Q}^{-1}(\boldsymbol{\theta}_2)), \\ \boldsymbol{\theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2] &\sim \pi(\boldsymbol{\theta}). \end{aligned} \tag{2.10}$$

With the SPDE approach in mind, we will for the functions $\mathbf{f}_k$ insert the spatial field $u(\mathbf{s})$ represented as in Equation (2.5). Furthermore, the parameter vector $\boldsymbol{\theta}$ will be given by the range ρ and the marginal variance σ_{s}^2 from the GMRF, as well as a nugget variance σ_N^2. The priors for these must now be specified.

2.3.2 Priors and their distributions

We need priors for the parameters range ρ and marginal standard deviation σ_{s} in the Matérn covariance function. In addition we want priors on the nugget effect σ_{N}. In this work, we will use the Penalized Complexity (PC) priors by Fuglstad et al. (2019). The joint PC prior that corresponds to a base model with infinite range and zero variance is

$$\begin{aligned}\pi(\rho, \sigma_{\mathrm{s}}) =& \frac{d}{2}\tilde{\lambda}_1\tilde{\lambda}_2\rho^{-d/2-1}\exp(-\tilde{\lambda}_1\rho^{-d/2} - \tilde{\lambda}_2\sigma_{\mathrm{s}}),\\ & \sigma_{\mathrm{s}} > 0, \rho > 0.\end{aligned} \tag{2.11}$$

The hyperparameters $\tilde{\lambda}_1$ and $\tilde{\lambda}_2$ are given by

$$\tilde{\lambda}_1 = -\log(\alpha_1)\rho_0^{d/2} \quad \text{and} \quad \tilde{\lambda}_2 = -\frac{\log(\alpha_2)}{\sigma_{\mathrm{s},0}}, \tag{2.12}$$

so that the prior in (2.11) satisfies $P(\rho < \rho_0) = \alpha_1$ and $P(\sigma_{\mathrm{s}} > \sigma_{\mathrm{s},0}) = \alpha_2$. Note that we limit the range with a lower limit ρ_0 and the marginal standard deviation with an upper limit $\sigma_{\mathrm{s},0}$.

The joint prior in (2.11) can be divided into $\pi(\rho)$ and $\pi(\sigma_{\mathrm{s}})$ by

$$\begin{aligned}\pi(\rho, \sigma_{\mathrm{s}}) &= \pi(\rho)\cdot\pi(\sigma_{\mathrm{s}}),\\ &= \tilde{\lambda}_1\rho^{-2}e^{-\tilde{\lambda}_1/\rho}\cdot\tilde{\lambda}_2 e^{-\tilde{\lambda}_2\sigma_{\mathrm{s}}}\\ &= \mathcal{IG}(1, \tilde{\lambda}_1)\cdot\mathcal{E}(\tilde{\lambda}_2), \quad \rho, \sigma_{\mathrm{s}} \in [0, \infty),\end{aligned} \tag{2.13}$$

where $\mathcal{IG}$ is the inverse Gamma distribution and $\mathcal{E}$ is the exponential distribution. We can look at these distributions to find a 95% interval for both quantities ρ and σ_{s} with fixed limits (ρ_0 and $\sigma_{\mathrm{s},0}$) and tail probabilities (α_1 and α_2).

The density $\pi(\rho)$ is an inverse Gamma distribution with unit shape and rate $\tilde{\lambda}_1$. Thus to easily find the interval boundaries, we can instead transform the boundaries of $\pi(1/\rho)$, which is an exponential distribution with rate parameter $\tilde{\lambda}_1$, since the interval boundaries are invariant to transformation. The cumulative distribution of $1/\rho$ is given by $F(1/\rho) = 1 - \exp(-\tilde{\lambda}_1/\rho)$. Thus we get the interval

$$\begin{aligned}[I_{\rho,\mathrm{L}}, I_{\rho,\mathrm{U}}] =& \big[(I_{1/\rho,\mathrm{U}})^{-1}, (I_{1/\rho,\mathrm{L}})^{-1}\big]\\ =& \big[(-\log(1-p_{\mathrm{U}})/\tilde{\lambda}_1)^{-1}, (-\log(1-p_{\mathrm{L}})/\tilde{\lambda}_1)^{-1}\big],\end{aligned} \tag{2.14}$$

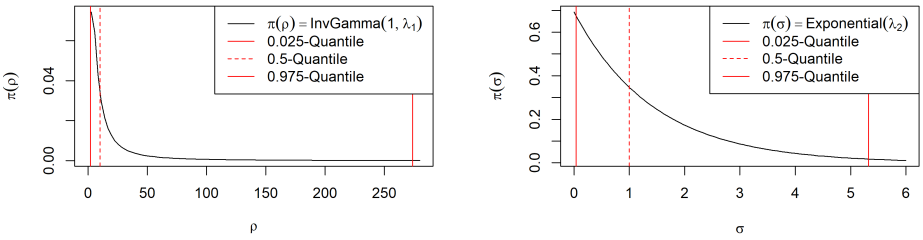
with $p_{\mathrm{L}} = 0.025$ and $p_{\mathrm{U}} = 0.975$. Note that the upper limit for $1/\rho$ becomes the lower limit for ρ and vice versa.

The density $\pi(\sigma_{\mathrm{s}})$ is an exponential distribution with parameter $\tilde{\lambda}_2$ as given in (2.12). The cumulative distribution function for σ_{s} is $F(\sigma_{\mathrm{s}}) = 1 - \exp(-\tilde{\lambda}_2\sigma_{\mathrm{s}})$, and the lower

and upper interval boundaries for σ_s are then given by

$$[I_{\sigma_s,\mathrm{L}}, I_{\sigma_s,\mathrm{U}}] = [-\log(1-p_\mathrm{L})/\tilde{\lambda}_2, -\log(1-p_\mathrm{U})/\tilde{\lambda}_2], \tag{2.15}$$

with $p_\mathrm{L} = 0.025$ and $p_\mathrm{U} = 0.975$.



(a) Probability distribution of ρ with corresponding 95% interval.

(b) Probability distribution of σ_s with corresponding 95% interval.

Figure 2.2: Probability distributions of ρ and σ with respective 95% credibility intervals.

Example 2.1. For an initial range $\rho_0 = 10$ with tail probability $\alpha_1 = 0.5$, the hyperparameter $\tilde{\lambda}_1 = 6.93$. A 95% credibility interval for $\pi(\rho)$ then becomes [1.88, 273.78] as shown in Figure 2.2a. An initial marginal standard deviation of $\sigma_{s,0} = 1$ with tail probability $\alpha_2 = 0.5$ yields a rate parameter of $\tilde{\lambda}_2 = 0.69$ and thus a 95% credibility interval with boundaries [0.04, 5.32], as shown in Figure 2.2b.

A complete Bayesian Hierarchical Model has now been defined, through the latent Gaussian models and appropriate priors. The next step is then to estimate the latent field $\mathbf{x}$ and parameters $\boldsymbol{\theta}$. Therefore, the posterior $\pi(\mathbf{x}, \boldsymbol{\theta} \mid \mathbf{y})$ is of interest. Unfortunately, this is often impossible to calculate analytically, and difficult to approximate. A classic approach to such cases is to use sampling-based methods, like Markov Chain Monte-Carlo (MCMC). These are, however, computationally expensive. Therefore, we will instead use Integrated Nested Laplace Approximations (INLA). This is a numerical method that can be used for performing fast approximate Bayesian inference on latent Gaussian models, and is often a computationally more efficient alternative to MCMC methods. A brief introduction will now be given.

2.3.3 Computationally efficient INLA

Integrated Nested Laplace Approximations (INLA) is a numerical method to do fast approximate Bayesian inference, see Rue et al. (2009). The aim is to approximate the posterior by a combination of analytical approximations and numerical algorithms, instead of using possibly high-dimensional sampling, which MCMC is based on.

In short, the INLA strategy consists of approximating the posterior marginals

$$\pi(\theta_i \mid \mathbf{y}) = \int \pi(\boldsymbol{\theta} \mid \mathbf{y}) d\boldsymbol{\theta}_{-i} \tag{2.16}$$

and

$$\pi(x_i \mid \mathbf{y}) = \int \pi(x_i \mid \boldsymbol{\theta}, \mathbf{y}) \pi(\boldsymbol{\theta} \mid \mathbf{y}) d\boldsymbol{\theta}, \tag{2.17}$$

and then solving the resulting integrals numerically. Hence, the INLA scheme is appropriate if $\pi(\theta_i \mid \mathbf{y})$ and $\pi(x_i \mid \mathbf{y})$ are of interest. Furthermore, the following elements in an LGM should be true for the INLA scheme to be appropriate:

1. $\mathbf{y}$ must be assumed conditionally independent given $\mathbf{x}$ and $\boldsymbol{\theta}_1$, such that the relation $\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}_1 \sim \prod_{i=1}^{n} \pi(y_i \mid x_i, \boldsymbol{\theta}_1)$ holds.

2. The size of the parameter vector $\boldsymbol{\theta}$ must be small, while the size of the latent field $\mathbf{x}$ can be large.

3. The precision matrix $\mathbf{Q}(\boldsymbol{\theta}_2)$ must be sparse.

This holds for the SPDE approach, since only the spatial field $u(\mathbf{s})$ is inserted for $\mathbf{f}_k$, the parameter vector is given by $\boldsymbol{\theta} = (\rho, \sigma_s^2, \sigma_N^2)$, and the precision matrix $\mathbf{Q}$ comes from a GMRF, which has the Markov property, and therefore is sparse. Finally, we are interested in determining $\mathbf{x}$ and $\boldsymbol{\theta}$ given $\mathbf{y}$, and for that the marginals $\pi(\theta_i \mid \mathbf{y})$ and $\pi(x_i \mid \mathbf{y})$ are sufficient. Hence, the INLA scheme is appropriate.

INLA solves the inference problem by first approximating the distributions $\pi(\boldsymbol{\theta}|\mathbf{y})$ and $\pi(x_i \mid \boldsymbol{\theta}, \mathbf{y})$, and then solving the integrals in (2.16) and (2.17) numerically. In particular, $\pi(\boldsymbol{\theta} \mid \mathbf{y})$ is approximated by using a Laplace approximation, while $\pi(x_i \mid \boldsymbol{\theta}, \mathbf{y})$ can be approximated by either a Gaussian approximation, a Laplace approximation or a simplified Laplace approximation.

The Laplace approximation for $\pi(\boldsymbol{\theta} \mid \mathbf{y})$ is based on the following relation:

$$\pi(\boldsymbol{\theta} \mid \mathbf{y}) = \frac{\pi(\mathbf{x}, \boldsymbol{\theta} \mid \mathbf{y})}{\pi(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{y})} \propto \frac{\pi(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}) \pi(\mathbf{x} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta})}{\pi(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{y})}. \tag{2.18}$$

Here, $\pi(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta})$ is available through the LGM, $\pi(\mathbf{x} \mid \boldsymbol{\theta})$ is available through the GMRF, and $\pi(\boldsymbol{\theta})$ is of course given. The denominator $\pi(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{y})$, however, is difficult to evaluate, and thus, an approximation is necessary. In the Laplace approximation, this is done by using a Gaussian approximation $\tilde{\pi}_G(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{y})$, which is built by using Taylor expansions to match the mode, and curvature around the mode, of $\pi(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{y})$. In this Gaussian approximation, it is necessary to calculate the Cholesky decomposition of the precision matrix. This is given by $\mathbf{Q} = \mathbf{L}\mathbf{L}^\top$, where $\mathbf{L}$ is a lower triangular matrix and $\mathbf{L}^\top$ is its conjugate transpose. This is decomposition normally a very computationally expensive step for dense matrices of this dimensionality. Hence, having a sparse precision matrix makes this Gaussian approximation feasible.

The approximation for $\pi(x_i \mid \boldsymbol{\theta}, \mathbf{y})$ is a little more complicated. It can be done efficiently from the Gaussian approximation $\tilde{\pi}_G(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{y})$. However, this method often results in location and skewness errors, and is therefore not recommended. A Laplace approximation can also be constructed, based on the similar relation to (2.18), given by

$$\pi(x_i \mid \boldsymbol{\theta}, \mathbf{y}) = \frac{\pi(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{y})}{\pi(\mathbf{x}_{-i} \mid x_i, \boldsymbol{\theta}, \mathbf{y})} \propto \frac{\pi(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}) \pi(\mathbf{x} \mid \boldsymbol{\theta}) \pi(\boldsymbol{\theta})}{\pi(\mathbf{x}_{-i} \mid x_i, \boldsymbol{\theta}, \mathbf{y})}. \tag{2.19}$$

Again, the nominator is easily calculatable, however, the denominator is difficult. Again, Laplace approximation can be applied by approximating the denominator by a Gaussian approximation $\tilde{\pi}_{GG}(\mathbf{x}_{-i} \mid x_i, \boldsymbol{\theta}, \mathbf{y})$, which can be built by matching the mode and curvature around the mode. However, this is very computationally expensive due to the high dimensionality of $\mathbf{x}$. The third and preferred approximation is the simplified Laplace approximation. This is based on correcting the Gaussian approximation in terms of location and skewness, by a Taylor expansion around the mode of the Laplace approximation. The simplified Laplace approximation will be used in this work.

Once the distributions $\pi(\boldsymbol{\theta} \mid \mathbf{y})$ and $\pi(x_i \mid \boldsymbol{\theta}, \mathbf{y})$ have been approximated, the integrals in (2.16) and (2.17) can be calculated numerically. This will be done by first exploring the space of $\boldsymbol{\theta}$ through the approximation for $\pi(\boldsymbol{\theta} \mid \mathbf{y})$. By locating the mode, and finding a set of high-density points $(\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^k)$, the integrals can be approximated by summing over the high-density area of $\pi(\boldsymbol{\theta} \mid \mathbf{y})$.

Note that since we work with an SPDE approximation, the estimated posterior marginals for ρ and σ_s^2 are not estimates of the true parameters of the Matérn covariance function for the underlying GRF. However, for a good approximation, the resulting parameters will in general be close to the true parameters.

