Contents

1 Introduction **1**

2 Corona Virus **2**

3 Introduction to Bayesian Inference **3**

3.1 Preliminaries 4

3.2 Basic Concepts of Bayesian Theory 8

3.2.1 Bayes' Theorem 8

3.2.2 Conditional Independence 8

3.2.3 Undirected Graphs 9

3.2.4 The Exponential Family 10

3.2.5 The Multivariate Normal Distribution 11

3.3 Prior Selection 13

3.3.1 Conjugate Priors 13

3.3.2 Jeffreys' Prior 13

3.3.3 Penalised Complexity Priors 14

3.4 Markov-Chain-Monte-Carlo-Methods 19

3.4.1 Monte Carlo Integration 19

3.4.2 Markov Chains 20

3.4.3 The Metropolis-Hastings Algorithm 23

3.4.4 The Gibbs Sampler 24

3.5 Latent Gaussian Models and INLA 26

3.5.1 Applications for Latent Gaussian Models 27

3.5.2 The MCMC Approach to Inference 28

3.5.3 Gaussian Random Fields 28

3.5.4 Gaussian Markov Random Fields 32

3.5.5 Integrated Nested Laplace Approximation 37

4 Analysis of Geospatial Health Data **40**

4.1 Geographic Data 41

4.1.1 Vector Data 41

4.1.2 Raster Data 43

4.2 Spatial Point Processes 46
4.2.1 Fundamentals of Point Processes 46
4.2.2 Poisson Processes 47
4.2.3 Random Measures and Cox Processes 48
4.3 Modeling and Visualising Health Data 50
4.3.1 Areal Data 50
4.3.2 Geostatistical Data 56

5 Dataset Collection **58**
5.1 Covid-19 Data 59
5.2 Demographic Data 60
5.3 Shapefiles 64
5.4 OpenStreetMap Data 65
5.5 Data Wrangling 66

6 Data Analysis **69**
6.1 Standardised Incidence Ratio 69
6.1.1 Standardised Incidence Ratio for Germany 69
6.1.2 Standardised Incidence Ratio for Norway 70
6.2 Spatial Models 72
6.2.1 Spatial Models for Germany 75
6.2.2 Spatial Models for Norway 82
6.3 Spatio-Temporal Models 89
6.3.1 Spatio-Temporal Models for Germany 89
6.3.2 Spatio-Temporal Models for Norway 89
6.4 Regression Models 89
6.4.1 Regression Models for Germany 90
6.4.2 Regression Models for Norway 90

7 Appendix **91**
7.1 The Negative Binomial Distribution 91
7.2 Code examples 91

Bibliography **97**

1 Introduction

1

Corona Virus

2

2

3 Introduction to Bayesian Inference

Bayesian Inference is a method of statistical inference that uses Bayes' theorem to update the probability of a hypothesis as more data are observed or more information becomes available. It is an essential technique in mathematical statistics and the polar opposite of the frequentist approach, which makes predictions based solely on data from an experiment. 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 $\pi\left(\boldsymbol{x}|\boldsymbol{\theta}, \sigma^2\right)$, and the observed data x, it is possible to calculate the probability distribution $\pi\left(\boldsymbol{\theta}, \sigma^2 \middle| \boldsymbol{x}\right)$ of $\boldsymbol{\theta}$ and σ^2 given the data $\boldsymbol{x}$. This distribution is called the *posterior* distribution of $\boldsymbol{\theta}$ and σ^2 an is used to make inferences about the parameters [BT11, p. 6].

3

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].

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 $(x_1, x_{i+1}, ..., x_j)^T$ 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 $\text{vec}(\boldsymbol{A}) = \left(\boldsymbol{A}_1^T, \boldsymbol{A}_2^2, ..., \boldsymbol{A}_m^T\right)$. 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].

