Contents

iv

Symbols

$\pi(\cdot)$	Density of its arguments
σ	Standard deviation
Var	Variance
Cov	Covariance
Prec	Precision
Corr	Correlation
$\mathbb{E}$	Expected value
$\mathbb{P}$	Probability
$\int$	Integral of its arguments
$\sum$	Sum of its arguments
$\prod$	Product of its arguments
$\exp$	Exponential function
$\log$	Logarithmic function
∂	The derivative
$\propto$	Proportional to
$\mathbb{R}$	Real numbers
$\mathbb{N}$	Natural numbers
$\mathbb{N}_0$	Natural numbers including 0
$\boldsymbol{I}$	Identity matrix

1

1 Introduction

2

2 Corona Virus

Viral diseases continue to pose a serious public health threat. Several viral epidemics have occurred in the last 20 years, including the SARS pandemic in 2002/3, H1N1 influenza in 2009, and more recently the Middle East Respiratory Syndrome Coronavirus (MERS-CoV), which was first detected in Saudi Arabia in 2012.
In late 2019, the first few cases of lower respiratory infections were detected in Wuhan, China. In February 2020, this viral disease was officially named "Covid-19", an acronym for "Coronavirus Disease 2019".
Due to the rapid spread of the virus, a Public Health Emergency of International Concern was declared at the end of January 2020, with 18 countries reporting cases and four countries reporting human-to-human transmission.
At the end of February 2020, the World Health Organisation (WHO) raised the risk of a Covid 19 epidemic to "very high" before declaring it a pandemic on 11 March. At that time, more than 118,000 cases in 114 countries and 4000 deaths had already been registered.
The first cases of the disease were linked to direct exposure at the Huanan Seafood Wholesale Market in Wuhan, with animal-to-human transmission suspected as the main mechanism. After subsequent cases could not be linked to this mechanism, human-to-human transmission was presumed to be the main transmission mechanism. Furthermore, sympotomatic individuals are thought to be the most common source of covid-19 spread. However, asymptomatic individuals can also transmit the virus, therefore isolation is the best way to contain this epidemic.
Similar to other respiratory diseases, e.g. influenza, transmission is thought to occur through respiratory droplets (particles $> 5 - 10\mu m$ in diameter) when coughing and sneezing. In closed rooms, transmission by aerosol is also possible.
Based on the data from the first cases in Wuhan, the incubation period is generally between 3 and 7 days, with a median of 5.1 days. According to the data, the number of infections doubled about every seven days and the basic reproductive number R is 2.2, which means that on average each infected individual infects another 2.2 individuals.
According to a report by the Chinese Centre for Disease Control, which studied 72,314 cases, the overall mortality rate of confirmed cases was 2.3%, with most of the fatal cases affecting people over 70 years of age.

Furthermore, the clinical manifestations of the disease can be divided into three groups according to their severity:

- Mild disease: non-pneumonia and mild pneumonia; this occurred in 81% of cases.
- Severe disease: dyspnea, respiratory rate ≥ 30 min, blood oxygen level $\leq 93\%$; this occurred in 14% of cases.
- Critical disease: respiratory failure, septic shock and/or multiple organ dysfunction or failure; this occurred in 5% of cases.

Subsequent reports indicate that the disease is asymptomatic or with very mild symptoms in 70% of patients, while the remaining 30% develop a respiratory syndrome with high fever, cough and even severe respiratory failure, which may require admission to the intensive care unit.
Most countries use some kind of clinical and epidemiological information to determine who should be tested. A molecular test, for example a PCR test, can be used to detect the disease.
The WHO recommends the collection of samples from both the upper and lower respiratory tract. In the laboratory, the genetic material extracted from the saliva or mucus sample is amplified by reverse polymerase chain reaction (RT-PCR), which synthesises a double-stranded DNA molecule from an RNA form. Once the genetic material is sufficient, the parts of the genetic code of the CoV that are conserved are searched for. The probes used are based on the original gene sequence published by the Shanghai Public Health Clinical Center & School of Public Health, Fudan University, Shanghai, China on Virological.org and subsequent confirmatory evaluation by other laboratories [Cas+21].

3 Introduction to Bayesian Inference

Bayesian inference is a branch of statistics that uses the Bayesian concept of probability and Bayes' theorem to investigate questions of stochastics.
Characteristic for Bayesian statistics is the consistent use of probability distributions or marginal distributions, whose form conveys the accuracy of the procedures or reliability of the data and the procedure.
The Bayesian concept of probability does not presuppose infinitely repeatable random experiments, so that Bayesian methods can be used even with small data sets. A small amount of data leads to a broad probability distribution, which is not strongly localised.
In the Bayesian approach, the parameters of interest are treated as random variables that are governed by their parameters, for instance the mean and standard deviation, and distributions.
Bayesian inference is an essential technique in mathematical statistics and the polar opposite of the frequentist approach, in which a hypothesis is tested without being assigned a probability.
In the Bayesian approach a *prior* distribution $\pi\left(\boldsymbol{\theta}, \sigma^2\right)$ is introduced as part of the model. This distribution is intended to express a state of knowledge or ignorance about $\boldsymbol{\theta}$ and σ^2 prior to obtaining the data. Using the prior distribution, the likelihood function $\pi\left(\boldsymbol{y}|\boldsymbol{\theta}, \sigma^2\right)$, and the observed data $\boldsymbol{y}$, it is possible to calculate the probability distribution $\pi\left(\boldsymbol{\theta}, \sigma^2\middle|\, \boldsymbol{y}\right)$ of $\boldsymbol{\theta}$ and σ^2 given the data $\boldsymbol{y}$. This distribution is called the *posterior* distribution of $\boldsymbol{\theta}$ and σ^2 and is used to make inferences about the parameters [BT11, p. 6].

3.1 Preliminaries

This work follows strict notation rules to easily represent different elements such as matrices or graphs and contains frequently used abbreviations. These and some other basic concepts used in this work are introduced below. The notation follows the one used by Rue and Held [RH05, pp. 14–19].

3.1.1 Matrices and Vectors

Vectors and matrices are indicated by bold notation, such as $\boldsymbol{x}$ and $\boldsymbol{A}$. The transpose of $\boldsymbol{A}$ is denoted by $\boldsymbol{A}^T$. The element in the *i*th row and *j*th column of $\boldsymbol{A}$ is referenced by A_{ij}. This notation is also used for vectors and x_i denotes the *i*th element of a vector. The vector $(x_i, x_{i+1}, ..., x_j)^T$ is abbreviated to $\boldsymbol{x}_{i:j}$. If the columns $\boldsymbol{A}_1, \boldsymbol{A}_2, ..., \boldsymbol{A}_m$ of a $n \times m$ matrix $\boldsymbol{A}$ are stacked on top of each other, this is denoted by $\text{vec}(\boldsymbol{A}) = \left(\boldsymbol{A}_1^T, \boldsymbol{A}_2^T, ..., \boldsymbol{A}_m^T\right)$. Deleting rows and/or columns from $\boldsymbol{A}$ creates a *submatrix*. If a submatrix of a $n \times n$ matrix $\boldsymbol{A}$ can be obtained by removing rows and columns of the same index, it is called a *principal submatrix*. If this matrix can be obtained by deleting the last $n - r$ rows and columns, it is called a *leading principal submatrix* of **A**.
A diagonal $n \times n$ matrix $\boldsymbol{A}$ is denoted by $\text{diag}(\boldsymbol{A})$ and has the following structure:

$$\text{diag}(\boldsymbol{A}) = \begin{pmatrix} A_{11} & & \\ & \ddots & \\ & & A_{nn} \end{pmatrix}.$$

The identity matrix is denoted by $\boldsymbol{I}$.
If $A_{ij} = 0$ for $i < j$ or $A_{ij} = 0$ where $i > j$, then $\boldsymbol{A}$ is called *upper triangular* and *lower triangular* respectively. The *bandwidth* of a matrix $\boldsymbol{A}$ is defined as $\max\{|i - j| : A_{ij} \neq 0\}$. The *lower bandwidth* is given by $\max\{|i - j| : A_{ij} \neq 0 \text{ and } i > j\}$. $|\boldsymbol{A}|$ denotes the *determinant* of a $n \times n$ matrix $\boldsymbol{A}$ and is equal to the product of the eigenvalues of $\boldsymbol{A}$. The *rank* of $\boldsymbol{A}$, referenced by $\text{rank}(\boldsymbol{A})$, is the number of linearly independent rows or columns of $\boldsymbol{A}$. The sum of the diagonal elements is called *trace* of $\boldsymbol{A}$, $\text{trace}(\boldsymbol{A}) = \sum_i A_{ii}$.
Finally, '$\odot$' denotes the element-wise multiplication of two matrices of size $n \times m$, '$\oslash$' denotes the element-wise division and raising each element of a matrix $\boldsymbol{A}$ to a scalar power uses the symbol '$\circledcirc$' [RH05, pp. 14–15].

3.1.2 Lattice and Torus

$\mathcal{I}_{\boldsymbol{n}}$ denotes a (regular) **lattice** (or grid) of size $\boldsymbol{n} = (n_1, n_2)$ (in the two-dimensional case). $\boldsymbol{y}$ can take values on $\mathcal{I}_{\boldsymbol{n}}$ and $y_{i,j}$ denotes the value of $\boldsymbol{y}$ at location ij, for $i = 1, ..., n_1$ and $j = 1, ..., n_2$. For easier reading this will be shortened to y_{ij}. On an *infinite lattice* $\mathcal{I}_{\boldsymbol{\infty}}$, ij are numbered as $i = 0, \pm 1, \pm 2, ...$, and $j = 0, \pm 1, \pm 2, ...$
A lattice with cyclic or toroidal boundary conditions is referred to as *torus* and is denoted by $\mathcal{I}_{\boldsymbol{\infty}}$. The dimension is $\boldsymbol{n} = (n_1, n_2)$ (in the two-dimensional case) and all indices are modulus $\boldsymbol{n}$ and run from 0 to $n_1 - 1$ or $n_2 - 1$. If a GMRF $\boldsymbol{y}$ is defined on $\mathcal{I}_{\boldsymbol{n}}$, the toroidal boundary conditions imply that y_{-2,n_2} is equal to $y_{n_1-2,0}$ since $-2 \mod n_1$ is equal to $n_1 - 2$ and $n_2 \mod n_2$ is equal to 0.
An *irregular lattice* refers to a spatial configuration of regions $i = 1, ..., n$ where the regions (mostly) have common boundaries, for instance the states of a nation [RH05, pp. 15–16]. An example of an irregular lattice is shown in Figure 3.1.

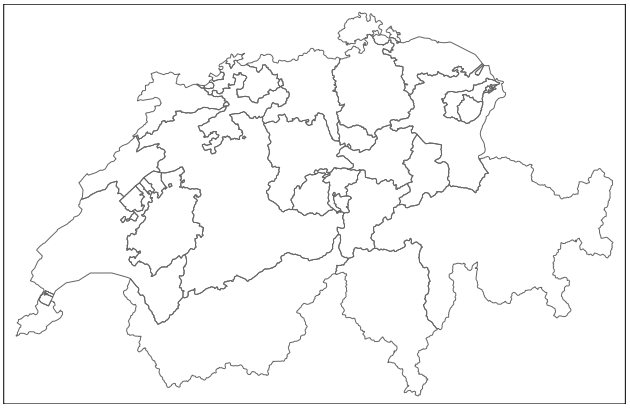

Fig. 3.1: The cantons of Switzerland, an example of an irregular lattice.

3.1.3 General Notation and Abbreviations

For $C \in \mathcal{I} = \{1, ..., n\}$ let $\boldsymbol{y}_C = \{y_i : i \in C\}$. $-C$ denotes the set $\mathcal{I} - \mathcal{C}$ such that $\boldsymbol{y}_{-C} = \{y_i : i \notin C\}$. For two sets A and B, $A \setminus B = \{i : i \in A \text{ and } i \notin B\}$.
$\pi(\cdot)$ denotes the density of its arguments, for example $\pi(\boldsymbol{y})$ for the density of $\boldsymbol{y}$ and $\pi(\boldsymbol{y}_A|\boldsymbol{y}_{-A})$ for the conditional density of $\boldsymbol{y}_A$, given $\boldsymbol{y}_{-A}$. '$\sim$' is used when a variable is 'distributed' according to the law l [RH05, p. 16].

3.1.4 Symmetric Positive Definite Matrices

An $n \times n$ matrix $\boldsymbol{A}$ is *positive definite* exactly if

$$\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} > 0, \qquad \forall \boldsymbol{x} \neq \boldsymbol{0}.$$

If $\boldsymbol{A}$ is also symmetric, it is called a symmetric positive definite (SPD) matrix. Only SPD matrices are considered and sometimes the notation '$\boldsymbol{A} > 0$' is used for an SPD matrix $\boldsymbol{A}$.
An SPD matrix $\boldsymbol{A}$ has some of the following properties.

1. $\text{rank}\,(\boldsymbol{A}) = n$.
2. $|\boldsymbol{A}| > 0$.
3. $A_{ii} > 0$.
4. $A_{ii}A_{jj} - A_{ij}^2 > 0$, for $i \neq j$.
5. $A_{ii} + A_{jj} - 2|A_{ij}| > 0$ for $i \neq j$.
6. $\max A_{ii} > \max_{i \neq j} |A_{ij}|$.
7. $\boldsymbol{A}^{-1}$ is SPD.
8. All principal submatrices of $\boldsymbol{A}$ are SPD.

If $\boldsymbol{A}$ and $\boldsymbol{B}$ are SPD, $\boldsymbol{A} + \boldsymbol{B}$ is also SPD, but the reverse is generally not true. Additionally, if $\boldsymbol{AB} = \boldsymbol{BA}$, $\boldsymbol{AB}$ is SPD.
The following conditions are all sufficient and necessary for a symmetric matrix $\boldsymbol{A}$ to be SPD:

1. All eigenvalues $\lambda_1, ..., \lambda_n$ of $\boldsymbol{A}$ are strictly positive.
2. There exists such a matrix $\boldsymbol{C}$ that $\boldsymbol{A} = \boldsymbol{C}\boldsymbol{C}^T$. If $\boldsymbol{C}$ is lower triangular, it is called the *Cholesky triangle* of $\boldsymbol{A}$.
3. All leading principal submatrices have strictly positive determinants.

A sufficient, but not necessary condition for a (symmetrical) matrix to be SPD is the criterion of *diagonal dominance*:

$$A_{ii} - \sum_{j:j \neq i} |A_{ij}| > 0, \qquad \forall i.$$

A $n \times n$ matrix $\boldsymbol{A}$ is called a *symmetric positive semidefinite* (SPSD) matrix. An SPSD matrix $\boldsymbol{A}$ is sometimes denoted '$\boldsymbol{A} \geq 0$' [RH05, pp. 18–19].

3.2 Basic Concepts of Bayesian Theory

3.2.1 Bayes' Theorem

At the heart of Bayesian inference is *Bayes' theorem*, which describes the probability of an event given prior knowledge of factors that might influence the event.
Let $\boldsymbol{y}^T = (y_1, ..., y_n)$ be a vector of n observations whose probability distribution $\pi(\boldsymbol{y}|\boldsymbol{\theta})$ depends on the values of k parameters $\boldsymbol{\theta}^T = (\theta_1, ..., \theta_k)$. Let $\pi(\boldsymbol{\theta})$ be the probability distribution of $\boldsymbol{\theta}$. Then

$$\pi(\boldsymbol{y}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{y}, \boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|\boldsymbol{y})\,\pi(\boldsymbol{y}). \tag{3.1}$$

Given the observed data $\boldsymbol{y}$, the conditional distribution of $\boldsymbol{\theta}$ is

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})\,\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{y})}. \tag{3.2}$$

This last statement is known as Bayes' theorem [Bay63]. The *prior* distribution $\pi(\boldsymbol{\theta})$ contains knowledge about $\boldsymbol{\theta}$ without knowledge of the data. $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ contains what is known about $\boldsymbol{\theta}$ given knowledge of the data and is the *posterior* distribution of $\boldsymbol{\theta}$ given $\boldsymbol{y}$.
If $\pi(\boldsymbol{y}|\boldsymbol{\theta})$ is considered as a function of $\boldsymbol{\theta}$ instead of $\boldsymbol{y}$, it is called the *likelihood function* of $\boldsymbol{\theta}$ given $\boldsymbol{y}$ and can be written as $l(\boldsymbol{\theta}|\boldsymbol{y})$. Thus Bayes' theorem can be written as

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto l(\boldsymbol{\theta}|\boldsymbol{y})\,\pi(\boldsymbol{\theta}). \tag{3.3}$$

It is evident that the posterior distribution of $\boldsymbol{\theta}$ given the data $\boldsymbol{y}$ is proportional to the product of the distribution of $\boldsymbol{\theta}$ prior to observing the data and the likelihood function of $\boldsymbol{\theta}$ given $\boldsymbol{y}$. Therefore,

$$\text{posterior distribution} \propto \text{likelihood} \times \text{prior distribution}.$$

The data $\boldsymbol{y}$ modifies the prior knowledge of $\boldsymbol{\theta}$ through the likelihood function, and thus can be be regarded as a representation of the information about $\boldsymbol{\theta}$ derived from the data [BT11].

3.2.2 Conditional Independence

In probability theory, two random variables x and y are *independent* given a third variable z if and only if the occurrence of x and y in their conditional probability

distribution given z are independent events. To calculate the conditional density of $\boldsymbol{x}_A$, given $\boldsymbol{x}_{-A}$, the following statement will repeatedly be used,

$$\pi\left(\boldsymbol{x}_A|\boldsymbol{x}_{-A}\right) = \frac{\pi\left(\boldsymbol{x}_A, \boldsymbol{x}_{-A}\right)}{\pi\left(\boldsymbol{x}_{-A}\right)} \propto \pi\left(\boldsymbol{x}\right). \tag{3.4}$$

It follows that x and y are independent precisely when $\pi\left(x, y\right) = \pi\left(x\right)\pi\left(y\right)$, which is expressed by $x \perp y$. x and y are conditionally independent for a given z if and only if $\pi\left(x, y|z\right) = \pi\left(x|z\right)\pi\left(y|z\right)$ [Daw79]. The conditional independence can be easily validated with the help of the following *factorisation criterion*,

$$x \perp y|z \Longleftrightarrow \pi\left(x, y, z\right) = f\left(x, z\right) g\left(y, z\right), \tag{3.5}$$

for some functions f and g, and for all z with $\pi\left(z\right) > 0$ [RH05, pp. 16–17].

3.2.3 Undirected Graphs

Undirected graphs are used to represent the conditional independence structure in a Gaussian Markov random field. An *undirected graph* $\mathcal{G}$ is defined as a tuple $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ contains all nodes in the graph and $\mathcal{E}$ is the set of edges $\{i, j\}$, with $i, j \in \mathcal{V}$ and $i \neq j$. For $\{i, j\} \in \mathcal{E}$ there exists an undirected edge from node i to node j in the other case such an edge does not exist. If $\{i, j\} \in \mathcal{E}\ \forall i, j \in \mathcal{V}$ with $i \neq j$ a graph is *fully connected*. Most often $\mathcal{V} = \{1, 2, ..., n\}$ will be assumed, which is referred to as *labelled*. A simple example of an undirected graph is shown in Figure 3.2.

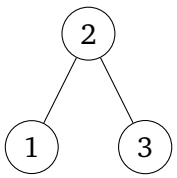

Fig. 3.2: An undirected labelled graph with 3 nodes, $\mathcal{V} = \{1, 2, 3\}$ and $\mathcal{E} = \{\{1, 2\}\{2, 3\}\}$.

The *neighbours* of node i are defined as all nodes in $\mathcal{G}$ with an edge to node i,

$$\text{ne}\left(i\right) = \{j \in \mathcal{V} : \{i, j\} \in \mathcal{E}\}.$$

This definition can be extended to a set $A \subset \mathcal{V}$, where the neighbours of A are defined as

$$\text{ne}(A) = \bigcup_{i \in A} \text{ne}(i) \setminus A.$$

A *path* from i_1 to i_m is defined as a sequence of certain nodes in $\mathcal{V}, i_1, i_2, ..., i_m$, for which $(i_j, i_{j+1}) \in \mathcal{E}$ for $j = 1, ..., m-1$. Two nodes $i \notin C$ and $j \notin C$ are *separated* by a subset $C \subset \mathcal{V}$, if every path from i to j contains at least one node from C. Two disjoint sets $A \subset \mathcal{V} \notin C$ and $B \subset \mathcal{V} \notin C$ are separated by C, if all $i \in A$ and $j \in B$ are separated by C, that is, it is not possible to "wander" on the graph from somewhere in A and end somewhere in B without crossing C.
If i and j are neighbours in $\mathcal{G}$, this can be expressed by $i \overset{\mathcal{G}}{\sim} j$ or $i \sim j$ for the case where the graph is implicit. It follows that $i \sim j \iff j \sim i$.
Let A be a subset of $\mathcal{V}$. A *subgraph* $\mathcal{G}^A$ is a graph restricted to A, i.e., the graph obtained after removing all nodes that do not belong to A and all edges where at least one node does not belong to A. $\mathcal{G}^A = \left\{\mathcal{V}^A, \mathcal{E}^A\right\}$, where $\mathcal{V}^A = A$ and

$$\mathcal{E}^A = \{\{i, j\} \in \mathcal{A} \text{ and } \{i, j\} \in A \times A\}.$$

Let $\mathcal{G}$ be the graph in Figure 3.2 and $\mathcal{A} = \{2, 3\}$, then $\mathcal{V}^A = \{2, 3\}$ and $\mathcal{E}^A = \{\{2, 3\}\}$ [RH05, pp. 17–18].

3.2.4 The Exponential Family

In statistics and probability theory, the *exponential family* is a parametric set of probability distributions of a specific form. The distribution of a random variable $\boldsymbol{y}$ belongs to the exponential family if the discrete or continuous density with respect to a σ-finite measure of $\boldsymbol{y}$ has the form

$$f(\boldsymbol{y}|\boldsymbol{\theta}, \lambda) = \exp\left(\frac{\boldsymbol{y}^T\boldsymbol{\theta} - b(\boldsymbol{\theta})}{\lambda} + c(\boldsymbol{y}, \lambda)\right), \tag{3.6}$$

with $c(\boldsymbol{y}, \lambda) \geq 0$. $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^q$ is the *natural* or *canonical* parameter of the exponential family, while $\lambda > 0$ is a *dispersion* or *nuisance* parameter [HL81]. The natural parameter space Θ is the set of all $\boldsymbol{\theta}$ satisfying

$$0 < \int \exp\left(\left(\boldsymbol{y}^T\boldsymbol{\theta} - b(\boldsymbol{\theta})\right)/\lambda + c(\boldsymbol{y}, \lambda)\right) d\boldsymbol{y} < \infty \tag{3.7}$$

Moreover, $b(\boldsymbol{\theta})$ is a twice differentiable function and all moments of $\boldsymbol{y}$ exist. Specifically,

$$\mathbb{E}_{\boldsymbol{\theta}}(\boldsymbol{y}) = \mu(\boldsymbol{\theta}) = \quad \frac{\partial b(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \tag{3.8}$$

$$\text{Cov}_{\boldsymbol{\theta}}(\boldsymbol{y}) = \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \lambda \frac{\partial^2 b(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}. \tag{3.9}$$

The covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is positive definite in Θ^0, therefore $\mu : \Theta^0 \to M = \mu\left(\Theta^0\right)$ is injective. By substituting the inverse function $\boldsymbol{\theta}(\mu)$ into $\frac{\partial^2 b(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$, the variance function

$$v(\mu) = \frac{\partial^2 b(\boldsymbol{\theta}(\mu))}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \tag{3.10}$$

is given and the covariance can be written as

$$\text{Cov}_{\boldsymbol{\theta}}(\boldsymbol{y}) = \lambda v(\mu). \tag{3.11}$$

Important members of the exponential family are the normal, binomial, Poisson, gamma and multivariate normal distribution [FT13, p. 433].

3.2.5 The Multivariate Normal Distribution

The density of a normally distributed random variable $\boldsymbol{y} = (y_1, ..., y_n)^T, n < \infty$ with mean vector $\boldsymbol{\mu}$ $(n \times 1)$ and SPD covariance matrix $\boldsymbol{\Sigma}$ $(n \times n)$ is

$$\pi(\boldsymbol{y}) = (2\pi)^{-n/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2} (\boldsymbol{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu})\right), \ \boldsymbol{y} \in \mathbb{R}^n \tag{3.12}$$

Here, $\mu_i = \mathbb{E}[y_i]$, $\Sigma_{ij} = \text{Cov}(y_i, y_j)$, $\Sigma_{ii} = \text{Var}(y_i) > 0$ and $\text{Corr}(y_i, y_j) = \Sigma_{ij} / (\Sigma_{ii} \Sigma_{jj})^{1/2}$. This is written as $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. For $n = 1$, $\mu = 0$ and $\Sigma_{11} = 1$ the standard normal distribution is obtained.
$\boldsymbol{y}$ is now split up into $\boldsymbol{y} = \left(\boldsymbol{y}_{\boldsymbol{A}}^T, \boldsymbol{y}_{\boldsymbol{B}}^T\right)$ and $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are divided accordingly:

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_{\boldsymbol{A}} \\ \boldsymbol{\mu}_{\boldsymbol{B}} \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{AA}} & \boldsymbol{\Sigma}_{\boldsymbol{AB}} \\ \boldsymbol{\Sigma}_{\boldsymbol{BA}} & \boldsymbol{\Sigma}_{\boldsymbol{BB}} \end{pmatrix}.$$

Some basic properties of the multivariate normal distribution are the following.

1. $\boldsymbol{y}_{\boldsymbol{A}} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{A}}, \boldsymbol{\Sigma}_{\boldsymbol{AA}})$.
2. $\boldsymbol{\Sigma}_{\boldsymbol{AB}} = \boldsymbol{0}$ precisely when $\boldsymbol{y}_{\boldsymbol{A}}$ and $\boldsymbol{y}_{\boldsymbol{B}}$ are independent.

3. The conditional distribution $\pi\left(\boldsymbol{y_A}|\boldsymbol{y_B}\right)$ is $\mathcal{N}\left(\boldsymbol{\mu_{A|B}}, \boldsymbol{\Sigma_{A|B}}\right)$, where

$$\boldsymbol{\mu_{A|B}} = \boldsymbol{\mu_A} + \boldsymbol{\Sigma_{AB}}\boldsymbol{\Sigma_{BB}^{-1}}\left(\boldsymbol{y_B} - \boldsymbol{\mu_B}\right) \text{ and}$$
$$\boldsymbol{\Sigma_{A|B}} = \boldsymbol{\Sigma_{AA}} - \boldsymbol{\Sigma_{AB}}\boldsymbol{\Sigma_{BB}^{-1}}\boldsymbol{\Sigma_{BA}}.$$

4. If $\boldsymbol{y} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ and $\boldsymbol{y'} \sim \mathcal{N}\left(\boldsymbol{\mu'}, \boldsymbol{\Sigma'}\right)$ are independent,

   then $\boldsymbol{y} + \boldsymbol{y'} \sim \mathcal{N}\left(\boldsymbol{\mu} + \boldsymbol{\mu'}, \boldsymbol{\Sigma} + \boldsymbol{\Sigma'}\right)$ [RH05, pp. 19–20].

3.2.6 Calculation Summary Statistics

As the posterior mean and the credibility intervals of coefficient are of interest, calculation of these is performed later on. This allows a better interpretation of the results.
To compute the posterior mean of the coefficients, the code shown in Listing 7.3 can be used.
To get the expected value given a marginal function $\pi\left(x\right)$, the expected value of a function $f\left(x\right)$ is calculated, i.e.

$$\int f\left(x\right)\pi\left(x\right)dx \tag{3.13}$$

[Ait91].
If necessary, e.g. if the target variable follows a (negative) binomial distribution, the value must be transformed to its original scale, as in these cases the log-likelihood is modelled. Therefore, in these cases, the expected value would have to be exponentiated to allow a clear interpretation.
To obtain a credibility interval of the fixed effects on the transformed scale, the code in Listing 7.4 can be used. In practice, the marginals are first transformed to their original scale, if necessary, and then the 2.5% quantile and the 97.5% quantiles are calculated.

3.2.7 Skewness

Skewness is a statistical indicator that describes the type and strength of the asymmetry of a probability distribution. It shows whether and how strongly the distribution is skewed to the right (right-skewed, left-skewed, negative skewness) or to the left (left-skewed, right-skewed, positive skewness). Any non-symmetrical distribution is

called skewed.
The skewness of a random variable X is the third standardized moment, defined as

$$\text{Skew}[X] = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mathbb{E}\left[(X-\mu)^3\right]}{\left(\mathbb{E}\left[(X-\mu)^{3/2}\right]\right)^2} = \frac{\mu_3}{\sigma^3}, \tag{3.14}$$

with μ_3 the third central moment and σ the standard deviation [DS11; Wil44].

3.2.8 Kurtosis

Kurtosis is a measure of the slope of a probability distribution of a random variable. Distributions with low kurtosis scatter relatively evenly; for distributions with high kurtosis, the scatter results more from extreme but rare events.
The kurtosis of a random variable X is the fourth standardized moment, defined as

$$\text{Kurt}[X] = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mathbb{E}\left[(X-\mu)^4\right]}{\left(\mathbb{E}\left[(X-\mu)^2\right]\right)^2} = \frac{\mu_4}{\sigma^4}, \tag{3.15}$$

with μ_4 the fourth central moment and σ the standard deviation [DeC97; Wil44].

3.3 Prior Selection

A key question in Bayesian analysis is the effect of the prior on the posterior, and how that effect can be measured. Do posterior distributions derived with different priors become very similar as more and more data is collected? It has been formally proven that under certain regularity conditions, the impact of the prior decreases with increasing sample size. From a practical point of view, it is more important to know what happens when the sample size n is finite. In this section, different types of priors are introduced.

3.3.1 Conjugate Priors

One property of exponential families is that they have conjugate priors [DY79], which is an important property in Bayesian statistics. If the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ and the prior distribution $\pi(\boldsymbol{\theta})$ belong to the same probability distribution family, the prior and posterior distributions are called *conjugate* distributions. Furthermore, the prior for the likelihood function $\pi(\boldsymbol{y}|\boldsymbol{\theta})$ is called the *conjugate prior*. These priors were first discussed and formalised by Raifa and Schlaifer in 1961 [RS61].

The construction of a conjugate prior is done by factorising the likelihood function into two parts. One part must be independent of the parameter(s) of interest but can be dependent on the data, while the other factor is a function that depends on the parameter(s) of interest and is dependent on the data only through the sufficient statistics. The family of conjugate priors is by definition proportional to the second factor. The posterior distribution resulting from the conjugate prior is itself a member of the same family as the conjugate prior [RS61]. In cases where the prior and posterior distributions are part of the same family, the prior is said to be closed under sampling. Furthermore, since the data are only incorporated into the posterior distribution through the sufficient statistics, there exist relatively simple formulas for updating the prior into the posterior.