2.4 Model Assessment and Prediction Scores

To evaluate how well the SPDE model we use fits the data and to which extent the SPDE model succeeds in predicting at some given locations, we need a model assessment score and prediction scores.

2.4.1 Model Assessment - WAIC

The Widely Applicable Information Criteria (WAIC) introduced by Watanabe and Opper (2010), also called "Watanabe-Akaike information criteria", is an extension of AIC and is a Bayesian approach for estimating a log pointwise predictive density and corrects for the effective number of parameters. We will use the approach described by Gelman et al. (2014). The reported score for a given model is

$$\text{WAIC} = -2(\text{lppd} - p_{\text{WAIC}}). \tag{2.20}$$

lppd means the log pointwise posterior predictive density and is given by

$$\begin{aligned} \text{lppd} &= \sum_{i=1}^{n} \log \int p(y_i \mid \theta) p(\theta \mid y) \mathrm{d}\theta \\ &\approx \sum_{i=1}^{n} \log \Big(\frac{1}{S} \sum_{s=1}^{S} p(y_i \mid \theta^s) \Big), \end{aligned} \tag{2.21}$$

for S simulations of the posterior $p(\theta \mid y)$ labeled θ^s. The symbol $\approx$ means the computed version.

p_{WAIC} is a correction for effective number of parameters in the model to adjust for overfitting. It is computed by summing the posterior variance of the log predictive density for each data point y_i as in the following

$$\begin{aligned} p_{\text{WAIC}} &= \sum_{i=1}^{n} \text{Var}_{\text{post}}(\log p(y_i \mid \theta)) \\ &\approx \sum_{i=1}^{n} \frac{1}{S-1} \sum_{s=1}^{S} \big(\log p(y_i \mid \theta^s) - \mu_i \big)^2, \end{aligned} \tag{2.22}$$

where $\mu_i = \sum_{s=1}^{S} \log p(y_i \mid \theta^s)$. In practice we sum over the sample variance for the posterior log predictive density.

We can now compare SPDE models with different mesh resolutions to see which model fits the data best. We want low values for WAIC.

The WAIC score does not explicitly say how big the difference between two models in practice is. It is therefore challenging to determine how big the difference between two models should be before, we can conclude that one is better than the other.

2.4.2 Prediction Scores

To evaluate the predictive performance of the SPDE model, we hold out some of the observations when fitting the models. These held-out observations are unobserved for the model, but the true response is known to us and the scoring methods. When predicting at a held-out location $\mathbf{s}_i$, we want to compare the predictive distribution with the true response y_i to measure how accurate the predictions are. We will use the root mean square error (RMSE) and the continuous ranked probability score (CRPS). Both of these scoring rules are negatively orientated, which means that a smaller value indicates a better prediction.

RMSE

In general, RMSE is given by

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2}, \tag{2.23}$$

where $\hat{y}_i$ is the prediction and y_i is the true response. This applies to both continuous and count data. In a repeated k-fold cross-validation, we compute the MSE for all predictions in a given fold, take the average of all K folds and then take the square root to get the mean RMSE for all repetitions. When predicting on a hold-out set, we need a weighted scoring rule, since the number of prediction locations inside a region varies.

$$\text{RMSE}_{\text{hold-out}} = \sqrt{\sum_{j=1}^{\text{No. of states}} MSE_j \times w_j}, \tag{2.24}$$

where $w_j = \frac{\text{No. of locations in state } j}{\text{Total number of locations in dataset}}$, and $\text{MSE}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} (\hat{y}_i - y_i)^2$ is the mean scoring for state j. When computing the RMSE, only point predictions are taken into

account, so that the prediction obtained at location $\mathbf{s}_i$, is only compared to the true response y_i at the same location.

CRPS

In contrast to RMSE, the CRPS uses the predictive cumulative distribution function and the predictive density function to evaluate how well the true value fits with the predictive distribution. As defined in Gneiting et al. (2007), CRPS measures *sharpness* which is the concentration or spread of the predicted distribution, and *calibration* which is the statistical consistency between the predictive distribution and the true response y_i.

From Gneiting and Raftery (2007), the CRPS is defined as

$$\text{CRPS}(F, x) = \int_{-\infty}^{\infty} (F(y) - \mathbb{I}\{y \geq x\})^2 \text{d}y, \tag{2.25}$$

where F is the cumulative distribution function for the predictive distribution, x is the observed value and $\mathbb{I}$ is the indicator function which equals 1 if $y \geq x$ and 0 if not. CRPS is a *proper scoring rule*, which means that the true distribution is the one that minimizes the expected score. A more formal definition is as follows

Definition 2.4. *(Proper Scoring Rules (Gneiting and Raftery, 2007)) Let $S(F, x) \in \mathbb{R}$ be a scoring rule, where F is the predictive distribution and x is an observation drawn from a distribution G. The expected score value of $S(F, x)$ is denoted $S(F, G)$. The scoring rule is proper if $S(G, G) \leq S(F, G)$ for all F and G. If we have equality $S(F, G) = S(G, G)$ only when $F = G$, the scoring rule is strictly proper.*

Note that the definition of CRPS in Equation (2.25) is multiplied by -1 compared to the one provided by Gneiting and Raftery (2007) since we want a negatively oriented score. In this work, we will use the CRPS for both continuous and count data, which means that we need both a continuous and a discrete version of CRPS. For the count data, we will have a Binomial likelihood, while for the continuous version, we will have a likelihood that is Gaussian, which means that the predictive distribution will be Gaussian with a mean μ and variance σ^2.

For an observation y_i in the Gaussian case, the CRPS can be written as

$$\text{crps}_{\text{c}}(\mathcal{N}(\hat{\mu}_i, \hat{\sigma}_i^2), y_i) = \sigma\Big[2\phi(z_i) + z_i(2\Phi(z_i) - 1 - \frac{1}{\sqrt{\pi}})\Big], \tag{2.26}$$

where $z_i = (y_i - \hat{\mu}_i)/\hat{\sigma}_i$ with predicted mean $\hat{\mu}_i$ and variance $\hat{\sigma}_i^2$. $\phi(\cdot)$ is the probability density function (pdf) and $\Phi(\cdot)$ is the cumulative distribution (cdf) of a standard Gaussian variable. When using the INLA approach, we can for a given location $\mathbf{s}_i$ get posterior summary statistics like the mean, standard deviation and median with others, for the prediction. For the predicted mean, $\hat{\mu}_i$, we can use the posterior mean from the SPDE model, but for the variance $\hat{\sigma}_i^2$, we can not use the posterior standard deviation. This is because it does not contain both the marginal variance σ_{s}^2 and the nugget variance σ_{N}^2. We will therefore generate samples and compute the variance $\hat{\sigma}_i^2$ using the following procedure:

- Generate S samples from the (approximate) joint posterior of the latent field

- For each of these samples:
  1. extract the estimated nugget variance $\tilde{\sigma}^2_{\text{N}}$
  2. draw a random effect $\tilde{\epsilon}_i \sim \mathcal{N}(0, \tilde{\sigma}^2_{\text{N}})$
  3. add $\tilde{\epsilon}_i$ to the latent field sample
- Take the variance of these S transformed samples, which will be the predicted variance $\hat{\sigma}^2_i$.

The average CRPS in the continuous case for a set of locations $\{\mathbf{s}_i\}_{i=1,\dots,n}$ with true responses y_i is then

$$\overline{\text{CRPS}}_{\text{c}} = \frac{1}{n}\sum_{i=1}^{n} \text{crps}_{\text{c}}(\mathcal{N}(\hat{\mu}_i, \hat{\sigma}^2_i), y_i). \tag{2.27}$$

For the discrete version of CRPS, when we want to evaluate count data, the approach is a bit different than for the continuous version. For the continuous CRPS, we use samples of the joint posterior of the latent field to create the variance $\hat{\sigma}^2_i$, which again is used as a parameter into the CRPS function in (2.26). For the discrete CRPS we will draw samples from the joint posterior of the latent field and use these to draw new samples from a Binomial distribution, and use the new samples themselves to create the cumulative distribution used in the general equation for CRPS, Equation (2.25). Our observations are probabilities on the form $p^*_i = y_i/n_i$, where y_i is the number of successes and n_i is the number of trials, which means that we want to compare these predictions at probability scale. The samples are created the following way.

- Generate S samples from the (approximate) joint posterior of the latent field, $\{\tilde{\eta}^s_i\}_{s=1\dots,S}$
- For each of these samples:
  1. Draw a new sample from the Binomial distribution using the latent field samples, $\tilde{y}^s_i \sim \mathcal{B}(n = n_i, p = \mathit{logit}^{-1}(\tilde{\eta}^s_i))$
- Create an empirical cumulative distribution of the new samples $\{\tilde{y}^s_i\}_{s=1,\dots,S}$

The empirical cumulative distribution is then for a location $\mathbf{s}_i$ given by

$$\hat{F}_i(t_i) = \frac{\text{Number of samples} \leq t_i}{S} = \frac{1}{S}\sum_{s=1}^{S} \mathbb{I}(\tilde{y}^s_i \leq t_i), \tag{2.28}$$

where $t_i = [0, \frac{1}{n_i}, \frac{2}{n_i}, \dots, \frac{n_i-1}{n_i}, 1]$ and $\mathbb{I}$ is the indicator function. After this, we can define the discrete CRPS for a location $\mathbf{s}_i$ as

$$\text{crps}_{\text{d}}(\hat{F}_i, p^*_i) = \frac{1}{n_i}\sum_{j=0}^{n_i} \left(\hat{F}_i(j/n_i) - \mathbb{I}(p^*_i \leq j/n_i)\right)^2. \tag{2.29}$$

The average discrete CRPS for a set of locations $\{\mathbf{s}_i\}_{i=1,\dots,n}$ is then given by

$$\overline{\text{CRPS}}_{\text{d}} = \frac{1}{n}\sum_{i=1}^{n} \text{crps}_{\text{d}}(\hat{F}_i, p^*_i). \tag{2.30}$$

As with RMSE, we will use CRPS to evaluate predictions for a repeated K-fold cross-validation and a hold-out set. The average CRPS for a repetition r will therefore be

$$\overline{\text{CRPS}}_r = \frac{1}{K}\sum_{k=1}^{K}\Big[\frac{1}{n_k}\sum_{i=1}^{n_k}\text{crps}(F_i, x_i)\Big], \tag{2.31}$$

with n_k being the number of observation/prediction locations in fold k and $\text{crps}(F_i, x_i)$ being the CRPS value for the observation at location $\mathbf{s}_i$ in fold k. This will either be the continuous CRPS, $\text{crps}_\text{c}(\mathcal{N}(\hat{\mu}_i, \hat{\sigma}_i^2), y_i)$, or the discrete, $\text{crps}_\text{d}(\hat{F}_i, p_i^*)$. For a hold-out region, the CRPS is

$$\text{CRPS}_\text{total} = \sum_{j=1}^{\text{No. of states}} \text{CRPS}_j \times w_j, \tag{2.32}$$

where $w_j = \frac{\text{No. of locations in state } j}{\text{Total number of locations in dataset}}$, and $\text{CRPS}_j = \frac{1}{n_j}\sum_{i=1}^{n_j}\text{crps}(F_i, x_i)$ is the mean scoring for state j with $\text{crps}(F_i, x_i)$ as above.

Chapter 3

Investigating the Influence of Mesh Resolution for Gaussian and Non-Gaussian Responses

In this chapter, we will perform two case studies to investigate how the SPDE model is influenced by different mesh configurations. We first explain the purpose of these case studies. Then a description of the studies and assessment criteria follows. Finally, the two case studies of datasets introduced in Sections 1.3 and 1.4 are presented. Here we give the specific model choices for the continuous and count data.

3.1 Aim of the Studies

We want to investigate the influence on estimation and prediction of mesh resolution for Gaussian and Non-Gaussian responses using the SPDE approach for GRFs in 2D (Lindgren et al., 2011) described in Section 2.2. This is interesting and important because the mesh in the SPDE approach determines both the accuracy and precision of our results as well as the computational cost. We also believe that the mesh has a different influence on continuous data, compared to count data. We will therefore study the influence of the mesh on two datasets, one with a Gaussian likelihood and one with a Binomial likelihood.

Throughout this study, we want to answer three questions; How well does the SPDE model with different mesh resolutions fit our data? How does the mesh resolution influence parameter estimates? How does the mesh resolution influence the prediction quality both in terms of accuracy and the computational complexity measured in running time?

3.2 Design of Case Studies

In this section, we will describe the general setup for our case studies, which includes the spatial model, the mesh setup and how to answer the three questions above. This setup

applies to both continuous and count data, while the specific details for each case will be explained in Section 3.3 and 3.4.

3.2.1 Spatial Model

Let $y_1, \ldots, y_n$ be observations at locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$. We assume that the underlying model for our observations is a Gaussian random field (GRF) with zero mean and a covariance matrix with Matérn covariance function, as defined in Definition 2.1. The Matérn covariance function has two parameters, the range ρ and the marginal variance σ_s^2.

We can therefore let our observations y_i follow a likelihood $p(\cdot)$ given a mean μ_i which is linked to a linear predictor η_i through a link function $g(\cdot)$. The linear predictor consists of an intercept β_0, covariates with coefficients, $\mathbf{X}\boldsymbol{\beta}$, and a spatial effect $u(\mathbf{s})$. In addition, we introduce a nugget parameter, σ_{N}^2. For the Gaussian model, σ_{N}^2 will be a likelihood parameter, while for the Binomial model, this is a parameter of an iid random effect ϵ_i in the linear predictor η_i.

As presented in Chapter 1 - Introduction, the GRF that we assume as an underlying model for the observations, can be computationally expensive to work with. Therefore, we will use the SPDE approach by Lindgren et al. (2011) described in Section 2.2. In this approach, we need the model specifications which include the likelihood $p(\cdot)$ and priors for the parameters defined above, and we must also decide on the mesh structure.