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). $\boldsymbol{x}$ can take values on $\mathcal{I}_{\boldsymbol{n}}$ and $x_{i,j}$ denotes the value of $\boldsymbol{x}$ at location ij, for $i = 1, ..., n_1$ and $j = 1, ..., n_2$. For easier reading this will be shortened to x_{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 $\boldsymbol{x}$ is defined on $\mathcal{I}_{\boldsymbol{n}}$, the toroidal boundary conditions imply that x_{-2,n_2} is equal to $x_{n_1-2,0}$ 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].

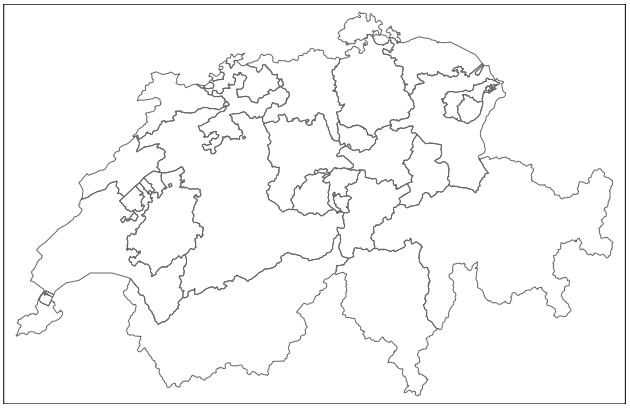

Fig. 3.1: The cantons of Switzerland, an example of an irregular lattice.

General Notation and Abbreviations

For $C \in \mathcal{I} = \{1, ..., n\}$ let $\boldsymbol{x}_C = \{x_i : i \in C\}$. $-C$ denotes the set $\mathcal{I} - \mathcal{C}$ such that $\boldsymbol{x}_{-C} = \{x_i : i \in C\}$. 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 $\pi(\boldsymbol{x})$ for the density of $\boldsymbol{x}$ and $\pi(\boldsymbol{x}_A|\boldsymbol{x}_{-A})$ for the conditional density of $\boldsymbol{x}_A$, given $\boldsymbol{x}_{-A}$. '$\sim$' is used when a variable is 'distributed' according to the law $\mathcal{L}$.
The expected value is denoted by $\mathbb{E}[\cdot]$, the variance by $\text{Var}(\cdot)$, the covariance by

$\mathrm{Cov}(\cdot)$, the precision by $\mathrm{Prec}(\cdot) = \mathrm{Cov}(\cdot)^{-1}$, the correlation by $\mathrm{Corr}(\cdot,\cdot)$ and a probability by $\mathbb{P}(\cdot)$ [RH05, p. 16].

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. $\mathrm{rank}(\boldsymbol{A}) = n$.
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 triangle, 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 $\boldsymbol{x}^T = (x_1, ..., x_n)$ be a vector of n observations whose probability distribution $\pi(\boldsymbol{x}|\boldsymbol{\theta})$ 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

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

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

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

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. $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ contains what is known about $\boldsymbol{\theta}$ given knowledge of the data and is the *posterior* distribution of $\boldsymbol{\theta}$ given $\boldsymbol{x}$.
If $\pi(\boldsymbol{x}|\boldsymbol{\theta})$ is considered as a function of $\boldsymbol{\theta}$ instead of $\boldsymbol{x}$, it is called the *likelihood function* of $\boldsymbol{\theta}$ given $\boldsymbol{x}$ and can be written as $l(\boldsymbol{\theta}|\boldsymbol{x})$. Thus Bayes' theorem can be written as

$$\pi(\boldsymbol{\theta}|\boldsymbol{x}) = l(\boldsymbol{\theta}|\boldsymbol{x})\,\pi(\boldsymbol{\theta})\,. \tag{3.3}$$

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

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

The data $\boldsymbol{x}$ 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.

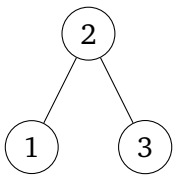

Fig. 3.2: An undirected labelled graph with 3 nodes, $\mathcal{V} = \{1, 2, 3\}$ and $\mathcal{E} = \{\{1, 2\}\{2, 3\}\}$.