For an example of the construction of a conjugated prior, see Fink 1997 [Fin97].

A drawback of conjugated priors is that the a priori known information about μ may be insufficient for determining both parameters or may be inconsistent with the structure imposed by conjugacy [RMR10]. Moreover, these priors can be too restrictive and not every belief about the prior can be described [Irw05].

Thus, although conjugate priors are easy to handle both mathematically and computationally [Irw05], they are not often used in practice because of these drawbacks.

3.3.2 Penalised Complexity Priors

One issue when selecting the prior distribution of a particular parameter is that it is not always intuitive when it comes to understanding and interpreting this distribution, something that is essential to ensure that it behaves as intended by the user. This problem can be addressed by using *penalised complexity priors*, which is a methodology that penalises the complexity of model components in relation to deviation from simple base model formulations.
PC priors provide a systematic and unified approach to calculating prior distributions for parameters of model components by using an inherited nested structure. This structure contains two models, the base model and a flexible version of the model. The first of the two is generally characterised by a fixed value of the relevant parameter, while the second version is considered a function of the random parameter. By penalising the deviation from the flexible model to the fixed base model, the PC prior is calculated [Mar+14].

3.3.2.1 The Principles Behind PC Priors

Four main principles should be followed to calculate priors in a consistent way and to understand their properties.

Support to Occam's Razor

Let $\pi(x|\xi)$ denote the density of a model component x and ξ the parameter to which a prior distribution is to be assigned. The base model is characterised by a density $\pi(x|\xi = \xi_0)$, where ξ_0 is a fixed value. The prior for ξ should be such that proper shrinkage is given to ξ_0. The simplicity of the model is therefore prioritised over the complexity of the model, preventing overfitting [Mar+14].

Penalisation of Model Complexity

Let $f_1 = \pi(x|\xi)$ and $f_0(x|\xi = \xi_0)$ denote the flexible model and the base model respectively. The complexity of f_1 compared to f_0 is characterised using the Kullback-Leibler divergence [KL51] to calculate a measure of complexity between the two models,

$$\mathrm{KLD}(f_1||f_2) = \int f_1(x) \log\left(\frac{f_1(x)}{f_0(x)}\right) dx. \tag{3.16}$$

This can be used to measure the information that is lost when f_1 is approximated by the simpler model f_0. For multinormal densities with zero mean, the calculation simplifies to

$$\text{KLD}\left(f_1||f_0\right) = \frac{1}{2}\left(\text{trace}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}_1\right) - n - \log\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_0|}\right)\right), \tag{3.17}$$

where $f_i \sim \mathcal{N}\left(0, \boldsymbol{\Sigma}_i\right), i = 0, 1$, while n represents the dimension. For easier interpretation, the Kullback-Leibler divergence is transformed into a unidirectional distance measure

$$d\left(\xi\right) = d\left(f_1||f_0\right) = \sqrt{2\text{KLD}\left(f_1||f_0\right)} \tag{3.18}$$

which can be interpreted as a measure of distance from f_1 to f_0 [Mar+14].

Constant Rate Penalisation

The derivation of the PC prior is based on a system of constant rate penalisation, given by

$$\frac{\pi_d\left(d\left(\xi\right) + \delta\right)}{\pi_d\left(d\left(\xi\right)\right)} = r^{\delta}, \qquad d\left(\xi\right), \delta \geq 0. \tag{3.19}$$

$r \in (0, 1)$ represents the constant decay rate and thus implies that the relative change in the prior distribution for $d\left(\xi\right)$ is independent of the actual distance. Therefore, $d\left(\xi\right)$ is exponentially distributed with density $\pi\left(d\left(\xi\right)\right) = \lambda \exp\left(-\lambda d\left(\xi\right)\right)$ and rate $\lambda = -\ln\left(r\right)$. By a standard variable change transformation, the corresponding PC prior for ξ is given [Mar+14].

User-Defined Scaling

Since λ characterises the shrinkage properties of the prior, it is important that the rate can be chosen in an intuitive and interpretable way. One possibility is to determine λ by including a probability statement of tail events, for example

$$\mathbb{P}\left(Q\left(\xi\right) > U\right) = \alpha, \tag{3.20}$$

where U represents an assumed upper bound for an interpretable transformation $Q\left(\xi\right)$ and α denotes a small probability [Mar+14].

3.3.2.2 Example: PC Prior for the Precision

A PC prior can be used to adjust the smoothness of a spatial field in an intuitive way by specifying such a prior for the precision τ. This makes it possible to adjust the smoothness of the spatial field in an intuitive way. In this case, the penalised complexity prior is defined by the parameter σ_0. Equation 3.20 therefore looks like this,

$$\mathbb{P}(\sigma > \sigma_0) = \alpha. \tag{3.21}$$

The actual expression of the prior is given by

$$\pi(\tau) = \frac{\lambda}{2}\tau^{-3/2}\exp\left(-\lambda\tau^{-1/2}\right), \qquad \tau > 0 \tag{3.22}$$

and is a type-2-Gumbel distribution.
The prior for τ corresponds to an exponential distribution with rate λ for the standard deviation.
λ quantifies the size of the penalty for deviation from the base model and is increased with higher values for it.
Here, $\lambda = \frac{-\log(\alpha)}{\sigma_0}$ [Mar+14].

3.4 Markov-Chain-Monte-Carlo-Methods

Markov chain Monte Carlo methods, also referred to as MCMC methods, are a set of algorithms that enable sampling from probability distributions based on the construction of Markov chains. After a sufficient number of iterations, the stationary distribution of a Markov chain can be taken as the desired distribution, with the quality of this distribution improving as the number of iterations increases. Most of the time, the construction of such a chain is relatively simple; the real challenge is to determine how many steps are needed before convergence towards the stationary distribution is achieved. MCMC methods are mostly used to compute numerical approximations of multidimensional integrals, for instance in Bayesian statistics or computational biology. The two main concepts used in MCMC methods are Monte Carlo integration and the aforementioned Markov chains, hence the name Markov Chain Monte Carlo.

3.4.1 Monte Carlo Integration

Monte Carlo integration is a technique that uses the generation of random numbers for numerical computation of definite integrals and is especially useful for higher-dimensional integrals. The problem the method addresses is the computation of the integral

$$\mathbb{E}_f\left[h\left(X\right)\right] = \int_{\chi} h(x) f(x) dx. \tag{3.23}$$

The integral can be approximated by using a sample $(X_1, ..., X_m)$ generated from f and calculating the arithmetic mean

$$\overline{h}_m = \frac{1}{m} \sum_{j=1}^{m} h\left(x_j\right). \tag{3.24}$$

According to the Strong Law of Large Numbers, $\overline{h}_m$ is likely to converge to $\mathbb{E}_f\left[h\left(X\right)\right]$. When the expectation of h^2 under f is finite, the convergence speed of $\overline{h}_m$ can be assessed. The variance too can be estimated from the sample $(X_1, ..., X_n)$ through

$$v_m = \frac{1}{m^2} \sum_{j=1}^{m} \left[h\left(x_j\right) - \overline{h}_m\right]^2. \tag{3.25}$$

For m large,

$$\frac{\overline{h}_m - \mathbb{E}_f\left[h\left(X\right)\right]}{\sqrt{v_m}} \tag{3.26}$$

is approximately distributed as a $\mathcal{N}(0,1)$ variable. This can be used for constructing a convergence test and to calculate confidence bounds for the approximation of $\mathbb{E}_f[h(X)]$ [RC13, pp. 83–84]. The term Monte Carlo was first used in 1949 by Metropolis and Ulam to describe a method dealing with problems related to "integro-differential equations that occur in various branches of the natural sciences" [MU49].

3.4.2 Markov Chains

Markov chains are stochastic processes that aim to provide the probability of the occurrence of future events. A Markov chain is defined by the fact that even if only a limited history is known, predictions about future developments can be made just as reliably as if the entire history of a process were known. Thus, the probability of moving from the current state to any state depends only on the current state of the chain. These probabilities are defined by a *transition kernel*, which is a function K on $\mathcal{X} \times \mathcal{B}(\mathcal{X})$, such that

i. $\forall x \in \mathcal{X}, K(x,\cdot)$ is a probability measure;

ii. $\forall A \in \mathcal{B}(\mathcal{X}), K(\cdot, A)$ is measureable.

In the discrete case, the transition kernel is a matrix $\boldsymbol{K}$ with elements

$$\mathbb{P}_{xy} = \mathbb{P}(X_n = y | X_{n-1} = x), \qquad x, y \in \mathcal{X}.$$

If $\mathcal{X}$ is continuous, the kernel denotes the conditional density $K\left(x, x^T\right)$ of the transition $K(x,\cdot)$,

$$\mathbb{P}(X \in A|x) = \int_A K\left(x, x^T\right) dx^T.$$

Given a transition kernel K, a sequence $X_0, X_1, ..., X_t$ of random variables is a *Markov chain* (X_n), if, for any t, the conditional distribution of X_t given the previous states is the same as the distribution of X_t given the last state, x_{t-1},

$$\begin{aligned}\mathbb{P}(X_{t+1} \in A | x_0, x_1, x_2, ..., x_t) &= \mathbb{P}(X_{t+1} \in A | x_t) \\ &= \int_A K(x_t, dx). \end{aligned} \tag{3.27}$$

These chains were first introduced by Markov in 1906 [Mar06]. Markov chains can have certain properties that affect their long-term behaviour and are of particular importance for MCMC algorithms. In sections 3.4.2.1– 3.4.2.5, some of them will be introduced.

3.4.2.1 Irreducibility

Irreducibility is critical to the construction of Markov chain Monte Carlo algorithms, as it ensures the convergence of such an algorithm. A Markov chain is *irreducible* if all states communicate, that is, for all states i and j the probability of getting from i to j in finite time is true positive.
Formally speaking, given a measure φ, a Markov chain (X_n) with transition kernel $K(x, y)$ is *φ-irreducible*, if, for every $A \in B(\mathcal{X})$ with $\varphi(A) > 0$, there exists n such that $K^n(x, A) \forall x \in \mathcal{X}$. The chain is *strongly φ-irreducible* if $n = 1 \forall$ measurable A [RC13, pp. 213–214].

3.4.2.2 Periodicity

The behaviour of a Markov chain can sometimes be limited by deterministic constraints on the transitions from X_n to X_{n+1}. For discrete chains, the *period* of a state $w \in \mathcal{X}$ is defined as.

$$d(w) = \text{g.c.d. } \{m \geq 1; K^m(w, w) > 0\},$$

with g.c.d the greatest common denominator. If a Markov chain is irreducible, the transition matrix can be written as a block matrix

$$\boldsymbol{P} = \begin{pmatrix} 0 & \boldsymbol{D}_1 & 0 & \dots & 0 \\ 0 & 0 & \boldsymbol{D}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & & \\ \boldsymbol{D}_d & 0 & 0 & & 0 \end{pmatrix}, \tag{3.28}$$

It is evident that at every d-th step there is a return to the initial group. There exists only one value for the period when a chain is irreducible. If this value is 1, the irreducible chain is *aperiodic* [RC13, pp. 217–218].

3.4.2.3 Transience and Recurrence

To guarantee an acceptable approximation of a simulated model, a Markov chain needs to have good stability properties. Irreducibility is not strong enough to ensure that the trajectory of (X_n) enters A often enough. This leads to the formalisation of *recurrence* and *transience*.
In a finite space $\mathcal{X}$, a state $w \in \mathcal{X}$ is *transient* if it is finitely often visited and *recurrent* if it is almost certainly infinitely often visited.

For irreducible chains, these two properties are properties of the chain, not of a particular state [RC13, pp. 218–219].

3.4.2.4 Ergodicity

When looking at a Markov chain (X_n) from a temporal point of view, it is essential to establish to what the chain is converging. A natural candidate for the limiting distribution is the stationary distribution π which leads to the need to define sufficient conditions on (X_n) for X_n to be asymptotically distributed according to π. There are several conditions that can be imposed on the convergence of P^n, the distribution of X_n to π. The most fundamental and important is that of *ergodicity*, that is, independence of initial conditions.
If a Markov chain (X_n) is both aperiodic and positive recurrent, it is called an *ergodic* Markov chain [RC13, pp. 231–234].

3.4.2.5 Stationary distribution

A chain (X_n) is more stable if the marginal distribution of X_n is independent of n. This is a requirement for the existence of a probability distribution π such that $X_{n+1} \sim \pi$ if $X_n \sim \pi$. Markov chain Monte Carlo methods rely on the fact that this condition can be satisfied.
A σ-finite measure π is *invariant* for the transition kernel $K(\cdot, \cdot)$ if

$$\pi(B) = \int_{\mathcal{X}} K(x, B)\, \pi(dx), \qquad \forall B \in \mathcal{B}(\mathcal{X}).$$

This distribution is referred to as *stationary* if π is a probability measure, as $X_0 \sim \pi$ implies that $X_n \sim \pi$ is $\forall n$. An irreducible Markov chain has a stationary distribution precisely if it is positively recurrent. The distribution is then given by

$$\pi_x = \left(\mathbb{E}_x[\tau_x]\right)^{-1}, \qquad x \in \mathcal{X}, \tag{3.29}$$

where $\mathbb{E}_x[\tau_x]$ can be interpreted as the average number of transitions between two passages in x.
In practice, the stationary distributions are often of special interest. If these distributions are defined as the starting distribution of X_0, then all following distributions of the states X_n for any n are equal to the starting distribution. The interesting question here is when such distributions exist and when any distribution converges against a stationary distribution of this kind [RC13, pp. 223–224].

3.4.3 The Metropolis-Hastings Algorithm

Having established the basics of MCMC methods, one of the best known MCMC algorithms, the Metropolis-Hastings algorithm, is introduced next. The algorithm is based on the Metropolis algorithm, which was developed to simulate the states of a system according to the Boltzmann distribution, with the newest state always depending on the previous state [Met+53].
The Metropolis-Hastings algorithm is a procedure for drawing random samples from a probability distribution from which direct sampling is difficult if a function proportional to the *target density* f is known. This function $q\left(\boldsymbol{y}|\boldsymbol{x}\right)$ is called the *proposal density* and must be easy to simulate in order for the Metropolis-Hastings algorithm to be implementable. Moreover, it must be either explicitly present or *symmetric*, meaning $q\left(\boldsymbol{x}|\boldsymbol{y}\right) = q\left(\boldsymbol{y}|\boldsymbol{x}\right)$.
The Metropolis-Hastings algorithm of a target density f and proposal density q produces a Markov chain $\left(X^{(t)}\right)$ by the following transition.

Algorithm 1 The Metropolis-Hastings Algorithm

Given $f\left(\boldsymbol{x}\right)$ and $q\left(\boldsymbol{y}|\boldsymbol{x}\right)$
1: Initialisation: Choose arbitrary x_t as the first sample
2: **for** each iteration t **do**
3: Generate $Y_t \sim q\left(\boldsymbol{y}|x^{(t)}\right)$
4: Take

$$X^{(t+1)} = \begin{cases} Y_t & \text{with probability } \mathbb{P}\left(x^{(t)}, Y_t\right) \\ x^{(t)} & \text{with probability } 1 - \mathbb{P}\left(x^{(t)}, Y_t\right) \end{cases}$$

where

$$\mathbb{P}\left(x, y\right) = \min\left\{\frac{f\left(\boldsymbol{y}\right)}{f\left(\boldsymbol{x}\right)}\frac{q\left(\boldsymbol{x}|\boldsymbol{y}\right)}{q\left(\boldsymbol{y}|\boldsymbol{x}\right)}, 1\right\}. \tag{3.30}$$

$\mathbb{P}\left(x, y\right)$ is the *Metropolis-Hastings acceptance probability*.
The algorithm always accepts values y_t that lead to an increase in the ratio $f\left(y_t\right)/q\left(y_t|x^{(t)}\right)$ compared to the previous value $f\left(x^{(t)}\right)/q\left(x^{(t)}|y_t\right)$. In the symmetric case, the acceptance probability simplifies to

$$\mathbb{P}\left(x, y\right) = \min\left\{\frac{f\left(\boldsymbol{y}\right)}{f\left(\boldsymbol{x}\right)}, 1\right\}$$

[Has70].
If the Markov chain starts with a value $x^{(0)} > 0$, then $f\left(x^{(t)}\right) > 0\ \forall t \in \mathbb{N}$ since the

values of y such that $f\left(y_t\right) = 0$ will all be rejected by the algorithm. As the number of iterations t increases, the distribution of saved states $x_0, ..., x_t$ will converge towards the target density $f(\boldsymbol{x})$ [RC13, pp. 270–275].

3.5 Latent Gaussian Models and INLA

In recent years, a growing amount of georeferenced data has become available, leading to an increased need for appropriate statistical modeling to handle large and complex datasets. Bayesian hierarchical models have proven to be effective in capturing complex stochastic structures in spatial processes. A large proportion of these models are based on latent Gaussian models, a subclass of structured additive regression models. These models include Integrated Nested Laplace Approximations (INLA), which are a class of models used to approximate the posterior marginals of a latent Gaussian field. These approximations work by reformulating the regression model as a three-part hierarchical model. The parts are as follows:

1. Approximation of the posterior marginal of $\boldsymbol{\theta}$ by using the Laplace approximation.
2. Computation of the (simplified) Laplace approximation for selected values of $\boldsymbol{\theta}$ to improve the Gaussian approximation.
3. Combination of the first two parts by numerical integration.

$\boldsymbol{\theta}$ is a vector of hyperparameters. The hyperparameters $\boldsymbol{\theta}$ can be, for example, the variance in the Gaussian likelihood or the shape parameter in the likelihood of the gamma distribution. In the case of latent fields, they can be, for instance, dispersion parameters or spatial correlation parameters. [RMC09]

3.5.1 Notation and Basic Properties

For structured additive regression models, the distribution of the response variable y_i is assumed to be a member of the exponential family, with the mean μ_i linked to a structured additive predictor η_i by a link function $g(\cdot)$ such that $g(\mu_i) = \eta_i$. The predictor η_i takes into account the effect of multiple covariates in an additive way,

$$\eta_i = \alpha + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}) + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \epsilon_i, \qquad i = 1, ..., n \tag{3.31}$$

[Sto85].
The $\left\{f^{(j)}(\cdot)\right\}$s are unknown functions of the covariates u, while the $\{\beta_k\}$s represent the linear effect of the covariates z and the ϵ_is are unstructured terms. Latent Gaussian models assign a Gaussian prior to α, $\left\{f^{(j)}(\cdot)\right\}$ and $\{\epsilon_i\}$. In the following $\boldsymbol{x}$ shall denote the vector of all latent Gaussian variables ($\{\eta_i\}$, α, $\left\{f^{(j)}\right\}$ and $\{\beta_k\}$)

and $\boldsymbol{\theta}$ the vector of hyperparameters.
The conditional density $\pi(\boldsymbol{x}|\theta_1)$ is Gaussian with an assumed zero mean and precision matrix $\boldsymbol{Q}(\theta_1)$. The Gaussian density $\mathcal{N}(\mu, \boldsymbol{\Sigma})$ with mean μ and covariance $\boldsymbol{\Sigma}$ at configuration $\boldsymbol{x}$ is denoted by $\mathcal{N}(\boldsymbol{x}; \mu, \boldsymbol{\Sigma})$. For simplicity, $\{\eta_i\}$ has been included instead of $\{\epsilon_i\}$.
The distribution for the n_d observational variables $y = \{y_i : i \in \mathcal{I}\}$ is denoted by $\pi(\boldsymbol{y}|\boldsymbol{x}, \theta_2)$ and is assumed conditionally independent given $\boldsymbol{x}$ and θ_2. Let $\boldsymbol{\theta} = \left(\theta_1^T, \theta_2^T\right)^T$ with $\dim(\boldsymbol{\theta}) = m$. For non-singular $\boldsymbol{Q}(\boldsymbol{\theta})$ the posterior is given by

$$\begin{aligned}\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) &\propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\prod_{i\in I}\pi(y_i|x_i, \boldsymbol{\theta}) \\ &\propto \pi(\boldsymbol{\theta})|\boldsymbol{Q}(\boldsymbol{\theta})|^{1/2}\exp\left[-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}(\boldsymbol{\theta})\boldsymbol{x} + \sum_{i\in I}\log\{\pi(y_i|x_i, \boldsymbol{\theta})\}\right]. \qquad (3.32)\end{aligned}$$

Most latent Gaussian models satisfy two basic properties:

1. The latent field $\boldsymbol{x}$ is of large dimension, $n \approx 10^2 - 10^5$. Therefore, the latent field is a Gaussian Markov random field with sparse precision matrix $\boldsymbol{Q}(\boldsymbol{\theta})$.
2. The number of hyperparameters, m, is small, $m \leq 6$.

In most cases, both properties are required to produce fast inference, and thus these will be assumed to be true for the remainder of this work [RMC09].

3.5.2 Applications for Latent Gaussian Models

Latent Gaussian models can be employed in a vast range of different domains, in fact most structured Bayesian models are of this particular form. Some of these domains are presented below.

3.5.2.1 Regression Models

Bayesian generalised linear models correspond to the linear relationship $\eta_i = \alpha + \sum_{k=1}^{n_\beta}\beta_k z_{ki}$ [DGM00]. Either the linear relationship of the covariates can be relaxed through the $f(\cdot)$ terms [FT13], random effects can be introduced through them or both. Smooth covariate effects are frequently modeled using penalised spline models [LB04] or random walk models [FT13], continuous indexed spline models [RH05] or Gaussian processes [CGW05]. The incorporation of random effects allows for the consideration of overdispersion caused by unobserved heterogeneity or correlation

in longitudinal data and can be introduced by defining $f(u_i) = f_i$ and $\{f_1\}$ to be independent, zero mean and Gaussian [FL01].

3.5.2.2 Dynamic Models

Temporal dependence can be introduced by using i in (3.31) as temporal index t and defining $f(\cdot)$ and $\boldsymbol{u}$ such that $f(u_t) = f_t$. Both a discrete-time and a continuous-time autoregressive model can be modeled by $\{f_t\}$. Furthermore, a seasonal effect or the latent process of a structured time series model can be modeled [KG96]. Alternatively, a smooth temporal function in the same sense as for regression models can be represented by $\{f_t\}$.

3.5.2.3 Spatial and Spatio-Temporal Models

Similar to the previous type of model, spatial dependence can be modeled by a spatial covariate $\boldsymbol{u}$ such that $f(u_s) = f_s$, where s denotes the spatial location or region s. The stochastic model for f_s is constructed to promote spacial smooth realisations of some sort. Popular models of this type include the Besag-York-Mollié [BYM91] model with extensions for regional data, continuous indexed Gaussian models [BCG14] and texture models [Mar+01]. The dependence between spatial and temporal covariates can be achieved either by using a spatio-temporal covariate (s, t) or a corresponding spatio-temporal Gaussian field [KW03].
Often the final model consists of a sum of several components, e.g. a spatial component, random effects and both linear and smooth effects of some covariates. In order to separate the effects of the different components in (3.31), sometimes linear or sum-to-zero constraints can be imposed [RMC09, pp. 319–321].

3.5.3 The MCMC Approach to Inference

The usual approach to inference for latent Gaussian models involves the previously introduced Markov chain Monte Carlo methods. Due to several factors, these methods may perform poorly when applied to such models. One factor is the interdependence of the components of the latent field $\boldsymbol{x}$ while another is that $\boldsymbol{\theta}$ and $\boldsymbol{x}$ are highly dependent on each other, especially for large n. The first of these problems can potentially be overcome by constructing a joint proposal based on a Gaussian approximation of the full conditional of $\boldsymbol{x}$ [Gam97], while the second problem requires, at least in part, a joint update of $\boldsymbol{\theta}$ and $\boldsymbol{x}$. There are several

proposals to solve these shortcomings, but MCMC sampling continues to show poor computational speed [RMC09, p. 322].

3.5.4 Gaussian Random Fields

Let $\boldsymbol{s} = (s_1, ..., s_n)^T$ be a vector of locations. A *Gaussian random field* (GRF)

$$\left\{ Z(s) : s \in D \subset \mathbb{R}^2 \right\} \tag{3.33}$$

is a set of random variables where the observations occur in a continuous domain and where each finite set of random variables follows a multivariate normal distribution. A random process $Z(\cdot)$ is strictly stationary if it is invariant to shifts, i.e., if for each set of locations and each $h \in \mathbb{R}^2$ the distribution of $\boldsymbol{Z(s)} = (Z(s_1), ..., Z(s_n))$ is equal to that of $\boldsymbol{Z(s+h)} = (Z(s_1 + h), ..., Z(s_n + h))$. A less constraining requirement is given by second-order stationarity. Under this condition, the process has a constant mean value

$$\mathbb{E}[\boldsymbol{Z(s)}] = \mu, \qquad \forall s \in D, \tag{3.34}$$

and the covariances depend only on the differences between locations

$$\text{Cov}(\boldsymbol{Z(s)}, \boldsymbol{Z(s+h)}) = C(\boldsymbol{h}), \qquad \forall \boldsymbol{s} \in D, \forall \boldsymbol{h} \in \mathbb{R}^2. \tag{3.35}$$

Furthermore, if the covariances depend only on the distances between the locations and not on the directions, the process is called isotropic. Else, the process is anisotropic. An intrinsically stationary process has a constant mean value and satisfies

$$\text{Var}(Z(s_i) - Z(s_j)) = 2\gamma(s_i - s_j), \qquad \forall s_i, s_j. \tag{3.36}$$

$2\gamma(\cdot)$ is the variogram and $\gamma(\cdot)$ is called the semivariogram [Cre15]. Under the assumption of intrinsic stationarity, the constant-mean assumption implies

$$2\gamma(\boldsymbol{h}) = \text{Var}(\boldsymbol{Z(s+h)} - \boldsymbol{Z(s)}) = \mathbb{E}\left[(\boldsymbol{Z(s+h)} - \boldsymbol{Z(s)})^2 \right],$$

and the estimation of the semivariogram can be obtained using the empirical semivariogram as follows:

$$2\widehat{\gamma}(\boldsymbol{h}) = \frac{1}{|N(\boldsymbol{h})|} \sum_{N(\boldsymbol{h})} (Z(s_i) - Z(s_j))^2, \tag{3.37}$$

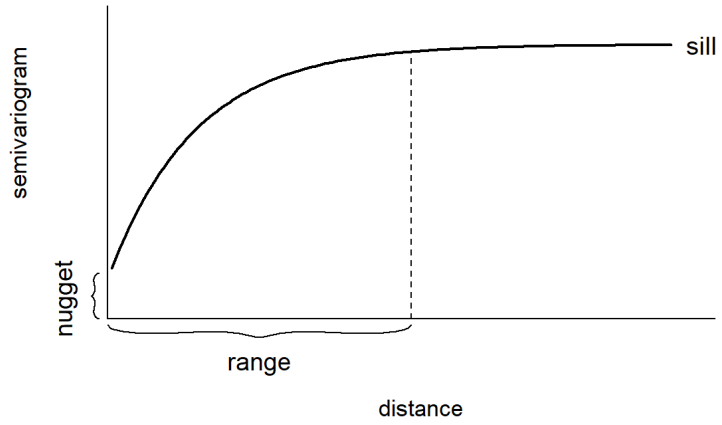

Fig. 3.3: A typical semivariogram

where $N(\boldsymbol{h}) = \{(s_1, s_j) : s_i - s_j = \boldsymbol{h}, i, j = 1, ..., n\}$ denotes the number of pairs and $|N(\boldsymbol{h})|$ the number of distinct pairs. For isotropic processes, the semivariogram is a function of distance $h = ||\boldsymbol{h}||$.

Plotting the empirical semivariogram against the separation distance conveys essential information regarding the continuity and spatial variability of the process. Given relatively short distances, the semivariogram tends to be small but increases with distance, indicating the similarity of observations in close proximity. The semivariogram levels off to a nearly constant value, also called the sill, as the separation distance increases, indicating a decrease in spatial dependence with distance within the range and no spatial correlation outside the range, which is reflected in a nearly constant variance. If there is a discontinuity or a vertical jump at the origin, the process has a nugget effect, which is often due to a measurement error, but may also be indicative of a spatially discontinuous process. The empirical semivariogram is an exploratory tool useful for assessing whether data exhibit spatial correlation. Furthermore, it can be compared to a Monte Carlo envelope of empirical semivariograms calculated from random permutations of the data while keeping the locations fixed [DRC03]. If the empirical semivariogram lies outside the Monte Carlo envelope with increasing distance, this is an indication of spatial correlation.

These graphs are again taken from "Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny" by Paula Moraga [Mor19]. An example of a semivariogram is shown in Figure 3.3. This graph is taken from "Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny" by Paula Moraga [Mor19].

The dependence structure of a GRF is given by the covariance matrix, which is constructed from a covariance function. Matérn models and exponential functions are conventionally used for this purpose [Gel+10]. For the locations $s_i, s_j \in \mathbb{R}^2$ the exponential covariance function is given by

$$\text{Cov}\left(Z\left(s_i\right), Z\left(s_j\right)\right) = \sigma^2 \exp\left(-\kappa \left|\left|s_i - s_j\right|\right|\right), \tag{3.38}$$

where the distance between the locations s_i and s_j is denoted by $||s_i - s_j||$, the variance of the spatial field is given by σ^2, while $\kappa > 0$ controls the rate at which the correlation decays as the distance increases.
The Matérn family represents a flexible class of covariance functions that arises naturally in a variety of scientific fields [GG06]. The Matérn covariance function is written as

$$\text{Cov}\left(Z\left(s_i\right), Z\left(s_j\right)\right) = \frac{\sigma^2}{2^{\nu-1}\Gamma\left(\nu\right)} \left(\kappa \left|\left|s_i - s_j\right|\right|\right)^{\nu} K_{\nu}\left(\kappa \left|\left|s_i - s_j\right|\right|\right). \tag{3.39}$$

σ^2 denotes the marginal variance of the spatial field, $K_{\nu}\left(\cdot\right)$ represents the modified Bessel function of second kind and order $\nu > 0$, where ν is an integer. The mean square differentiability of the process is determined by ν and is usually fixed since it is difficult to identify in applications. For $\nu = 0.5$, this covariance function is the equivalent of the exponential covariance function. $\kappa > 0$ is related to the range ρ, which is defined as the distance at which there is approximately no correlation between two given points, $\rho = \sqrt{8\nu}/\kappa$ to be exact [Cam+13]. Examples of these two covariance functions are shown in Figure 3.4. These graphs are again taken from "Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny" by Paula Moraga [Mor19].