3.2.2 Mesh Setup

In Section 2.2.2, the mesh on which the SPDE approach works was defined. To investigate how the prediction quality varies with the mesh resolution, we will change two mesh parameters; h, the largest allowed triangle edge length, and r, a factor that adds an outer extension to the domain boundary. It is possible and common to use different mesh resolutions in the inner and the outer domain, where the latter can have a lower resolution to save running time. We choose to use only one resolution on the whole domain since we are only interested in resolution effects and boundary effects. Besides, more parameters would give more complex results. The mesh will be made independent of the observation locations. This is because we are interested in the general performance for different mesh resolutions, and we therefore do not want the mesh to be influenced by the observation locations of these particular datasets.

3.2.3 Setup of Case Studies

In the first question introduced in Aim of the Studies, we are interested in finding out how well the chosen model fits with the datasets for different mesh configurations. A configuration in this context is given by the largest allowed edge length, h, in the mesh triangulation and the outer boundary extension, r, which will vary for different configurations. We therefore begin with a model assessment, where we calculate the model score WAIC defined in Section 2.4. We also report the computation time of fitting the model.

To answer the second question, we fit the spatial model for different mesh resolutions on both datasets and report the parameter estimates. These parameters are estimated by maximizing the log-likelihood of the SPDE-approximation. We consider fixed parameters

such as the intercept β_0 and covariate coefficients $\boldsymbol{\beta}$, parameters for the spatial model, range ρ, marginal variance σ_s^2, and the nugget effect σ_N^2. Specifically, we will investigate which mesh resolution the parameter estimates start to stabilize on, and how different the estimated posteriors are compared to their priors. We do not know the values of the true parameters, thus we can not check whether the parameter estimates really are close, we can only look at their behaviour.

For answering the third question, the mesh's impact on predictive power, we will perform repeated ten-fold cross-validation, as well as a hold-out analysis where one region is held out at a time. Firstly, we evaluate predictive performance by calculating the scoring rules CRPS and RMSE as described in Section 2.4, and report running times for different mesh resolutions. The motivation behind doing the repeated cross-validation is to investigate how much better the model predicts when increasing the mesh resolution. For each repetition, the dataset will be randomly split into ten folds of equal size. Observations from nine folds will be used to fit the model, and the last fold will be predicted on and assigned a score by the scoring rules CRPS and RMSE. This will be done for all ten folds so that every observation location is predicted on once. The average score of the held-out observations in one fold is then calculated, and we report the average for a repetition. This process is repeated ten times so that the choice of folds does not affects the final results.

We are also interested in seeing how the prediction quality changes with increasing mesh resolution when holding out larger areas at a time. We will therefore, for each mesh configuration, hold out one region (states in the US / counties in Kenya), use the remaining observations to fit the model, and then predict on the held out locations. This will say something about how well the model predicts over longer distances when the distance between the point to be predicted is possibly at a long distance from other observations.

Lastly, we visualize the model estimates for both datasets on a fine grid and report the runtimes for a selection of mesh configurations. This is to visually inspect our results and give a qualitative summary. Do we see any visual differences when increasing the mesh resolution?

3.3 Case Study 1 - Precipitation in the Conterminous US

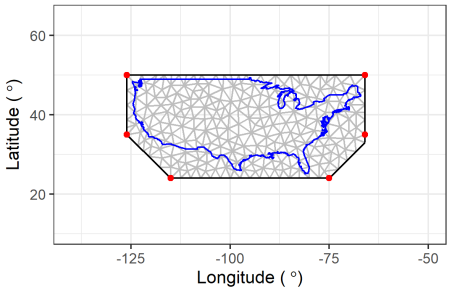
We will now examine the dataset "Precipitation in the Conterminous US" which was introduced in Section 1.3. The objective is to solve the questions raised in Section 3.1. Let $y_1, \ldots, y_n$ be observations of precipitation on $\log_2$-scale at locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$, with $n = 7040$ in this particular case. These observations are shown in Figure 1.1. We also have elevation data, x_i, at these locations. We assume that y_i can be described by the model defined in the previous section, with a Gaussian likelihood and the identity link function.

The hierarchical model for y_i then becomes

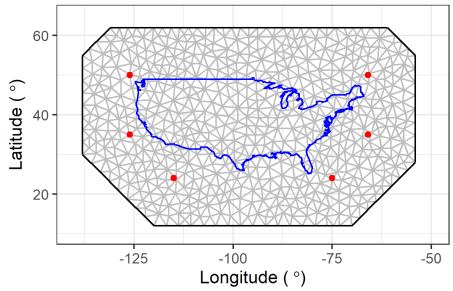
$$\begin{aligned} &\textit{Stage 1:} && y_i \mid \eta_i \sim \mathcal{N}(y_i \mid \eta_i, \sigma_{\text{N}}^2), \\ & && \eta_i = \beta_0 + x_i\beta_1 + u(\mathbf{s}_i) \\ &\textit{Stage 2:} && u_i \mid \boldsymbol{\theta} \sim \text{GRF}(0, \Sigma(\boldsymbol{\theta})), \quad i = 1, \ldots, n \\ &\textit{Stage 3:} && \boldsymbol{\theta}, \sigma_{\text{N}}^2 \sim \pi(\boldsymbol{\theta}, \sigma_{\text{N}}^2). \end{aligned} \tag{3.1}$$

As presented in Section 2.3.2, we will use PC-priors for ρ, σ_{s}^2 and σ_{N}^2. From an initial look, we choose the prior for the range parameter to have a median of 10, which is about 1/5 of the diameter of our study region. For the marginal standard deviation, σ_{s}, we choose a median of 1, and we set the nugget variance prior to be 10 % of the marginal variance, which yields a median of $\sqrt{0.1}$ for σ_{N}. As shown in Example 2.1 in Section 2.3.2, this gives 95% credible intervals of [1.9, 273.8] for ρ, [0.04, 5.32] for σ_{s} and [0.01,1.68] for σ_{N}.

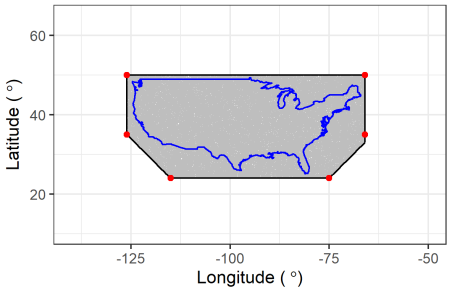
We will run the different analyses described in Section 3.1 with the following values of maximum edge length h: [8, 4, 2, 1, 0.5, 0.25, 0.125]. The offset values will be [0, -0.05, -0.1, -0.2, -0.4], which corresponds to an increase of [0, 5, 10, 20, 40] % of the existing domain. In Figure 3.1, six meshes are shown with the boundary of the conterminous US (blue line) and our location domain (red points) for different combinations of maximum edge length h and offset r.



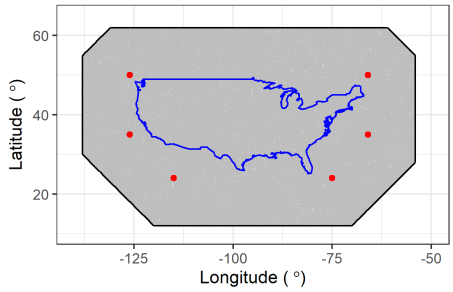
(a) $h = 4$ and $r = 0\%$. Number of triangles are 254.



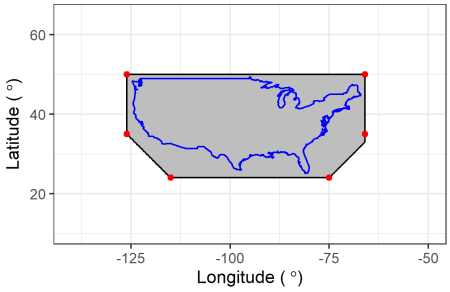
(b) h = 4 and r = 20%. Number of triangles are 649.



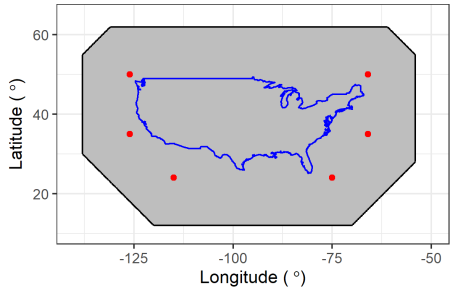
(c) h = 1 and r = 0 %. Number of triangles are 3676.



(d) h = 1 and r = 20%. Number of triangles are 9873.



(e) h = 0.25 and r = 0%. Number of triangles are 58430.



(f) h = 0.25 and r = 20%. Number of triangles are 154154.

Figure 3.1: Meshes with different resolution. Grey line is the mesh, blue line is the US boundary and red points are the predefined domain.

3.3.1 Model Assessment and Parameter Evaluation.

We start by investigating the model fit for different maximum edge lengths, h. The model score WAIC is shown in Table 3.1 with running times and the number of triangles. Increasing the mesh resolution improves the WAIC score for all resolutions, but at a lower rate for the finest resolutions. $h = 0.125$ gives the best WAIC score of -7868. The runtime, on the other hand, increases slowly in the beginning (with a rate of 2 to 3) and faster for the higher resolutions (rate of 5). For $h = 0.125$, the running time is almost an hour, compared to about 10 seconds for the lowest resolution ($h = 8$ to $h = 2$), and a few minutes for $h = 1$ to $h = 0.25$. Note that the running time includes making the mesh, fitting the SPDE model and estimating parameters.

Table 3.1: Table of model fitness and running times for estimating the model and creating the meshes, for increasing mesh resolution in the first case study.

r	h	WAIC	Est. time (s) + mesh time (s)	no. of triangles
10%	8	1285	8.3 + 0.3	115
	4	-1065	7.5 + 0.2	441
	2	-4044	13 + 0.2	1675
	1	-6506	40 + 0.3	6495
	0.5	-7394	114 + 0.9	25521
	0.25	-7779	602 + 5.7	101459
	0.125	**-7868**	3183 + 56	404994

The increased runtime is caused by the higher number of triangles that are needed to cover the domain when reducing h. With more triangles in the mesh, there are more nodes on which the SPDE model works. The dimensionality m of the projector matrix $\mathbf{A}$ and the Gaussian weights $\mathbf{w}$ increases, and thus the total CPU time for running the model increases as well. In addition to a larger estimation time for finer mesh, the time it takes to build the mesh also increases.

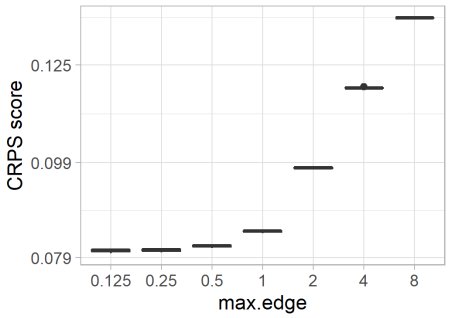
To answer the second question for this case, we continue with investigating how the mesh resolution influences the parameter estimates. Table 3.2 show estimates for the fixed parameters intercept ($\hat{\beta}_0$) and elevation ($\hat{\beta}_1$), the spatial parameters range ($\hat{\rho}$) and marginal standard deviation ($\hat{\sigma}_{\mathrm{s}}$), and the nugget effect ($\hat{\sigma}_{\mathrm{N}}$). The intercept $\hat{\beta}_0$ stabilizes at h = 0.5 with value $\hat{\beta}_0 = 6.2$ and the elevation $\hat{\beta}_1$ stabilizes at h = 1 with value $\hat{\beta}_1 = 0.7$. The

Table 3.2: Tables of estimated parameters for increasing mesh resolution with offset $r = 10\%$. The values shown are the 50% quantiles with corresponding 2.5% and 97.5% quantiles in brackets.

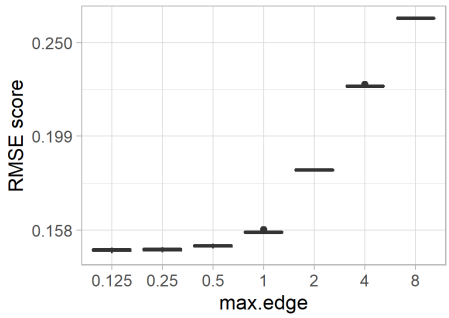
h	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\rho}$	$\hat{\sigma}_{\mathrm{s}}$	$\hat{\sigma}_{\mathrm{N}}$
8	6.62 [4.86, 8.40]	0.52 [0.50, 0.54]	26.3 [16.0, 52.9]	1.09 [0.78, 1.73]	0.264 [0.259, 0.268]
4	6.50 [5.47, 7.63]	0.65 [0.63, 0.67]	18.6 [12.6, 37.2]	1.03 [0.76, 1.70]	0.222 [0.218, 0.225]
2	6.37 [5.84, 6.95]	0.66 [0.64, 0.68]	10.9 [8.6, 15.1]	0.86 [0.70, 1.15]	0.174 [0.171, 0.178]
1	6.24 [5.91, 6.58]	0.69 [0.68, 0.71]	7.2 [6.1, 8.7]	0.76 [0.65, 0.89]	0.138 [0.135, 0.140]
0.5	6.21 [5.92, 6.52]	0.70 [0.68, 0.72]	6.4 [5.6, 7.6]	0.73 [0.64, 0.85]	0.122 [0.120, 0.125]
0.25	6.20 [5.92, 6.50]	0.70 [0.68, 0.72]	6.2 [5.1, 7.2]	0.74 [0.62, 0.85]	0.116 [0.113, 0.120]
0.125	6.19 [5.94, 6.46]	0.70 [0.68, 0.72]	6.2 [5.3, 8.4]	0.73 [0.63, 0.94]	0.115 [0.111, 0.118]

size of the 95% credibility interval for $\hat{\beta}_0$ is at its smallest for $h = 0.125$, while for $\hat{\beta}_1$ the smallest interval is at $h = 1$. The parameters $\hat{\rho}$ and $\hat{\sigma}_{\mathrm{N}}$ stabilize at $h = 0.25$, while $\hat{\sigma}_{\mathrm{s}}$ stabilizes at $h = 0.5$. The 95% credibility intervals are at the smallest for $h = 0.5$ for $\hat{\rho}$, $\hat{\sigma}_{\mathrm{s}}$ and $\hat{\sigma}_{\mathrm{N}}$.