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 $\mathcal{E}^A = \{\{1, 2\}\}$ [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 $c(\boldsymbol{y}, \lambda) \geq 0$ and measurable. $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^q$ 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

$0 < \int \exp\left(\left(\boldsymbol{y}^T\boldsymbol{\theta} - b(\boldsymbol{\theta})\right)/\lambda + c(\boldsymbol{y}, \lambda)\right) d\boldsymbol{y} < \infty$. Moreover, $b(\boldsymbol{\theta})$ is a twice differentiable function and all moments of $\boldsymbol{y}$ exist. Specifically,

$$\mathbb{E}_{\boldsymbol{\theta}}(\boldsymbol{y}) = \mu(\boldsymbol{\theta}) = \frac{\partial b(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \tag{3.7}$$

$$\text{Cov}_{\boldsymbol{\theta}}(\boldsymbol{y}) = \boldsymbol{\Sigma}(\boldsymbol{\theta}) = \lambda \frac{\partial^2 b(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}. \tag{3.8}$$

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 $\theta(\mu)$ into $\frac{\partial^2 b(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$, the variance function

$$v(\mu) = \frac{\partial^2 b(\boldsymbol{\theta}(\mu))}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \tag{3.9}$$

is given and the covariance can be written as

$$\text{Cov}_{\boldsymbol{\theta}}(\boldsymbol{y}) = \lambda v(\mu). \tag{3.10}$$

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 $\boldsymbol{x} = (x_1, ..., x_n)^T, n < \infty$ with mean vector $\boldsymbol{\mu}$ $(n \times 1)$ and SPD covariance matrix Σ $(n \times n)$ is

$$\pi(\boldsymbol{x}) = (2\pi)^{-n/2} |\boldsymbol{\Sigma}|^{-1/2} \exp \pi\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right), \ \boldsymbol{x} \in \mathbb{R}^n \tag{3.11}$$

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

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu_A} \\ \boldsymbol{\mu_B} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma_{AA}} & \boldsymbol{\Sigma_{AB}} \\ \boldsymbol{\Sigma_{BA}} & \boldsymbol{\Sigma_{BB}} \end{pmatrix}.$$

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

1. $\boldsymbol{x_A} \sim \mathcal{N}(\boldsymbol{\mu_A}, \boldsymbol{\Sigma_{AA}})$.
2. $\boldsymbol{\Sigma_{AB}} = \boldsymbol{0}$ precisely when $\boldsymbol{x_A}$ and $\boldsymbol{x_B}$ are independent.

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

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

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

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

3.3 Prior Selection

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 $\pi(\boldsymbol{\theta}|\boldsymbol{x})$ 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 $\pi(\boldsymbol{x}|\boldsymbol{\theta})$ 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 [Fin97]. An example of the construction of a conjugated prior is provided in the appendix.
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 priorities can be too restrictive and not every belief about the priori 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 Jeffreys' Prior

As conjugate priors depend on information from the data, they are so-called *informative priors*. *Jeffreys' prior* [Jef46] is characterised in comparison to other a priori distributions by the fact that it is invariant to a reparametrisation of the model parameters. Due to this invariance, the Jeffreys' a priori distribution is also called *non-informative*.

Let $\phi = \phi(\theta)$ be a one-to-one transformation of θ and $I(\cdot)$ be the Fisher information, then

$$I_\phi(\phi) = I_\theta(\theta)\left(\frac{\partial\theta}{\partial\phi}\right)^2. \tag{3.12}$$

If some principle of choice led to $\pi(\theta)$ as a non-informative prior for θ, the same principle should lead to

$$\pi_\phi(\phi) = \pi_\theta(\theta)\left|\frac{\partial\theta}{\partial\phi}\right| \tag{3.13}$$

as a non-informative prior for ϕ. The desired "invariance" is then given by

$$\pi_\phi(\phi) \propto \sqrt{I_\phi(\phi)} = \sqrt{I_\theta(\theta)}\left|\frac{\partial\theta}{\partial\phi}\right| \propto \pi_\theta(\theta)\left|\frac{\partial\theta}{\partial\phi}\right| \tag{3.14}$$

[BT11, pp. 42–43]. While Jeffreys' priors work well for single parameter models, they struggle when it comes to models with multidimensional parameters as these priors can have poor convergence properties [Jor10].