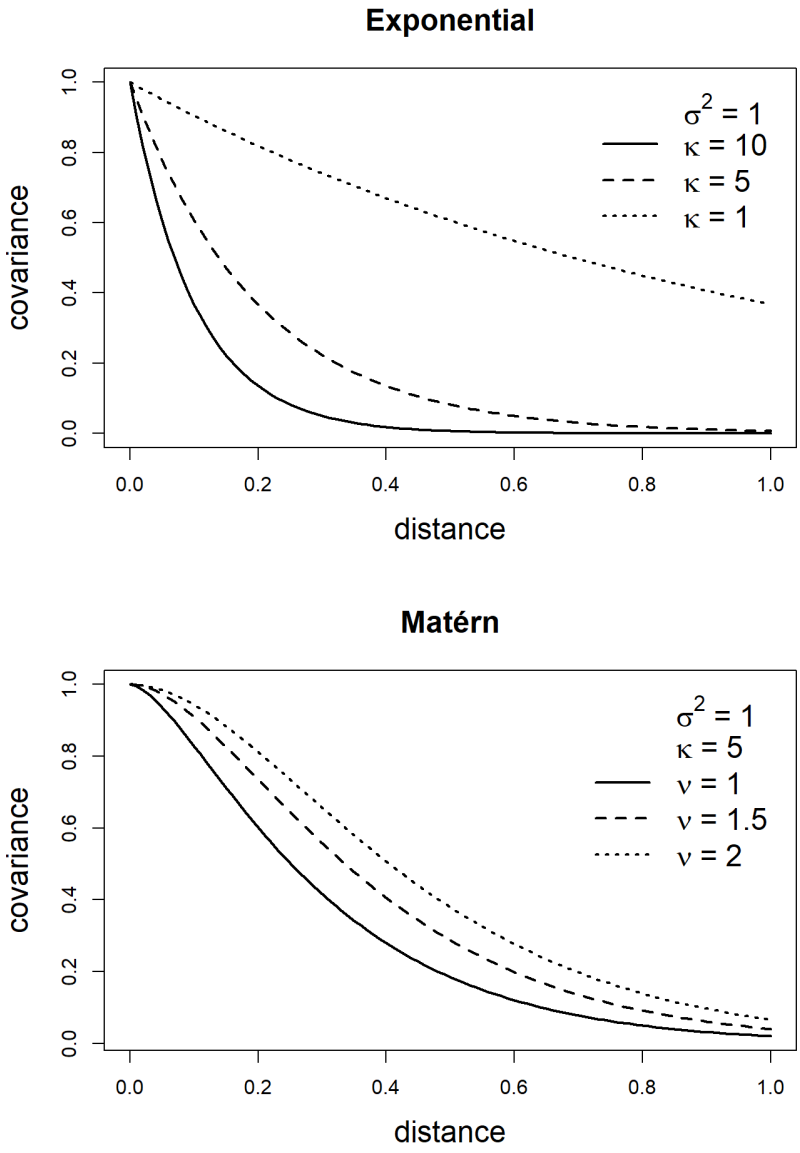

Fig. 3.4: Covariance functions corresponding to exponential and Matérn models.

3.5.5 Gaussian Markov Random Fields

3.5.5.1 Definition of GMRFs

Let $\boldsymbol{x} = (x_1, ..., x_n)^T$ be normally distributed with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, ..., \}$ and $\mathcal{E}$ be such that there is no edge between nodes i and j exactly when $x_i \perp x_j | \boldsymbol{x}_{ij}$. Then $\boldsymbol{x}$ is a *Gaussian Markov random field* (GMRF) with respect to $\mathcal{G}$.

Since $\boldsymbol{\mu}$ does not affect the pairwise conditional independence properties of $\boldsymbol{x}$, this information is 'hidden' in $\boldsymbol{\Sigma}$. Hence,

$$x_i \perp x_j | x_{ij} \iff Q_{ij} = 0.$$

Therefore, the non-zero pattern of $\boldsymbol{Q}$ determines $\mathcal{G}$, i.e. whether x_i and x_j are conditionally independent, and can be derived from $\boldsymbol{Q}$. If $\boldsymbol{Q}$ is a fully dense matrix, then $\mathcal{G}$ is fully connected, implying that any normal distribution with SPD covariance matrix is a GMRF and vice versa.
The elements of $\boldsymbol{Q}$ are used for conditional interpretations. For any GMRF with respect to $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with mean $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{Q} > 0$,

$$\mathbb{E}\left[x_i | \boldsymbol{x}_{-i}\right] = \mu_i - \frac{1}{Q_{ii}} \sum_{j:j\sim i} Q_{ij}\left(x_j - \mu_j\right), \tag{3.40}$$

$$\text{Prec}\left(x_i | \boldsymbol{x}_{-i}\right) = Q_{ii} \qquad \text{and} \tag{3.41}$$

$$\text{Corr}\left(x_i, x_j | \boldsymbol{x}_{ij}\right) = -\frac{Q_{ij}}{\sqrt{Q_{ii}Q_{jj}}}, \qquad i \neq j \tag{3.42}$$

[RH05, p. 21].
On the main diagonal of $\boldsymbol{Q}$ are the conditional precisions of x_i given $\boldsymbol{x}_{-i}$ are placed, while the other elements, when scaled appropriately, provide information about the conditional correlation between x_i and x_j given $\boldsymbol{x}_{ij}$. Since $\text{Var}\left(x_i\right) = \Sigma_{ii}$ and $\text{Corr}\left(x_i, x_j\right) = \Sigma_{ij}/\sqrt{\Sigma_{ii}\Sigma_{jj}}$, the information about the marginal variance of x_i and the marginal correlation between x_i and x_j is given by $\boldsymbol{\Sigma}$. The marginal interpretation provided by the correlation matrix is intuitive and informative, as the scope of the interpretation is reduced from an n-dimensional distribution to a one- or two-dimensional distribution. $\boldsymbol{Q}$ is difficult to interpret marginally because either $\boldsymbol{x}_{-i}$ or $\boldsymbol{x}_{ij}$ would have to be integrated out of the joint distribution parameterized with respect to $\boldsymbol{Q}$. $\boldsymbol{Q}^{-1} = \boldsymbol{\Sigma}$ by definition, and in general Σ_{ii} depends on each element in $\boldsymbol{Q}$ and vice versa [RH05, pp. 20–23].

3.5.5.2 Markov Properties of GMRFs

One property of GMRFs is that more information regarding conditional independence can be extracted from $\mathcal{G}$. The following three properties are equivalent.
The *pairwise Markov property*:

$$x_i \perp x_j | \boldsymbol{x}_{ij} \qquad \text{if } \{i, j\} \notin \mathcal{E} \text{ and } i \neq j.$$

The *local Markov property*:

$$x_i \perp \boldsymbol{x}_{-\{i,\text{ne}(i)\}} | \boldsymbol{x}_{\text{ne}(i)} \qquad \forall i \in \mathcal{V}.$$

The *global Markov property*:

$$\boldsymbol{x}_A \perp \boldsymbol{x}_B | \boldsymbol{x}_C$$

for all disjoint sets A, B and C where A and B are non-empty and separated by C. Illustrations for these properties are shown in Figure 3.5, Figure 3.6 and Figure 3.7. These illustrations are taken from Rue and Held [RH05, pp. 23–24].

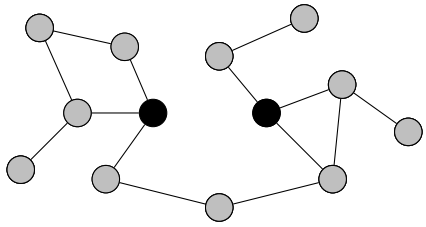

Fig. 3.5: The pairwise Markov property; the black nodes are conditionally independent given the light gray nodes.

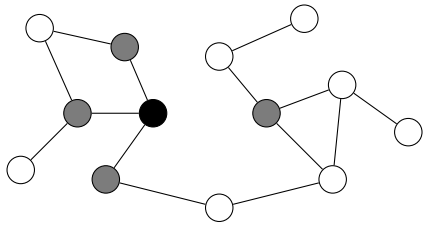

Fig. 3.6: The local Markov property; the black nodes and white nodes are conditionally independent given the dark gray nodes.

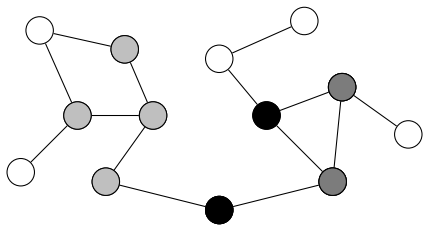

Fig. 3.7: The global Markov property; the dark gray and light gray nodes are globally independent given the black nodes.

3.5.5.3 Conditional Properties of GMRFs

An essential result of GMRFs is the conditional distribution for a subset $\boldsymbol{x}_a$ given $\boldsymbol{x}_{-A}$. Here the canonical parameterisation proves useful, since by definition it can be easily updated by successive conditioning.
By splitting the indices into the non-empty sets A and B, of which the latter is equal to -A,

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_A \\ \boldsymbol{x}_B \end{pmatrix}. \tag{3.43}$$

The mean and the precision are divided accordingly,

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_A \\ \boldsymbol{\mu}_B \end{pmatrix}, \qquad \text{and} \qquad \boldsymbol{Q} = \begin{pmatrix} \boldsymbol{Q}_{AA} & \boldsymbol{Q}_{AB} \\ \boldsymbol{Q}_{BA} & \boldsymbol{Q}_{BB} \end{pmatrix}. \tag{3.44}$$

The conditional distribution of $\boldsymbol{x}_A|\boldsymbol{x}_B$ is then a GMRF with respect to the subgraph $\mathcal{G}^A$ with mean $\boldsymbol{\mu}_{A|B}$ and precision matrix $\boldsymbol{Q}_{A|B} > 0$, where

$$\boldsymbol{\mu}_{A|B} = \boldsymbol{\mu}_A - \boldsymbol{Q}_{AA}^{-1}\boldsymbol{Q}_{AB}\left(\boldsymbol{x}_B - \boldsymbol{\mu}_B\right) \tag{3.45}$$

and

$$\boldsymbol{Q}_{A|B} = \boldsymbol{Q}_{AA}.$$

Thus, the explicit knowledge of $\boldsymbol{Q}_{A|B}$ is available through $\boldsymbol{Q}_{AA}$, i.e. no calculation is required to obtain the conditional precision matrix. Moreover, the conditional mean depends only on the values of $\boldsymbol{\mu}$ and $\boldsymbol{Q}$ in $A \cup \text{ne}\left(A\right)$, since $Q_{ij} = 0 \, \forall j \not\in \text{ne}\left(i\right)$.
For successive conditioning, the canonical parameterisation for GMRF is useful.
A GMRF $\boldsymbol{x}$ with respect to $\mathcal{G}$ and canonical parameters $\boldsymbol{b}$ and $\boldsymbol{Q} > 0$ has the density

$$\pi\left(\boldsymbol{x}\right) \propto \exp\left(-1\frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{x}\right).$$

The precision matrix is $\boldsymbol{Q}$ and the mean is $\boldsymbol{\mu} = \boldsymbol{Q}^{-1}\boldsymbol{b}$. The canonical parameterisation is written as

$$\boldsymbol{x} \sim \mathcal{N}_C\left(\boldsymbol{b}, \boldsymbol{Q}\right).$$

Furthermore,

$$\mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{Q}^{-1}\right) \Longleftrightarrow \mathcal{N}_C\left(\boldsymbol{Q}\boldsymbol{\mu}, \boldsymbol{Q}\right).$$

If the indices are partitioned into two non-empty sets A and B and $\boldsymbol{x}$, $\boldsymbol{b}$ and $\boldsymbol{Q}$ are partitioned as in (3.43) and (3.44), then

$$\boldsymbol{x}_A|\boldsymbol{x}_B \sim \mathcal{N}_C\left(\boldsymbol{b}_A - \boldsymbol{Q}_{AB}\boldsymbol{x}_B, \boldsymbol{Q}_{AA}\right). \tag{3.46}$$

Let $\boldsymbol{y}|\boldsymbol{x} \sim \mathcal{N}\left(\boldsymbol{x}, \boldsymbol{P}^{-1}\right)$ and $\boldsymbol{x} \sim \mathcal{N}_C\left(\boldsymbol{b}, \boldsymbol{Q}\right)$, then

$$\boldsymbol{x}|\boldsymbol{y} \sim \mathcal{N}_C\left(\boldsymbol{b} + \boldsymbol{P}\boldsymbol{y}, \boldsymbol{Q} + \boldsymbol{P}\right). \tag{3.47}$$

This allows the calculation of conditional densities with multiple sources of conditioning, e.g. conditioning on observed data and a subset of variables. Therefore, the canonical parameterisation can be repeatedly updated without explicitly calculating the mean until it is actually needed. The computation of the mean requires the solution of $\boldsymbol{Q}\boldsymbol{\mu} = \boldsymbol{b}$, but only matrix-vector products are needed for updating the canonical parameterisation [RH05, pp. 25–27].

3.5.5.4 Specification Through Full Conditionals

Alternatively, a GMRF can be specified by the full conditionals $\{\pi\left(x_i|\boldsymbol{x}_{-i}\right)\}$ in place of $\boldsymbol{\mu}$ and $\boldsymbol{Q}$. Suppose the full conditionals are given as normals with

$$\mathbb{E}\left[x_i|\boldsymbol{x}_{-i}\right] = \mu_i - \sum_{j:j\sim i} \beta_{ij}\left(x_j - \mu_j\right) \qquad \text{and} \tag{3.48}$$

$$\text{Prec}\left(x_i|\boldsymbol{x}_{-i}\right) = \kappa_i > 0 \tag{3.49}$$

for $i = 1, ..., n$, for $\boldsymbol{\mu}$, $\boldsymbol{\kappa}$ and some $\{\eta_{ij}, i \neq j\}$. Evidently, $\sim$ is implicitly defined by the non-zero terms of $\{\beta_{ij}\}$. For there to exist a joint density $\pi\left(\boldsymbol{x}\right)$ leading to these full conditional distributions, these full conditionals must be consistent. Since $\sim$ is symmetric, it follows that if $\beta_{ij} \neq 0$, then $\beta_{ji} \neq 0$. If the entries of the precision matrix are chosen such that

$$Q_{ii} = \kappa_i, \qquad \text{and} \qquad Q_{ij} = \kappa_i\beta_{ij}$$

and $\boldsymbol{Q}$ must be symmetrical, i.e.,

$$\kappa_i\beta_{ij} = \kappa_j\beta_{ji},$$

then $\boldsymbol{x}$ is a GMRF with respect to a labelled graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with mean $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{Q} = (Q_{ij})$ [RH05, p. 27].

3.5.5.5 Multivariate GMRFs

A *multivariate GMRF* (MGMRF) is a multivariate extension of a GMRF that has proven useful in applications. Let $\boldsymbol{x}$ be a GMRF with respect to $\mathcal{G}$, then the Markov property implies that

$$\pi\left(x_i|\boldsymbol{x}_{-i}\right) = \pi\left(x_i|\left\{x_j : j \sim i\right\}\right).$$

x_i is the value related to node i. Often the nodes have physical interpretations such as an administrative region of a country, which can be used to define the neighbours of node i. Let each of the n nodes have an associated vector $\boldsymbol{x}_i$ of dimension p, resulting in a GMRF of size np. Such a GMRF is denoted by $\boldsymbol{x} = \left(\boldsymbol{x}_1^T, ..., \boldsymbol{x}_n^T\right)^T$. The Markov property with respect to the nodes is preserved, i.e.,

$$\pi\left(\boldsymbol{x}_i|\boldsymbol{x}_{-i}\right) = \pi\left(\boldsymbol{x}_i|\left\{\boldsymbol{x}_j : j \sim i\right\}\right),$$

where $\sim$ is with respect to *the same graph* $\mathcal{G}$. Let $\boldsymbol{\mu} = \left(\boldsymbol{\mu}_1^T, ..., \boldsymbol{\mu}_n^T\right)^T$ be the mean of $\boldsymbol{x}$, where $\mathbb{E}\left[\boldsymbol{x}_i\right] = \boldsymbol{\mu}_i$, and $\widetilde{\boldsymbol{Q}} = \left(\widetilde{\boldsymbol{Q}}_{ij}\right)$ its precision matrix, where each element of the matrix is a $p \times p$ matrix.
It follows that

$$\boldsymbol{x}_i \perp \boldsymbol{x}_j|\boldsymbol{x}_{-ij} \Longleftrightarrow \widetilde{\boldsymbol{Q}}_{ij} = \boldsymbol{0}.$$

Formally, a random vector $\boldsymbol{x} = \left(\boldsymbol{x}_1^T, ..., \boldsymbol{x}_n^T\right)^T$ with $\dim\left(\boldsymbol{x}_i\right) = p$, is called a MGMRF$_p$ with respect to $\mathcal{G} = \left(\mathcal{V} = \{1, ..., n\}, \mathcal{E}\right)$ with mean $\boldsymbol{\mu}$ and precision matrix $\widetilde{\boldsymbol{Q}} > 0$, exactly when its density has the form

$$\begin{aligned}
\pi\left(\boldsymbol{x}\right) &= \left(\frac{1}{2\pi}\right)^{np/2} \left|\widetilde{\boldsymbol{Q}}\right|^{1/2} \exp\left(-\frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{\mu}\right)^T \widetilde{\boldsymbol{Q}}\left(\boldsymbol{x} - \boldsymbol{\mu}\right)\right) \\
&= \left(\frac{1}{2}\right)^{np/2} \left|\widetilde{\boldsymbol{Q}}\right|^{1/2} \exp\left(-\frac{1}{2}\sum_{ij}\left(\boldsymbol{x}_i - \boldsymbol{\mu}_i\right)^T \widetilde{\boldsymbol{Q}}_{ij}\left(\boldsymbol{x}_j - \boldsymbol{\mu}_j\right)\right)
\end{aligned}$$

and

$$\widetilde{\boldsymbol{Q}}_{ij} \neq 0 \Longleftrightarrow \{i, j\} \in \mathcal{E}\, \forall\, i \neq j.$$

A MGMRF$_p$ is equivalent to a GMRF of dimension np with identical mean vector and precision matrix. Therefore, all results valid for a GMRF are also valid for a MGMRF$_p$, with modifications, since the graph for a MGMRF$_p$ has size n and is defined with respect to $\{\boldsymbol{x}_i\}$, while for a GMRF it has size np and is defined with respect to $\{x_i\}$.

The interpretation of $\widetilde{\boldsymbol{Q}}_{ii}$ and $\widetilde{\boldsymbol{Q}}_{ij}$ can be derived from the full conditional $\pi(\boldsymbol{x}_i|\boldsymbol{x}_{-i})$. The extensions of (3.40) and (3.41) are

$$\mathbb{E}[\boldsymbol{x}_i|\boldsymbol{x}_{-i}] = \boldsymbol{\mu}_i - \widetilde{\boldsymbol{Q}}_{ii}^{-1} \sum_{j:j\sim i} \widetilde{\boldsymbol{Q}}_{ij}(\boldsymbol{x}_j - \boldsymbol{\mu}_j) \tag{3.50}$$

$$\text{Prec}(\boldsymbol{x}_i|\boldsymbol{x}_{-i}) = \widetilde{\boldsymbol{Q}}_{ii}. \tag{3.51}$$

In some applications, the full conditionals

$$\mathbb{E}[\boldsymbol{x}_i|\boldsymbol{x}_{-i}] = \boldsymbol{\mu}_i - \sum_{j:j\sim i} \boldsymbol{\beta}_{ij}(\boldsymbol{x}_j - \boldsymbol{\mu}_j) \tag{3.52}$$

$$\text{Prec}(\boldsymbol{x}_i|\boldsymbol{x}_{-i}) = \boldsymbol{\kappa}_i > 0, \tag{3.53}$$

In some applications, the full conditionals are used to define the MGMRF$_p$, for given $p \times p$-matrices $\{\boldsymbol{\beta}_{ij}, i \neq j\}$, $\{\boldsymbol{\kappa}_i\}$, and vectors $\boldsymbol{\mu}_i$. Again, $\sim$ is implicitly defined by the non-zero matrices $\{\boldsymbol{\kappa}_i\}$. Similar requirements as for $p = 1$ apply to the existence of the joint density: $\boldsymbol{\kappa}_i\boldsymbol{\beta}_{ij} = \boldsymbol{\beta}_{ij}^T\boldsymbol{\kappa}_j$ for $i \neq j$ and $\widetilde{\boldsymbol{Q}} > 0$. The $p \times p$ elements of $\widetilde{\boldsymbol{Q}}$ are

$$\widetilde{\boldsymbol{Q}}_{ij} = \begin{cases} \boldsymbol{\kappa}_i\boldsymbol{\beta}_{ij} & i \neq j \\ \boldsymbol{\kappa}_i & i = j \end{cases};$$

therefore $\widetilde{\boldsymbol{Q}} > 0 \iff (\boldsymbol{I} + (\boldsymbol{\beta}_{ij})) > 0$ [RH05, pp. 29–30].

3.5.6 Integrated Nested Laplace Approximation

An alternative to MCMC methods that is both less computationally intensive and suitable for performing approximate Bayesian inference in latent Gaussian models is *Integrated nested Laplace Approximation* (INLA). The basis of INLA is the use of a combination of analytical approximations and numerical algorithms for sparse matrices to approximate the posterior distribution using closed-form expressions. This speeds up inference and circumvents problems of sample convergence and mixing, making it suitable for fitting large data sets or exploring other models [RMC09].

INLA can be used for all models of the following form,

$$\begin{aligned} y_i|\boldsymbol{x},\boldsymbol{\theta} &\sim \pi(y_i|x_i,\boldsymbol{\theta}), \qquad i = 1,...,n, \\ \boldsymbol{x}|\boldsymbol{\theta} &\sim \mathcal{N}\left(\mu(\boldsymbol{\theta}), \boldsymbol{Q}(\boldsymbol{\theta})^{-1}\right), \\ \boldsymbol{\theta} &\sim \pi(\boldsymbol{\theta}). \end{aligned}$$

As introduced in subsection 3.5.1, $\boldsymbol{y}$ are the observed data, $\boldsymbol{x}$ is a Gaussian field, $\boldsymbol{\theta}$ represents the hyperparameters, while $\mu(\boldsymbol{\theta})$ and $\boldsymbol{Q}(\boldsymbol{\theta})$ denote the mean and precision matrix respectively. To ensure fast inference, the dimension of the hyperparameter vector $\boldsymbol{\theta}$ should be small, since the approximations are computed by numerical integration over the hyperparameter space.
In most cases, the observations y_i are assumed to belong to the exponential family with mean $\mu_i = g^{-1}(\eta_i)$. As shown in equation (3.31), η_i accounts for the effects of several covariates in an additive way, which makes it suitable for a wide range of models, including spatial and spatio-temporal models, since $\left\{f^{(j)}\right\}$ can take very different forms.
Let $\boldsymbol{x} = \left(\alpha, \{\beta_k\} | \theta \sim \mathcal{N}\left(\mu(\theta), \boldsymbol{Q}(\theta)^{-1}\right)\right)$ be the vector of latent Gaussian variables, and let $\boldsymbol{\theta}$ be the vector of hyperparameters, which are not required to be Gaussian. INLA calculates accurate and fast approximations for the posterior marginals of the components of the latent Gaussian variables

$$\pi(x_i|\boldsymbol{y}), \qquad i = 1, ..., n,$$

as well as the posterior marginals for the hyperparameters of the latent Gaussian model

$$\pi(\theta_j|\boldsymbol{y}), \qquad j = 1, ..., \dim(\boldsymbol{\theta}).$$

For each element x_i of $\boldsymbol{x}$ the posterior marginals are given by

$$\pi(x_i|\boldsymbol{y}) = \int \pi(x_i|\boldsymbol{\theta}, \boldsymbol{y}) \pi(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}, \tag{3.54}$$

and the posterior marginal for the hyperparameters can be expressed by

$$\pi(\theta_j|\boldsymbol{y}) = \int \pi(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}_{-j}. \tag{3.55}$$

$\pi(x_i|\boldsymbol{y})$ is approximated by combining analytical approximations to the full conditionals $\pi(x_i|\boldsymbol{\theta}, \boldsymbol{y})$ and $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ and numerical integration routines to integrate out $\boldsymbol{\theta}$. Similarly, $\pi(\theta_j|\boldsymbol{y})$ is approximated by approximating $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ and integrating out $\boldsymbol{\theta}_{-j}$. In particular, the posterior density of $\boldsymbol{\theta}$ is obtained through Gaussian approximation for the posterior of the latent field, $\widetilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$, evaluated at the posterior mode, $x^*(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{x}} \pi_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$,

$$\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \left.\frac{\pi(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{y})}{\widetilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})}\right|_{\boldsymbol{x}=x^*(\boldsymbol{\theta})}. \tag{3.56}$$

Next, the following nested approximations are constructed,

$$\widetilde{\pi}\left(x_i|\boldsymbol{y}\right) = \int \widetilde{\pi}\left(x_i|\boldsymbol{\theta},\boldsymbol{y}\right)\widetilde{\pi}\left(\boldsymbol{\theta}|\boldsymbol{y}\right)d\boldsymbol{\theta}, \qquad \widetilde{\pi}\left(\theta_j|\boldsymbol{y}\right) = \int \widetilde{\pi}\left(\boldsymbol{\theta}|\boldsymbol{y}\right)d\boldsymbol{\theta}_{-j}. \tag{3.57}$$

Finally, these approximations are numerically integrated with respect to $\boldsymbol{\theta}$

$$\widetilde{\pi}\left(x_i|\boldsymbol{y}\right) = \sum_k \widetilde{\pi}\left(x_i|\theta_k,\boldsymbol{y}\right)\widetilde{\pi}\left(\theta_k|\boldsymbol{y}\right)\times\Delta_k, \tag{3.58}$$

$$\widetilde{\pi}\left(\theta_j|\boldsymbol{y}\right) = \sum_l \widetilde{\pi}\left(\theta_l^*|\boldsymbol{y}\right)\times\Delta_l^*, \tag{3.59}$$

with Δ_k and Δ_l^* representing the area weights corresponding to θ_k and θ_l^*.
To obtain the approximations for the posterior marginals for the x_i's conditioned on selected values of θ_k and $\widetilde{\pi}\left(x_i|\theta_k,\boldsymbol{y}\right)$, a Gaussian, Laplace or simplified Laplace approximation can be used. Using a Gaussian approximation derived from $\widetilde{\pi}_G\left(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y}\right)$ is the simplest and fastest solution, but in some situations it produces errors in the location and is unable to capture skewness behaviour. Therefore, the Laplace approximation is favoured over the Gaussian approximation, although it is relatively expensive. The simplified Laplace approximation is associated with lower costs and addresses inaccuracies of the Gaussian approximation in terms of location and skewness in a satisfactory manner [Mor19].

3.6 Bayesian Spatial Models

In general, it can be assumed that areas in close proximity to each other have a more frequent burden of disease than areas that are further away from each other. By setting up a neighbourhood structure, this "proximity" can be defined. It is assumed that i and j are neighbours if they share a common boundary, denoted $i \sim j$. The set of neighbours of region i is denoted by δ_i and its size is given by n_{δ_i}.

3.6.1 Besag Spatial Models

3.6.1.1 Besags' Improper Spatial Model

A commonly used approach to modelling spatial correlation is the Besag model, also known as an intrinsic GMRF model. The conditional distribution for a random vector $\boldsymbol{x} = (x_1, ..., x_n)^T$ is given by

$$x_i|\boldsymbol{x}_{-i}, \tau_x \sim \mathcal{N}\left(\frac{1}{n_{\delta_i}} \sum_{j \in \delta_i} x_j, \frac{1}{n_{\delta_i}\tau_x}\right), \tag{3.60}$$

with τ_x as a precision parameter. The mean of the effects over all neighbours is given by the mean of x_i, while the precision is proportional to the number of neighbours. The joint distribution for $\boldsymbol{x}$ is given by

$$\pi\left(\boldsymbol{x}|\tau_x\right) \propto \exp\left(-\frac{\tau_x}{2} \sum_{i \sim j} (x_i - x_j)^2\right) \propto \exp\left(-\frac{\tau_x}{2} \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x}\right). \tag{3.61}$$

The precision matrix $\boldsymbol{Q}$ is given by

$$Q_{ij} = \begin{cases} n_{\delta_i} & i = j, \\ -1 & i \sim j, \\ 0 & \text{else.} \end{cases} \tag{3.62}$$

Q is a singular matrix, i.e. it has a non-empty null space $\boldsymbol{V}$, hence the model is called intrinsic or Besags' improper spatial model.
The Besag model for spatial effects has one hyperparameter, the precision τ_x, which is represented as

$$\theta_1 = \log \tau_x. \tag{3.63}$$

The prior is defined on θ_1 [Bes74; Rie+16].

3.6.1.2 Besags' Proper Spatial Model

To overcome this impropriety, the precision matrix has to be redefined as follows,

$$Q_{ij} = \begin{cases} \tau_x \left(n_{\delta_i} + d\right) & i = j, \\ -\tau & \text{else.} \end{cases} \tag{3.64}$$

$d > 0$ is an additional term added to the diagonal to control the "properness". The conditional distribution for the proper version of the Besag model is then given by

$$x_i | \boldsymbol{x}_{-i}, \tau_x, d \sim \mathcal{N}\left(\frac{1}{d + n_{\delta_i}} \sum_{i \sim j} x_j \frac{1}{\tau_x \left(d + n_{\delta_i}\right)}\right). \tag{3.65}$$

The proper version of the Besag model for spatial effects has two hyperparameters, the precision τ_x, which is represented as

$$\theta_1 = \log \tau_x \tag{3.66}$$

and the diagonal parameter d, which is represented as

$$\theta_2 = \log d. \tag{3.67}$$

The priors are defined on θ_1 and θ_2, respectively [Bes74; Rie+16].

3.6.2 The Besag-York-Mollié Model

The Besag-York-Mollié (BYM) model is a lognormal Poisson model that is a combination of a Besag model u and an ordinary random effect component v for non-spatial heterogeneity. It combines the regional spatial effect $\boldsymbol{x}$ into the sum of an unstructured and a structured spatial component, so that $\boldsymbol{x} = \boldsymbol{v} + \boldsymbol{u}$.
$\boldsymbol{v} \sim \mathcal{N}\left(0, \tau_v^{-1} \boldsymbol{I}\right)$ accounts for pure overdispersion, while $\boldsymbol{u} \sim \mathcal{N}\left(\boldsymbol{0}, \tau_u^{-1} \boldsymbol{Q}^-\right)$ is the Besag model.
By using a spatial and a non-spatial error term, the overdispersion that is not modelled by the Poisson variables is taken into account. Thus, if the observed variance is not fully explained by the spatial structure of the data, the error terms explain the rest of the variance.
The resulting covariance matrix of $\boldsymbol{x}$ is given by

$$\text{Var}\left(\boldsymbol{x} | \tau_u, \tau_v\right) = \tau_v^{-1} \boldsymbol{I} + \tau_u^{-1} \boldsymbol{Q}^-, \tag{3.68}$$

where $\boldsymbol{Q}^{-}$ denotes the generalised inverse of $\mathbf{Q}$.
The hyperparameters of the model are the precision τ_u of the Besag model u and the precision τ_v of the iid model v. They are represented as

$$\boldsymbol{\theta} = (\theta_1, \theta_2) = (\log \tau_v, \log \tau_u) \tag{3.69}$$

and the prior is defined on $\boldsymbol{\theta}$ [BYM91; Rie+16].