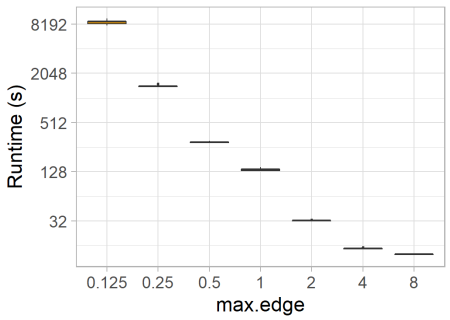
To summarize this initial analysis, we see that by increasing the mesh resolution, the model score WAIC improves, and the parameters converge, to a cost of longer running time.



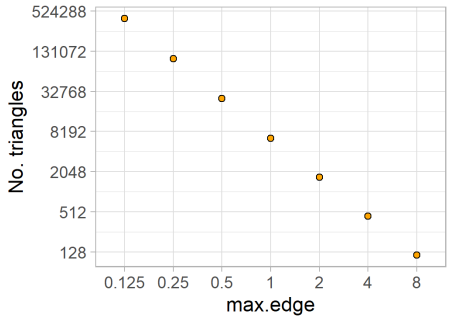
(a) Boxplot of CRPS scores for cross-validation.



(b) Boxplot of RMSE scores for cross-validation.



(c) Total CPU time for estimation.



(d) Number of triangles in the mesh for each h.

Figure 3.2: Cross-validation scores, running time in seconds and number of triangles for varying h and constant $r = 10\%$.

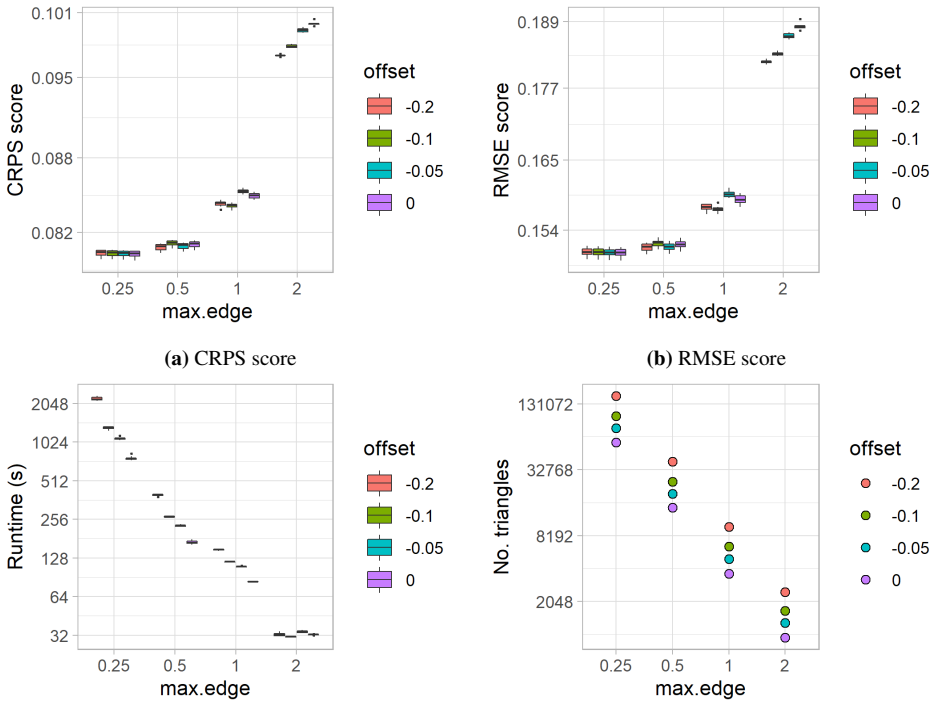
3.3.2 Repeated 10-fold Cross-Validation

In the third question, we want to investigate the predictive power as function of mesh resolution. In Figure 3.2a and 3.2b, boxplots with mean CRPS score and mean RMSE score for the ten repetitions are shown for decreasing maximum edge lengths. Within the boxplots, the variations are very small, which means that the choice of prediction locations almost not affects the scores. We see that both prediction scores stabilize around $h = 0.5$ with a value of about 0.08 for CRPS and 0.149 for RMSE. Boxplots of total computation time for estimation are shown in Figure 3.2c with the corresponding number of triangles in the mesh in Figure 3.2d. We see a linear relationship on $\log_2$-scale between the running

time and h, as well as the number of triangles and h. When halving h, the number of triangles is quadrupled.

As described in Section 2.4.2, we need to sample the estimated values at prediction locations to compute the variance which is needed in the CRPS score. This sampling procedure takes some time in addition to the time it takes to fit the model and estimate values at prediction locations. The mean computation time for sampling is reported in the Appendix, in Table A.1, together with mean RMSE, mean CRPS and mean estimation time.

Note that the running times in Figure 3.2c are larger than the running times shown in Table 3.1. The implementation in R-INLA is parallellized for 10-fold CV and hold-out region analysis, so the above difference is because only one thread is used in Figure 3.2c, while 10 threads are used in Table 3.1.



(a) CRPS score

(b) RMSE score

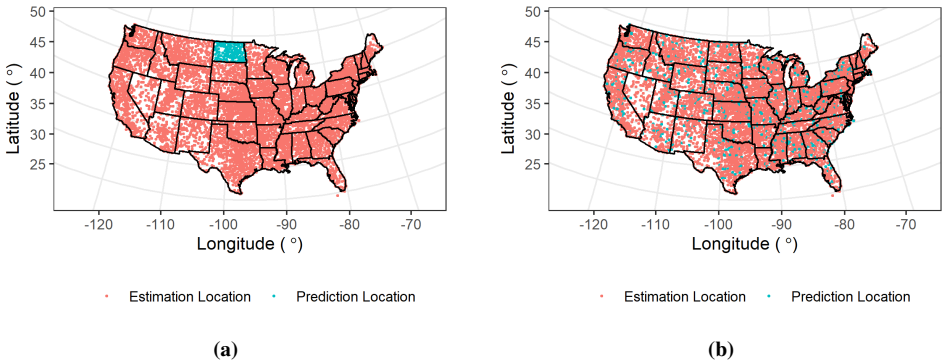
(c) Total CPU time in seconds.

(d) Number of triangles in mesh.

Figure 3.3: Scores, runtime and number of triangles for combinations of h and r for cross-validation. The order of boxes within each h group coincides in all plots with r = 20% to the left (red), r = 0% to the right (purple).

Until now, we have only run the SPDE model with $r = 10\%$, which is an increase of 10% of the original domain. To check whether there is any further connection between a low maximum edge length and a larger outer boundary extension, combinations of h = [2, 1, 0.5, 0.25] and r = [0%, 5%, 10%, 20%] are run.

The results are shown in Figure 3.3. We clearly see that the maximum edge length



(a) **(b)**

Figure 3.4: (a) The state North Dakota as hold-out region with prediction locations in blue. The other observations (red) will be used for estimation. (b) Example of a 10-fold CV splitting. One fold with locations shown in blue will be used for predictions while the nine remaining are for estimation.

parameter h, influences the prediction quality in terms of RMSE and CRPS score more than the offset parameter r. For h between 0.25 and 1, a change in r has very little effect on the RMSE and CRPS score, it only increases the running time for estimating at prediction locations. For both the scores and the runtimes in Figure 3.3c, there is little variation between the repetitions within each h group, which give the small boxplots.

The runtime increases with a larger r, expect for the runs with $h = 2$, which is natural as there are more triangles for lower values of h. For $h = 2$, which is an exception from this trend, there is a difference of 960 triangles ($r = 0\%$) to 2508 triangles ($r = 20\%$), which is a rather small difference. There are other factors in the estimation process which do not depend on the number of triangles, and we therefore do not observe the same trend here.

From this analysis we can justify continuing using our default offset value $r = 10\%$, corresponding to a 10% increase of the mesh.

3.3.3 Hold-Out Regions

In our study region, which is the conterminous US, there are 48 states. Some are close to the boundary and some are in the middle of the mesh. In addition to the former analyses, we are interested in seeing how the prediction quality changes with increasing mesh resolution when holding out larger areas at a time, like a state.

In Figure 3.4, two plots of the conterminous US with estimation and prediction locations are shown for (a) a hold-out region and (b) one prediction fold in the 10-fold CV. In comparison to using a hold-out region for prediction, the 10-fold CV randomly splits all observations in the study region into ten folds, where one is a prediction fold and the remaining nine are for estimation. Thus, most prediction locations in the 10-fold CV have closely related points that are used for fitting the model, while most prediction locations in the hold-out region have a long distance to the closest observations.

We therefore leave observations of one state out at a time. We then predict the log annual precipitation at these hold-out locations, and compare the predictions with the ob-

served values through the weighted scoring rules RMSE and CRPS,

In Table 3.3, the prediction scores with mean estimation time and mean sampling times are presented. From these results we see that by reducing the maximum edge length, h, the prediction quality improves. $h = 0.5$ yields the best prediction qualities both with RMSE and CRPS. This corresponds to a mean estimation time of 324 seconds and a prediction time of 136 seconds, in total about 8 minutes on average.

Table 3.3: Table of prediction scores and runtimes including estimation time and sampling time for computing CRPS, for increasing mesh resolution for hold-out regions analysis.

r	h	RMSE	CRPS	Estimation time + Prediction time (s)
10%	8	0.372	0.184	14 + 15
	4	0.304	0.154	17 + 18
	2	0.288	0.146	35 + 29
	1	0.280	0.141	148 + 52
	0.5	**0.270**	**0.138**	324 + 136
	0.25	0.274	0.139	1598 + 438
	0.125	0.274	0.139	8767 + 1756

As a final task, we keep h constant with a value of 0.5 and vary the offset r for one hold-out region located near the boundary of the US, specifically North Dakota. By doing this, we can check whether the model suffers from boundary effects. In Table 3.4 RMSE and CRPS values are shown for varying r. There is no clear trend for the prediction quality, especially not in the CRPS score. Thus, it does not seem that the model suffers from boundary effects, at least not when the maximum edge length is as short as $h = 0.5$.

Table 3.4: Table of prediction scores and runtimes including estimation time and sampling time for computing CRPS, for increasing boundary region when holding out the state North Dakota.

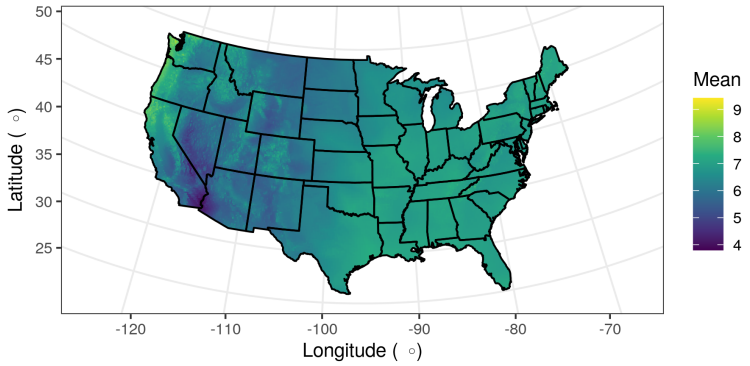
r	h	RMSE	CRPS	Estimation time + Prediction time (s)	No. Triangles
0%	0.5	**0.163**	0.115	89 + 37	14706
5%	0.5	0.165	**0.113**	124 + 45	19796
10%	0.5	0.171	0.115	151 + 53	25521
20%	0.5	0.171	0.114	238 + 72	38706

3.3.4 Predictions on Grid

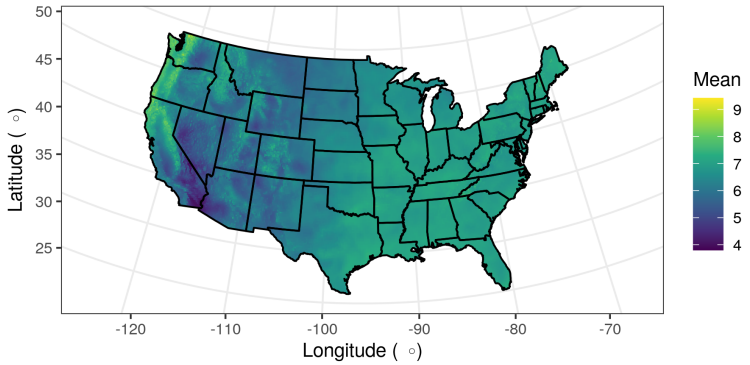
To visualize what we have seen, we predict the model on a grid of size 400×200. As we want to predict the posterior mean and the posterior standard deviation on the grid locations, we need elevation data at all these 80 000 locations. We use elevation data from Hastings et al. (1999). In Figure 3.5, the posterior mean is shown for three different values of h; 4, 1 and 0.25. The elevation covariate is clearly seen in all three plots.

When only looking at the random field, as in Figure A.1 in Appendix, a lower value of h, i.e., a finer mesh, gives more variability in the predicted mean.

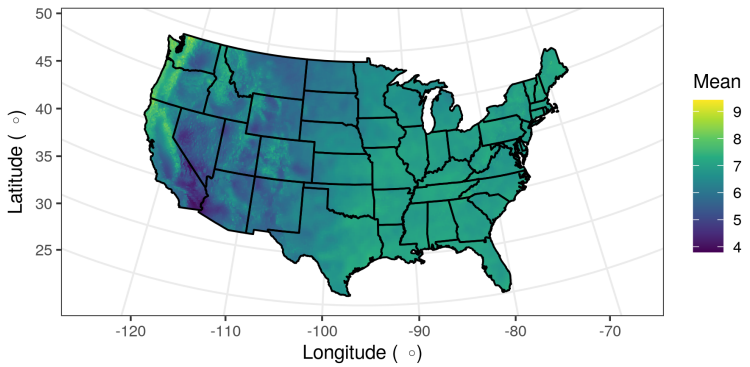
Figure 3.6 shows the posterior standard deviation on the predicted locations. The colorbar is shown on log-scale. For the low resolution case, with maximum edge length



(a) Parameter $h = 4$.