3.3.3 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 priority 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.

The Principles Behind PC Priors

Four main principles should be followed to calculate priorities 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 [Sim+17].

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,

$$\text{KLD}(f_1||f_2) = \int f_1(x) \log\left(\frac{f_1(x)}{f_0(x)}\right) dx. \tag{3.15}$$

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

$$\text{KLD}(f_1||f_0) = \frac{1}{2}\left(\text{trace}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}_1\right) - n - \ln\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_0|}\right)\right), \tag{3.16}$$

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

$$d(\xi) = d(f_1||f_0) = \sqrt{2\text{KLD}(f_1||f_0)} \tag{3.17}$$

which can be interpreted as a measure of distance from f_1 to f_0 [Sim+17].

Constant Rate Penalisation

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

$$\frac{\pi_d(d(\xi) + \delta)}{\pi_d(d(\xi))} = r^\delta, \qquad d(\xi), \delta \geq 0. \tag{3.18}$$

$r \in (0, 1)$ represents the constant decay rate and thus implies that the relative change in the priority distribution for $d(\xi)$ is independent of the actual distance. Therefore, $d(\xi)$ is exponentially distributed with density $\pi(d(\xi)) = \lambda \exp(-\lambda d(\xi))$ and rate

$\lambda = -\ln(r)$. By a standard variable change transformation, the corresponding PC prior for ξ is given [Sim+17].

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

$$\mathbb{P}(Q(\xi) > U) = \alpha, \tag{3.19}$$

where U represents an assumed upper bound for an interpretable transformation $Q(\xi)$ and α denotes a small probability [Sim+17].

PC Priors for AR(1)

The first-order AR process is given by

$$x_t = \phi x_{t-1} + \epsilon_t, \qquad \epsilon \sim \mathcal{N}\left(0, \kappa^{-1}\right), \ t = 2, ..., n, \tag{3.20}$$

where x_1 is assumed to follow a normal distribution with mean 0 and marginal precision $\tau = \kappa(1-\phi^2)$. The variables $\{\epsilon_t\}_{t=1}^n$ are independent and follow a $\mathcal{N}(0, \kappa)$ distribution. The AR(1) model represents an important special case of AR processes where the autocorrelation coefficient ϕ specifies the complete dependence structure. For the precision parameter τ a type 2 Gumbel distribution is used as PC prior:

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

This prior is derived using infinite precision as the base model, following Simpson et. al (2017) [Sim+17]. The rate λ is inferred through $\mathbb{P}(1/\sqrt{\tau} > U) = \alpha$, with α denoting a small probability [SR17].

Base Model: No Dependency in Time

The correlation matrix of an AR(1) is generally defined as $\mathbf{\Sigma}_1 = \left(\phi^{|i-j|}\right)$. In the case of no autocorrelation, white noise results and the correlation matrix is equal to the identity matrix, $\mathbf{\Sigma}_0 = \boldsymbol{I}$. The distance function is defined as $d(\phi) = \sqrt{(1-n)\log(1-\phi^2)}$. According to the constant rate penalty principle, $d(\phi)$ is

assigned an exponential prior with rate $\theta/\sqrt{n-1}$. This leads to a prior distribution that is invariant to n, and the PC for the one-lag autocorrelation is given by

$$\pi(\phi) = \frac{\theta}{2}\exp\left(-\theta\sqrt{-\ln(1-\phi^2)}\right)\frac{|\phi|}{(1-\phi^2)\sqrt{-\ln(1-\phi^2)}}, \qquad |\phi|<1, \theta>0. \tag{3.21}$$

The rate parameter θ influences at what rate the prior shrinks towards the white noise base model. To infer θ, a tail event is used. In the case of $\phi = 0$ a tail event can be defined by the fact that large absolute correlations are less likely, i.e..,

$$\mathbb{P}(|\phi| > U) = \alpha.$$

This implies that $\theta = -\ln(\alpha)/\sqrt{-\ln(1-U^2)}$ [SR17].