3.6.3 The Leroux Model

One problem with the BYM model is that the structured and unstructured components are not identifiable because they cannot be considered independently. Moreover, τ_v and τ_u do not represent variability at the same level, which makes the choice of hyperpriors difficult. The Leroux model is formulated in such a way that the compromise between the two variations is made more explicit. It is assumed that $\boldsymbol{x}$ follows a normal distribution with zero mean and covariance matrix

$$\text{Var}\left(\boldsymbol{x}|\tau_x, \phi\right) = \tau_x^{-1}\left((1-\phi)\,\boldsymbol{I} + \phi\boldsymbol{Q}\right)^{-1}, \tag{3.70}$$

with $\phi \in [0, 1]$ as mixing parameter. For $\phi = 0$ the model reduces to pure overdispersion and for $\phi = 1$ to the Besag model. The conditional expected value of x_i for all other random effects is the weighted mean of the unstructured model with zero mean and the mean of the Besag model, while the conditional variance is the weighted mean of τ_x^{-1} and $(\tau_x \cdot\ n_{\delta_i})^{-1}$ [LLB00; Rie+16].

3.6.4 The BYM2 Model

One problem that all the aforementioned models have is the lack of scaling of the spatially structured component. Scaling facilitates the assignment of hyperpriors and ensures that the interpretation of hyperpriors remains the same across different areas.
Another problem is that the marginal standard deviations of the commonly used IGMRF priors can vary greatly, a fact that should be taken into account by assigning hyperpriors to the precision parameters of these models.
Since the Besag model penalises a local deviation from its null space, the hyperprior will control this local deviation and thus affect the smoothness of the estimated spatial effects. If the estimate of the field is too smooth, the precision will be large and the spatial variation may be blurred. On the other hand, if the precision is too

small, the model could overfit due to the large local variability.
The marginal variances $\tau_x^{-1}\left[\boldsymbol{Q}^{-}\right]_{ii}$ depend on the structure of the graph, which is reflected in the structure matrix $\boldsymbol{Q}$. A generalised variance can be calculated as the geometric mean of the marginal variance as follows

$$\sigma^2_{\text{GV}}(\boldsymbol{u}) = \exp\left(\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{1}{\tau_x}\left[\boldsymbol{Q}^{-}\right]_{ii}\right)\right) = \frac{1}{\tau_x}\exp\left(\frac{1}{n}\sum_{i=1}^{n}\log\left(\left[\boldsymbol{Q}^{-}\right]_{ii}\right)\right). \quad (3.71)$$

In order to unify the interpretation of a chosen prior for τ_x and make it transferable across domains, the structured effect must be scaled such that $\sigma^2_{\text{GV}}(\boldsymbol{x}) = \tau_x^{-1}$. This implies that τ_x denotes the accuracy of the (marginal) deviation from a constant level, independent of the underlying graph.
A modification of the BYM model that addresses this scaling problem is the BYM2 model. It uses a scaled structured component $\boldsymbol{u}_*$, where $\boldsymbol{Q}_*$ denotes the precision matrix of the Besag model, scaled with the marginal variance σ^2_{GV} as a factor. The random effect is then given by

$$\boldsymbol{x} = \frac{1}{\tau_x}\left(\sqrt{1-\phi}\boldsymbol{v} + \sqrt{\phi}\boldsymbol{u}_*\right), \quad (3.72)$$

with covariance matrix

$$\text{Var}(\boldsymbol{x}|\tau_x,\phi) = \frac{1}{\tau_x}\left((1-\phi)\boldsymbol{I} + \phi\boldsymbol{Q}_*^{-}\right). \quad (3.73)$$

Equation 3.72 emphasises the trade-off between pure overdispersion and spatially structured correlation, where $0 \leq \phi \leq 1$ measures the fraction of the marginal variance explained by the structured effect. For $\phi = 0$ the model reduces to pure overdispersion, while for $\phi = 1$ it becomes a Besag model [Mar+14; Rie+16].

3.7 Goodness-of-Fit indicators

The goodness of fit indicates "how well" an estimated model can explain a set of observations. Measures of goodness of fit allow a statement to be made about the discrepancy between the theoretical values of the random variables under investigation, which are expected or predicted on the basis of the model, and the values actually measured.
The goodness of fit of a model to available data can be assessed with the help of statistical tests or suitable ratios.

3.7.1 The Akaike Information Criterion

The historically oldest criterion was proposed in 1973 by Hirotsugu Akaike (1927-2009) as an information criterion and is known today as the Akaike information criterion (AIC). The AIC is one of the most frequently used criteria for model selection in the context of likelihood-based inference.
Let the population contain the distribution of a variable with unknown density function p. The maximum likelihood estimation assumes a known distribution with an unknown parameter θ, hence the density function can be written as $q(\theta)$. The Kullback-Leibler divergence is used as a distance measure between p and $q\left(\widehat{\theta}\right)$ with $\widehat{\theta}$ the estimated parameter from the maximum likelihood estimation. The better the maximum likelihood model, the small the Kullback-Leibler divergence $D(P||Q)$.
For a maximum likelihood model with a p-dimensional parameter vector $\widehat{\boldsymbol{\theta}}$, the Akaike information criterion is defined as

$$\text{AIC} = -2l\left(\widehat{\boldsymbol{\theta}}_{ML}\right) + 2p, \tag{3.74}$$

with l the log-likelihood function [Aka74].

3.7.2 The Deviance Information Criterion

In statistics, the deviance information criterion, or DIC for short, is a measure (criterion) for the prediction error of a model. This measure is an information criterion and belongs to the environment of the Bayesian method for model comparisons. The smaller the deviance information criterion, the better the model fit. The deviance information criterion can be regarded as the Bayesian equivalent of the Akaike

information criterion.
The deviance is defined as

$$D(\boldsymbol{\theta}) = -2\log\left(l\left(\boldsymbol{y}|\boldsymbol{\theta}\right)\right) + C, \tag{3.75}$$

with $\boldsymbol{y}$ the data, $\boldsymbol{\theta}$ the unknown parameters of the model and l the likelihood function. C is a constant that cancels out in all calculations that compare different models and therefore it does not need to be known [NW72].
The DIC is given by

$$\text{DIC} = D\left(\bar{\boldsymbol{\theta}}\right) + 2p_D, \tag{3.76}$$

with

$$p_D = \overline{D\left(\boldsymbol{\theta}\right)} - D\left(\bar{\boldsymbol{\theta}}\right), \tag{3.77}$$

where $\bar{\boldsymbol{\theta}}$ is the expected value of $\boldsymbol{\theta}$ [Spi+14].

3.7.3 The Watanabe-Akaike Information Criterion

The Watanabe-Akaike information criterion (WAIC) is the generalised AIC onto singular statistical models.
The WAIC is given by

$$\text{WAIC} = -2\text{LLPD} + 2p_{\text{WAIC}}, \tag{3.78}$$

with the log pointwise predictive density (LLPD) given by

$$\text{LLPD} = \sum_{i=1}^{n} \log\left(\int \pi\left(y_i|\boldsymbol{\theta}\right)\pi_{\text{post}}\left(\boldsymbol{\theta}\right)\right) d\boldsymbol{\theta}. \tag{3.79}$$

LLPD can be seen as the Bayesian analogue of $l\left(\widehat{\boldsymbol{\theta}}_{ML}\right)$ in the calculation of the AIC. The penalty term of the WAIC is also fully Bayesian and given by

$$p_{\text{WAIC}} = \sum_{i=1}^{n} \text{Var}_{\text{post}}\left(\log\left(\pi\left(y_i|\boldsymbol{\theta}\right)\right)\right), \tag{3.80}$$

where the term represents the variance of the individual terms in the LLPD over all data points [WO10; Yon18].

3.7.4 The Conditional Predictive Ordinate

The conditional predictive ordinate (CPO) is a Bayesian diagnostic that can be used to detect surprising observations. It is often used in the context of univariate sampling, the multivariate normal distribution and regression models.
The conditional predictive ordinate is given by

$$\text{CPO} = \pi\left(y_i|\boldsymbol{y}_{-i}\right) \tag{3.81}$$

with $\boldsymbol{y}$ the data, $\boldsymbol{y}_{-i}$ the data without the i-th observation, and $\pi\left(\cdot|\boldsymbol{y}_{-i}\right)$ the predictive distribution of a new observation at $\boldsymbol{y}_{-i}$. Low values of CPO are an indication that y_i is surprising given prior knowledge and the other observations [Pet90; Cox80].

3.8 Model Issues

One problem that plagues these models is that they cannot be directly compared due to their different parameterisations and the fact that the precision in these models is interpreted differently. Since neither a Besag model nor a BYM model nor a Leroux model is scaled, the precision parameter is not representative of the marginal precision but is confounded with the mixing parameter. Therefore, the effect of a prior assigned to the precision parameter is dependent on the graph structure of the application. Thus, a given prior is not transferable between different applications if the underlying graph changes. Furthermore, the goal of the BYM2 model is not to optimise goodness-of-fit indicators, but to provide a meaningful model formulation where all parameters have a clear meaning. By mapping the precision parameter to the marginal standard deviation, the model parameters are flexible and the assignment of meaningful hyperpriors is made easier [Rie+16].
Additionally, the goodness-of-fit indicators introduced in Section 3.7 have their own problems. The DIC, for example, produces unreasonable results if the posterior distribution is not well summarised by its mean, while the WAIC is based on a data partition that would create difficulties for structured models, such as for spatial or network data [GHV14].
Finally, the choice of the prior also affects the value of these criteria and depending on the values chosen for the PC priors used in this work, overfitting of the models may occur, which is then reflected in these criteria, but more on this later.

3.9 The Variance Inflation Factor

The Variance Inflation Factor (VIF) is a measurement that can be used to avoid multicollinearity between covariates. The VIF quantifies the severity of multicollinearity in a generalised linear model. It provides an index that measures the extent to which the variance of an estimated regression coefficient is increased due to collinearity. For $p-1$ independent variables,

$$\text{VIF}_i = \frac{1}{1-R^2}, \qquad i = 1, ..., p-1, \tag{3.82}$$

with R^2 the coefficient of determination. In most literature, a value of at least 5 is suggested as too high and is therefore used as the threshold in this work [CS02].

4 Analysis of Geospatial Health Data

Healthcare data provides information for detecting public health problems and reacting adequately when they occur. With this information, prevention and control of a multitude of health conditions including infectious diseases, non-communicable diseases, injuries and health-related behaviours can be achieved. To analyse and interpret health data, the process involves a wide variety of system designs, analytical methods, modes of presentation and interpretive uses[TC+00]. Descriptive methods generally form the basis of routine reporting of surveillance data. Rather than focusing on observed patterns in the data, these may also attempt to compare the relative occurrence of health outcomes in different subgroups. More specific hypotheses can be explored using inferential methods. The aim of these methods is to draw statistical inferences about patterns or outcomes of health.

The increasing availability of geo-referenced health data, population data, satellite imagery of environmental factors influencing levels of disease activity, and the development of geographic information systems (GIS) and address geocoding software, the rise of studies of spatial and spatio-temporal variation in disease has been facilitated. John Snow's investigation of the cholera outbreak in London in 1854 offers one of the most well-known examples of spatial analysis. By using a map, Snow illustrated how cholera deaths seemed to accumulate around a public water pump. Evaluating the spatial pattern of cholera cases was essential to identify the source of infection and supported the theory of cholera transmission through drinking water [Sno57].

A broad range of spatial and spatio-temporal methods exist for disease surveillance, including methods for disease mapping, clustering and geographic correlation studies. These methods can be used to identify areas of high risk, risk factors, evaluate spatial variations in temporal trends, measure excess disease risk near a suspected source and detect outbreaks at an early stage.

49

4.1 Geographic Data

In spatial statistics, two fundamental types of geographic data exist, namely *vector data* and *raster data*. In the vector data model, the world is represented by points, lines and polygons with discrete, well-defined boundaries, which tends to result in high accuracy. Raster data, on the other hand, divides the surface into cells of uniform size, and raster datasets are used as the basis for background images in web mapping.
Determining which data type to use depends on the domain of the application. Vector data dominates in the social sciences because human settlements typically have discrete boundaries, while raster data are commonly used in many environmental sciences because they are based on remote sensing data. Naturally, there is also some overlap and both types can be used together or one form can be converted into the other [LNM19].

4.1.1 Vector Data

The geographic vector data model is based on points located within a *coordinate reference system* (CRS), in which points either represent self-standing features or form more complex geometric shapes, i.e. lines and polygons. Using this system, Trondheim can be represented by the coordinates $(10.4, 63.4)$, meaning 10.4 degrees east of the prime meridian and 63.4 degrees north of the equator. It could also be written as $(1157722.70, 9199010.75)$, which is the position of Trondheim using the Web Mercator projection, the de facto standard for web mapping applications. More will be said about CRS later, but for now it is sufficient to know that it is possible to display coordinates in various ways. An example of a CRS is shown in Figure 4.1.

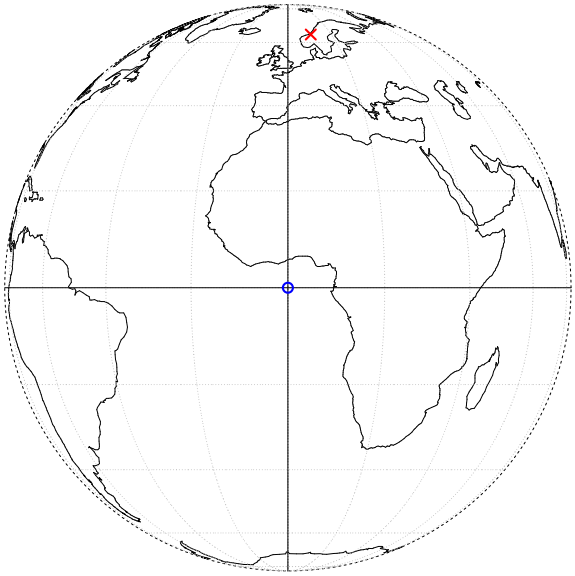

Fig. 4.1: A geographic CRS with an origin at 0° longitude and latitude. The red X denotes the location of Trondheim.

Different Types of Vector Data

As mentioned earlier, there are different types of vector data. There are 17 different geometry types in the standard *simple features*, but there are seven core types that can be used in most analysis software. These types are visualised in Figure 4.2.

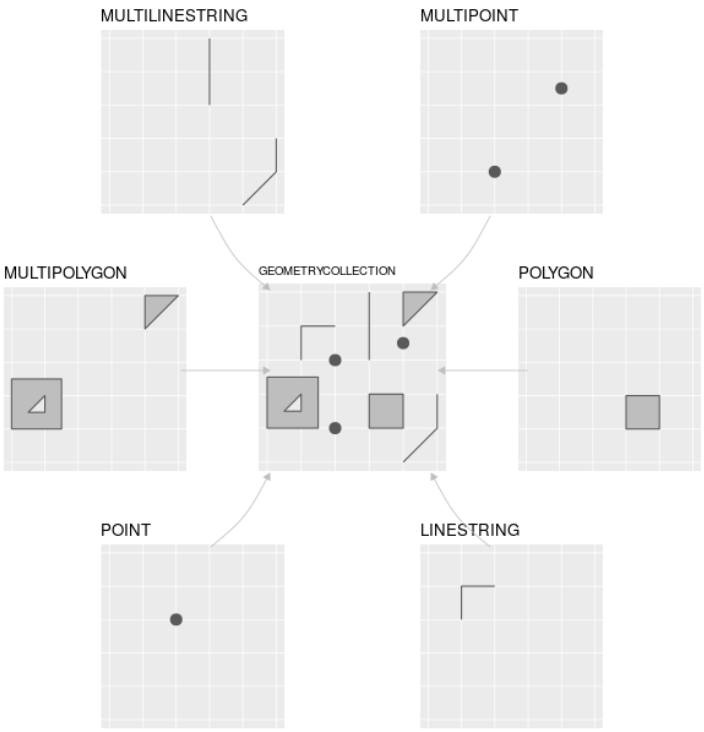

Fig. 4.2: The most commonly used simple feature types.

Simple Features was developed by the Open Geospatial Consortium and is an open, standardised, hierarchical data model that represents a wide range of geometry types. The use of this data model ensures that scientific work can be transferred to other institutions, e.g. when importing from and exporting to spatial databases [LNM19].

4.1.2 Raster Data

The geographic raster data model consists in most cases of a raster header and a matrix representing uniformly distributed cells/pixels. The raster header defines the CRS, the origin (starting point) and the extent. Since the number of columns and rows and the resolution of the cell size are stored in the extent, starting from the origin, it is easy to access and change each cell by its ID or by specifying the row and column number. In this type of representation, the coordinates of the four vertices of each cell are not explicitly stored, instead only the origin is stored. This speeds up data processing and makes it more efficient, but each raster layer can only contain a single value, which can be either numeric or categorical. Typically, raster maps are

used to depict continuous features such as elevation or temperature, but categorical variables, for example soil or land cover, as shown in Figure 4.3 [LNM19].

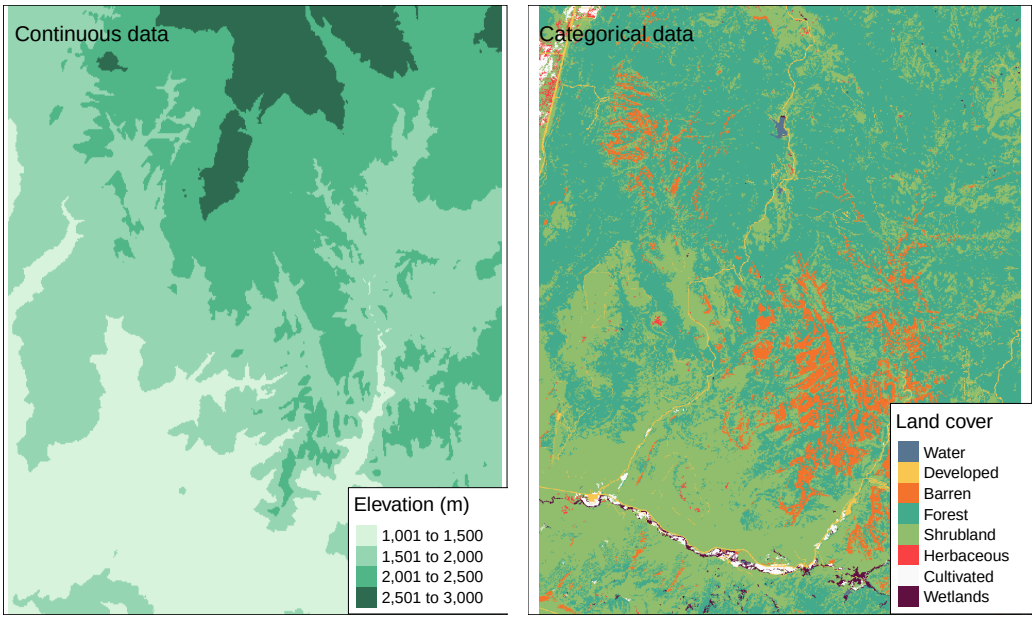

Fig. 4.3: An example of continuous and categorical raster data

Coordinate Reference Systems

A common denominator of vector and raster data are that both use the coordinate reference system (CRS), which defines how spatial elements relate to the surface of the Earth. The CRS can be either geographic or projected.

Geographic Coordinate Systems

Geographic coordinate systems use two values, *longitude* and *latitude*, to identify any location on Earth. Longitude is defined as the east-west location at an angular distance from the prime meridian plane, while latitude is the angular distance north or south of the equator. Consequently, distances in geographic CRS are not measured in metres.
The Earth's surface is typically represented in geographical coordinate systems by

a spherical or ellipsoidal surface. The former assumes that the Earth is a perfect sphere of a certain radius, which has the advantage of being a simplistic model, but is associated with inaccuracies owing to the fact that the Earth is not a sphere. Ellipsoidal models are defined by the equatorial radius and the polar radius, providing a better model since the equatorial radius is approximately 11.5 km longer than the polar radius.
The *datum* is a broader component of CRS that contains information about which ellipsoid to use and the exact relationship between Cartesian coordinates and the location on the Earth's surface. The notation *proj4string* is used to store these additional details. It allows for local variations of the Earth's surface, such as large mountain ranges, to be taken into account in local CRS. Datum can again be divided into two categories, *local* and *geocentric*, the difference being that in the local datum the ellipsoidal surface is shifted to match the surface at a particular location, whereas in the geocentric datum the centre of gravity of the Earth is the centre and the accuracy of the projections is not optimised for any particular location [LNM19].

Projected Coordinate Systems

Projected CRS are based on Cartesian coordinates on an implicitly flat surface and have an origin, x and y axes, and a linear unit of measurement, metres for instance. They are based on geographic CRS and rely on map projections to convert between the three-dimensional surface of the Earth and the east/north values (x and y) in a projected CRS.
This transition always entails some distortion, skewing some of the properties of the earth's surface, such as area, direction, distance and shape. Generally, the name of a projection is based on a property it preserves, e.g. equal area projection preserves area, equidistant projection preserves distance and conformal projection preserves local shape.
Again, subgroups exist in projection coordinate systems, *conic*, *cylindrical* and *planar* projections. In a conic projection, the earth's surface is projected onto a cone along one or two tangent lines. Along these lines the distortions are minimised and increase with the distance to the lines. The projection is therefore best suited for maps of mid-latitude areas. Cylindrical projections map the surface onto a cylinder. These types of projections can be created by touching the surface of the Earth along one or two tangent lines. They are often used to map the entire Earth. A planar projection projects data onto a flat surface that touches the globe at a point or along a tangent line, and is typically used in mapping polar projections [LNM19].

4.2 Spatial Point Processes

A stochastic process that describes the location of particular events/points that occur in a region is known as a point process. The number of points as well as the location of the points are random. An example of a point process would be the number of earthquakes and their locations.

4.2.1 Fundamentals of Point Processes

Let Z be a random, at most countable set of points in a space $\mathbb{X}$, for example $\mathbb{R}^d$. Ignoring measurability issues, Z can be thought of as a mapping $\omega \mapsto Z(\omega)$ from Ω into the set of countable subsets of $\mathbb{X}$, where $(\Omega, \mathcal{F}, \mathbb{P})$ defines an underlying probability space. Z can then be identified with the family of mappings

$$\omega \mapsto \eta(\omega, B) := \text{card}\,(z(\omega) \cap B), \qquad B \subset \mathbb{X}, \tag{4.1}$$

which counts the number of points from Z in B. For any fixed $\omega \in \Omega$, $\eta(\omega, \cdot)$ is the counting measure supported by $Z(\omega)$ [CI80].
For a general definition of a point process, let $(\mathbb{X}, \mathcal{X})$ be a measurable space and let $N_{<\infty}(\mathbb{X}) \equiv N_{<\infty}$ be the space of all measures μ on $\mathbb{X}$ such that $\mu(B) \in \mathbb{N}_0 := \mathbb{N} \cup \{0\} \forall B \in \mathcal{X}$. Let $N(\mathbb{X}) \equiv N$ be the space of all measures describable as a countable sum of measures from $N_{<\infty}$, for example the *zero measure* 0 which is equal to 0 on $\mathcal{X}$. In general, any sequence $(x_n)_{n=1}^k$ of elements of $\mathbb{X}$, where $k \in \overline{\mathbb{N}} := \mathbb{N} \cup \{\infty\}$ denotes the number of terms in the sequence, can be used to define a measure

$$\mu = \sum_{n=1}^{k} \delta_{x_n}. \tag{4.2}$$
$$\Rightarrow \mu(B) = \sum_{n=1}^{k} \mathbf{1}_B(x_n), \qquad B \in \mathcal{X}.$$

More generally, for any measurable $f : \mathbb{X} \to [0, \infty]$,

$$\int f d\mu = \sum_{n=1}^{k} f(x_n) \tag{4.3}$$

For $k = 0$ in (4.2), μ is equal to the zero measure. The point set $\boldsymbol{x} = (x_1, ..., x_n)^T$ is said to be not pairwise different and if $x_i = x_j$ with $i \neq j$, μ is said to have multiplicities. The multiplicity of x_i is equal to the number

$$\text{card}\left\{j \leq k : x_j = x_i\right\}.$$

Any μ of the form (4.2) is interpreted as a counting measure with possible multiplicities, but in general it cannot be guaranteed that every $\mu \in N$ can be written in this particular form.
A point process η on $\mathbb{X}$ is called *proper point process* if random elements $X_1, X_2, ...$exist $\in \mathbb{X}$ and a $\overline{\mathbb{N}}_0$-valued random variable κ such that almost surely

$$\eta = \sum_{n=1}^{\kappa} \delta_{X_n}. \tag{4.4}$$

For $\kappa = 0$ this is the zero measure on $\mathbb{X}$.
This terminology is motivated by the intuition that a point process is a (random) set of points, rather than an integer measure. A proper point process fits this intuition better, since it can be interpreted as a countable set of points in $\mathbb{X}$ [LP17, pp. 9–12].

4.2.2 Poisson Processes

Poisson processes are defined by the fact that the number of points in a given set follows a Poisson distribution. Furthermore, the numbers of points in disjoint sets are stochastically independent.
In application, Poisson processes are used in a wide range of fields, including biology, economics and image processing.
Let λ be an s-finite measure on $\mathbb{X}$. Let a *Poisson process* with intensity measure λ be defined as a point process η on $\mathbb{X}$ with the following two properties:

1. $\forall B \in \mathcal{X} : \eta(B) \sim \text{Po}(\lambda(B); k)\, \forall k \in \mathbb{N}_0 \Longleftrightarrow \mathbb{P}(\eta(B) = k)$
2. $\forall m \in \mathbb{N}$ and all pairwise disjoint sets $B_1, ..., B_m \in \mathcal{X}$: the random variables $\eta(B_1), ..., \eta(B_m)$ are independent.

A point process satisfying the second of these conditions is called *completely independent*. If η is a Poisson process with intensity measure λ, then

$$\mathbb{E}[\eta(B)] = \lambda(B). \tag{4.5}$$

For the zero measure,

$$\mathbb{P}\left(\eta(\mathbb{X}) = 0\right) = 1,$$

with $\lambda = 0$ [LP17, p. 19].

4.3 Modeling and Visualising Health Data

4.3.1 Areal Data

Areal or lattice data are the result of segmenting a fixed domain into a finite number of sub-regions where results are aggregated, e.g. the number of infections with a specific disease in districts or the number of overweight people in provinces. Often the aim of disease risk models is to assess the risk within the same areas for which data are available. This can be done with a simple measure such as the *standardised incidence ratio* (SIR) or by using a Bayesian hierarchical model, which allows information to be drawn from neighbouring areas and incorporates covariates, thereby smoothing and reducing extreme values.
A widely used model is the *Besag-York-Mollié* (BYM) [BYM91], which takes spatial correlation and the potential for observations in neighbouring areas to be more similar than those in distant regions into account. It includes a spatial random effect that smoothes the data according to a neighbourhood structure, and an unstructured exchangeable component that models uncorrelated noise. In settings where disease numbers are monitored over time, spatio-temporal models account for temporal correlations in addition to spatial correlation, while also accounting for spatio-temporal interactions [Mor19].

4.3.1.1 Spatial Neighbourhood Matrices

Spatial or proximity matrices are useful for exploratory analysis of area data. Let w_{ij} denote the (i, j) element of a *spatial neighbourhood matrix* $\boldsymbol{W}$. w_{ij} connects the two areas in some spatial way. The neighbourhood structure over the complete study region is defined by $\boldsymbol{W}$, and the elements of the matrix can be considered as weights. The closer j is to i, the more weight is associated with it. The simplest neighbourhood definition is given by the binary matrix

$$w_{ij} = \begin{cases} 1 & \text{if regions } i \text{ and } j \text{ share a border} \\ 0 & \text{else} \end{cases} \tag{4.6}$$

Since a region cannot share a boundary with itself, $w_{ii} = 0$ [Mor19].
In Figure 4.4, the number of shared borders of each canton in Switzerland are mapped.

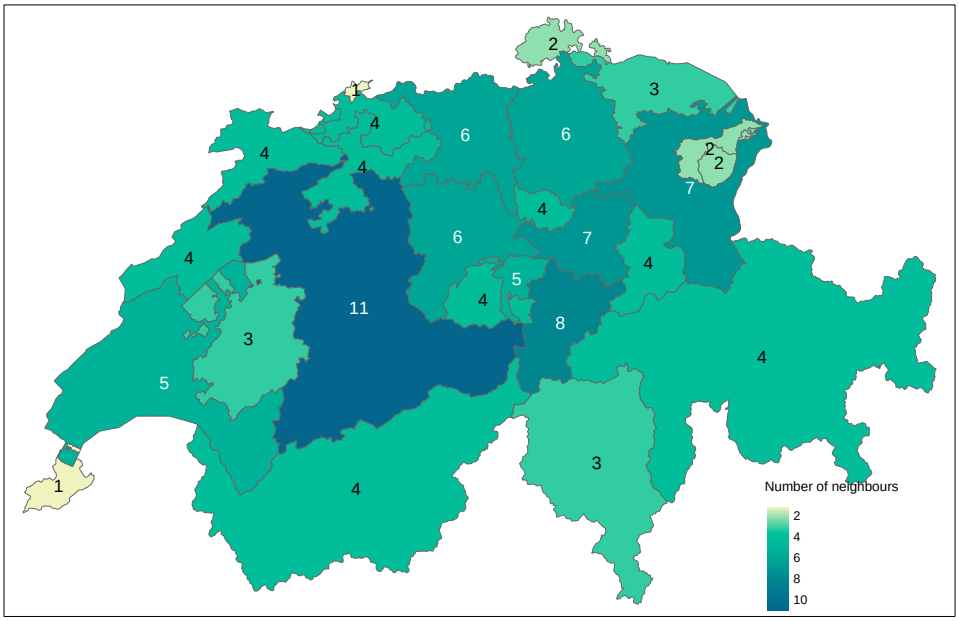

Fig. 4.4: The number of shared borders of cantons in Switzerland

4.3.1.2 Moran's I

Moran's I is a measure of spatial autocorrelation developed by Patrick Moran. Spatial autocorrelation is characterised by a correlation in a signal between close locations in space. Spatial autocorrelation is inherently more complex than one-dimensional autocorrelation due to the fact that spatial correlation is multidimensional (i.e. 2 or 3 spatial dimensions) and multi-directional. The formula for Moran's I is given by

$$I = \frac{n}{W} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \left(x_i - \overline{x}\right) \left(x_j - \overline{x}\right)}{\sum_{i=1}^{n}}, \tag{4.7}$$

with n denoting the number of spatial units indexed by i and j, x the parameter of interest, $\boldsymbol{w}$ a spatial neighbourhood matrix and W the sum of all w_{ij}.
Using Moran's I, a test for spatial autocorrelation can be constructed with the following hypotheses:

$$H_0: \text{ No spatial autocorrelation vs. } H_1: \text{ Spatial autocorrelation.} \tag{4.8}$$

Under H_0 the expected value is given by

$$\mathbb{E}\left[I\right] = \frac{-1}{n-1}. \tag{4.9}$$

As n approaches infinity, the expected value therefore approaches 0 [Mor50].