(b) Parameter $h = 1$.



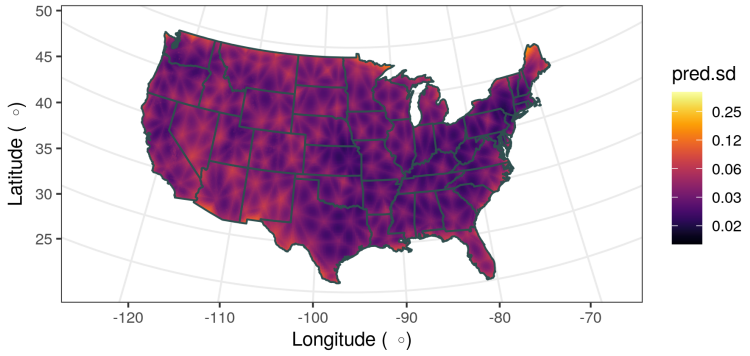
(c) Parameter $h = 0.25$.

Figure 3.5: Predicting annual log-precipitation on 400×200 grid.

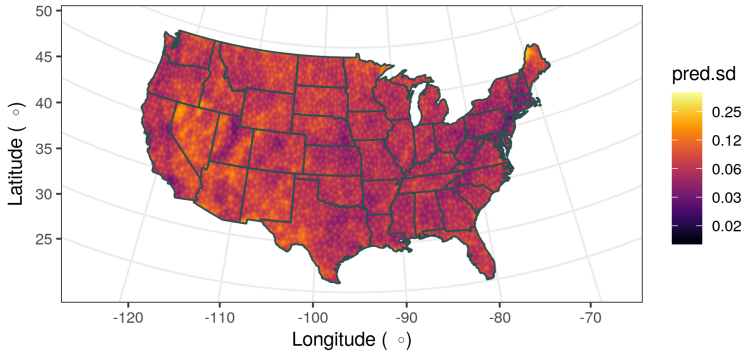
$h = 4$, we clearly see the triangles in the mesh as the triangle edges and intersections have higher standard deviation. This effect was explained in Section 2.2.2.

For shorter maximum edge lengths in Figure 3.6b and 3.6c, the standard deviation is generally higher. If we compare the observation locations in Figure 1.1 with the standard deviation, we see that areas with low density of observations have higher standard deviation, which is natural. Thus, if the posterior standard deviation is of interest, one should increase the mesh resolution.

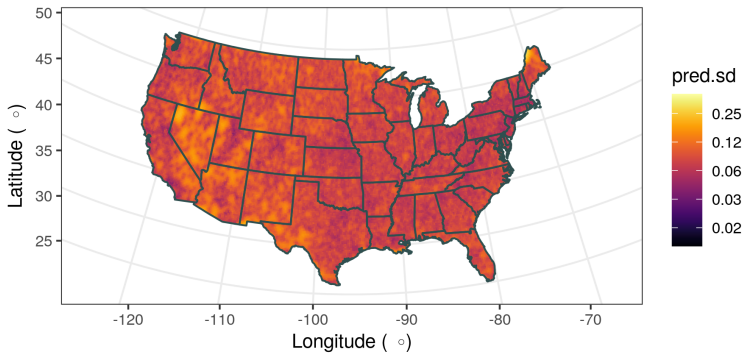
Predicting on the grid with $h = 4$ took about 88 seconds, with $h = 1$ it took 267 seconds and with $h = 0.25$ it took 891 seconds ($\approx$ 15 minutes). In comparison, estimation in the 10-fold CV took respectively about 16, 128 and 1448 seconds for each fold, in each repetition, in all 100 times.



(a) Parameter h=4.



(b) Parameter h=1.



(c) Parameter h=0.25.

Figure 3.6: Standard deviation of predicted annual log-precipitation on 400 $\times$ 200 grid.

3.4 Case Study 2 - Secondary Education Prevalence for Women in Kenya

We will now perform a similar analysis as we did in Section 3.3 on the dataset of secondary education prevalence for women in Kenya introduced in Section 1.4. Let $n_1, \ldots, n_n$ be the number of interviewed women at observation locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$. Furthermore, let $y_1, \ldots, y_n$ denote the secondary education prevalence at these locations, i.e., the number of interviewed women that have completed their secondary education. The observed probability of completing secondary education at location $\mathbf{s}_i$ is then given by $p_i^* = y_i/n_i$. The dataset is shown in Figure 1.2.

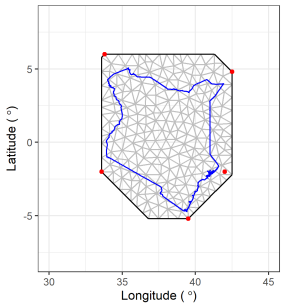
We assume that our data y_i have a Binomial likelihood which can be expressed by

$$\begin{aligned} y_i \mid \eta_i &\sim \mathcal{B}(n_i, p_i), \\ \eta_i &= \text{logit}(p_i) = \log \frac{p_i}{1-p_i} \\ &= \beta_0 + \beta_{\text{URB}} \mathbb{I}_{\text{urban}}(\mathbf{s}_i) + u(\mathbf{s}_i) + \epsilon_i, \quad i = 1, \ldots, n \end{aligned} \tag{3.2}$$

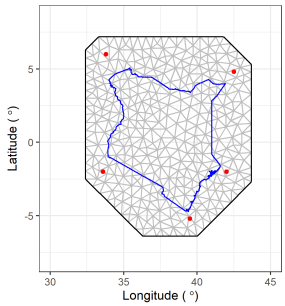
where $\mathcal{B}(n_i, p_i)$ is the Binomial likelihood with n_i trials and a probability p_i of success. The logit of the prevalence p_i is denoted η_i and is a sum of the intercept β_0, the indicator covariate "urbanicity" which is 1 if the location is urban and 0 if it is rural, with parameter β_{URB}, a spatial field $u(\mathbf{s}_i)$ defined by our GRF and an iid Gaussian random effect, ϵ_i, with zero mean and a nugget variance σ_{N}^2.

Our spatial field $u(\mathbf{s}_i)$ is a GRF with a Matérn covariance function, as defined in Section 2.1. This function has two parameters, the range ρ and the marginal variance σ_{s}^2. The Gaussian noise ϵ_i in the linear predictor η_i has parameter σ_{N}^2. As we are in a Bayesian setting, we want priors on these hyperparameters, and we use the PC-priors defined in Section 2.3.2. Following the approach of Paige et al. (2019), we set the median of the prior for the range ρ to be about one fifth of the diameter of the spatial domain. On longitude-latitude scale, this corresponds to a prior of $P(\rho < 2) = 0.5$. For the marginal standard deviation and the nugget effect we set $P(\sigma_{\text{s}} > 1) = P(\sigma_{\text{N}} > 1) = 0.01$. This yields 95% credibility intervals of [0.4, 54.8] for ρ and [0.006, 0.801] for σ_{N} and σ_{s}.

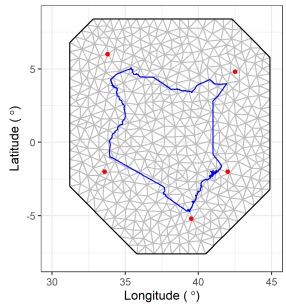
As a final step, we need to decide on the mesh structure, in terms of its size and resolution. As earlier described, we will vary the maximum edge length, h, and the offset, r. We have set the median of the prior to be 2, thus we start with a maximum edge length of $h = 4$, and halve this until $h = 0.125$ with a boundary offset of $r = 10\%$. In Figure 3.7, we see different configurations of the mesh structure to be used. The boundary of Kenya is shown in blue together with the chosen boundary domain with red points. A finer resolution gives more triangles.



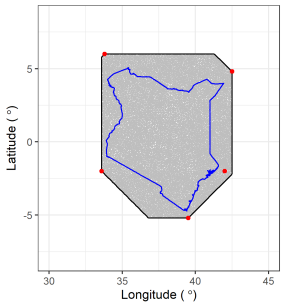
(a) $h = 1$ and $r = 0\%$. Number of triangles are 240.



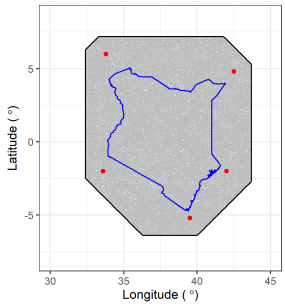
(b) $h = 1$ and $r = 10\%$. Number of triangles are 376.



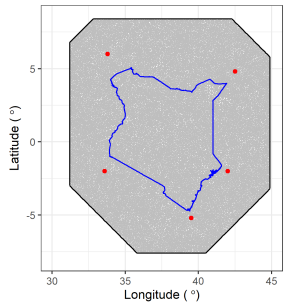
(c) $h = 1$ and $r = 20\%$. Number of triangles are 520.



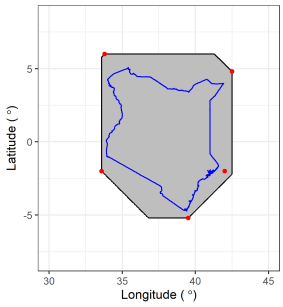
(d) $h = 0.25$ and $r = 0\%$. Number of triangles are 3674.



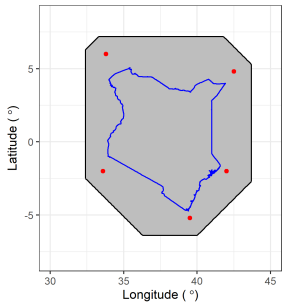
(e) $h = 0.25$ and $r = 10\%$. Number of triangles are 5590.



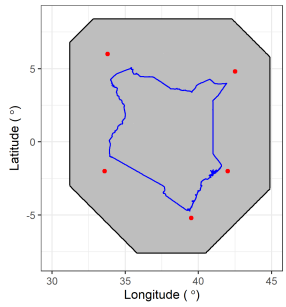
(f) $h = 0.25$ and $r = 20\%$. Number of triangles are 7926.



(g) $h = 0.125$ and $r = 0\%$. Number of triangles are 14325.



(h) $h = 0.125$ and $r = 10\%$. Number of triangles are 21985.



(i) $h = 0.125$ and $r = 20\%$. Number of triangles are 31111.

Figure 3.7: Meshes with different resolution. Grey line is the mesh, blue line is the US boundary and red points are the predefined domain.

Table 3.5: Table of model fitness and running times for estimating the model and creating the meshes, for increasing mesh resolution in the seond case study.

r	h	WAIC	Estimation time + Mesh time (s)	No. of triangles
10%	4.00	4856	9.1 + 0.3	32
	2.00	4836	10.7 + 0.3	103
	1.00	4801	9.6 + 0.3	376
	0.50	4784	11.1 + 0.3	1441
	0.25	4785	51.8 + 0.4	5590
	0.125	4783	296.0 + 1.3	21985
	0.0625	**4783**	6092.9 + 5.7	87379

Table 3.6: Table of fixed parameters and hyperparameters with median and [2.5%, 97.5%] credibility interval in brackets. Offset r = 10%.

h	$\hat{\beta}_0$	$\hat{\beta}_{\text{URB}}$	$\hat{\rho}$	$\hat{\sigma}_{\text{s}}$	$\hat{\sigma}_{\text{N}}$
4.0000	-2.42 [-3.34, -1.53]	1.09 [0.97, 1.22]	2.81 [1.33, 5.82]	1.01 [0.65, 1.61]	0.77 [0.85, 0.70]
2.0000	-2.47 [-3.14, -1.87]	1.04 [0.91, 1.17]	2.08 [1.07, 3.94]	0.92 [0.65, 1.33]	0.77 [0.82, 0.67]
1.0000	-2.55 [-3.14, -2.03]	1.01 [0.88, 1.14]	1.80 [1.12, 2.97]	0.92 [0.70, 1.21]	0.69 [0.77, 0.62]
0.5000	-2.57 [-3.12, -2.10]	1.00 [0.87, 1.13]	1.67 [1.14, 2.54]	0.90 [0.70, 1.17]	0.66 [0.74, 0.58]
0.2500	-2.56 [-3.14, -2.07]	0.99 [0.86, 1.13]	1.82 [1.26, 2.74]	0.88 [0.68, 1.15]	0.66 [0.74, 0.58]
0.1250	-2.56 [-3.14, -2.08]	0.99 [0.86, 1.13]	1.79 [1.24, 2.68]	0.88 [0.68, 1.15]	0.65 [0.73, 0.58]
0.0625	-2.56 [-3.13, -2.08]	0.99 [0.86, 1.13]	1.77 [1.24, 2.67]	0.88 [0.68, 1.15]	0.65 [0.73, 0.58]

3.4.1 Model Assessment and Parameter Evaluation

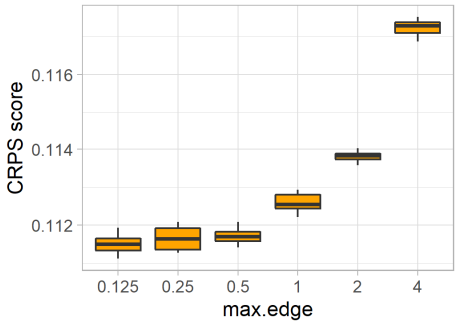
Similarly to the Gaussian data case, we start off with a model assessment and a parameter evaluation. In Table 3.5, the results of an initial analysis are shown. The model score WAIC stabilizes at a maximum edge length of $h = 0.5$, with a value of 4784. This corresponds to an estimation time of about 11 seconds and a mesh creation time of 0.3 seconds. Reducing h two steps further, to $h = 0.125$, only improves the WAIC-score to 4783 but increases the runtime to about 5 minutes.