Base Model: No Change in Time

As an alternative to the base model for the AR(1) process, it can be assumed that the process remains constant in time ($\phi = 1$), thus representing a limiting random walk case, which is a non-stationary and singular process. To derive the PC prior for ϕ, let $\mathbf{\Sigma}_1 = \left(\phi^{|i-j|}\right)$ and $\mathbf{\Sigma}_0 = \left(\phi_0^{|i-j|}\right)$, where ϕ_0 is close to 1 and $\phi < \phi_0$. The Kullback-Leibler divergence is

$$\begin{aligned}&\text{KLD}(f_1(\phi)||f_0) = \\ &\frac{1}{2}\left(\frac{1}{1-\phi_0^2}\left(n - 2(n-1)\phi_0\phi + (n-2)\phi_0^2\right) - n - (n-1)\ln\left(\frac{1-\phi^2}{1-\phi_0^2}\right)\right).\end{aligned}$$

Considering the limit as $\phi_0 \to 1$, the distance is

$$\begin{aligned}d(\phi) &= \lim_{\phi_0\to 1}\sqrt{2\text{KLD}(f_1(\phi)||f_0)} \\ &= \lim_{\phi_0\to 1}\sqrt{\frac{2(n-1)(1-\phi)}{1-\phi_0^2}} = c\sqrt{1-\phi}, \qquad |\phi|<1,\end{aligned}$$

constant for c, independent of ϕ. Since $0 \le d(\phi) \le c\sqrt{2}$, $d(\phi)$ is assigned a truncated exponential distribution with rate θ/c, resulting in the following PC prior,

$$\pi(\phi) = \frac{\theta\exp\left(-\theta\sqrt{1-\phi}\right)}{\left(1-\exp\left(-\sqrt{2}\theta\right)\right)2\sqrt{1-\phi}}, \qquad |\phi|<1. \tag{3.22}$$

To scale the prior in terms of θ, (U, α) is determined in terms of $\mathbb{P}(\phi > U) = \alpha$. This equation is solved by

$$\frac{1 - \exp\left(-\theta\sqrt{1-U}\right)}{1 - \exp\left(-\sqrt{2}\theta\right)} = \alpha,$$

provided that α is larger than the lower limit $\sqrt{(1-U)/2}$ [SR17].

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 $X_0, X_1, ..., X_n$ 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},

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

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. Next, some of them will be introduced.

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].

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].

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].

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 ans 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].

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 $f(y_t) = 0$ 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].

3.4.4 The Gibbs Sampler

Gibbs-Sampling is a special case of the Metropolis-Hastings Algorithm, that is used to generate a sequence of samples of the joint probability distribution of two or more random variables. The aim of the method is to approximate this unknown joint probability distribution. Gibbs sampling is especially suitable when the joint distribution of a random vector is unknown, but the conditional distribution of each random variable is known. The underlying principle is to repeatedly select a variable and generate a value according to its conditional distribution, depending on the values of the other variables. During this iteration step, the values of the other variables remain unchanged. A Markov chain can be derived from the resulting sequence of sample vectors, and it can be shown that the stationary distribution of this Markov chain is precisely the sought joint distribution of the random vector.
The Gibbs sampler was first introduced in a 1984 paper by Stuart and Donald Geman [GG84] and is named after the physicist Josiah Willard Gibbs.

The Two-Stage Gibbs Sampler

A general introduction to Gibbs sampling is the two-stage Gibbs sampler, which is applicable to a wide range of statistical models that do not demand the generality of the multi-stage Gibbs sampler.
Implementing the algorithm is straightforward. If the random variables X and Y have a joint density $f(\boldsymbol{x}, \boldsymbol{y})$, the two-stage Gibbs sampler generates a Markov chain (X_t, Y_t) as shown below. $f_{Y|X}$ and $f_{X|Y}$ represent the conditional densities

Algorithm 2 The Two-Stage Gibbs Sampler

Take $X_0 = x_0$
1: **for** each iteration t **do**
2: Generate $Y_t \sim f_{Y|X}(\cdot|x_{t-1})$
3: Generate $X_t \sim f_{X|Y}(\cdot|y_t)$

associated with f. It is worth noting that not only (X_t, Y_t) is a Markov chain, but also the subsequences (X_t) and (Y_t) are. [RC13, p. 339].