4.3.1.3 Standardised Incidence Ratio

A basic measure of disease risk is the *standardised incidence ratio*, which yields an estimate in each of the areas that form a partition of the study region. It is defined as the ratio of observed counts to expected counts

$$\text{SIR}_i = \frac{Y_i}{E_i}. \tag{4.10}$$

E_i represents the sum of the expected number of cases of a given area i that behave according to the way the standard population behaves. It is calculated using indirect standardisation as

$$E_i = \sum_{j=1}^{m} r_j^{(s)} n_j^{(i)}, \tag{4.11}$$

with $r_j^{(s)}$ the rate in stratum j in the standard population and $n_j^{(i)}$ the population in stratum j of area i. If the stratum information is unavailable, the expected counts can be calculated as follows

$$E_i = r^{(s)} n^{(i)},$$

where $r^{(s)}$ denotes the rate in the standard population and $n^{(i)}$ is the population of area i. If the standardised incidence rate is greater than 1, area i has a higher risk than expected from the standard population, while for $\text{SIR}_i = 1$ the risk is the same and for $\text{SIR}_i < 1$ it is lower than expected. The ratio is also called the standardised mortality ratio when applied to mortality data [Mor19].

4.3.1.4 Spatial Small Area Disease Risk Estimation

While SIRs may prove useful in some situations, in areas with low population sizes or rare diseases, expected counts may be low, making SIRs insufficiently reliable for reporting. It is therefore preferable to assess disease risk using models that allow information to be borrowed from neighbouring areas and incorporate information from covariates, thus smoothing or shrinking extreme values due to small sample sizes [Gel+10].
The observed counts Y_i in area i are typically modeled with a Poisson distribution with mean $E_i\theta_i$, where E_i is the expected counts and θ_i denotes the relative risk in area i. To account for extra Poisson reliability, the logarithm of the relative risk is expressed as the total of the intercept and the random effects. θ_i quantifies whether area i has a higher $(\theta_i > 1)$ or lower $(\theta_i < 1)$ risk than the average risk in

the standard population. If the risk of an area i is half the average risk, then $\theta_i = 0.5$. The general model for spatial data is formulated as follows:

$$Y_i \sim \text{Po}\left(E_i\theta_i\right), \qquad i = 1, ..., n, \tag{4.12}$$

$$\log\left(\theta_i\right) = \alpha + u_i + v_i. \tag{4.13}$$

The overall risk in the region of study is represented by α, u_i is a random effect specific to each area to model the spatial dependence between relative risks, and v_i is an unstructured exchangeable component that models uncorrelated noise, $v_i \sim \mathcal{N}\left(0, \sigma_v^2\right)$. Covariates are often included to measure risk factors and other random effects to deal with different sources of variability. For example,

$$\log\left(\theta_i\right) = \boldsymbol{d_i\beta} + u_i + v_i,$$

with $\boldsymbol{d}_i = (1, d_{i1}, ..., d_{ip})$ a vector of the intercept and p covariates corresponding to the area i and $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)^T$ the vector of coefficients. An increase in d_j $(j = 1, ..., p)$ by one unit, leads to an increase in the relative risk by a factor of $\exp\left(\beta_j\right)$, provided that all other covariates remain constant.
In the Besag-York-Mollié (BYM) [BYM91] model, this spatial random effect u_i is assigned a conditional autoregressive (CAR) distribution that smooths the data according to a given neighbourhood structure that defines two areas as neighbours if they share a common boundary [Mor19].

4.3.1.5 Spatio-Temporal Small Area Disease Risk Estimation

When disease counts are monitored over time, spatio-temporal models are useful as they take into account not only the spatial structure but also temporal correlations and spatio-temporal interactions [MLB08]. Let Y_{ij} be the counts observed in area i and at time j, θ_{ij} be the relative risk, E_{ij} be the expected number of cases in area i and at time j, then

$$Y_{ij} \sim \text{Po}\left(E_{ij}\theta_{ij}\right), \qquad i = 1, ..., I,\ j = 1, ..., J. \tag{4.14}$$

$\log\left(\theta_{ij}\right)$ is written as the sum of several components, including spatial and temporal structures, to consider that neighbouring areas and successive times may have similar risk. Spatio-temporal interactions can be included to account for the fact that temporal trends may differ from area to area but may be more alike in neighbouring areas.

Bernardinelli et al. [Ber+95], for example, propose a spatio-temporal model with parametric time trends that expresses the logarithm of relative risks as

$$\log(\theta_{ij}) = \alpha + u_i + v_i + (\beta + \delta_i) \times t_j. \tag{4.15}$$

The intercept is denoted by α, $u_i + v_i$ is a random area effect, β represents a global linear trend effect and δ_i is an interaction between space and time which is the difference between β and the area-specific trend. For modeling u_i and δ_i, a CAR distribution is used and v_i is i.i.d.. This specification allows each of the areas to have its individual time trend, where the spatial intercept is given by $\alpha + u_i + v_i$ and the slope by $\beta + \delta_i$. δ_i is referred to as the differential trend of the i-th area and represents the amount by which the time trend of area i deviates from the overall time trend β. If $\delta_i \neq 0$, then area i has a time trend with a slope that is either steeper or less steep than the overall time trend β.
For models that do not demand linearity of the time trend, non-parametric models such as the one proposed by Knorr-Held [Kno00] can be used. This specific model incorporates spatial effects, temporal random effects and an interaction between space and time as follows:

$$\log(\theta_{ij}) = \alpha + u_i + v_i + \gamma_j + \phi_j + \delta_{ij}. \tag{4.16}$$

The intercept is again denoted by α, $u_i + v_i$ is a spatial random effect defined as before, i.e. u_i follows a CAR distribution and v_i is i.i.d.. $\gamma_j + \phi_j$ represents a temporal random effect and γ_j follows either a first order random walk in time (RW1)

$$\gamma_j | \gamma_{j-1} \sim \mathcal{N}\left(\gamma_{j-1}, \sigma_\gamma^2\right), \tag{4.17}$$

or second order random walk in time (RW2)

$$\gamma_j | \gamma_{j-1}, \gamma_{j-2} \sim \mathcal{N}\left(2\gamma_{j-1} - \gamma_{j-2}, \sigma_\gamma^2\right). \tag{4.18}$$

The unstructured temporal effect is given by $\phi_j \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma_\phi^2\right)$. The interaction between space and time, δ_{ij}, can be specified in a number of ways by combining the structure of the random effects that interact. The interactions proposed by Knorr-Held are those between the effects (u_i, γ_j), (u_i, ϕ_j), (v_i, γ_j) and (v_i, ϕ_j) [Kno00].
Using the last of these interactions leads to the assumption that there is no spatial or temporal structure on δ_{ij}. Thus, the interaction term can be modeled as
$\delta_{ij} \sim \mathcal{N}(0, \sigma_\delta^2)$ [Mor19].

4.3.1.6 Issues With Areal Data

The analysis of spatially aggregated data is subject to the "misaligned data problem" (MIDP), which arises when the data to be analysed is at a different scale from that at which it was collected [BCG14]. This may be solely due to the fact that the aim is to obtain the spatial distribution of a variable at a new spatial level of aggregation, e.g. if predictions are to be made at the county level with data that was originally collected at the postcode level. Another objective may be to try to find an association between variables available at different spatial scales, e.g. determining whether the risk of an unfavourable outcome provided at the country level correlates with exposure to an environmental pollutant measured at different stations, taking into account the population at risk and other demographic information available at the postcode level.
The Modifiable Area Unit Problem (MAUP) [Ope84] describes a problem where the inference may differ when the same underlying data are grouped at a new spatial level of aggregation. It consists of two interrelated effects, the first of which is the scale/aggregation effect. It relates to the different conclusions obtained when the same data are grouped into larger and larger areas. The other effect is the grouping/zoning effect, which accounts for the variability in results due to alternative formations of the areas, resulting in differences in area shape given the same or similar scales.
Ecological studies are defined by their reliance on aggregated data [Rob09] and the inherent potential for ecological fallacies. This phenomenon occurs when estimated associations obtained from the analysis of variables measured at the aggregate level lead to conclusions that differ from analyses based on the same variables measured at the individual level. This can be considered a special case of MAUP and the resulting so-called ecological bias is composed of two effects similar to the aggregation and zoning effects in MAUP. Namely, the aggregation bias caused by the aggregation of individuals and the specification bias due to the different distribution of confounding variables that results from the aggregation [GY02; Mor19].

4.3.2 Geostatistical Data

Geostatistical data are measurements of one or more spatially continuous features collected at specific locations. They can be a disease risk measured by a survey in different villages, the level of a pollutant recorded at several monitoring stations, or the density of mosquitoes responsible for disease transmission measured by traps set at different locations [WG04]. Let $Z(s_1), ..., Z(s_n)$ be the observations of a spatial

variable Z at locations $s_1, ..., s_n$. Geostatistical data are often assumed to be partial realisations of a random process

$$\left\{Z(s) : s \in D \subset \mathbb{R}^2\right\}, \tag{4.19}$$

where D denotes a fixed subset of $\mathbb{R}^2$ and the spatial index $\boldsymbol{s}$ varies continuously over D. For practical reasons, it is only possible to observe $Z(\cdot)$ at a finite set of locations. The inference of the characteristics, e.g. mean and variability of the process, of the spatial process is based on this partial realisation. Using these characteristics, it is possible to predict the process at unobserved locations and construct a spatially continuous surface of the variable of interest.

4.3.2.1 Stochastic Partial Differential Equation Approach

With geostatistical data, an underlying spatially continuous variable can often be assumed and modelled using a Gaussian random field. A spatial model can be fitted using the stochastic partial differential equation (SPDE) approach and the variable of interest can be predicted at new locations. A GRF with a Matérn covariance matrix can be written as a solution to the following continuous domain SPDE [Whi63]:

$$\left(\kappa^2 - \Delta\right)^{\alpha/2} (\tau x(\boldsymbol{s})) = \mathcal{W}(\boldsymbol{s}). \tag{4.20}$$

The GRF is represented by $x(\boldsymbol{s})$, where smoothness is controlled by α, while $\mathcal{W}(s)$ denotes a Gaussian spatial white noise process. $\kappa > 0$ is a scale parameter and Δ denotes the Laplacian given by $\sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}$, where d is the dimension of the spatial domain D.
The smoothness parameter ν of the Matérn covariance function is linked to the SPDE by

$$\nu = \alpha - \frac{d}{2}$$

while the marginal variance σ^2 is related to the SPDE by

$$\sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\alpha)(4\pi)^{d/2}\kappa^{2\nu}\tau^2}.$$

For $d = 2$ and $\nu = 0.5$ this corresponds to the exponential function.
The SPDE can be solved approximately using the *finite element* method, which partitions the spatial domain D into a set of non-intersecting triangles, resulting in a triangulated mesh with n vertices and n basis functions $\psi_k(\cdot)$. These functions are piecewise linear functions on each triangle, equal to 1 at vertex k and 0 other-

wise. The continuously indexed Gaussian field x is thus represented as a discretely indexed Gaussian Markov random field by the finite basis functions defined on the triangulated mesh

$$x(\boldsymbol{s}) = \sum_{k=1}^{n} \psi_k(\boldsymbol{s}) x_k, \tag{4.21}$$

with n the number of vertices of the triangulate, $\psi_k(\cdot)$ the piecewise linear basis functions and $\{x_k\}$ zero-mean Gaussian distributed weights.
The joint distribution of the weight vector follows a Gaussian distribution, $\boldsymbol{x} = (x_1, ..., x_n) \sim \mathcal{N}(0, \boldsymbol{Q}^{-1}(\tau, \kappa))$, which approximates the solution $x(\boldsymbol{s})$ of the SPDE in the mesh nodes, and the basis functions transform $x(\boldsymbol{s})$ from the mesh nodes to the other spatial locations of interest [LRL11].

5 Dataset Collection

As is often the case with statisticians, the construction of the dataset used to analyse a research question is an essential task and frequently involves the merging of multiple data sources to create a dataset. This was the case in this thesis and in the following chapter a brief overview of the data sources used, their pre-processing and how they were combined is given.

66

5.1 Covid-19 Data

5.1.1 Covid-19 Data for Norway

The Covid-19 data for Norway comes from a dataset made available to the public via the a repository on the website Github.com, created by the user thohan88. The repository contains a daily updated dataset that is the result of combining several data sources, which include the Institute of Public Health and the Norwegian Directorate of Health. According to the author of the repository, the project is "an open-source effort to make data about the Covid-19 situation in Norway available to the public in a timely and coherent manner" [Han20].
A few sample data points from this dataset are displayed in Table 5.1.

Tab. 5.1: An excerpt from the Covid-19 data for Norway. Does not contain all variables.

kommune_no	kommune_name	population	2020-03-26	2020-03-27
1103	Stavanger	143574	87	88
1507	Ålesund	66258	20	20
4601	Bergen	283929	231	248
5001	Trondheim	205163	113	136

5.1.2 Covid-19 Data for Germany

In Germany, the Robert Koch Institute publishes daily situation reports in which the number of new cases is published at NUTS 3 level. These reports are available as pdf files via the Institute's website. They can be downloaded and grouped via the R package `covid19germany`[Sch+21], as was done for this work.
A few sample data points from this dataset are displayed in Table 5.2.
The variable *CumNumberTestedIll* contains the cumulative number of people that have tested positive for Covid-19.

Tab. 5.2: An excerpt from the Covid-19 data for Germany. Does not contain all variables.

Landkreis	Date	CumNumberTestedIll	population
SK München	2020-01-29	1	1471508
SK München	2020-02-03	2	1471508
SK München	2020-02-11	3	1471508
LK Rosenheim	2020-02-29	1	260983
LK Rosenheim	2020-03-08	2	260983
LK Rosenheim	2020-03-10	6	260983

5.2 Demographic Data

As demographics tend to differ between different geographic units, the decision was made to include demographic variables in the analysis of the research question to see if the risk for infection may be higher when a certain characteristic is present in the population.

5.2.1 Demographic Data for Norway

The demographic data collected for Norway comes from Statistisk Sentralbyrå and is made available to the public through their online database, StatBank[@Sen16].
The first characteristic collected was the age of the population in a given municipality. For each age, starting at 0 and ending at 105, the number of people of that age is known.
Next, unemployment data were collected for a given municipality. For each municipality, the percentage of all people out of work is known, as well as the percentage of all immigrants out of work.
Other data collected include data related to the number of workers in a particular industry, as well as immigration data. Since there is discussion about whether workers from certain industries, in this case construction, contribute to the spread of Covid-19, the decision was made to collect this type of data. For each community, the number of workers across all industries is known, as well as the number of workers in the construction industry. Workers, in this case, are individuals employed in a given municipality who are between the ages of 20 and 66. It is also known how many people work full-time and how many work part-time.
Finally, for immigration data, it is known how many immigrants live in a given municipality and how many Norwegians were born to immigrant parents. These figures are known in terms of the percentage of the population in 2020.

5.2.2 Demographic Data for Germany

The demographic data collected for Germany comes from the federal and state statistical offices and is made available to the public through their online database, Regionaldatenbank Deutschland [@Bun20].
The first characteristic collected was unemployment data at the NUTS 3 level. For each municipality, the number of unemployed people as well as the number of unemployed foreigners was collected.

Next, data related to the European elections in 2019 were collected. In each municipality, it is known how many people voted in total, how many people voted for the six largest parties, and how many votes the remaining parties received combined.
Data was also collected in relation to people seeking protection, welfare recipients and in relation to asylum seeker benefits. It is known how many people sought protection in Germany, how many received social welfare and how many received asylum seeker benefits.
Finally, trade tax, income tax, and payroll tax data were collected for each municipality.

5.3 Shapefiles

In addition to numeric variables, the dataset also contains a geographic variable containing the geographic boundaries of a given municipality or city/district.

5.3.1 Shapefiles for Norway

The data for the Norwegian shapefiles comes from Geonorge [@Geo21] and is downloaded from a Github repository, as the data there was in a cleaner state [@Smi20]. In addition to the geographic shape, the dataset also includes a variable that contains the ID of each municipality.

5.3.2 Shapefiles for Germany

The data for the German shapefiles comes from Esri Germany [@Deu20].

5.4 OpenStreetMap Data

OpenStreetMap (OSM) is a free project that collects, structures and stores freely usable geodata in a database for use by anyone (Open Data). This data is available under a free license, the Open Database License. The core of the project is therefore an openly accessible database of all contributed geoinformation [@Ope17].
In R, the OpenStreetMap API can be queried using the R package `osmdata` [Pad+17]. To download all locations of a given type in a given region, a shape or bounding box must be specified along with a key and optionally a value. These key-value pairs are used to specify the type of location, for example, the "amenity" key is used for all facilities used by visitors and residents. If you use the "biergarten" value together with the "amenity" key, the locations of all beer gardens in a given geographic region will be downloaded.
OpenStreetMap users have the option to map a location as either `POINT`, `POLYGON`, `MULTIPOLYGON`, `LINESTRING`, or `MULTILINESTRING`. Conventionally, the first three are used. Therefore, only sites mapped as one of these were used for this work. If a location was mapped as either `POLYGON` or `MULTIPOLYGON`, the centroid of the location was calculated.
A complete list of all key-value pairs used for this work can be found in the Appendix.

5.5 Data Wrangling

The final step before analyzing the research question at hand is to combine all of these data sources into one dataset. This section will show how this was achieved.

5.5.1 Data Wrangling for Norway

The initial step in creating the final dataset was to convert the data from a wide format, as seen in Table 5.1, to a long format. This was done using the function `melt()` from the R package `reshape2` [Wic07]. The long version of the dataset is shown in Table 5.3.

Tab. 5.3: An excerpt from the long version of the Norwegian Covid-19 data. Does not contain all variables.

kommune_no	kommune_name	population	date	value
1507	Ålesund	66258	2020-03-26	20
5001	Trondheim	205163	2020-03-26	113
1507	Ålesund	66258	2020-03-27	20
5001	Trondheim	205163	2020-03-27	136

Next, the demographic data for Norway was loaded and processed. Since the age data contains the number of people of a certain age, the median age was calculated for each region based on how many people of each age group live in each region. The other demographic variables are left unchanged. To combine the demographic data with the Covid-19 data, the municipality IDs were extracted using the `str_extract()` function from the `stringr` [Wic19] R package using the regular expression `[0-9]{4}`. Next, all demographic datasets and the Covid-19 dataset were merged using the `merge()` function.
Using the `st_intersects()` function from the `sf` [Peb18] R package, the number of points of interest downloaded via OpenStreetMap was calculated for each municipality. Since the shapefiles contain the ID for each community, these data were then merged with the data containing the demographic and Covid-19 data.
For each numeric variable, e.g. the number of schools or the number of employees, this number was scaled.
If there were missing values in the covariates, these values were imputed using the median of the respective variable.
Finally, seven new variables were created:

1. expected_count, which is the expected number of cases in each municipality
2. sir, which is the standardised incidence ratio in each municipality
3. idarea_1, which is a unique ID given to each municipality
4. higher_education, which counts the number of universities and colleges in a given area.
5. sex, which gives the proportion of females living in a given area.
6. pop_dens, i.e. the number of people per square kilometre in a given area.
7. urb_dens, i.e. the number of residential buildings per square kilometre in a given area.

The final dataset contains the variables shown in Table 5.4.

Tab. 5.4: The variables contained in the final dataset.

Variable Name	Explanation	Scale
kommune_no	The municipality ID	None
kommune_name	The municipality name	None
population	Population in a municipality	None
date	The date of the data used	None
value	The number of infected people	None
median_age	The median age	None
unemp_tot	The proportion of unemployed people	scaled
unemp_immg	The proportion of unemployed immigrants	scaled
workers_ft	The number of full-time workers	scaled
workers_pt	The number of part-time workers	scaled
construction_ft	The number of full-time construction workers	scaled
construction_pt	The number of part-time construction workers	scaled
immigrants_total	The proportion of immigrants	scaled
marketplace	The number of marketplaces	scaled
entertainment	The number of entertainment venues	scaled
sport	The number of sports amenities	scaled
clinic	The number of clinics	scaled
hairdresser	The number of hairdresser	scaled
shops	The number of shops	scaled
place_of_worship	The number of places of worship	scaled
retail	The number of retail stores	scaled
nursing_home	The number of nursing homes	scaled
restaurant	The number of restaurants	scaled
aerodrome	The number of aerodromes	scaled
office	The number of offices	scaled
platform	The number of public transport platforms	scaled
kindergarten	The number of kindergartens	scaled
schools	The number of schools	scaled
bakeries	The number of bakeries	scaled
residential	The number of residential buildings	None
higher_education	The number of colleges and universities	scaled
expected_count	The expected number of infections	None
sir	The standardised incidence ratio	None
idarea_1	A unique ID	None
area	The area in km^2	None
pop_dens	People per km^2	scaled
urb_dens	Residential buildings per km^2	scaled
sex	The proportion of females	scaled

5.5.2 Data Wrangling for Germany

The data processing procedure for Germany is identical to that for Norway. First, all demographic variables were loaded and left unchanged before being merged with the Covid-19 data prior to calculating the spatial intersections between the points of interest and the NUTS-3 areas. After merging all the data, the scaled numbers were calculated for the numeric variables. For the variables containing the number of people who voted for a particular political party, the relative percentage of votes the party received was calculated. Again, missing values were imputed using the median. Finally, the same seven new variables were created. The final dataset contains the variables shown in Table 5.5.

Tab. 5.5: The variables contained in the final dataset.

Variable Name	Explanation	Scale
municipality_id	The municipality ID	None
municipality	The municipality name	None
population	Population in a municipality	None
date	The date of the data used	None
value	The number of infected people	None
trade_tax	The trade tax in Euros	scaled
income_tax	The income tax in Euros	scaled
income_total	The income and payroll tax in Euros	scaled
asyl_benefits	The number of people receiving asylum seeker benefits	scaled
welfare_recipients	The number of welfare recipients	scaled
unemployed_total	The number of unemployed people	scaled
unemployed_foreigners	The number of unemployed foreigners	scaled
protection_seekers	The number of protection seekers	scaled
Union	Percentage of vote for Union	scaled
SPD	Percentage of vote for SPD	scaled
Gruene	Percentage of vote for Gruene	scaled
FDP	Percentage of vote for FDP	scaled
die_linke	Percentage of vote for die Linke	scaled
afd	Percentage of vote voted for AfD	scaled
marketplace	The number of marketplaces	scaled
entertainment	The number of entertainment venues	scaled
sport	The number of sports amenities	scaled
clinic	The number of clinics	scaled
hairdresser	The number of hairdresser	scaled
shops	The number of shops	scaled
place_of_worship	The number of places of worship	scaled
retail	The number of retail stores	scaled
nursing_home	The number of nursing homes	scaled
restaurant	The number of restaurants	scaled
aerodrome	The number of aerodromes	scaled
office	The number of offices	scaled
platform	The number of public transport platforms	scaled
kindergarten	The number of kindergartens	scaled
schools	The number of schools	scaled
bakeries	The number of bakeries	scaled
residential	The number of residential buildings	None
higher_education	The number of colleges and universities	scaled
expected_count	The expected number of infections	None
sir	The standardised incidence ratio	None
idarea_1	A unique ID	None
area	The area in km^2	None
pop_dens	People per km^2	scaled
urb_dens	Residential buildings per km^2	scaled
sex	The proportion of females	scaled

6 Data Analysis

In this chapter the models fitted for each country are reviewed. First, a look at the standardised incidence rate for each country is taken, before spatial models, spatio-temporal models and finally predictive models are discussed.

6.1 Standardised Incidence Ratio (SIR)

This section takes a brief look at the SIR for the countries of interest. Recall from Equation 4.10, that the SIR is defined as the ratio of observed counts to expected counts.

6.1.1 SIR for Germany

When looking at the SIR for Germany in Figure 6.1, it is noticeable that the actual number of infections in the eastern parts of Germany, especially in Saxony, is considerably higher than the expected number of infections. Furthermore, parts of Bavaria have an increased SIR compared to the rest of Germany, excluding Saxony. This could be due to the fact that the regions share a border with the Czech Republic, a country that is substantially more affected by Covid-19 than Germany. The northern parts of Germany show the lowest SIR which is possibly due to the fact that this region is sparsely populated.

78

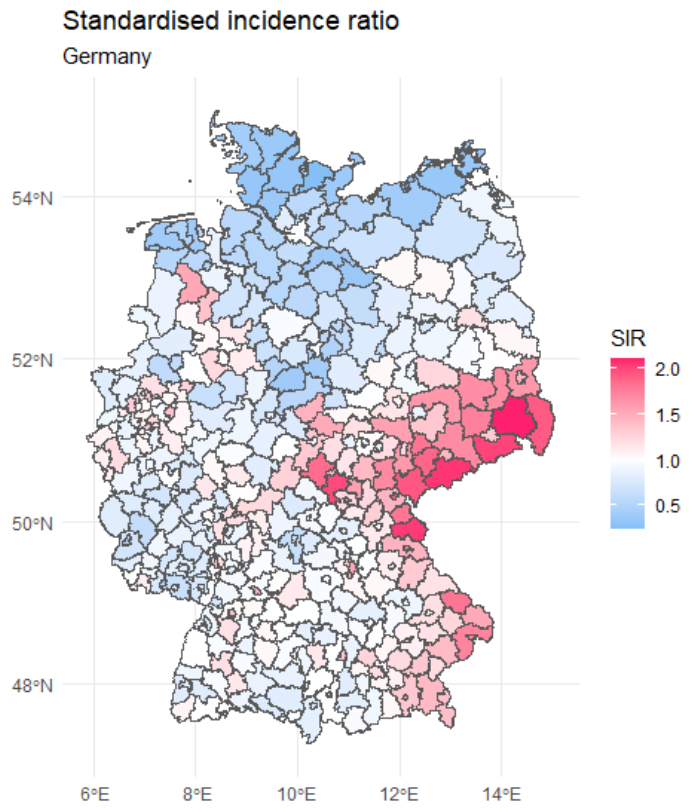

Fig. 6.1: The SIR for Germany based on the data of the 24th of March 2021

6.1.2 SIR for Norway

Looking at the standardised incidence rate for Norway in Figure 6.2, a standardised incidence rate of less than 1 can be seen for most municipalities north of Trondheim. In the southern parts of Norway there are several municipalities with a rate above 1, for example the standardised incidence rate around the capital Oslo is around 2. However, the two small municipalities, Hyllestad and Ulvik, have the highest standardised incidence rate in Norway. In Hyllestad, 95 of 1328 people have been infected with Covid-19 so far, while in Ulvik, 134 of 1080 people have been infected so far.

The SIR in Hyllestad is around 4.5, following an outbreak in a shipyard in autumn 2020 [Kor20], while Ulvik has a ratio of around 8, following an outbreak of the UK variant of Covid-19. According to the head of the municipality, Hans Petter Thorbjørnsen, the infections are thought to have spread through children [NTB21].

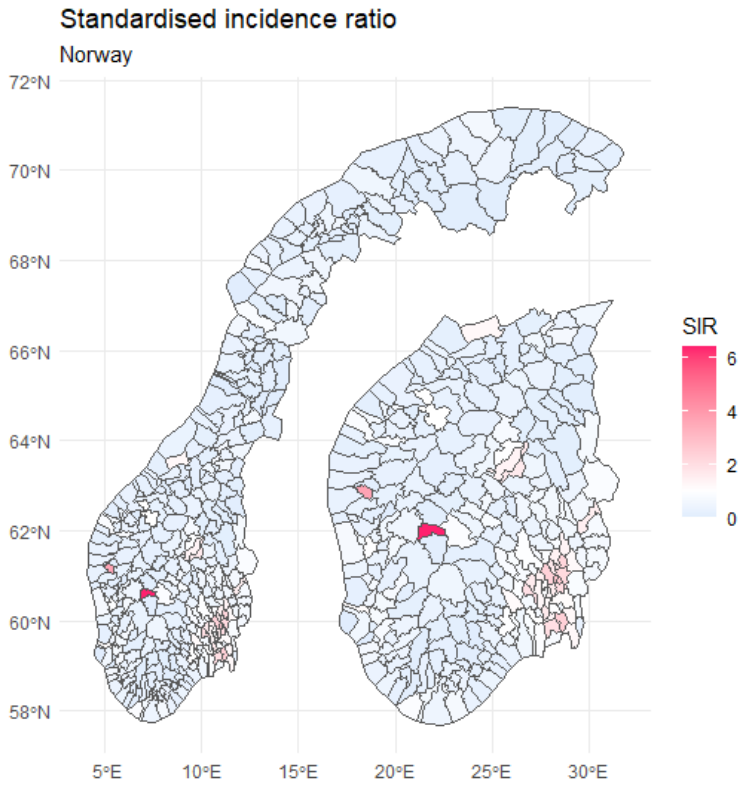

Fig. 6.2: The SIR for Norway based on the data of the 24th of March 2021

Because the high numbers from two small municipalities complicate the interpretation of Figure 6.2, Figure 6.3 shows the SIR on a log10 scale. On this scale, a value of 0 means that the risk of infection in a given municipality is neither lower nor higher. Values below 0 mean that the risk of infection in a municipality is lower than average, while values above 1 mean that the risk of infection in a municipality is higher than average. It is now clearer that the standardized incidence ratio is below 1 in most parts of Norway, but that there is a higher risk in the region around Oslo.