We are also interested in looking at the behaviour of our estimated parameters, both the fixed parameters $\hat{\beta}_0$ and $\hat{\beta}_{\text{URB}}$, and the hyperparameters $\hat{\rho}$, $\hat{\sigma}_{\text{s}}$ and $\hat{\sigma}_{\text{N}}$. The median estimates of these parameters are shown in Table 3.6 with corresponding 2.5% and 97.5% - quantiles inside brackets. Most model parameters stabilize around $h = 0.5$, and thus increasing the resolution more than this will not change the estimated parameters further. Keep in mind that the spatial hyperparameters, $\hat{\rho}$ and $\hat{\sigma}_{\text{s}}$ are not necessarily the true parameters of the underlying GRF, but the parameters of the GRF approximation.

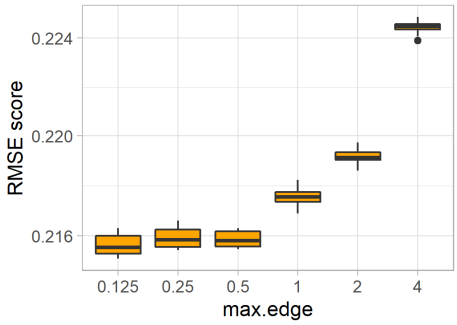
3.4.2 Repeated 10-fold Cross-Validation

For this dataset as well, we have performed a repeated 10-fold cross-validation. In contrast to the Gaussian case, we now need to calculate the CRPS in a discrete way, since we are working with Binomial data. These calculations are described in Equation 2.30 in Section 2.4.

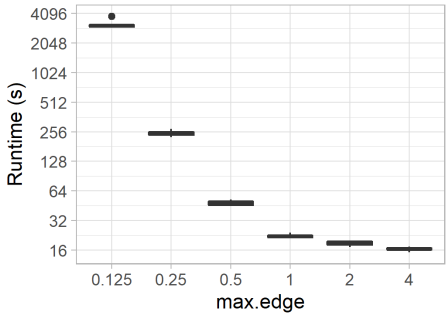
Figure 3.8 shows the results of the repeated cross-validation. Both the CRPS score and the RMSE score start to stabilize at $h = 0.5$, with values of about 0.112 and 0.216



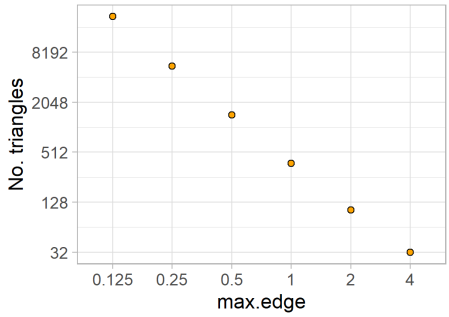
(a) Boxplot of CRPS score CV repetitions.



(b) Boxplot of RMSE score for CV repetitions.



(c) Total CPU estimation time for CV repetitions.

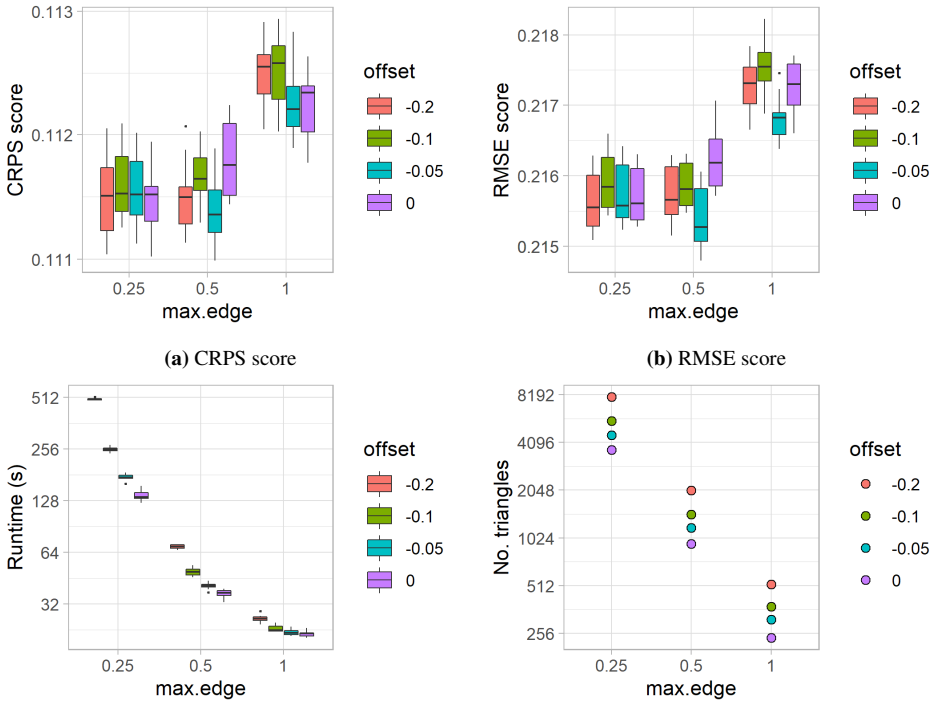


(d) Number of triangles in the mesh for each h.

Figure 3.8: Repeated cross-validation scores for case 2. Varying maximum edge length h with constant offset r = 10%. Running time in seconds and number of triangles in the mesh.

respectively. The running time in Figure 3.8c seems to increase more than linearly, while the number of triangles increases linearly with a factor of four when halving h. The running time in this case only consists of estimation time. The prediction time used to sample predictions to compute CRPS is included in Table A.2 in Appendix. Note again that the running times in Figure 3.8c are larger than the running times shown in Table 3.5 due to the parallellization structure when predicting.

As in the Gaussian case, we have until now only varied the largest allowed edge length, h. We also want to investigate the effect of the boundary offset, and therefore vary this as well. Specifically, we use $h = [1, 0.5, 0.125]$ and $r = [0\%, 5\%, 10\%, 20\%]$. The results are shown in Figure 3.9. Here we see that the best scores with $h = 1$ and $h = 0.5$ are obtained when using $r = 5\%$. For higher mesh resolution there is almost no change. When varying r, the locations of the mesh nodes might change in order for the mesh adapt to the mesh parameters, and thus small changes in the scoring rules can occur because of the mesh node locations rather than the extended boundary.



(a) CRPS score

(b) RMSE score

(c) Total CPU time in seconds.

(d) Number of triangles in mesh.

Figure 3.9: Scores, runtime and number of triangles for combinations of h and r for cross-validation. The order of boxes within each h group coincides in all three plots with r = 20% to the left (red), r = 0% to the right (purple).

3.4.3 Hold-Out Regions

There are 47 counties in our study region, Kenya. As with the Gaussian case, we are interested in seeing how the prediction quality changes with increasing mesh resolution when holding out larger areas at a time. Table 3.7 shows scoring results from holding out one county at a time and predicting the prevalence of secondary education at these locations. For this analysis we use the weighted versions of the scoring rules CRPS and RMSE, which depend on number of locations in the county, as explained in Equation 2.24 and 2.32 in Section 2.4.2.

To check how the offset parameter r influences boundary effects for hold-out regions, we hold out the region Narok and keep a fixed maximum edge length $h = 0.25$ and vary the boundary offset r. Table 3.8 shows that the best scores are obtained using $r = 5\%$, which also gives the lowest runtime.

Table 3.7: Table of prediction scores and runtimes including estimation time and sampling time for computing CRPS, for increasing mesh resolution for hold-out regions analysis.

r	h	CRPS	RMSE	Estimation + Prediction time (s)
10%	4.000	0.121	0.229	18 + 4
	2.000	0.116	0.223	21 + 5
	1.000	0.117	0.224	23 + 5
	0.500	0.115	0.221	54 + 6
	0.250	0.115	0.221	252 + 14
	0.125	**0.115**	**0.220**	3360 + 49

Table 3.8: Table of prediction scores and runtimes including estimation time and sampling time for computing CRPS, for holding out the county Narok.

r	h	CRPS	RMSE	Estimation + Prediction time (s)
0%	0.25	0.096	0.188	129 + 9
5%		0.095	**0.185**	122 + 11
10%		0.095	0.186	183 + 12
20%		**0.095**	0.186	278 + 15

3.4.4 Predictions on Grid

Lastly, we visualize our predictions on a 200×200 grid for different mesh resolutions, given by $h = [1, 0.5, 0.25, 0.125]$ and $r = 10\%$. To be able to visualize the posterior mean, we need information about the covariate "urbanicity" on each of the 40 000 grid locations, i.e., whether or not the particular location is an urban or rural area. Unfortunately, we do not have this information. We therefore choose to visualize the effect that the posterior median of the random field, $\hat{u}_{0.5\text{q}}(\mathbf{s})$, projected onto the 200×200 grid, has on the odds. On each grid location inside the border of Kenya, we plot this median effect, given by

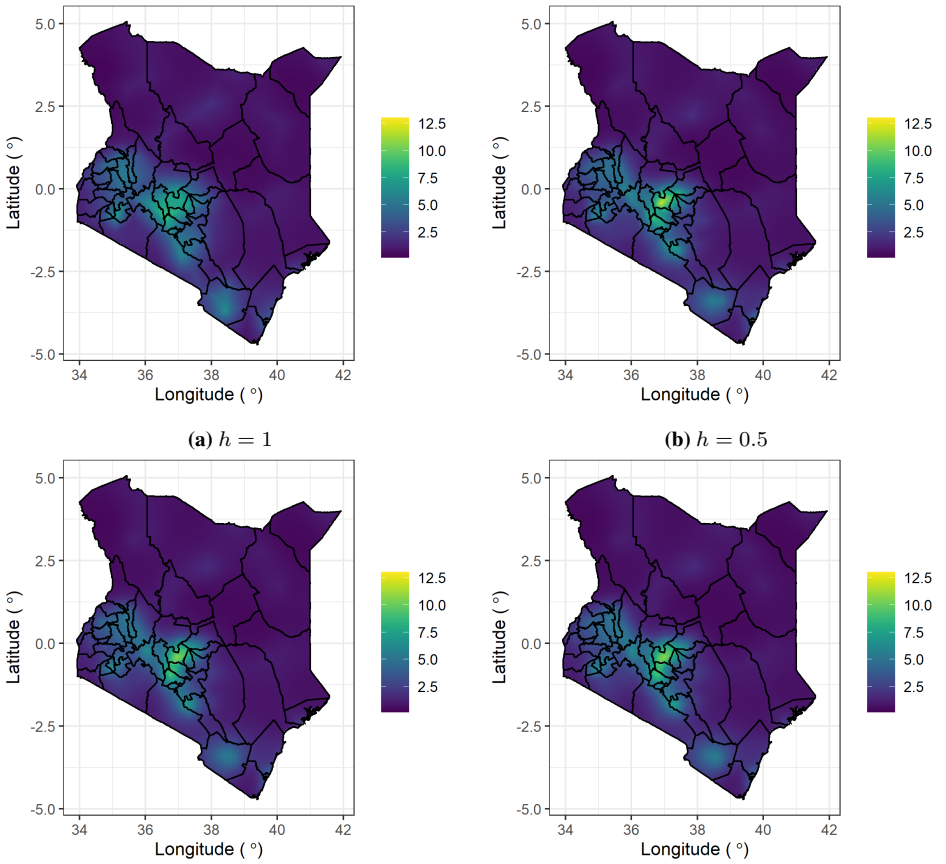
$$\exp\{\hat{u}_{0.5\text{q}}(\mathbf{s})\}. \tag{3.3}$$

The projection is shown in Figure 3.10. We clearly see the connection between higher odds and the observations with high probability in Figure 1.2. For the lowest resolution, we see traces of the mesh structure in the upper middle part of Kenya. For higher resolutions, the estimates are smoother.

In Figure 3.11, four plots of the posterior 95% credibility interval proportion projected onto the 200×200 grid is shown. This proportion is calculated by

$$\frac{\exp\{\hat{u}_{0.975\text{q}}(\mathbf{s})\}}{\exp\{\hat{u}_{0.025\text{q}}(\mathbf{s})\}}, \tag{3.4}$$

where $\hat{u}_{0.975\text{q}}(\mathbf{s})$ and $\hat{u}_{0.025\text{q}}(\mathbf{s})$ denote the upper and lower 95% quantiles of the estimated random field $\hat{u}(\mathbf{s})$ projected onto the grid locations. For the lowest mesh resolution, we see the mesh structure in the upper part of Kenya. For finer resolutions, the estimates are smoother, but they are otherwise similar.



(a) $h = 1$ **(b)** $h = 0.5$

(c) $h = 0.25$ **(d)** $h = 0.125$

Figure 3.10: Median posterior effect on odds for the second case study with different mesh resolutions.

3.4.5 Analysis using Non-Spatial Model

In the analysis of this section, we have seen that the results from running models with different mesh resolutions do not vary that much. The model fit score, WAIC, improves with an absolute value of 73 from the coarsest to the finest mesh, and the CRPS and RMSE scores only have changes in the third decimals. It is therefore tempting to check whether the spatial part of the model really contributes to improving the model fit and predictive power.