Normal Bivariate Gibbs Sampler

In the case of the bivariate normal density

$$(X, Y) \sim \mathcal{N}_2 \left(0, \begin{pmatrix} 1 & p \\ p & 1 \end{pmatrix}\right)$$

the Gibbs sampler reads as follows:

Algorithm 3 The Two-Stage Gibbs Sampler for a normal distribution

Given y_t

1: **for** each iteration t **do**

2: Generate $X_{t+1}|y_t \sim \mathcal{N}\left(py_t, 1 - p^2\right)$

3: Generate $Y_{t+1}|x_{t+1} \sim \mathcal{N}\left(px_{t+1}, 1 - p^2\right)$

[RC13, p. 340]

The Multi-Stage Gibbs Sampler

Let $p > 1$, then the random variable $X \in \mathcal{X}$ can be written as $X = (X_1, ..., X_p)$, where the X_i's are either one-dimensional or multidimensional. Moreover, assume that a simulation is possible from the corresponding univariate conditional densities $f_1, ..., f_p$, i.e.,

$$X_i|\boldsymbol{x}_{-i} \sim f_i\left(x_i|\boldsymbol{x}_{-i}\right)$$

can be simulated for $i = 1, ..., p$. The Gibbs sampler is then specified by the following transition from $X^{(t)}$ to $X^{(t+1)}$:

Algorithm 4 The Multi-Stage Gibbs Sampler

Given $x^{(t)} = \left(x_1^{(t)}, ..., x_p^{(t)}\right)$, generate

1: $X_1^{(t+1)} \sim f_1\left(x_1|x_2^{(t)}, ..., x_p^{(t)}\right)$;

2: $X_2^{(t+1)} \sim f_2\left(x_2|x_1^{(t)}, x_3^{(t)}, ..., x_p^{(t)}\right)$;

$\vdots$

$X_p^{(p+1)} \sim f_p\left(x_p|\boldsymbol{x}_{-p}\right)$

$f_1, ..., f_p$ are referred to as the *full conditionals* and these are the only densities used for simulation. Hence, all simulations can be univariate, even for a high-dimensional problems [RC13, pp. 371–373].

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.

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,

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

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 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

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

Most latent Gaussian models satisfy two basic properties:

1. The latent field $\boldsymbol{x}$ is of large dimension, $n \approx 10^2 - 10^5$. Therefor, 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].

3.5.1 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.

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].

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 $\{f_t\}$.

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].

3.5.2 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 performance from the end user's point of view [RMC09, p. 322].

3.5.3 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:

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

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

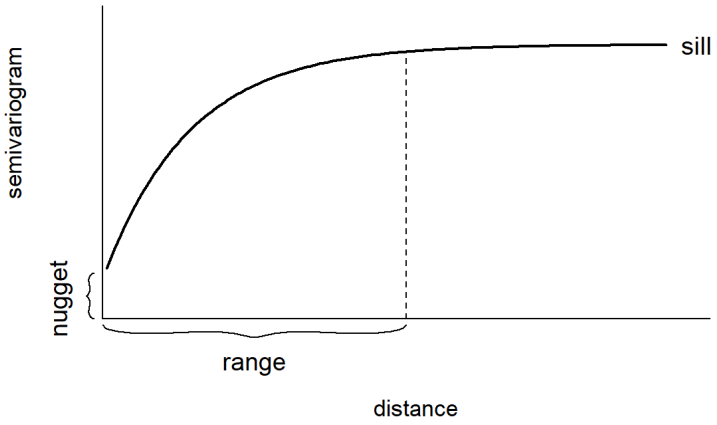

Fig. 3.3: A typical semivariogram

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.
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 below [Mor19].