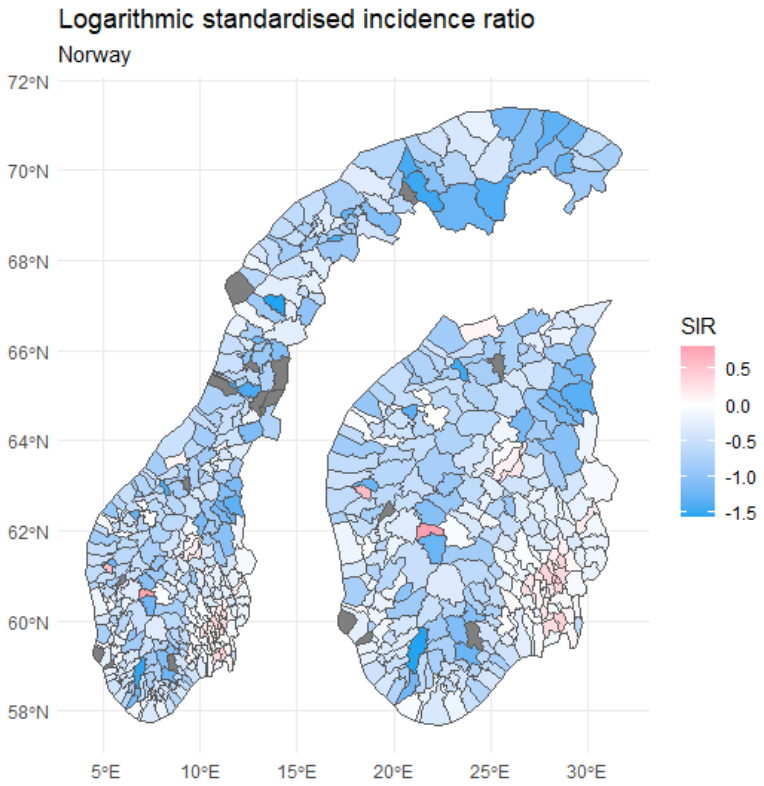

Fig. 6.3: The log10 SIR for Norway based on the data of the 24th of March 2021

6.2 Data Modelling

After looking at the standardised incidence rates for the countries of interest, the next step is to take a closer look at the current figures for the respective countries. Spatial models are used to try to extract the factors that cause some populations to be at higher risk than populations in other geographical regions. Three different types of models are used for each country:

1. Besags Proper Spatial Model
2. A Leroux Model
3. A BYM2 Model

All of these models are computed using the INLA [BGR15] R package.
To specify each type of model, the code shown in Listing 7.1 can be used.
The measures introduced in Section 3.7, namely the DIC, the WAIC, the CPO and the mean absolute error (MAE), are used to compare the models.
For all countries, the models are computed with

1. only the demographic variables as covariates
2. only the infrastructural variables as covariates
3. both, demographic and infrastructural variables, as covariates

In addition to specifying what type of spatial model to use, if any, there is also the option of specifying a prior.
As can be seen in Section 3.3.2, a pc prior can be specified for the precision parameter τ, which is what is done here.
For the parameters σ_0 and α in Equation 3.22 the values 1 and 0.01 are chosen.
The models are compared using the mean absolute error. For this, 20% of the observations are removed from the training set and used for testing instead. The predicted number of infections for these municipalities is then compared to the actual numbers.
A list of all calculated models along with their performance measures is provided in the appendix.

6.2.1 Choice of Likelihood

Before the models are computed, however, the distribution that fits the number of cases must first be found. One way to do this, the function `descdist()` from the `fitdistrplus` R package is used. The Cullen and Frey graph illustrates how "close" a sample is to a theoretical distribution based on the kurtosis and the square of the skewness, defined in Equation 3.15 and Equation 3.14. It can be used to get a preliminary idea of which distributions fit the data, in this case the number of infections, reasonably well.

The plots for Germany and Norway can be seen in Figure 6.4 and Figure 6.5. The blue dot represents the data, the star a theoretical normal distribution, the dashed line a theoretical Poisson distribution and the grey area a theoretical negative binomial distribution. In both cases, the blue dot is relatively far from the star and lies in the region of a negative binomial distribution. For the Norwegian sample shown in Figure 6.5, the sample is closer to a Poisson distribution than is the case for the German sample in Figure 6.4.

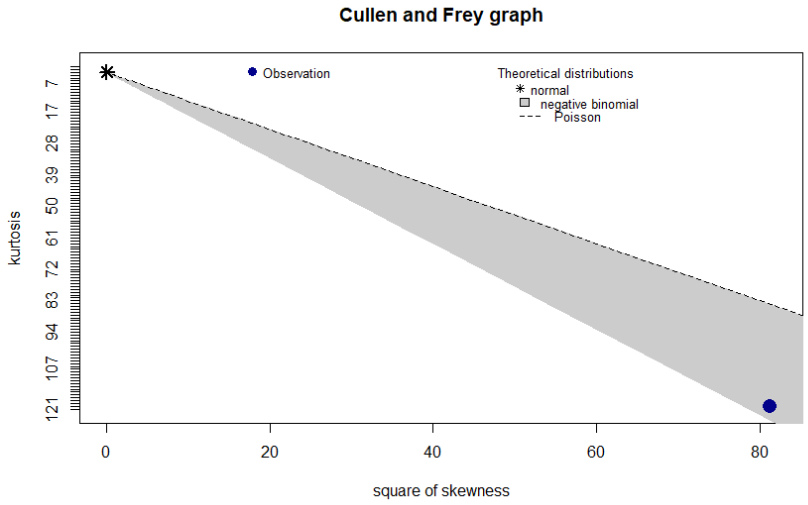

Fig. 6.4: The Cullen and Frey graph for Germany

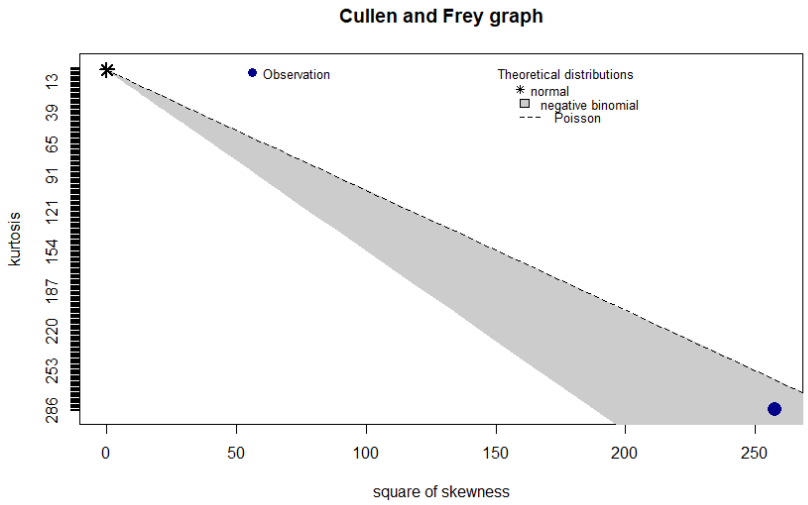

Fig. 6.5: The Cullen and Frey graph for Norway

Next, a negative binomial distribution, a normal distribution, and a Poisson distribution are fitted to the data using the maximum likelihood method. The negative binomial fits for both countries can be seen in Figure 6.6 and Figure 6.7. The fits for the normal and Poisson distribution for both countries, are shown in the Appendix in Figure 7.1, Figure 7.2, Figure 7.3 and Figure 7.4.
The QQ-plot for Germany and Norway looks quite similar, as there appears to be a linear relationship between the theoretical quantile and the sample quantiles, up to a certain point where the sample quantiles have a higher value than the theoretical quantiles, indicating that the distribution is right skewed. Since there are many municipalities with relatively few cases and few municipalities with a large number of cases, this is to be expected. It can also be seen that the empirical cumulative density function closely follows the theoretical cumulative density function.

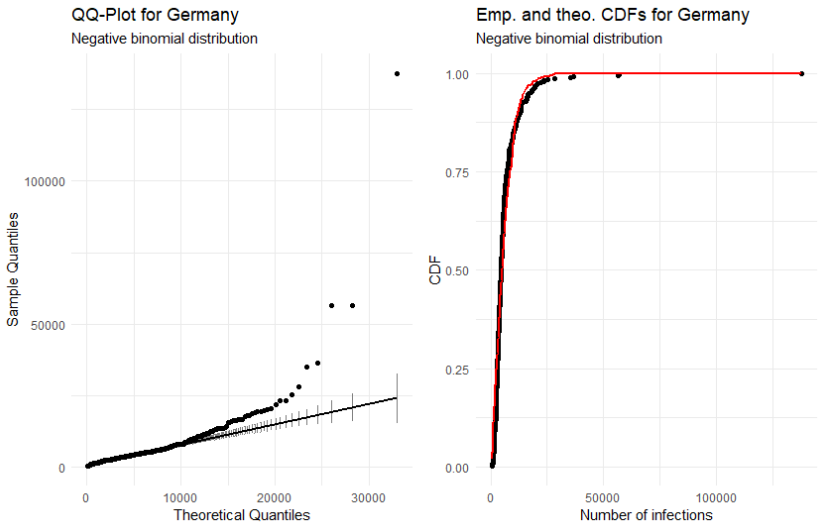

Fig. 6.6: A negative binomial fit to the number of cases in German municipalities

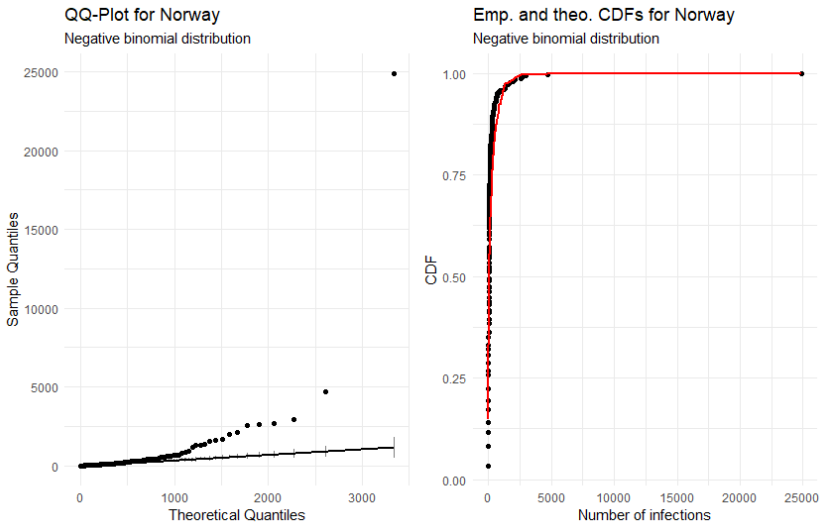

Fig. 6.7: A negative binomial fit to the number of cases in Norwegian municipalities

Lastly, the AIC is calculated for fitting a normal distribution to the data, a Poisson distribution to the data and a negative binomial distribution to the data. The values can be seen in Table 6.1. Afterwards, the negative binomial distribution is chosen as the distribution of the target variable in both cases.

Tab. 6.1: The AIC for different distributions for Germany and Norway

Country	Distribution	AIC
Germany	Normal	8360
Germany	Poisson	2148100
Germany	Negative Binomial	7731
Norway	Normal	6166
Norway	Poisson	366181
Norway	Negative Binomial	4086

The poor fit for the Poisson distribution can be explained by looking at the range of the number of confirmed cases in a given municipality. For Germany, this number ranges from 508 to 137634 (as of March 18, 2021), while for Norway, the number ranges from 0 to 24905 (as of March 20, 2021). This results in a mean and standard deviation for Germany of 6617 and 9014, respectively. For Norway, the values for these metrics are 236 and 1389. This is problematic because, as shown in Equation 7.5 and Equation 7.6, for a Poisson distribution the expected value and the variance should be equal.

Looking at a histogram for the confirmed number of cases and overlaying the densities of a normal, Poisson and a negative binomial distribution helps to confirm the choice of a negative binomial distribution as the distribution that the data most closely resembles. Figure 6.8 and Figure 6.9 both show that a negative binomial distribution fits the data better than a normal distribution. Due to the high values for the AIC, the Poisson distribution is excluded from these graphics.

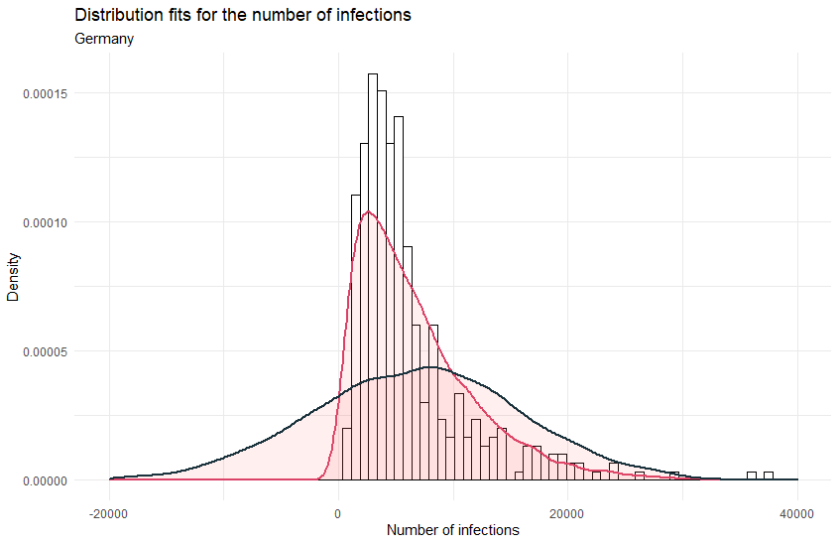

Fig. 6.8: Histogram for the number of cases in German municipalities with a normal and a negative binomial distribution overlayed.

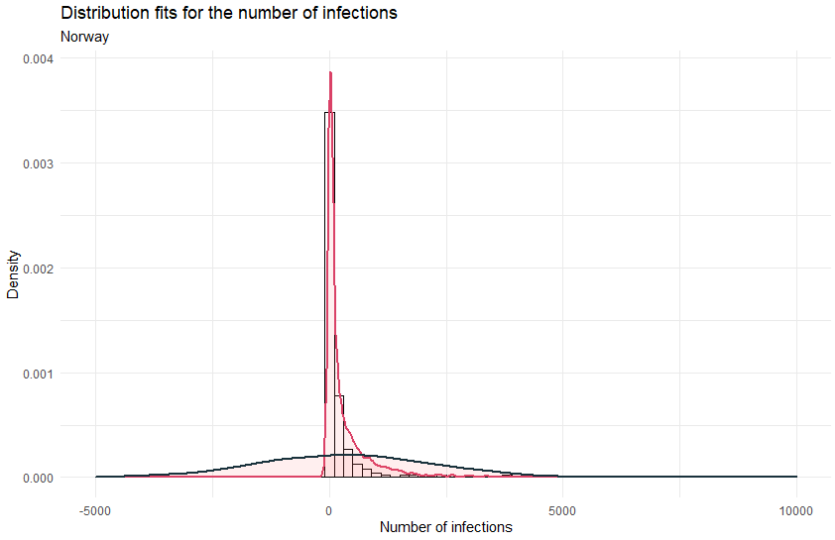

Fig. 6.9: Histogram for the number of cases in Norwegian municipalities with a normal and a negative binomial distribution overlayed.

6.3 Models without a Spatial Component

To establish a baseline, a look is first taken at models that do not include a spatial effect. This way, it can be observed how the means and credibility intervals of the covariates change when a spatial effect is added to a model and how the performance of the model changes with respect to the goodness-of-fit indicators introduced in Section 3.7.

Before computing the models, using the VIF introduced in Section 3.9, predictors are removed if their VIF is above 5. To do this, an OLS regression is first run on all variables, followed by the calculation of VIF. Then the variable with the highest value is removed before running the regression model again with all remaining variables. This process is repeated until only variables with a VIF of less than 5 remain.

In sections 6.3.1 and 6.3.2 the results of these models are presented. Their goodness-of-fit indicators as well as the coefficients together with their credibility intervals calculated as in section 3.2.6 are reported.

These models are based on data from 24 March 2021, when 2,695,037 people were infected with Covid-19 in Germany, while 87,537 people were infected in Norway. The five municipalities with the most infections in Germany are shown in Table 6.2 and in Table 6.3 for Norway.

Tab. 6.2: The German municipalities with the most infections as of March 24th 2021.

Municipality	Population	Number of infections
SK Berlin	3644826	141639
SK Hamburg	1841179	58661
SK Munich	1471508	57690
SK Cologne	1085664	37592
Region Hannover	1157624	36241

Tab. 6.3: The Norwegian municipalities with the most infections as of March 24th 2021.

Municipality	Population	Number of infections
Oslo	693494	26151
Bergen	283929	4738
Drammen	101386	3149
Bærum	127731	2834
Lillestrøm	85983	2779

6.3.1 Models without a Spatial Component for Germany

Table 6.4 contains the performance measures for the baseline model for Germany, while Table 6.5 contains the posterior mean, the exponentiated posterior mean and the credibility intervals of the coefficients. It can be seen that the intercept as well as six of the coefficients are significant.

Tab. 6.4: The performance measures for the model without a spatial component.

DIC	WAIC	CPO	MAE
5504	5506	-2778	179234

Tab. 6.5: The fixed effects for the model. Values are rounded. A * denotes a significant effect.

Variable	mean_p	$\exp(\text{mean}_p)$	$\exp(\text{q0025}_p)$	$\exp(\text{q0975}_p)$	sig.
(Intercept)	-0.07213	0.9553	0.9291	0.9821	*
pop_dens	0.1566	1.170	1.115	1.227	*
log trade_tax	0.08821	1.093	1.041	1.146	*
sex	0.04363	1.045	1.009	1.081	*
platform	0.03324	1.034	0.9796	1.091	
FDP	0.03242	1.033	0.9945	1.073	
higher_ education	0.01918	1.020	0.9787	1.063	
clinic	0.01448	1.015	0.9608	1.073	
urb_dens	0.009294	1.010	0.9699	1.051	
aerodrome	-0.0009688	0.9991	0.9731	1.028	
nursing_ home	-0.006835	0.9933	0.9609	1.027	
place_of_ worship	-0.03459	0.9662	0.9256	1.008	
marketplace	-0.03482	0.9663	0.9080	1.027	
office	-0.04362	0.9576	0.9140	1.004	
die_linke	-0.0721	0.9306	0.8948	0.9676	*
SPD	-0.1316	0.8768	0.8476	0.9068	*
Gruene	-0.2816	0.7548	0.7197	0.7913	*

6.3.2 Models without a Spatial Component for Norway

Table 6.6 contains the performance measures for the baseline model for Germany, while Table 6.7 contains the posterior mean, the exponentiated posterior mean and

the credibility intervals of the coefficients. It can be seen that the intercept as well as five of the coefficients are significant.

Tab. 6.6: The performance measures for the model without a spatial component.

DIC	WAIC	CPO	MAE
2713	2718	-1623	10170

Tab. 6.7: The fixed effects for the model. Values are rounded. A * denotes a significant effect.

Variable	mean_p	$\exp(\text{mean}_p)$	$\exp(\text{q0025}_p)$	$\exp(\text{q0975}_p)$	sig.
(Intercept)	-0.8721	0.4185	0.3832	0.4569	*
immigrants_total	0.2773	1.322	1.165	1.497	*
unemp_immg	0.2086	1.236	1.062	1.435	*
urb_dens	0.1790	1.200	1.027	1.423	*
unemp_tot	0.07675	1.085	0.8969	1.304	
platform	0.07141	1.078	0.9139	1.271	
higher_education	0.01850	1.020	0.9375	1.125	
nursing_home	0.01501	1.016	0.9377	1.117	
median_age	-0.004360	0.9969	0.9020	1.098	
marketplace	-0.006536	0.9959	0.8735	1.145	
place_of_worship	-0.04057	0.9634	0.8233	1.130	
office	-0.1372	0.8749	0.7434	1.030	
sex	-0.1535	0.8591	0.7691	0.9568	*
aerodrome	-0.1954	0.8247	0.7010	0.9335	*

6.4 Spatial Models

Under the null hypothesis of no spatial autocorrelation, a p-value greater than 0.05 would be expected. The results of the test are presented in Table 6.8. Looking at the p-value for both countries, it can be seen that there is a spatial correlation for the number of infections in a municpality.

Tab. 6.8: Results of the Moran test for Germany and Norway.

Country	Moran's I	$\mathbb{E}[I]$	p-Value
Germany	0.1047	-0.002500	< 0.01
Norway	0.1085	-0.002817	< 0.01

Therefore, after the models without spatial effect have been calculated and established as baseline models, a spatial term is added to the models calculated in Section 6.3 , in order to model this spatial correlation.

6.4.1 Spatial Models for Germany

Looking at the performance of the spatial models and the model with the spatial component shown in Table 6.9, it can be seen that the spatial models perform better in terms of the DIC, WAIC and MAE, while they perform equally well or better in terms of the CPO.
The best performance of all models, in terms of MAE, was observed for the BYM2 model, which slightly outperformed the Besag model.

Tab. 6.9: The performance measures for the best performing demographic + infrastructure model of each type.

Model	DIC	WAIC	CPO	MAE
No spatial	5504	5506	-2778	179234
Besag	4706	4685	-2738	176201
BYM2	4628	4612	-2729	176017
Leroux	5118	5108	-2998	179150

Figure 6.10 shows the differences between the coefficients in the model without the spatial component and the BYM2 model. Excluding the intercept, only three effects are significant in the BYM2 model compared to six in the model without the spatial component. Moreover, the coefficients of the BYM2 model are closer to 1 and have

smaller credibility intervals. The reason for this is that when using a spatial field in combination with a PC prior, higher values for σ_0 make the spatial field larger, which in turn causes the posterior mean to shrink towards 1.

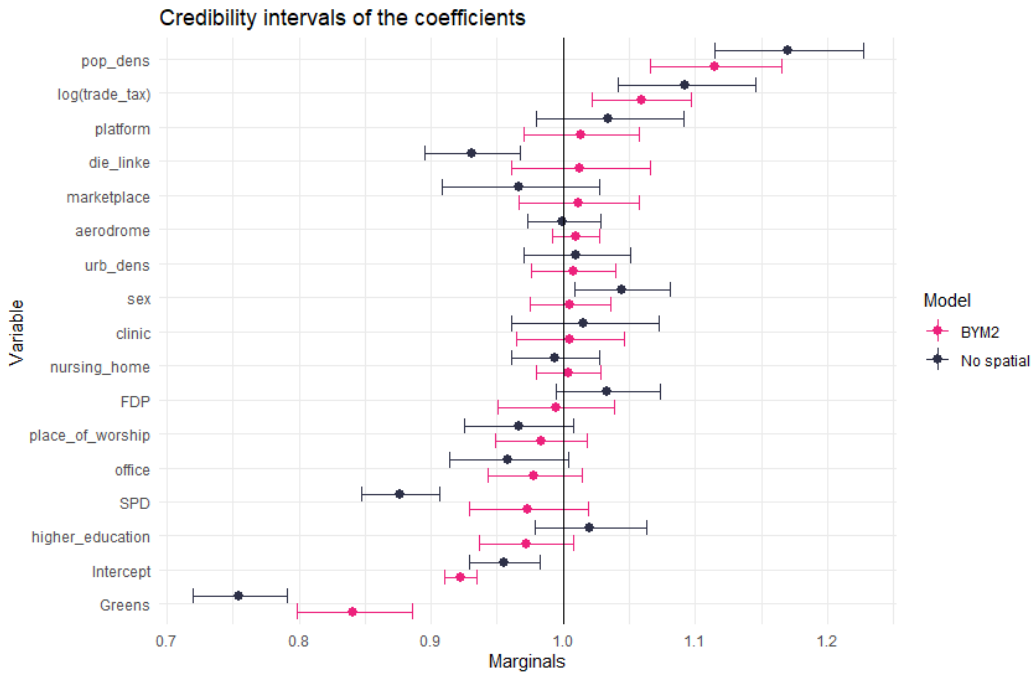

Fig. 6.10: The posterior mean and credibility intervals of the coefficients

The values of the coefficients and credibility intervals are shown in Table 6.10.

Tab. 6.10: The fixed effects for the model. Values are rounded. A * denotes a significant effect.

Variable	mean_p	$\exp(\text{mean}_p)$	$\exp(\text{q0025}_p)$	$\exp(\text{q0975}_p)$	sig.
(Intercept)	-0.08087	0.9223	0.9105	0.9345	*
pop_dens	0.1084	1.115	1.066	1.165	*
log trade_tax	0.5714	1.059	1.022	1.097	*
platform	0.01315	1.013	0.9705	1.058	
die_linke	0.01160	1.012	0.9605	1.066	
marketplace	0.01117	1.011	0.9669	1.057	
aerodrome	0.009694	1.010	0.9919	1.028	
urb_dens	0.007543	1.008	0.9760	1.040	
sex	0.004908	1.005	0.9746	1.036	
clinic	0.004868	1.005	0.9648	1.047	
nursing_ home	0.003748	1.004	0.9799	1.028	
FDP	-0.006087	0.9942	0.9507	1.039	
place_of_ worship	-0.01738	0.9829	0.9483	1.018	
office	-0.02228	0.9781	0.9428	1.014	
SPD	-0.02744	0.9732	0.9290	1.019	
higher_ education	-0.02879	0.9718	0.9368	1.008	
Gruene	-0.1735	0.8410	0.7980	0.8856	*

For the hyperparameters, a value of 14.22 is reported for the precision and a value of 0.9172 for ϕ. Hence, 91.72% of the marginal variance is explained by the structured effect. Therefore, this model is far from reducing to pure overdispersion and comes close to a Besag model, which is also reflected in the similar values of the goodness-of-fit indicators in Table 6.9.

6.4.2 Spatial Models for Norway

Comparing the performance of the models, the spatial models again showed better performance in terms of DIC and WAIC and this time also significantly better in terms of CPO. However, the model without the spatial component showed better predictive performance, as indicated by the lowest value for the MAE in Table 6.11. This could indicate that neighbourhood effects are not as strong in Norway as in Germany.

Tab. 6.11: The performance measures for the best performing demographic + infrastructure model of each type.

Model	DIC	WAIC	CPO	MAE
No spatial	2713	2718	-1623	10170
Besag	2452	2468	-3782	11266
BYM2	2278	2265	-5627	11520
Leroux	2272	2230	-8261	11434

Looking at the differences between the coefficients and credibility intervals in Figure 6.11, the picture is similar to Figure 6.10. In the BYM2 model, fewer effects are significant, two compared to five in the model without the spatial component, the coefficients are closer to 1 and the credibility intervals are narrower.

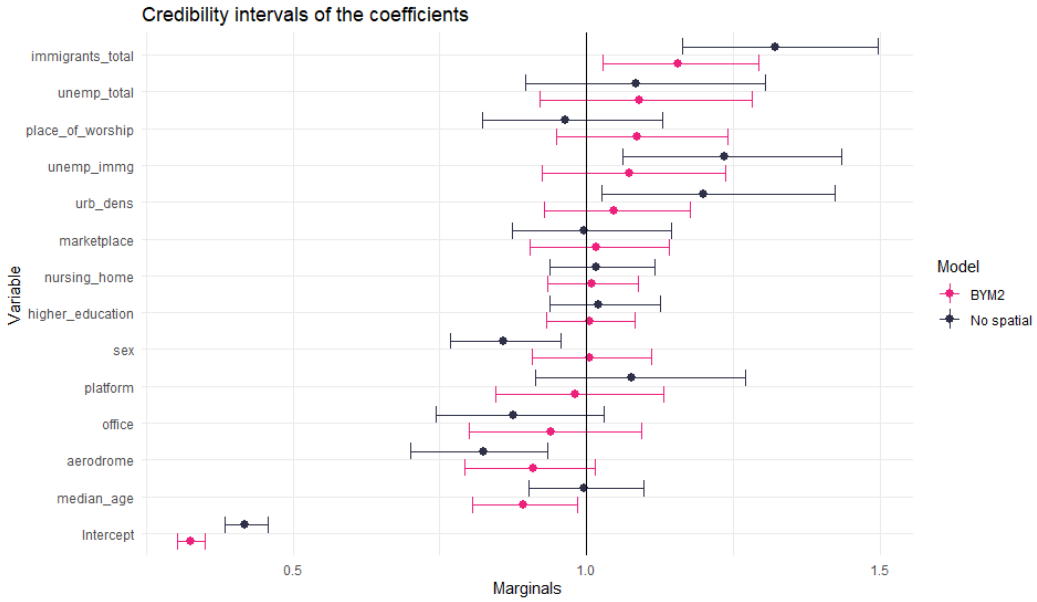

Fig. 6.11: The posterior mean and credibility intervals of the coefficients

The values of the coefficients and credibility intervals are shown in Table 6.12.

Tab. 6.12: The fixed effects for the model. Values are rounded. A * denotes a significant effect.

Variable	mean_p	$\exp(\text{mean}_p)$	$\exp(\text{q0025}_p)$	$\exp(\text{q0975}_p)$	sig.
(Intercept)	-1.121	0.3260	0.3026	0.3504	*
immigrants_total	0.1432	1.156	1.029	1.294	*
unemp_tot	0.08298	1.090	0.9203	1.281	
place_of_worship	0.08127	1.087	0.9488	1.242	
unemp_immg	0.06734	1.073	0.9247	1.237	
urb_dens	0.04398	1.047	0.9284	1.178	
marketplace	0.01543	1.017	0.9041	1.142	
nursing_home	0.008396	1.009	0.9340	1.089	
higher_education	0.005382	1.006	0.9325	1.083	
sex	0.004273	1.006	0.9073	1.112	
platform	-0.02252	0.9805	0.8450	1.131	
office	-0.06488	0.9402	0.8004	1.095	
aerodrome	-0.09666	0.9097	0.7926	1.015	
median_age	-0.1151	0.8925	0.8052	0.9853	*

For the hyperparameters, a value of 1.733 is reported for the precision and a value of 0.6249 for ϕ. Hence, 62.49% of the marginal variance is explained by the structured effect. Therefore, this model lies somewhere between pure overdispersion and the Besag model, but closer to the Besag model than to overdispersion.

6.5 Choice of Hyperpriors

As can be seen in Equation 3.21, there is flexibility when it comes to choosing the values for the standard deviation σ_0 as well as the probability α. Therefore, an upper bound for the standard deviation can be chosen as well as the weight placed on this "tail event", describing how informative the resulting prior is.
Some of the issues that come with the choice of these hyperpriors were already discussed in Section 3.8.
In the following, an assessment is made of how the performance of a Besag model, a BYM2 model and a Leroux model changes when playing around with the value for the standard deviation σ_0. To create these plots, models were calculated with σ_0 values of $\boldsymbol{\sigma_0} = (0.1, 0.11, 0.12, ..., 5)$. Focus is only on the WAIC, as it is the only truly Bayesian performance measure, and the MAE to assess how the predictive performance changes.
In Figure 6.12 it can be seen that when choosing a higher value for σ_0, the WAIC is lower in the case of the Besag model and the BYM2 model. For the Leroux model, on the other hand, the WAIC gets lower until about 2 before it rises until $\sigma_0 = 2.5$ and then falls again. It is a positive sign that this is not the case with the BYM2 model, as it was designed to avoid exactly this kind of thing.
For the MAE in Figure 6.13, it can be seen for both the BYM2 and the Besag model that a higher value for σ_0 leads to a higher value for the MAE. The same is true for the Leroux model, but again there is a small interval where it continuously decreases even though the value for σ_0 increases.