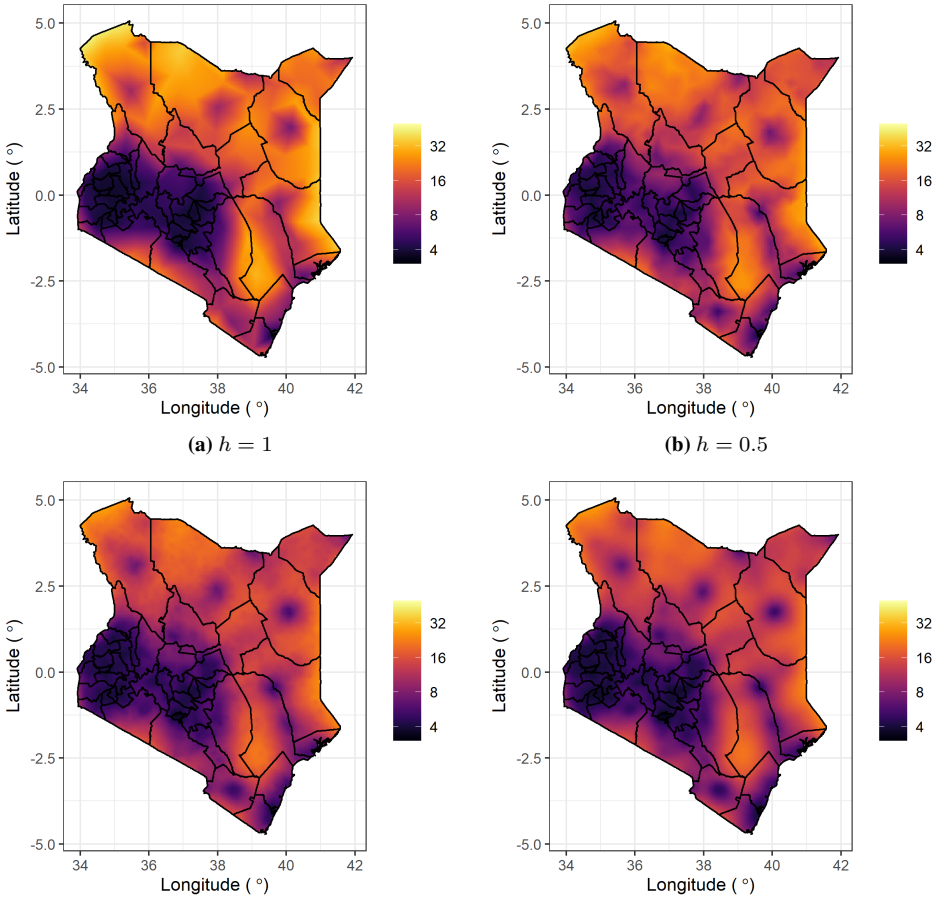
When removing the spatial field from the model in (3.2), the linear predictor η_i becomes

$$\eta_i = \text{logit}(p_i) = \beta_0 + \beta_{\text{URB}}\mathbb{I}_{\text{urban}}(\mathbf{s}_i) + \epsilon_i, \quad i = 1, \ldots, n. \tag{3.5}$$

When using this simplified model, we get a WAIC score of 5009, compared to 4856 for the poorest model with the spatial field included. A lower score means a better fit,

and thus, the non-spatial model noteworthy worse than the spatial model. The parameters when excluding the spatial model is only the intercept β_0, the urbanicity parameter β_{URB} and the nugget parameter σ_{N}. These are estimated to be $\hat{\beta}_0 = -1.63[-1.72, -1.53]$, $\hat{\beta}_{\text{URB}} = 1.18[1.05, 1.32]$ and $\hat{\sigma}_{\text{N}} = 0.92[0.84, 1.00]$ for the non-spatial model. Compared to the spatial model, the non-spatial model has a higher intercept, a higher urbanicity effect and a higher standard deviation. For the 10-fold cross-validation, the average CRPS is 0.128 and the average RMSE is 0.238, which is higher than when including the spatial part. The predictive power of the non-spatial model is therefore worse.

As we expect, running this non-spatial model is more efficient than running the spatial model since the spatial (and computationally heavy) part is excluded. Each run takes about 7 seconds plus a prediction time of 2 seconds in comparison with 48 and 6 seconds for the spatial model with $h = 0.5$ and $r = 10\%$. All in all, we can conclude that the non-spatial model performs worse than the spatial model.



(a) $h = 1$ **(b)** $h = 0.5$

(c) $h = 0.25$ **(d)** $h = 0.125$

Figure 3.11: Projected 95% credibility interval proportion for different mesh resolutions for the second case study.

Chapter 4

Discussion and Recommendations

In this chapter, we discuss the results presented in Chapter 3. From this discussion, we try to gather the insight we have gained on how the mesh resolution influences the results, and how important this influence is. Finally, we formulate practical recommendations on how to construct meshes.

4.1 Discussion

We have seen how different input meshes affect parameter estimates, the model fit and the predictive power for the SPDE model. The questions raised in Section 3.1 will now be discussed in turn. These questions were: How well does the SPDE model with different mesh resolutions fit our data? How does the mesh resolution influence parameter estimates? How does the mesh resolution influence the prediction quality both in terms of accuracy and the computational complexity measured in running time?

Model Fitness

The model fit is measured using the WAIC score. In both the Gaussian and Binomial cases, we see that the WAIC score improves with an increasing mesh resolution, which means that the model fits better to the data for higher mesh resolutions. This is the expected result.

For the Binomial case, this fitness improvement stops for resolutions higher than $h = 0.5$, while for the Gaussian case, the improvement continues for all resolutions, though at a slower pace for higher resolutions. Thus, it seems like there is more to gain by using a very fine mesh in the Gaussian case.

Parameter Estimates

The parameter estimates for the models stabilize at around $h = 0.25$ to $h = 0.5$ for both the Gaussian and Binomial case. Hence, we observe that there is no gain in using a finer mesh than $h = 0.25$ if the parameter estimates are of interest.

The estimated ranges for the models are $\hat{\rho}_{\mathcal{N}} = 6.2$ for the Gaussian case and $\hat{\rho}_{\mathcal{B}} = 1.8$ for the Binomial case, the estimated marginal standard deviations are $\hat{\sigma}_{\mathrm{s},\mathcal{N}} = 0.7$ and $\hat{\sigma}_{\mathrm{s},\mathcal{B}} = 0.9$, and the estimated nugget effects are $\hat{\sigma}_{\mathrm{N},\mathcal{N}} = 0.1$ and $\hat{\sigma}_{\mathrm{N},\mathcal{B}} = 0.7$.

With the estimated ranges above for the two cases, $h = [0.25, 0.5]$ corresponds to $h/\hat{\rho}_{\mathcal{N}} = [0.04, 0.08]$ of the estimated range for the Gaussian case, and $h/\hat{\rho}_{\mathcal{B}} = [0.14, 0.28]$ of the estimated range for the Binomial case. This indicates that there is more to gain by using a finer mesh with continuous data, compared to count data.

Predictive Power

The predictive power is measured by performing a 10-fold cross-validation and a hold-out region analysis for both the Gaussian and Binomial cases. The prediction quality, measured by CRPS and RMSE, improves little for meshes finer than $h = 0.5$. Thus, if the predictions are of interest, there is still no gain in using a mesh finer than this. It seems like the predictive power reaches more or less its maximum when the parameter estimates have converged.

This is natural since the parameters of the SPDE model become the same for all models with a sufficiently fine mesh. One could imagine that a finer mesh, which would yield a smoother finite element representation, would result in more accurate predictions, even though the underlying model was the same. However, we observe that this is not the case and that when the parameter estimates have converged, the finite element approximation is sufficiently precise to yield good predictions.

When predicting on hold-out regions, the CRPS and RMSE for both cases are in general larger than when predicting on randomly chosen folds in the cross-validation. This is natural since there is no information about the field in the entire hold-out region. Thus, the field can not be adapted to nearby observations in this case, and the error becomes larger.

In the previous analyses, we have worked with the same boundary extension, given by $r = 10\%$. We also varied the boundary extension, to investigate if there are any connections between boundary effects and predictive power. This is done both for the 10-fold cross-validation, and for the hold-out region case, where a region on the edge of the domain is held out. For the Gaussian case, we observe a slight improvement for larger extensions when the mesh resolution is low. This improvement is still small compared to the influence of the mesh resolution. Furthermore, for higher mesh resolutions, the size of the boundary extension does not influence the prediction quality. For the Binomial case, there is no clear trend for the influence of the boundary extension size. Thus, it seems like the boundary extension is not that important for the SPDE model, which means that the SPDE model almost does not suffer from boundary effects.

As we have seen, the computational complexity of running the models increases with higher mesh resolutions. For the Gaussian case, the running times increase linearly on the logarithmic scale when halving the maximum edge length h. In the Binomial case, however, it seems to be a higher than a linear relationship on a logarithmic scale between

the runtime and the maximum allowed edge length. The important thing to note is that fine meshes require long runtimes, up to hours for each model run. Thus when we are interested in performing several model runs, like for cross-validation, it is necessary to have computationally feasible runtimes.

General Comments

There can be many reasons why the Gaussian model and the non-Gaussian model differ in how fine the mesh resolution needs to be to obtain optimal results. First, when looking at the datasets, the Binomial case has data with a more clustered structure than the Gaussian case. This can make it difficult to predict in the less clustered part of the domain for the Binomial case, since there is a higher distance between prediction locations and observation locations, both in the 10-fold cross-validation and for hold-out regions.

It also seems to be more variation in the Binomial case, and the higher nugget effect shows that the responses deviate more from the model, compared to the Gaussian case. As presented in the Introduction, Section 1.1, the nugget effect for the Binomial case includes more elements in general than in the Gaussian case. In the Binomial case we only observe a discrete number, y successes out of n trials, and if n is low, $p = \frac{y}{n}$ and $p = \frac{y+1}{n}$ will be different numerical values. However, the Binomial likelihood is rather diffuse for low n, thus the model will not learn that much when observing different $p = \frac{y}{n}$ and $p = \frac{y+1}{n}$. Since the median number of trials n for the prevalence data is low, (≈ 7), we can expect the Binomial likelihood to be confounded with the nugget effect.

In the author's project thesis, (Røste, 2020), we saw that the amount of nugget effect present in the dataset affected at which point it is not necessary to increase mesh resolution any further. It seems like this is the case here as well since the Binomial data have a relatively higher nugget effect, and also requires a less fine mesh for optimal results.

Predictions on Grid

Often, when doing spatial modeling, we are interested in visualizing our results. We want to plot for example the posterior mean and the posterior standard deviation, and visually compare with the observations. We have done this for both the Gaussian and the Binomial case, for different mesh resolutions. For both cases, increasing the mesh resolution gives smoother plots, and we see more of the spatial field. For the posterior standard deviation plots in the Gaussian case, we see the mesh structure for low resolutions clearly. This structure effect is discussed in Section 2.2.2. In addition, higher mesh resolution yields higher standard deviation. For the Binomial case, we have shown the proportion of the 95% posterior credibility interval, thus a large interval yields high values, and therefore shows spread. We do not see the same trends for the Binomial case, which means that the sizes of the intervals do not change for increasing mesh resolution.

Comparing with Results of Righetto et al. (2020)

In the study of Righetto et al. (2020), they use the observation locations as mesh nodes. These observations are generated randomly on a unit square with a response $y_i = 1+2x_i+u_i+\epsilon_i$, where x_i is a uniform covariate, u_i is the SPDE approximation and $\epsilon_i \sim \mathcal{N}(0, 0.3)$

for locations i. In the SPDE approximation, they used $\alpha = 1.5$, which corresponds to a smoothness $\nu = 0.5$, and a log-gamma distribution for the prior on the parameters.

They found that the best meshes in terms of model fitness and prediction quality (MSE) had a cutoff of $c = 0.05$, $c = 0.03$, and $c = 0.01$ for the number of observations $n = 50$, $n = 100$ and $n = 300$ respectively. In addition, they found that for these cutoff values, the best value of h was given by $h = 0.05$. The simulated field had an approximate range of $\rho^* = 0.2$, thus their meshes had a resolution relative to the range, $[c/\rho^*, h/\rho^*]$, given by $[0.25, 0.25]$, $[0.15, 0.25]$ and $[0.05, 0.25]$ respectively, where the lower bound represents the shortest allowed edge length (c) and the upper bound represents the maximal edge length (h).

We have made the meshes independent of the distribution of the observation locations, thus the sizes of the triangles are approximately the same size. In the results we have obtained for the Gaussian case, which is the most similar case to their simulation study, we have found that a maximum edge length of $1/12$ of the spatial range should give the best results. This corresponds to $h = \rho^*/12 = 0.2/12 = 0.017$, which is much smaller than $h = 0.05$, which is the lowest value they used for h in the simulation design. The main challenge is that the chosen values for h are too large so that the maximum edge length h is not bounded by the size of the triangles from above. Instead, the density of the observations, together with the cutoff c, decides how the mesh becomes. Thus, when they find that the value of h does not seem to influence the results very much, we argue that this is because the observation locations, together with c, overrides the effect of h on the mesh.

When defining a shortest edge length, c, they obtain some triangles with shorter edge lengths when locations are located close, and thus their mesh resolution varies over the domain. This is a nice property when all locations one is interested in are know and used to create the mesh since it maintains a good model fit and reduces the overall computational complexity. On the other hand, it might lead to lower predictive power if one wants to predict on unobserved locations that are unknown for the mesh, like with cross-validation. This is why we, in this work, chose to create the mesh independently of the spatial distribution of the locations.

4.2 Recommendations

One of the main goals of this work is to collect the results and formulate recommendations on how to proceed when defining a mesh.

We first note that the range parameter ρ describes how the underlying GRF fluctuates relative to the scale we are on. In the SPDE approach, an approximated effective range is used, corresponding to a Matérn correlation of 0.1 when the distance between two observations is the range ρ. Since the mesh has to be sufficiently fine to accurately model these fluctuations, it makes sense to formulate maximum edge length recommendations relative to ρ. In practice, we therefore have to determine ρ before proceeding with the mesh, where an initial guess of 1/5 of the spatial domain often has been a good starting point.

Through these analyses, we have seen that using a maximum edge length of $h = 0.5$, gives good results with feasible runtime. For the Gaussian case, the estimated range is $\hat{\rho} = 6.2$ which corresponds to a maximum allowed edge length relative to the spatial range of $0.5/\hat{\rho}_{\mathcal{N}} = 0.08 \approx \frac{1}{12}$. In the Binomial case, the estimated range $\hat{\rho} = 1.8$, which

gives $0.5/\hat{\rho}_{\mathcal{B}} = 0.28 \approx \frac{1}{4}$ relative to the spatial range.