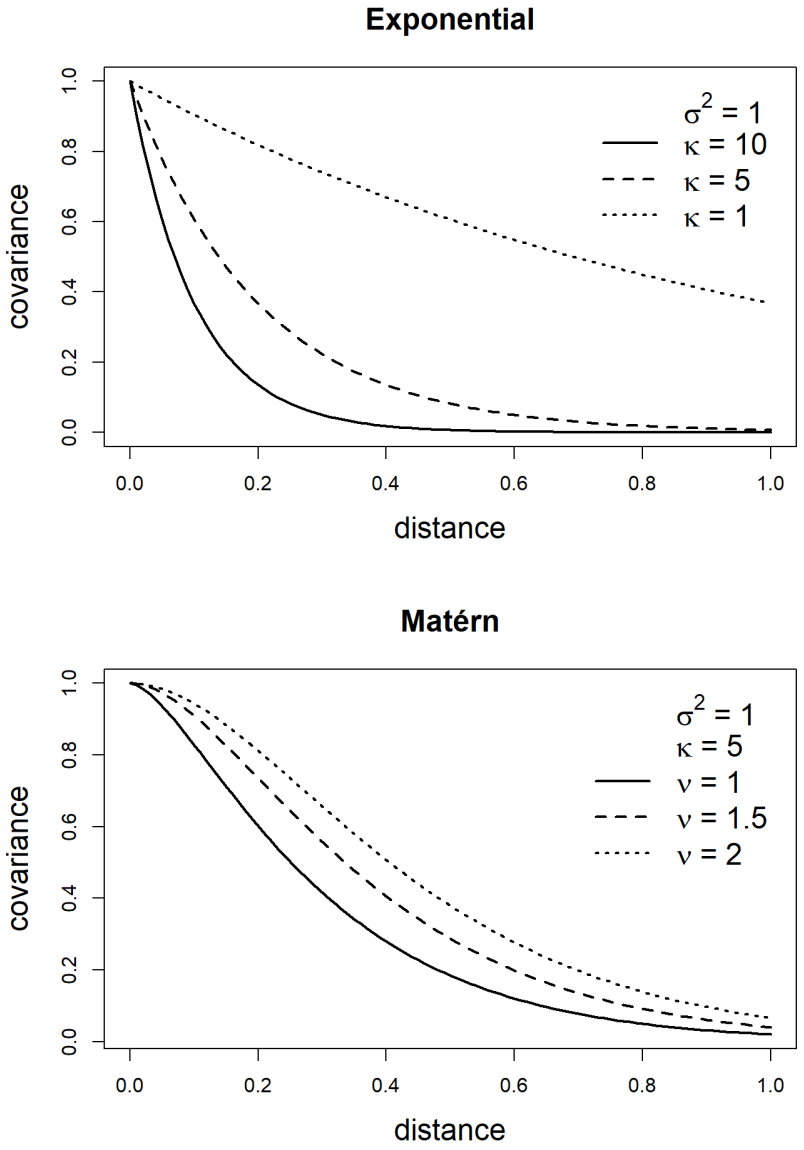

Fig. 3.4: Covariance functions corresponding to exponential and Matérn models.

3.5.4 Gaussian Markov Random Fields

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 $\boldsymbol{Q} > 0$,

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

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

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

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].

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 [RH05, pp. 23–24].

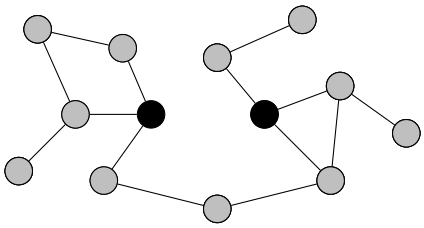

Fig. 3.5: The pairwise Markov property; the black nodes are conditionally independent given the light gray nodes.