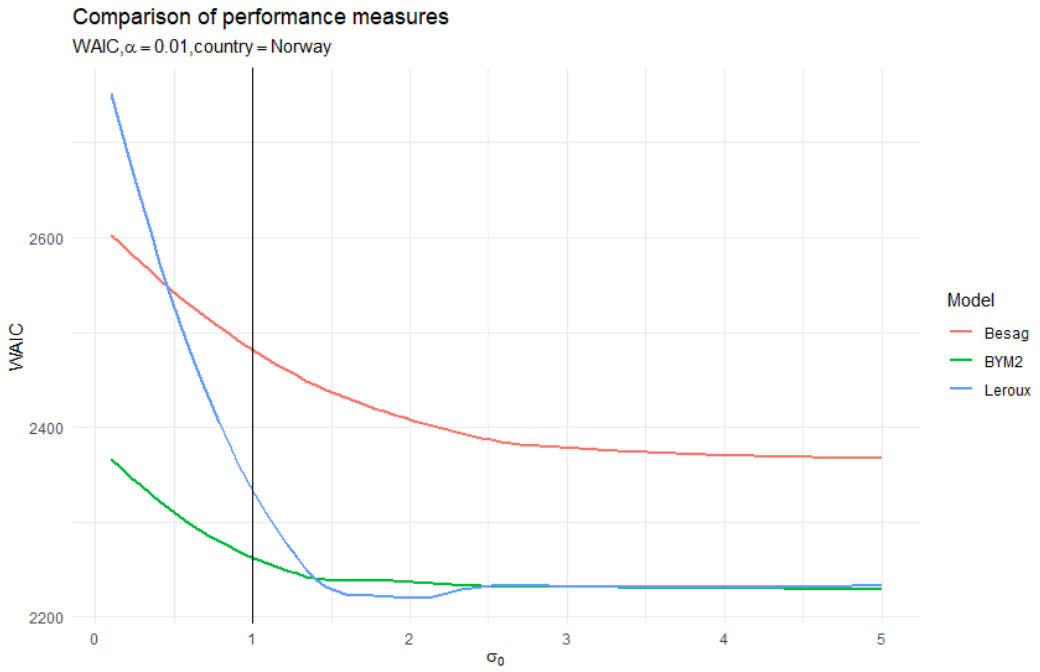

Fig. 6.12: Value of the WAIC when changing the value for σ_0. The black line highlights the values for $\sigma_0 = 1$.

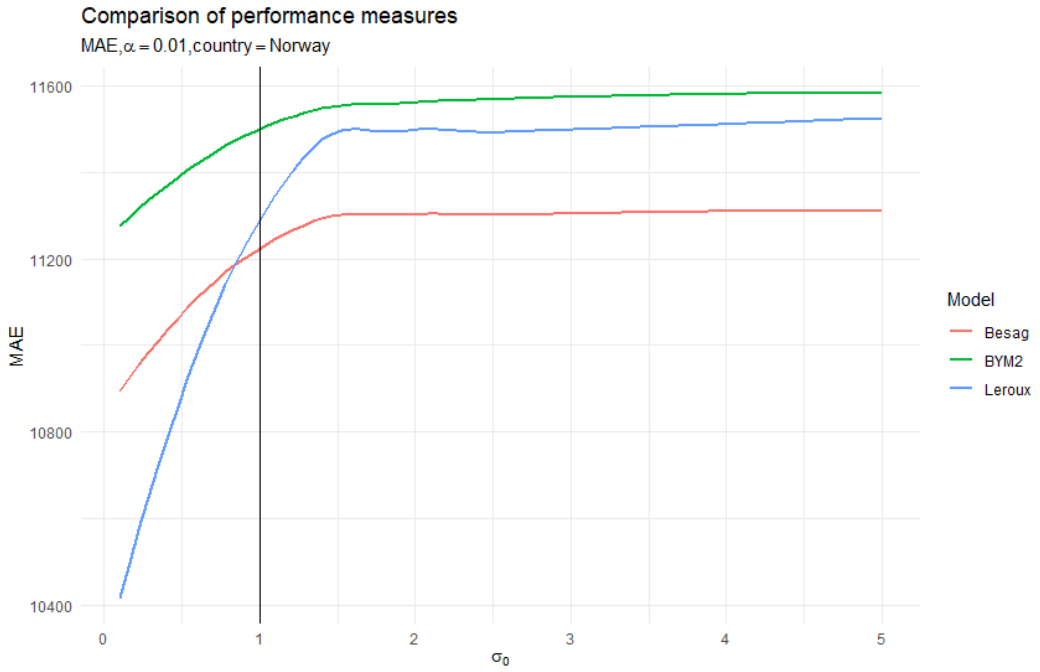

Fig. 6.13: Value of the MAE when changing the value for σ_0. The black line highlights the values for $\sigma_0 = 1$.

By allowing the precision to be greater, the variance is forced to be smaller. Hence, choosing a lower value for the precision leads to lower values for the WAIC. While this indicates a better fit to the training data, Figure 6.13 also shows that the MAE increases when a higher value for σ_0 is chosen, as the models overfit on the training data and therefore make worse predictions.
The corresponding figures for Germany are shown in Figure 7.5 and Figure 7.6 in the Appendix.
Figure 6.14 shows how the credibility intervals of the coefficients of a BYM2 model change when the value for σ_0 is increased. In general, the credibility intervals for $\sigma_0 = 1$ are narrower than those for $\sigma_0 = 0.1$, but no significant changes are seen between $\sigma_0 = 1$ and $\sigma_0 = 5$. The values of the coefficients tend to remain relatively similar most of the time, especially when the value of the coefficient is close to 1. However, a few times, for example for the variables immigrants_total, platform and unemp_immg, the values differ. Furthermore, the coefficients for $\sigma_0 = 1$ and $\sigma_0 = 5$ are more closer to 1 than the ones for $\sigma_0 = 0.1$ The reason for that is that a higher value for σ_0 makes the variance of the spatial field larger, therefore the posterior mean should shrink towards 1.

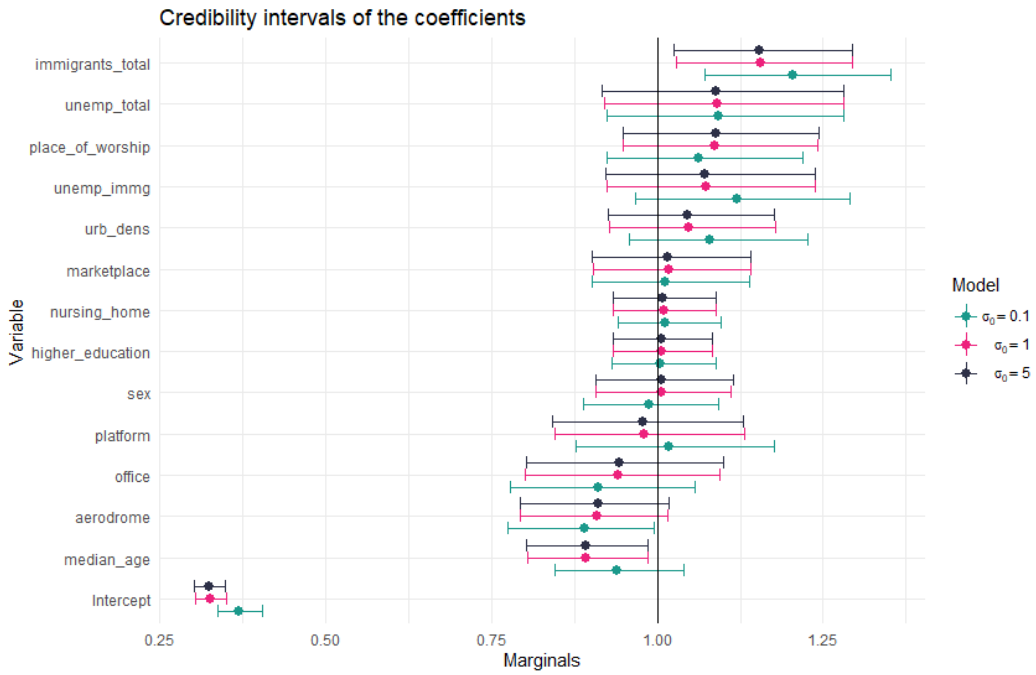

Fig. 6.14: Comparison of the credibility intervals of a BYM2 model for different values of σ_0.

The corresponding figure for Germany is shown in Figure 7.7 in the Appendix.
Figure 6.15 and Figure 6.16 underline the problem of the models not being compara-

ble with each other. Figure 6.15 already shows slight differences in the spatial field of the Besag Model and the Leroux Model, e.g. in the central and northern parts of Norway, where the posterior mean is higher for the Leroux Model. However, the problem becomes clear when comparing the spatial fields of the Besag model and the Leroux model with the spatial fields of the BYM2 model shown in Figure 6.16. In the left part of the figure 6.16 the values of the equation 3.72 are plotted, while in the right part the values of u_* are plotted.
For the spatial field of the unstructured random effect, the values of the posterior mean are similar to the values of the Besag model. For the structured component, however, there are more extreme values, with the posterior mean for northern Norway mostly around -2 and the posterior mean around the Oslo region around 1.5. For comparison, these values are around -1 and 0.8 for the Besag model and -0.8 and 0.6 for the Leroux model, respectively.

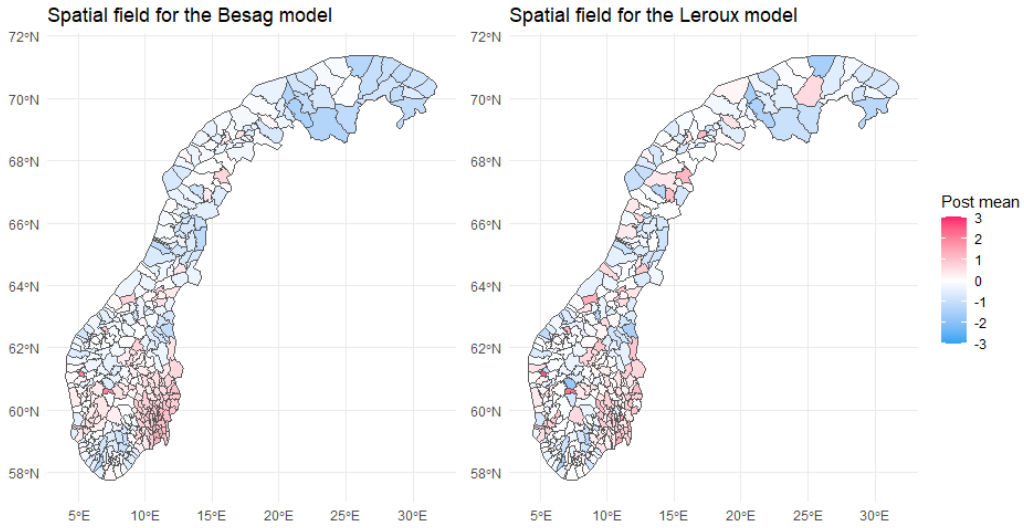

Fig. 6.15: Spatial field for a Besag model and a Leroux model.

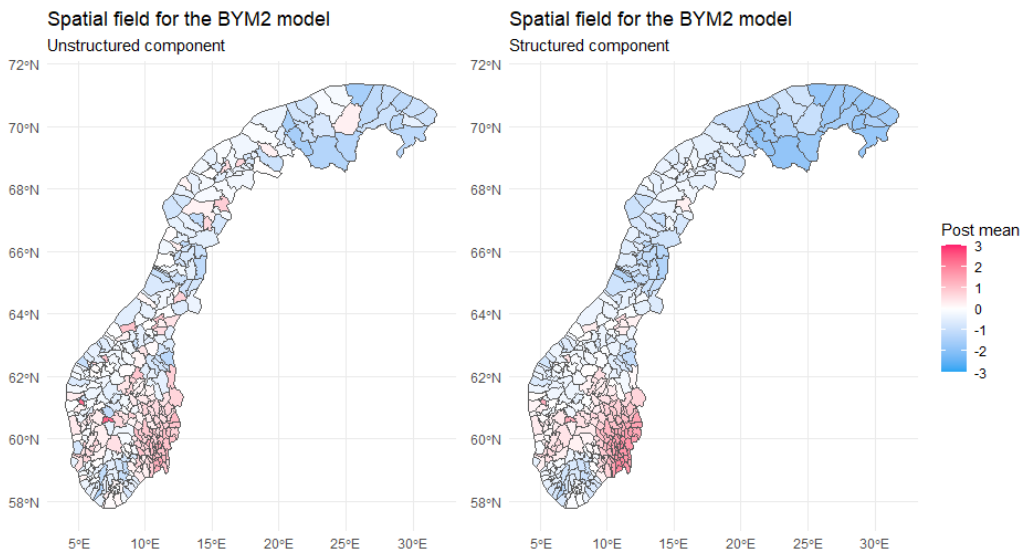

Fig. 6.16: Spatial fields for a BYM2 model.

Finally, looking at the spatial field of the structured component when changing the value for σ_0, as seen in Figure 6.17, it can be seen that for a small value like $\sigma_0 = 0.1$, the values of the posterior mean are mostly around 0, while for a higher value not much change can be seen. Interestingly, however, if the posterior mean is given its own scale for $\sigma_0 = 0.1$, as done in Figure 6.18, it can be seen that the higher up north a municipality lies, the lower the posterior mean becomes. The reason for this is that if σ_0 is small, the spatial field is only one plane. σ_0 specifies how much the spatial effect is allowed to deviate from this plane, and with a small value for σ_0 only a small deviation is allowed, which can then look similar to a linear effect.

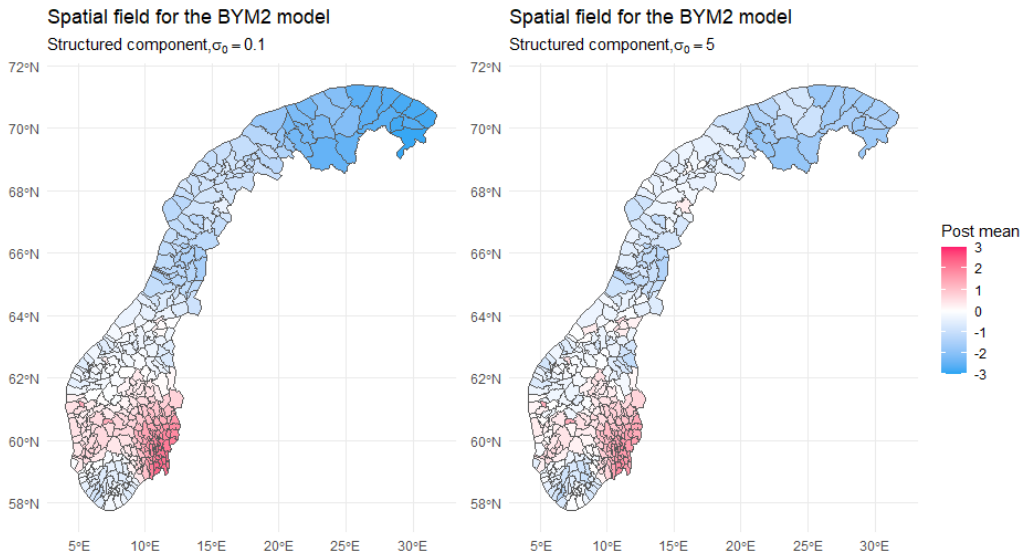

Fig. 6.17: Spatial fields for the structured component of a BYM2 model when changing the value for σ_0.

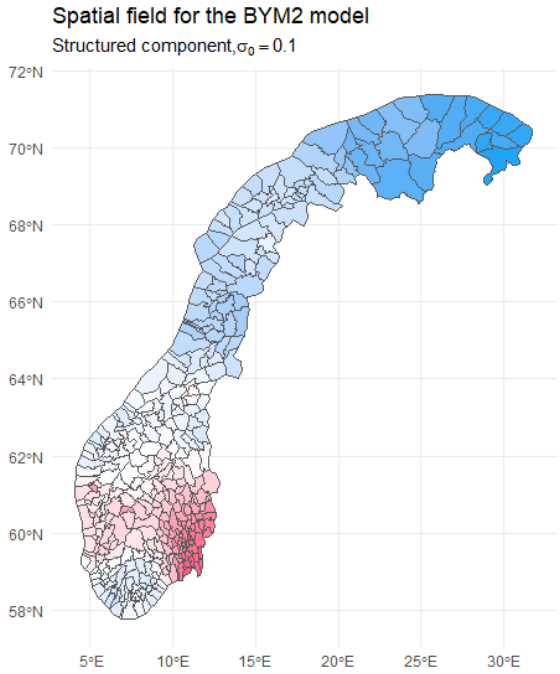

Fig. 6.18: Structured component of the spatial field for a BYM2 model.

The corresponding figures for Germany are shown in Figure 7.8, Figure 7.9, Figure 7.10 and Figure 7.11 in the Appendix.

6.6 Spatio-Temporal Models

6.6.1 Spatio-Temporal Models for Germany

6.6.2 Spatio-Temporal Models for Norway

6.7 Predictive Models

6.7.1 Predictive Models for Germany

6.7.2 Predictive Models for Norway

7 Appendix

7.1 Probability Distributions

7.1.1 The Normal Distribution

The normal distribution is an important type of continuous probability distribution in stochastics. The special significance of the normal distribution is based, among other things, on the central limit theorem, according to which distributions that result from the additive combination of a large number of independent influences are approximately normally distributed under weak conditions.
The density is given by

$$f\left(\boldsymbol{x}|\mu,\sigma\right)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{\boldsymbol{x}-\mu}{\sigma}\right)^{2}\right). \tag{7.1}$$

The first two moments of the distribution are given by

$$\mathbb{E}\left[X\right]=\mu \tag{7.2}$$

$$\text{Var}\left[X\right]=\sigma^{2}. \tag{7.3}$$

The graph of this density function has a "bell-shaped form" and is symmetrical with μ as the centre of symmetry [Fah+16, pp. 83–85].

7.1.2 The Poisson Distribution

The Poisson distribution is a discrete probability distribution that can be used to model the number of events that occur independently of each other at a constant mean rate in a fixed time interval or spatial area.
The density is given by

$$f\left(k\right)=\mathbb{P}\left(X=k\right)=\begin{cases}\frac{\lambda^{k}}{k!}\exp\left(-\lambda\right) & \text{for } x\in\{0,1,...\}\\ 0 & \text{else}\end{cases} \tag{7.4}$$

105

with λ representing the expected value of X.
The first two moments of the distribution are given by

$$\mathbb{E}[X] = \lambda \tag{7.5}$$

$$\text{Var}[X] = \lambda. \tag{7.6}$$

For $\lambda \geq 10$ the distribution becomes approximately symmetrical and can thus be approximated by a normal distribution [Fah+16, p. 243].

7.1.3 The Negative Binomial Distribution

The negative binomial distribution is a univariate probability distribution that belongs to the discrete probability distributions. It models the number of trials required to achieve a given number of successes in a Bernoulli process.
The density is given by

$$f(k, r, p) = \mathbb{P}(X = k) = \binom{k + r - 1}{r - 1} (1 - p)^k p^r, \tag{7.7}$$

with r the number of successes, k the number of failures, and p the probability of success.
The first two moments of the distribution are given by

$$\mathbb{E}[X] = \frac{pr}{1 - p} \tag{7.8}$$

$$\text{Var}[X] = \frac{pr}{(1 - p)^2}. \tag{7.9}$$

For large values of r, the negative binomial distribution can be approximated by a normal distribution [Hal41].

7.2 Distribution Fits

7.2.1 Distribution Fits for Germany

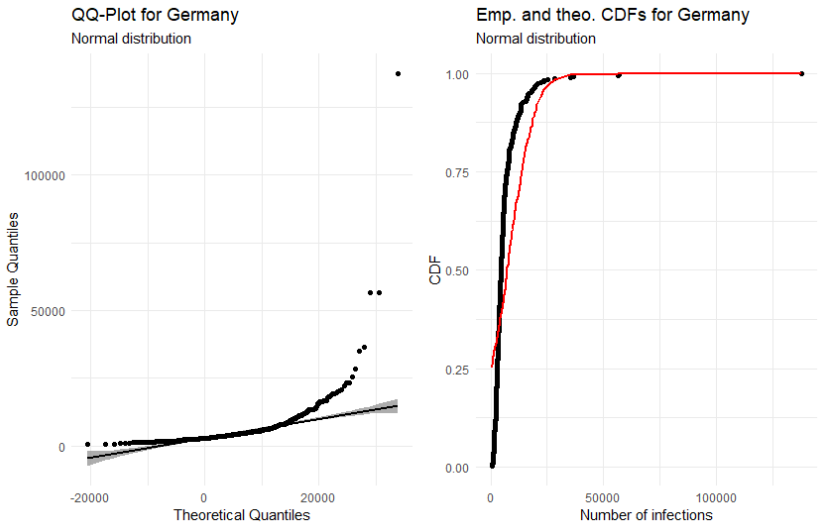

Fig. 7.1: A normal fit to the number of cases in German municipalities

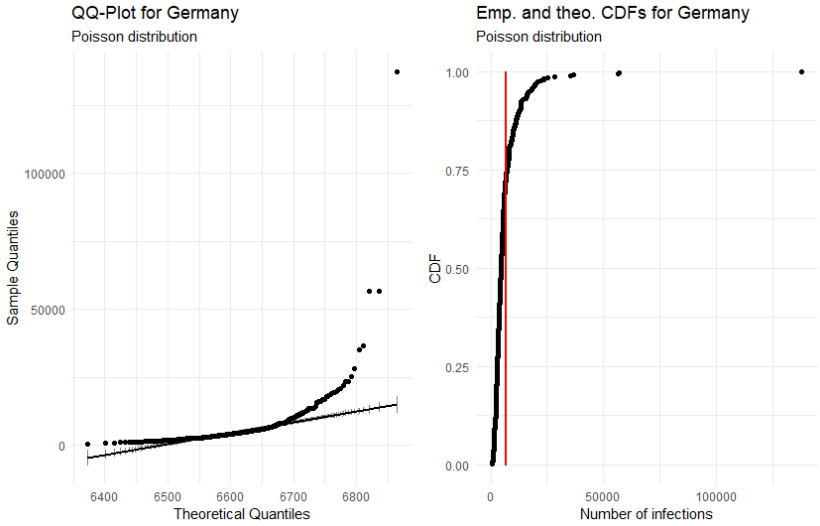

Fig. 7.2: A Poisson fit to the number of cases in German municipalities

7.2.2 Distribution Fits for Norway

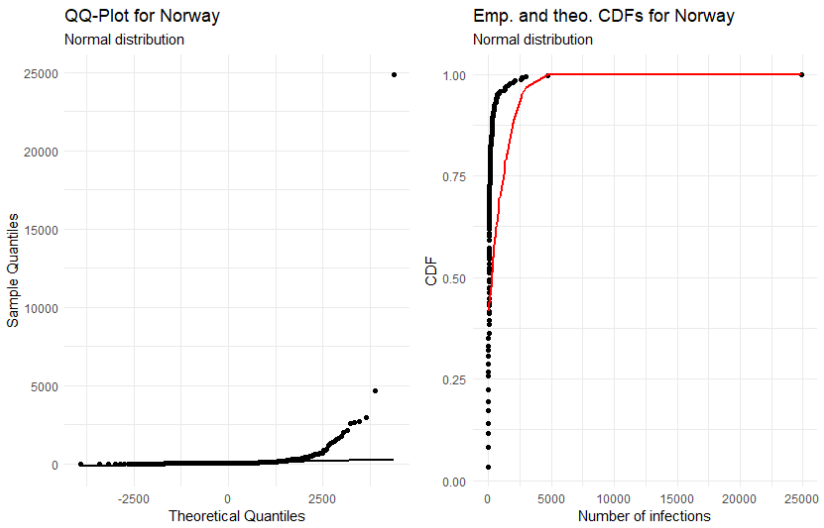

Fig. 7.3: A normal fit to the number of cases in Norwegian municipalities

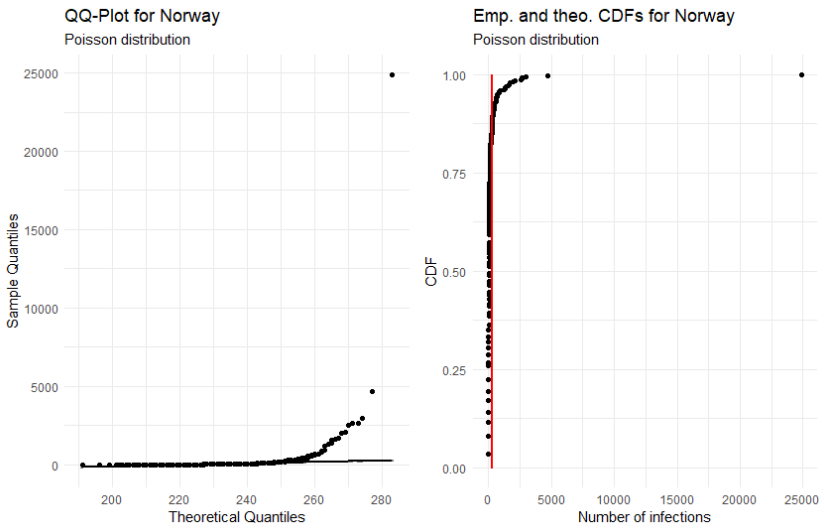

Fig. 7.4: A Poisson fit to the number of cases in Norwegian municipalities

7.3 Choice of Hyperpriors for Germany

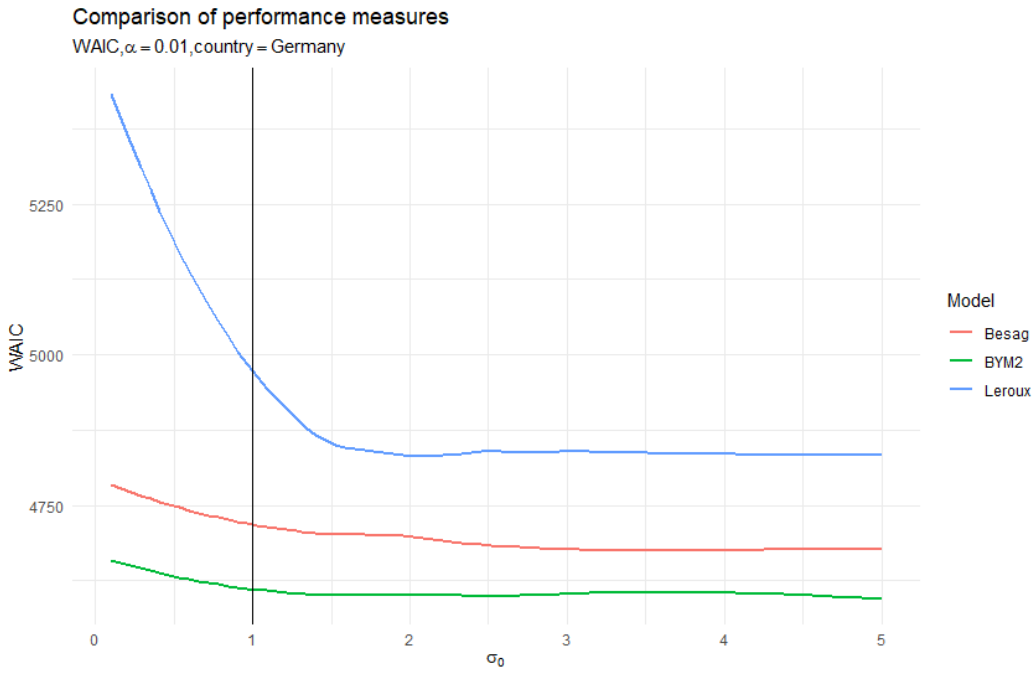

Fig. 7.5: Value the WAIC when changing the value for σ_0. The black line highlights the values for $\sigma_0 = 1$.

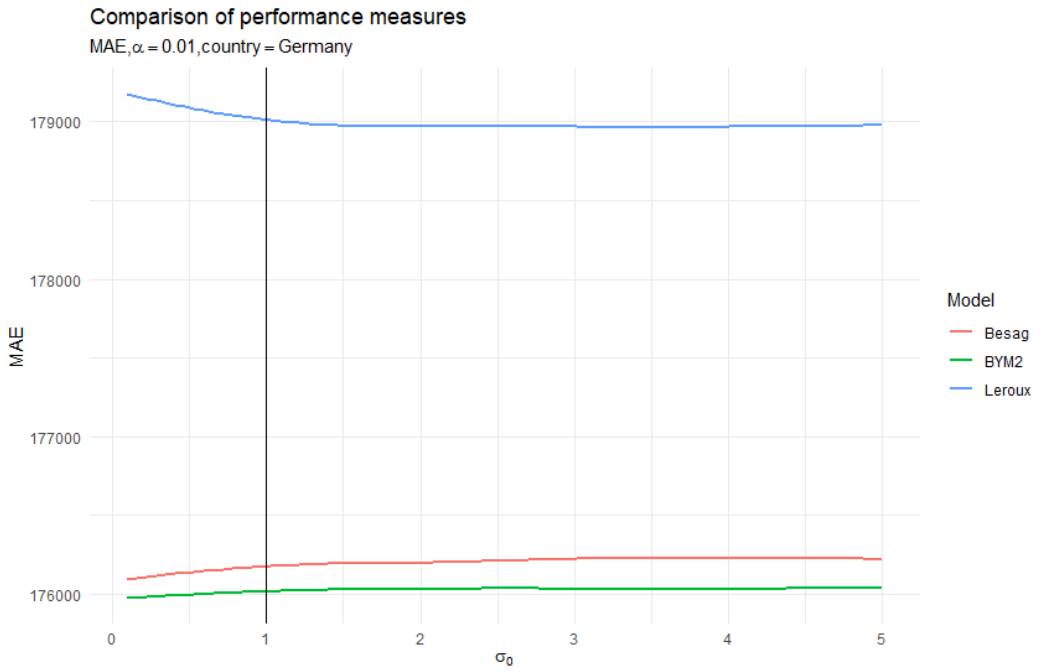

Fig. 7.6: Value the MAE when changing the value for σ_0. The black line highlights the values for $\sigma_0 = 1$.