These values are only indications of where to start when building the mesh. In general, increasing the mesh resolution improves the results, but if there is a large nugget effect present in the data, it is not certain that a finer mesh improves the results. We recommend that users of the SPDE approach try different meshes based on this work, like halving the maximum edge length h to see if their results change. If there is a significant change when increasing the mesh resolution, one has to consider whether it is worth the running time or not. If there was not a significant change, one should decrease the mesh resolution to save running time.

Chapter 5

Conclusion and Further Work

A thorough analysis of the behaviour of the SPDE approach, for various meshes, has now been performed. In particular, various maximum edge lengths h, and boundary extension sizes r have been applied on a dataset with continuous responses, as well as a dataset with count responses. For the various meshes, we have studied how the model fits the data, the parameter estimates, and the prediction quality of the model.

For both the Gaussian and the Binomial case, increasing the resolution of the mesh improves the model fit and predictive power, and makes the parameter estimates stabilize. However, the model only improves for mesh resolutions up until a certain point, which typically is around $h = 0.5$ for these particular datasets. It turns out that for most applications, there is really no need to have a finer mesh resolution than this. The computational time, on the other hand, keeps increasing for finer mesh resolutions. Thus, it is clear that we do not want to use a finer mesh than necessary.

From the analysis of this work, it seems like a maximum edge length of $h = \rho/12$ for the Gaussian case, and $h = \rho/4$ for the Binomial case is sufficient to obtain more or less optimal results. If computational time is a high priority, good results might be obtained with somewhat coarser meshes as well, but the results indicate that a too coarse mesh will result in a significantly worse model performance. These recommendations are suggestions on where to start when making a mesh. For a different dataset, one can possibly obtain optimal results in terms of model fit and prediction quality at a lower runtime cost than we have found here. It can also be that the model requires a finer mesh than what we have recommended to obtain optimal results.

The effect of varying the boundary extension size r has also been investigated, and the results have indicated that this setting plays an insignificant role on the resulting model performance.

There are several extensions to this work that could be applied. First of all, it would be interesting to do a simulation study based on the estimated parameters from the datasets used in this work. Then, the spatial parameters range ρ and marginal standard deviation σ_{s} would be input parameters to a simulated GRF. The estimated parameters from the approximated SPDE model could then be compared to the true, underlying parameters, and although the SPDE estimates cannot be interpreted as the true GRF parameters, they will generally be close for good approximations.

Furthermore, in a simulation study, it would be possible to vary the input parameters ρ and σ_s for the underlying GRF. By doing this, the mesh resolution could be analyzed for a wider range of different data types. This way, it would be possible to verify if the recommendations of this work hold, particularly for various ranges ρ.

This could also be analyzed more thoroughly in real-world applications, by studying a wide range of different datasets. By looking at different types of data for different areas, the true, underlying parameters would likely cover a higher range. This way, it would be easier to generalize the results and give mesh recommendations based on a higher collection of results.

Bibliography

H. Bakka, H. Rue, G.-A. Fuglstad, A. Riebler, D. Bolin, J. Illian, E. Krainski, D. Simpson, and F. Lindgren. Spatial Modeling with R-INLA: A Review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(6):e1443, 2018.

G.-A. Fuglstad and J. Beguin. Environmental Mapping using Bayesian Spatial Modelling (INLA/SPDE): A Reply to Huang et al.(2017). *The Science of the Total Environment*, 624:596, 2018.

G.-A. Fuglstad, D. Simpson, F. Lindgren, and H. Rue. Does Non-Stationary Spatial Data Always Require Non-Stationary Random Fields? *Spatial Statistics*, 14:505–531, 2015.

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

A. Gelman, J. Hwang, and A. Vehtari. Understanding Predictive Information Criteria for Bayesian Models. *Statistics and Computing*, 24(6):997–1016, 2014.

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

T. Gneiting, F. Balabdaoui, and A. E. Raftery. Probabilistic Forecasts, Calibration and Sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

D. A. Hastings, P. K. Dunbar, G. M. Elphingstone, M. Bootz, H. Murakami, H. Maruyama, H. Masaharu, P. Holland, J. Payne, N. A. Bryant, et al. The Global Land One-Kilometer Base Elevation (GLOBE) Digital Elevation Model, Version 1.0. *National Oceanic and Atmospheric Administration, National Geophysical Data Center*, 325:80305–3328, 1999.

Kenya National Bureau of Statistics, Ministry of Health/Kenya, National AIDS Control Council/Kenya, Kenya Medical Research Institute, and National Council for Population and Development/Kenya. *Kenya Demographic and Health Survey 2014*. Rockville, MD, USA, 2015. Available at `http://dhsprogram.com/pubs/pdf/FR308/FR308.pdf`.

E. T. Krainski, V. Gómez-Rubio, H. Bakka, A. Lenzi, D. Castro-Camilo, D. Simpson, F. Lindgren, and H. Rue. *Advanced Spatial Modeling with Stochastic Partial Differential Equations using R and INLA*. CRC Press, Boca Raton. Florida., 2018.

K. Kristensen, A. Nielsen, C. W. Berg, H. Skaug, and B. M. Bell. TMB: Automatic Differentiation and Laplace Approximation. *Journal of Statistical Software, Articles*, 70(5):1–21, 2016. ISSN 1548-7660. doi: 10.18637/jss.v070.i05. URL `https://www.jstatsoft.org/article/view/v070i05`.

F. Lindgren, H. Rue, and J. Lindström. An Explicit Link Between Gaussian Fields and Gaussian Markov Random Fields: The Stochastic Partial Differential Equation Approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.

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

J. Paige, G.-A. Fuglstad, A. Riebler, and J. Wakefield. Design and Model-Based Approaches to Small-Area Estimation in a Low and Middle Income Country Context: Comparisons and Recommendations. *arXiv preprint arXiv:1910.06512*, 2019.

A. J. Righetto, C. Faes, Y. Vandendijck, and P. J. Ribeiro Jr. On the Choice of the Mesh for the Analysis of Geostatistical Data using R-INLA. *Communications in Statistics-Theory and Methods*, 49(1):203–220, 2020.

H. Rue and L. Held. *Gaussian Markov Random Fields: Theory and Applications*. CRC Press, Boca Raton. Florida., 2005.

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian Inference for Latent Gaussian Models by using Integrated Nested Laplace Approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009.

J. Røste. Investigating the Influence of Mesh Resolution of the SPDE Approach for Gaussian Random Fields in 1D. 2020.

J. P. Snyder. *Map Projections–A Working Manual*, volume 1395. US Government Printing Office, 1987.

S. Watanabe and M. Opper. Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research*, 11(12), 2010.

Appendix

A.1 Additional Results for Case 1

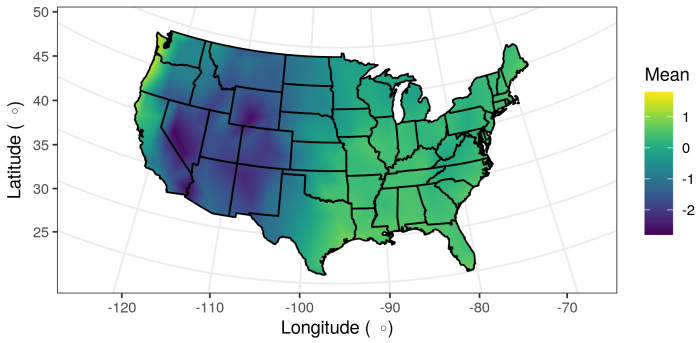
Table A.1: Table with results from 10-fold CV for the first dataset.

r	h	CRPS	RMSE	Estimation time + Prediction time (s)
10%	8.000	0.140	0.265	12 + 7
	4.000	0.118	0.225	14 + 9
	2.000	0.098	0.183	32 + 19
	1.000	0.084	0.157	125 + 36
	0.500	0.081	0.152	274 + 75
	0.250	0.080	0.151	1365 + 277
	0.125	0.080	0.150	7537 + 1201

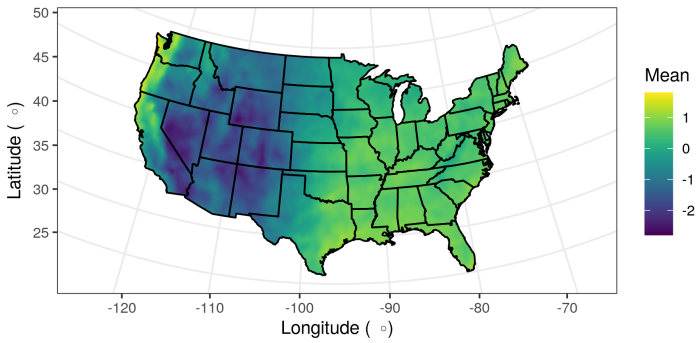
A.2 Additional Results for Case 2

Table A.2: Table with results from 10-fold CV for the second dataset.

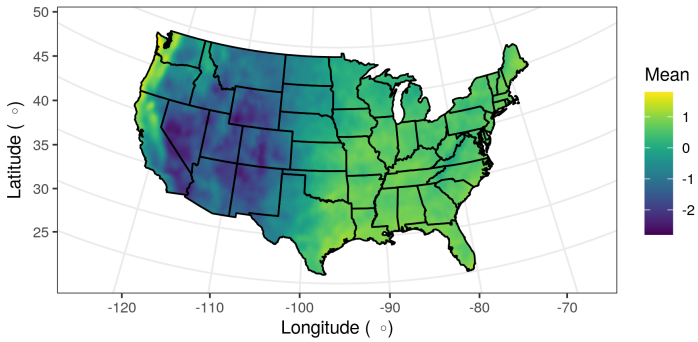
r	h	CRPS	RMSE	Estimation time + Prediction time (s)
10%	4.0000	0.117	0.224	16 + 4
	2.0000	0.114	0.219	19 + 4
	1.0000	0.113	0.218	22 + 4
	0.5000	0.112	0.216	48 + 6
	0.2500	0.112	0.216	247 + 14
	0.1250	0.112	0.216	3128 + 46



(a) Prediction on grid of size 400×200. Parameter $h = 4$.

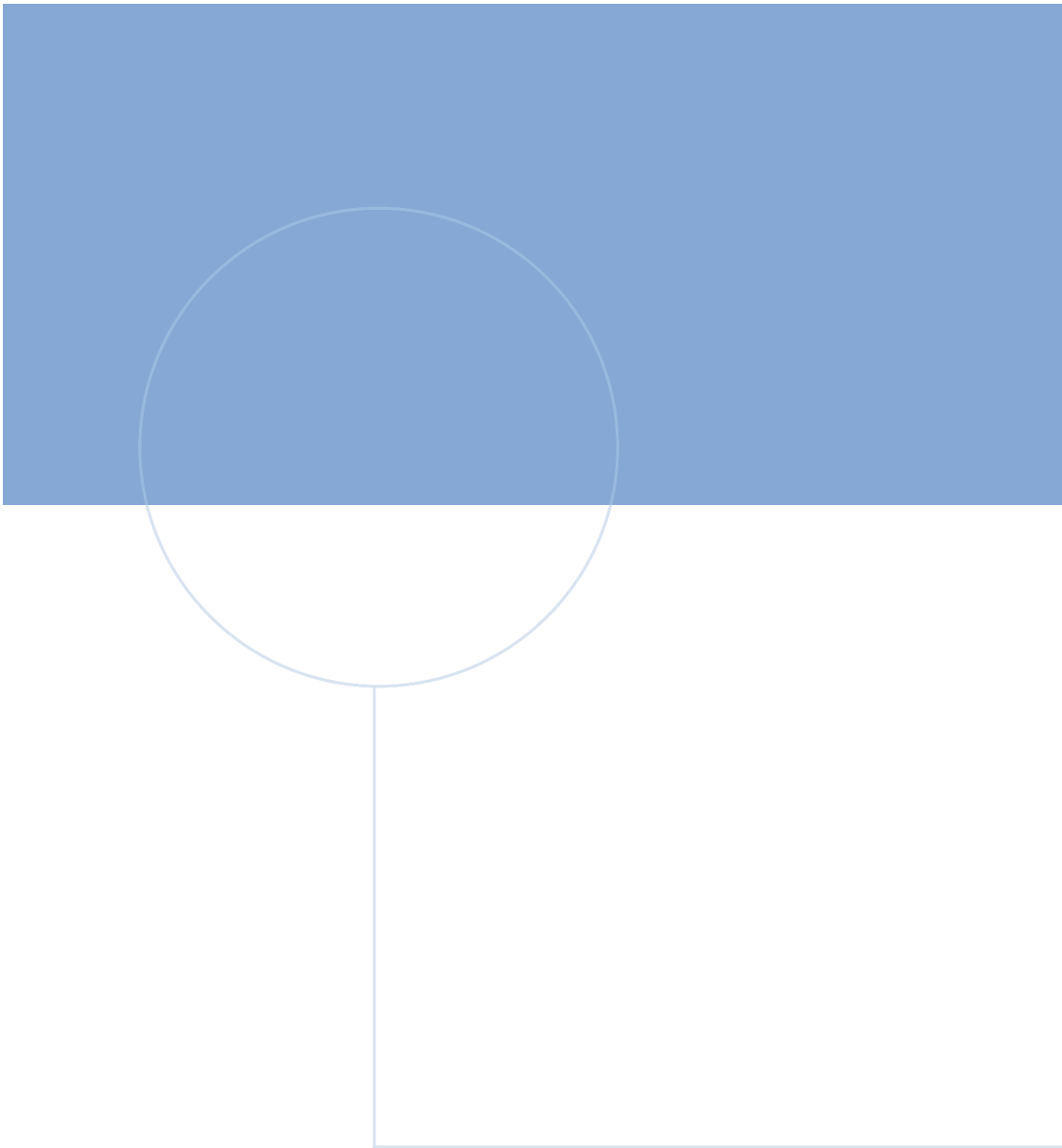


(b) Prediction on grid of size 400×200. Parameter $h = 1$.



(c) Prediction on grid of size 400×200. Parameter $h = 0.25$.

Figure A.1: Projection of random field onto 400×200 grid for case study 1.



NTNU

Norwegian University of
Science and Technology