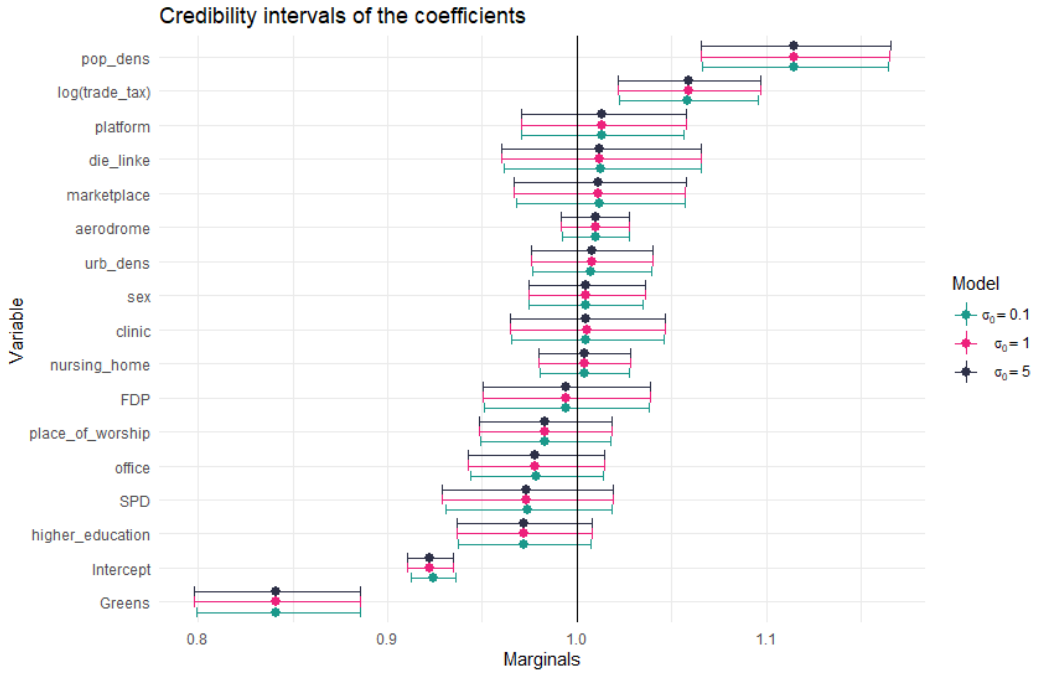

Fig. 7.7: Comparison of the credibility intervals of a BYM2 model for different values of σ_0

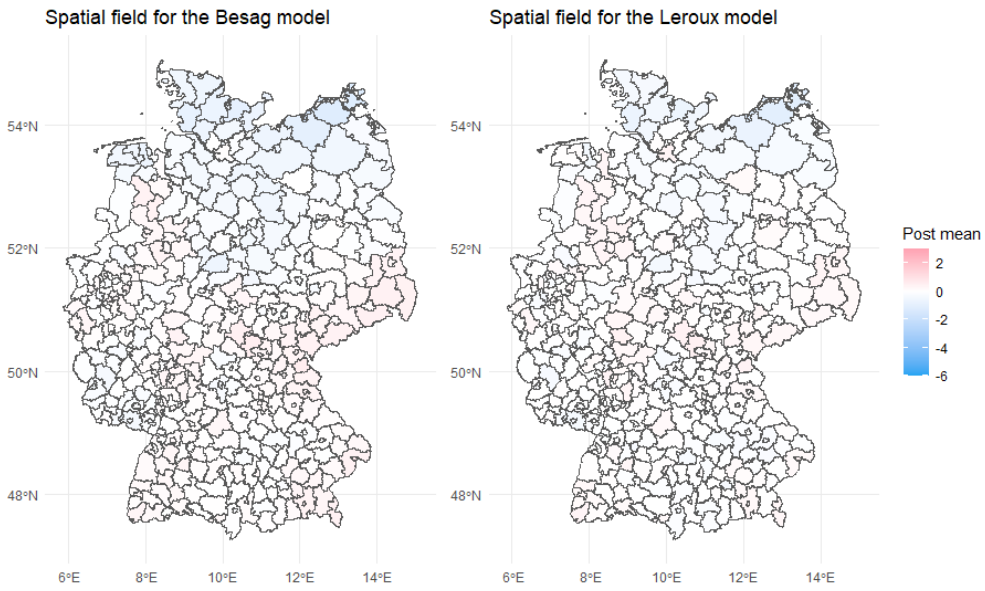

Fig. 7.8: Spatial field for a Besag model and a Leroux model.

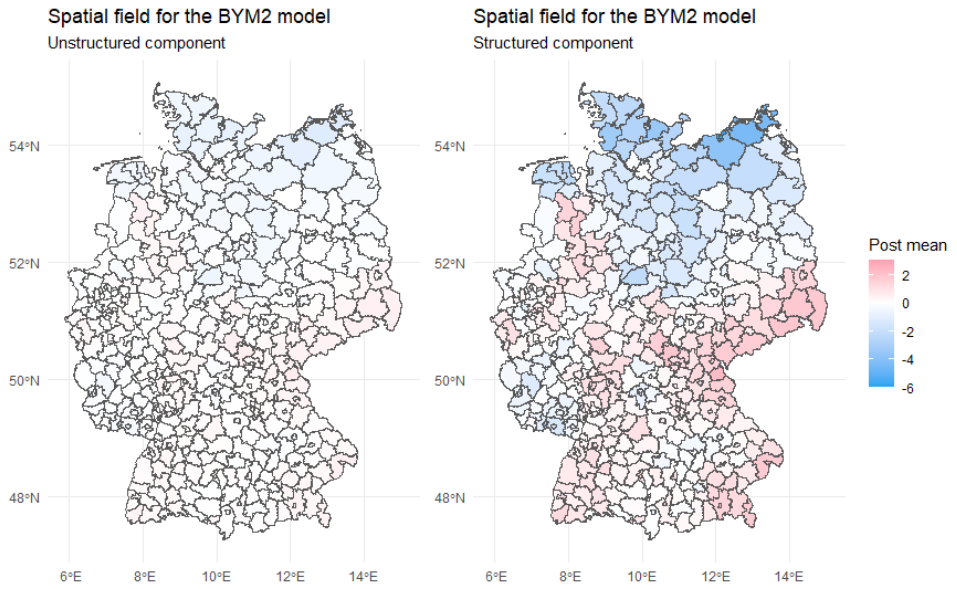

Fig. 7.9: Spatial fields for a BYM2 model.

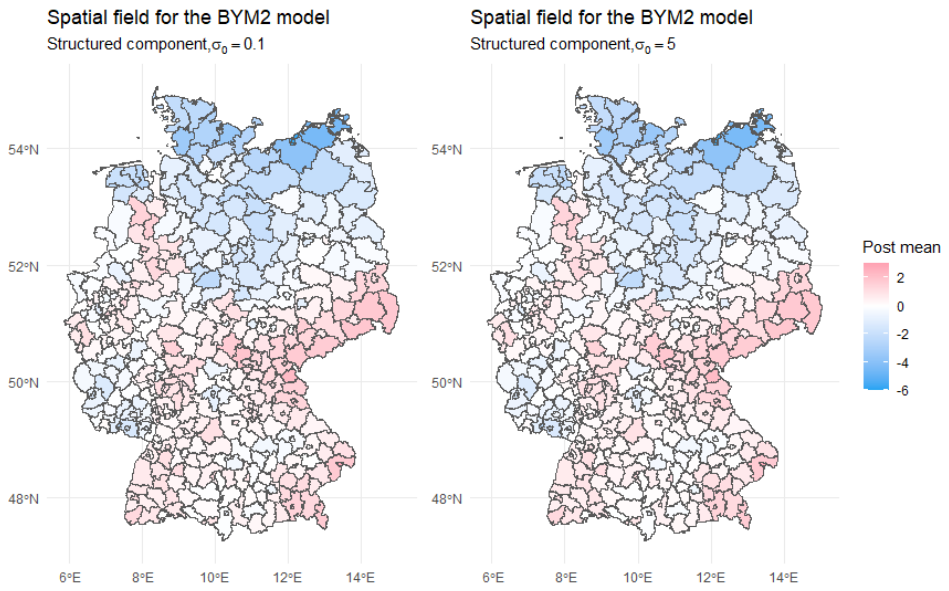

Fig. 7.10: Spatial fields for the structured component of a BYM2 model when changing the value for σ_0.

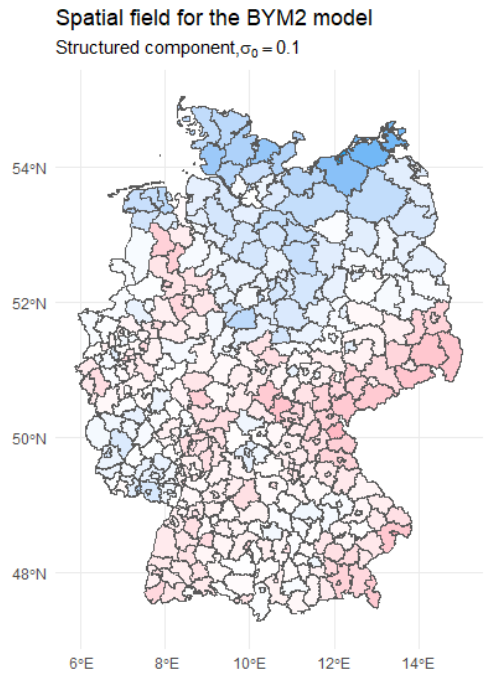

Fig. 7.11: Structured component of the spatial field for a BYM2 model.

7.4 Code Examples

7.4.1 Specifying the Different Types of Models

```
# set the seed
set.seed(420)
# draw a sample
test <- sample(
    seq_len(nrow(newest_numbers)),
    size = floor(0.2 * nrow(newest_numbers))
  )
# get the number of infections for the test data
test_value <- newest_numbers$value[test]
# set the number of infections to NA in the train data
newest_numbers$value[test] <- NA
# define the link function
link <- rep(NA, nrow(newest_numbers))
link[which(is.na(newest_numbers$value))] <- 1
# define the penalised prior
prior_1 <- list(
  prec = list(
    prior = "pc.prec",
    param = c(1, 0.01)
  )
)
# create the neighbordhood matrix
nb <- poly2nb(newest_numbers)
# save the matrix
nb2INLA("maps/map.adj", nb)
g <- inla.read.graph(filename = "maps/map.adj")
# define the C matrix for the Leroux model
Q <- Diagonal(x = sapply(nb, length))
for (i in 2:nrow(newest_numbers)) {
  Q[i - 1, i] <- -1
  Q[i, i - 1] <- -1
}

C <- Diagonal(x = 1, n = nrow(newest_numbers)) - Q
# define the formula for besags proper model
formula_besag <- value ~
  pop_dens + urb_dens + sex +
  f(idarea_1, model = "besagproper", graph = g, hyper = prior_1)
# define the formula for bym2 model
formula_bym2 <- value ~
  pop_dens + urb_dens + sex +
  f(
    idarea_1, model = "bym2", graph = g,
```

```
    scale.model = TRUE, hyper = prior_1
  )
# define the formula for leroux model
formula_leroux <- value ~
  pop_dens + urb_dens + sex +
  f(idarea_1, model = "generic1", Cmatrix = C, hyper = prior_1)
# compute the models
res_besag <- inla(
  formula_besag,
  family = "nbinomial",
  data = newest_numbers,
  E = expected_count,
  control.predictor = list(
    compute = TRUE,
    link = link
  ),
  Ntrials = newest_numbers$population,
  control.compute = list(dic = TRUE, waic = TRUE, cpo = TRUE)
)
res_bym2 <- inla(
  formula_bym2,
  family = "nbinomial",
  data = newest_numbers,
  E = expected_count,
  control.predictor = list(
    compute = TRUE,
    link = link
  ),
  Ntrials = newest_numbers$population,
  control.compute = list(dic = TRUE, waic = TRUE, cpo = TRUE)
)
res_leroux <- inla(
  formula_leroux,
  family = "nbinomial",
  data = newest_numbers,
  E = expected_count,
  control.predictor = list(
    compute = TRUE,
    link = link
  ),
  Ntrials = newest_numbers$population,
  control.compute = list(dic = TRUE, waic = TRUE, cpo = TRUE)
)
```

Listing 7.1: Specifying different models in INLA.

7.4.2 Making Predictions for the Test Data

```
# create a vector to save the predictions to
predicted <- c()
# now make the predictions
for(i in seq_len(nrow(newest_numbers))) {
  predicted[i] <- inla.emarginal(
    function(x) x * newest_numbers$population[i],
    res_bym2$marginals.fitted.values[[i]]
  )
}
# calculate the MAE for the test data
mean(abs(predicted[test] - test))
```

Listing 7.2: The code for making predictions in INLA.

7.4.3 Calculating the Posterior Mean

```
inla.emarginal(exp, model_leroux$marginals.fixed$trade_tax)
# calculate the increase in risk for an increase
# by 2 in the trade tax
inla.emarginal(
  exp,
  model_leroux$marginals.fixed$trade_tax
  ) ^ 2
```

Listing 7.3: Calculating the posterior mean of a coefficent.

7.4.4 Calculating a Credibility Interval

```
inla.qmarginal(
  c(0.025,0.975),
  inla.tmarginal(
    exp,
    model_leroux$marginals.fixed$trade_tax
  )
)
```

Listing 7.4: Extracting the credibility interval for a coefficient

7.4.5 Best Spatial Models For Germany

7.4.5.1 Best Spatial Model using Demographic Variables

```
prior_1 <- list(
  prec = list(
    prior = "pc.prec",
    param = c(1, 0.01)
  )
)
formula <- value ~
  pop_dens + urb_dens + SPD + Gruene + FDP + die_linke +
  f(
    idarea_1, model = "bym2", graph = g,
    scale.model = TRUE, hyper = prior_1
  )
```

Listing 7.5: The code for the demographic model.

7.4.5.2 Best Spatial Model using Infrastructural Variables

```
prior_1 <- list(
  prec = list(
    prior = "pc.prec",
    param = c(1, 0.01)
  )
)
formula <- value ~
  pop_dens + urb_dens + clinic + place_of_worship + retail +
  nursing_home + aerodrome + platform + higher_education +
  f(
    idarea_1, model = "bym2", graph = g,
    scale.model = TRUE, hyper = prior_1
  )
```

Listing 7.6: The code for the infrastructure model.

7.4.5.3 Best Spatial Model using Both Types of Variables

```
prior_1 <- list(
  prec = list(
    prior = "pc.prec",
    param = c(1, 0.01)
  )
)
formula <- value ~
  pop_dens + urb_dens + sex + log(trade_tax) + SPD +
  Gruene + FDP + die_linke + clinic + place_of_worship +
  retail + nursing_home + aerodrome + platform +
```

```
  higher_education +
  f(
    idarea_1, model = "bym2", graph = g,
    scale.model = TRUE, hyper = prior_1
  )
```

Listing 7.7: The code for the demographic + infrastructure model.

7.4.6 Best Spatial Models For Norway

7.4.6.1 Best Spatial Model using Demographic Variables

```
prior_1 <- list(
  prec = list(
    prior = "pc.prec",
    param = c(0.5 / 0.31, 0.01)
  )
)
formula <- value ~
  urb_dens + sex +
  f(
    idarea_1, model = "bym2", graph = g,
    scale.model = TRUE, hyper = prior_1
  )
```

Listing 7.8: The code for the demographic model.

7.4.6.2 Best Spatial Model using Infrastructural Variables

```
prior_1 <- list(
  prec = list(
    prior = "pc.prec",
    param = c(0.5 / 0.31, 0.01)
  )
)
formula <- value ~
  urb_dens + marketplace + place_of_worship +
  nursing_home + aerodrome + office + platform +
  higher_education +
  f(
    idarea_1, model = "bym2", graph = g,
    scale.model = TRUE, hyper = prior_1
  )
```

Listing 7.9: The code for the infrastructural model.

7.4.6.3 Best Spatial Model using Both Types of Variables

```
prior_1 <- list(
  prec = list(
    prior = "pc.prec",
    param = c(0.5 / 0.31, 0.01)
  )
)
formula <- value ~
  urb_dens + median_age + unemp_tot + unemp_immg +
  immigrants_total + sex + marketplace + place_of_worship +
  nursing_home + aerodrome + office + platform +
  higher_education +
  f(
    idarea_1, model = "bym2", graph = g,
    scale.model = TRUE, hyper = prior_1
  )
```

Listing 7.10: The code for the demographic + infrastructure model.

Bibliography

[Ait91] Murray Aitkin. “Posterior bayes factors”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 53.1 (1991), pp. 111–128 (cit. on p. 13).

[Aka74] Hirotugu Akaike. “A new look at the statistical model identification”. In: *IEEE transactions on automatic control* 19.6 (1974), pp. 716–723 (cit. on p. 44).

[BCG14] Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical modeling and analysis for spatial data*. CRC press, 2014 (cit. on pp. 27, 63).

[Bay63] Thomas Bayes. “LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFR S”. In: *Philosophical transactions of the Royal Society of London* 53 (1763), pp. 370–418 (cit. on p. 9).

[Ber+95] Luisa Bernardinelli, D Clayton, Cristiana Pascutto, et al. “Bayesian analysis of space—time variation in disease risk”. In: *Statistics in medicine* 14.21-22 (1995), pp. 2433–2443 (cit. on p. 62).

[Bes74] Julian Besag. “Spatial interaction and the statistical analysis of lattice systems”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 36.2 (1974), pp. 192–225 (cit. on pp. 40, 41).

[BYM91] Julian Besag, Jeremy York, and Annie Mollié. “Bayesian image restoration, with two applications in spatial statistics”. In: *Annals of the institute of statistical mathematics* 43.1 (1991), pp. 1–20 (cit. on pp. 27, 42, 58, 61).

[BGR15] Roger S. Bivand, Virgilio Gómez-Rubio, and Håvard Rue. “Spatial Data Analysis with R-INLA with Some Extensions”. In: *Journal of Statistical Software* 63.20 (2015), pp. 1–31 (cit. on p. 82).

[BT11] George EP Box and George C Tiao. *Bayesian inference in statistical analysis*. Vol. 40. John Wiley & Sons, 2011 (cit. on pp. 5, 9).

[Cam+13] Michela Cameletti, Finn Lindgren, Daniel Simpson, and Håvard Rue. “Spatio-temporal modeling of particulate matter concentration through the SPDE approach”. In: *AStA Advances in Statistical Analysis* 97.2 (2013), pp. 109–131 (cit. on p. 30).

[Cas+21] Marco Cascella, Michael Rajnik, Arturo Cuomo, Scott C Dulebohn, and Raffaela Di Napoli. “Features, evaluation, and treatment of coronavirus (COVID-19)”. In: *Statpearls [internet]* (2021) (cit. on p. 4).

[CGW05] Wei Chu, Zoubin Ghahramani, and Christopher KI Williams. “Gaussian processes for ordinal regression.” In: *Journal of machine learning research* 6.7 (2005) (cit. on p. 26).

[CI80] David Roxbee Cox and Valerie Isham. *Point processes*. Vol. 12. CRC Press, 1980 (cit. on p. 55).

[Cox80] DR Cox. "Discussion of 'Sampling and Bayes' inference in scientific modelling and robustness". In: *JR Stat Soc Ser A* 143 (1980), p. 410 (cit. on p. 46).

[CS02] Trevor A Craney and James G Surles. "Model-dependent variance inflation factor cutoff values". In: *Quality Engineering* 14.3 (2002), pp. 391–403 (cit. on p. 48).

[Cre15] Noel Cressie. *Statistics for spatial data*. John Wiley & Sons, 2015 (cit. on p. 28).

[Daw79] A Philip Dawid. "Conditional independence in statistical theory". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 41.1 (1979), pp. 1–15 (cit. on p. 10).

[DeC97] Lawrence T DeCarlo. "On the meaning and use of kurtosis." In: *Psychological methods* 2.3 (1997), p. 292 (cit. on p. 14).

[DGM00] Dipak K Dey, Sujit K Ghosh, and Bani K Mallick. *Generalized linear models: A Bayesian perspective*. CRC Press, 2000 (cit. on p. 26).

[DY79] Persi Diaconis and Donald Ylvisaker. "Conjugate priors for exponential families". In: *The Annals of statistics* (1979), pp. 269–281 (cit. on p. 15).

[DRC03] Peter J Diggle, Paulo J Ribeiro, and Ole F Christensen. "An introduction to model-based geostatistics". In: *Spatial statistics and computational methods*. Springer, 2003, pp. 43–86 (cit. on p. 29).

[DS11] David P Doane and Lori E Seward. "Measuring skewness: a forgotten statistic?" In: *Journal of statistics education* 19.2 (2011) (cit. on p. 14).

[Fah+16] Ludwig Fahrmeir, Christian Heumann, Rita Künstler, Iris Pigeot, and Gerhard Tutz. *Statistik: Der weg zur datenanalyse*. Springer-Verlag, 2016 (cit. on pp. 105, 106).

[FL01] Ludwig Fahrmeir and Stefan Lang. "Bayesian inference for generalized additive mixed models based on Markov random field priors". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 50.2 (2001), pp. 201–220 (cit. on p. 27).

[FT13] Ludwig Fahrmeir and Gerhard Tutz. *Multivariate statistical modelling based on generalized linear models*. Springer Science & Business Media, 2013 (cit. on pp. 12, 26).

[Fin97] Daniel Fink. "A compendium of conjugate priors". In: *See http://www. people. cornell. edu/pages/df36/CONJINTRnew% 20TEX. pdf* 46 (1997) (cit. on p. 15).

[Gam97] Dani Gamerman. "Sampling from the posterior distribution in generalized linear mixed models". In: *Statistics and Computing* 7.1 (1997), pp. 57–68 (cit. on p. 27).

[Gel+10] Alan E Gelfand, Peter Diggle, Peter Guttorp, and Montserrat Fuentes. *Handbook of spatial statistics*. CRC press, 2010 (cit. on pp. 30, 60).

[GHV14] Andrew Gelman, Jessica Hwang, and Aki Vehtari. "Understanding predictive information criteria for Bayesian models". In: *Statistics and computing* 24.6 (2014), pp. 997–1016 (cit. on p. 47).

[GY02] Carol A Gotway and Linda J Young. "Combining incompatible spatial data". In: *Journal of the American Statistical Association* 97.458 (2002), pp. 632–648 (cit. on p. 63).

[GG06] Peter Guttorp and Tilmann Gneiting. "Studies in the history of probability and statistics XLIX on the Matérn correlation family". In: *Biometrika* 93.4 (2006), pp. 989–995 (cit. on p. 30).

[Hal41] John BS Haldane. "The fitting of binomial distributions". In: *Annals of Eugenics* 11.1 (1941), pp. 179–181 (cit. on p. 106).

[Han20] Thomas Hansen. *Public COVID-19 Data for Norway (covid19data.no)*. `https://github.com/thohan88/covid19-nor-data`. 2020 (cit. on p. 67).

[Has70] W Keith Hastings. "Monte Carlo sampling methods using Markov chains and their applications". In: (1970) (cit. on p. 23).

[HL81] Paul W Holland and Samuel Leinhardt. "An exponential family of probability distributions for directed graphs". In: *Journal of the american Statistical association* 76.373 (1981), pp. 33–50 (cit. on p. 11).

[Irw05] ME Irwin. *Prior choice, summarizing the posterior*. 2005 (cit. on p. 15).

[KW03] EE Kammann and Matthew P Wand. "Geoadditive models". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 52.1 (2003), pp. 1–18 (cit. on p. 27).

[KG96] Genshiro Kitagawa and Will Gersch. *Smoothness priors analysis of time series*. Vol. 116. Springer Science & Business Media, 1996 (cit. on p. 27).

[Kno00] Leonhard Knorr-Held. "Bayesian modelling of inseparable space-time variation in disease risk". In: *Statistics in medicine* 19.17-18 (2000), pp. 2555–2567 (cit. on p. 62).

[Kor20] Astrid Iren Solheim Korsvoll. "Nye smittetilfelle i Hyllestad- no har sju personar korona". In: *Firda* (Oct. 1, 2020) (cit. on p. 79).

[KL51] Solomon Kullback and Richard A Leibler. "On information and sufficiency". In: *The annals of mathematical statistics* 22.1 (1951), pp. 79–86 (cit. on p. 16).

[LB04] Stefan Lang and Andreas Brezger. "Bayesian P-splines". In: *Journal of computational and graphical statistics* 13.1 (2004), pp. 183–212 (cit. on p. 26).

[LP17] Günter Last and Mathew Penrose. *Lectures on the Poisson process*. Vol. 7. Cambridge University Press, 2017 (cit. on pp. 56, 57).

[LLB00] Brian G Leroux, Xingye Lei, and Norman Breslow. “Estimation of disease rates in small areas: a new mixed model for spatial dependence”. In: *Statistical models in epidemiology, the environment, and clinical trials*. Springer, 2000, pp. 179–191 (cit. on p. 42).

[LRL11] Finn Lindgren, Håvard Rue, and Johan Lindström. “An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.4 (2011), pp. 423–498 (cit. on p. 65).

[LNM19] Robin Lovelace, Jakub Nowosad, and Jannes Muenchow. *Geocomputation with R*. CRC Press, 2019 (cit. on pp. 50, 52–54).

[Mar06] Andrey Andreyevich Markov. “Extension of the law of large numbers to dependent quantities”. In: *Izv. Fiz.-Matem. Obsch. Kazan Univ.(2nd Ser)* 15 (1906), pp. 135–156 (cit. on p. 20).

[Mar+01] Jose L. Marroquin, Fernando A. Velasco, Mariano Rivera, and Miguel Nakamura. “Gauss-Markov measure field models for low-level vision”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23.4 (2001), pp. 337–348 (cit. on p. 27).

[MLB08] Miguel A Martínez-Beneito, Antonio López-Quilez, and Paloma Botella-Rocamora. “An autoregressive approach to spatio-temporal disease mapping”. In: *Statistics in medicine* 27.15 (2008), pp. 2874–2889 (cit. on p. 61).

[Mar+14] Thiago G Martins, Daniel P Simpson, Andrea Riebler, et al. “Penalising model component complexity: A principled, practical approach to constructing priors”. In: *arXiv preprint arXiv:1403.4630* (2014) (cit. on pp. 16–18, 43).

[Met+53] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. “Equation of state calculations by fast computing machines”. In: *The journal of chemical physics* 21.6 (1953), pp. 1087–1092 (cit. on p. 23).

[MU49] Nicholas Metropolis and Stanislaw Ulam. “The monte carlo method”. In: *Journal of the American statistical association* 44.247 (1949), pp. 335–341 (cit. on p. 20).

[Mor19] Paula Moraga. *Geospatial health data: Modeling and visualization with R-INLA and shiny*. CRC Press, 2019 (cit. on pp. 29, 30, 39, 58, 60–63).

[Mor50] Patrick AP Moran. “Notes on continuous stochastic phenomena”. In: *Biometrika* 37.1/2 (1950), pp. 17–23 (cit. on p. 59).

[NW72] John Ashworth Nelder and Robert WM Wedderburn. “Generalized linear models”. In: *Journal of the Royal Statistical Society: Series A (General)* 135.3 (1972), pp. 370–384 (cit. on p. 45).

[NTB21] NTB. “Ordfører: Ulvik-utbruddet spredte seg trolig blant barna”. In: *Haugesunds Avis* (Feb. 8, 2021) (cit. on p. 79).

[Ope84] Stan Openshaw. *The modifiable areal unit problem. Norwich, UK*. 1984 (cit. on p. 63).

[Pad+17] Mark Padgham, Bob Rudis, Robin Lovelace, and Maëlle Salmon. “osmdata”. In: *The Journal of Open Source Software* 2.14 (June 2017) (cit. on p. 72).

[Peb18] Edzer Pebesma. “Simple Features for R: Standardized Support for Spatial Vector Data”. In: *The R Journal* 10.1 (2018), pp. 439–446 (cit. on p. 73).

[Pet90] LI Pettit. “The conditional predictive ordinate for the normal distribution”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 52.1 (1990), pp. 175–184 (cit. on p. 46).

[RS61] Howard Raiffa and Robert Schlaifer. “Applied Statistical Decision Theory, Division of Research, GraduateSchool of Business Administration, Harvard University, Boston, 1961”. In: *RaiffaApplied Statistical Decision Theory1961* (1961) (cit. on p. 15).

[Rie+16] Andrea Riebler, Sigrunn H Sørbye, Daniel Simpson, and Håvard Rue. “An intuitive Bayesian spatial model for disease mapping that accounts for scaling”. In: *Statistical methods in medical research* 25.4 (2016), pp. 1145–1165 (cit. on pp. 40–43, 47).

[RC13] Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013 (cit. on pp. 20–22, 24).

[RMR10] Christian P Robert, Jean-Michel Marin, and Judith Rousseau. “Bayesian inference”. In: *arXiv preprint arXiv:1002.2080* (2010) (cit. on p. 15).

[Rob09] William S Robinson. “Ecological correlations and the behavior of individuals”. In: *International journal of epidemiology* 38.2 (2009), pp. 337–341 (cit. on p. 63).

[RH05] Havard Rue and Leonhard Held. *Gaussian Markov random fields: theory and applications*. CRC press, 2005 (cit. on pp. 6–8, 10, 11, 13, 26, 32, 33, 35, 37).

[RMC09] Håvard Rue, Sara Martino, and Nicolas Chopin. “Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations”. In: *Journal of the royal statistical society: Series b (statistical methodology)* 71.2 (2009), pp. 319–392 (cit. on pp. 25–28, 37).

[Sch+21] Clemens Schmid, Stephan Schiffels, Johannes Boog, Martin Lange, and Moritz Aschoff. *covid19germany: Load, visualise and analyse daily updated data on the COVID-19 outbreak in Germany*. R package version 0.1.1. 2021 (cit. on p. 67).

[Sno57] John Snow. “Cholera, and the water supply in the south districts of London”. In: *British Medical Journal* 1.42 (1857), p. 864 (cit. on p. 49).

[Spi+14] David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van der Linde. “The deviance information criterion: 12 years on”. In: *Journal of the Royal Statistical Society: Series B: Statistical Methodology* (2014), pp. 485–493 (cit. on p. 45).

[Sto85] Charles J Stone. “Additive regression and other nonparametric models”. In: *The annals of Statistics* (1985), pp. 689–705 (cit. on p. 25).

[TC+00] Steven M Teutsch, R Elliott Churchill, et al. *Principles and practice of public health surveillance*. Oxford University Press, USA, 2000 (cit. on p. 49).

[WG04] Lance A Waller and Carol A Gotway. *Applied spatial statistics for public health data*. Vol. 368. John Wiley & Sons, 2004 (cit. on p. 63).

[WO10] Sumio Watanabe and Manfred Opper. "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory." In: *Journal of machine learning research* 11.12 (2010) (cit. on p. 45).

[Whi63] Peter Whittle. "Stochastic-processes in several dimensions". In: *Bulletin of the International Statistical Institute* 40.2 (1963), pp. 974–994 (cit. on p. 64).

[Wic07] Hadley Wickham. "Reshaping Data with the reshape Package". In: *Journal of Statistical Software* 21.12 (2007), pp. 1–20 (cit. on p. 73).

[Wic19] Hadley Wickham. *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.4.0. 2019 (cit. on p. 73).

[Wil44] J Ernest Wilkins. "A note on skewness and kurtosis". In: *The Annals of Mathematical Statistics* 15.3 (1944), pp. 333–335 (cit. on p. 14).

[Yon18] Luo Yong. "LOO and WAIC as model selection methods for polytomous items". In: *arXiv preprint arXiv:1806.09996* (2018) (cit. on p. 45).

Webpages

[@Bun20] Statistische Ämter des Bundes und der Länder. *Regionaldatenbank Deutschland*. 2020. URL: `https://www.regionalstatistik.de/genesis/online` (visited on Feb. 28, 2021) (cit. on p. 69).

[@Deu20] Esri Deutschland. *Kreisgrenzen 2017*. 2020. URL: `https://opendata-esri-de.opendata.arcgis.com/datasets/affd8ace4c204981b5d32070f9547eb9_0` (visited on Feb. 28, 2021) (cit. on p. 71).

[@Geo21] Geonorge. *Kartkatalogen*. 2021. URL: `https://www.geonorge.no/` (visited on Feb. 28, 2021) (cit. on p. 71).

[@Ope17] OpenStreetMap contributors. *Planet dump retrieved from https://planet.osm.org*. 2017 (cit. on p. 72).

[@Sen16] Statistisk Sentralbyrå. *StatBank Norway*. 2016. URL: `https://www.ssb.no/en/statbank/` (visited on Feb. 28, 2021) (cit. on p. 69).

[@Smi20] Erik Smistad. *konverter-norgeskart-projeksjon*. 2020 (cit. on p. 71).

List of Figures

125

List of Tables

List of Listings