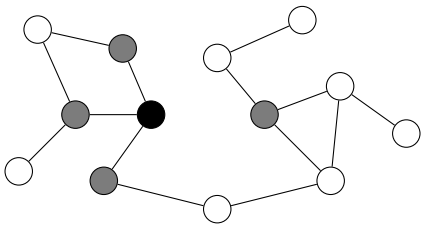

Fig. 3.6: The local Markov property; the black nodes and white nodes are conditionally independent given the dark gray nodes.

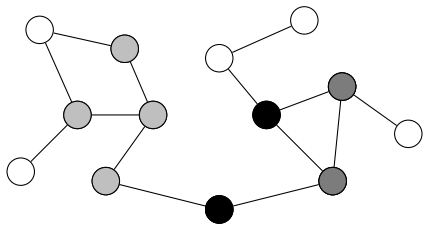

Fig. 3.7: The global Markov property; the dark gray and light gray nodes are globally independent given the black nodes.

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,

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

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 $A \cup \text{ne}(A)$, since $Q_{ij} = 0 \, \forall j \not\in \text{ne}(i)$.
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

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

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

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

Furthermore,

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

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

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

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

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

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].

Specification Through Full Conditionals

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

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

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

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 $\pi(\boldsymbol{x})$ 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

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

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].

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].

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

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

As introduced in section 3.5, $\boldsymbol{y}$ are the observed data, $\boldsymbol{x}$ is a Gaussian field, $\boldsymbol{\theta}$ represents the hyperparameters, while $\mu\left(\boldsymbol{\theta}\right)$ and $\boldsymbol{Q}\left(\boldsymbol{\theta}\right)$ 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 $\mu_i = g^{-1}\left(\eta_i\right)$. 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 $\boldsymbol{x} = \left(\alpha, \{\beta_k\}\,|\theta \sim \mathcal{N}\left(\mu\left(\theta\right),\boldsymbol{Q}\left(\theta\right)^{-1}\right)\right)$ 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

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

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

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

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

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

and the posterior marginal for the hyperparameters can be expressed by

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

$\pi\left(x_i|\boldsymbol{y}\right)$ is approximated by combining analytical approximations to the full conditionals $\pi\left(x_i|\boldsymbol{\theta},\boldsymbol{y}\right)$ and $\pi\left(\boldsymbol{\theta}|\boldsymbol{y}\right)$ and numerical integration routines to integrate out $\boldsymbol{\theta}$. Similarly, $\pi\left(\theta_j|\boldsymbol{y}\right)$ is approximated by approximating $\pi\left(\boldsymbol{\theta}|\boldsymbol{y}\right)$ 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,

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

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

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

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

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 $\widetilde{\pi}(x_i|\theta_k,\boldsymbol{y})$, a Gaussian, Laplace or simplified Laplace approximation can be used. Using a Gaussian approximation derived from $\widetilde{\pi}_G(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})$ 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].

4 Analysis of Geospatial Health Data

40

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.

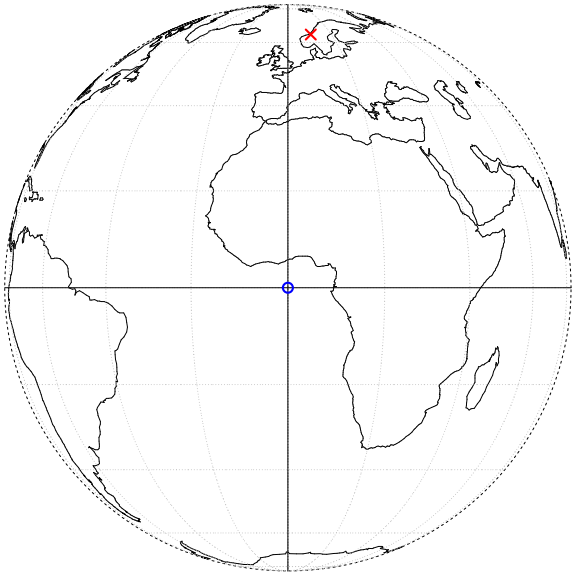

Fig. 4.1: A geographic CRS with an origin at 0° longitude an latitude. The red X denotes the location of Trondheim.

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 the following graphic.

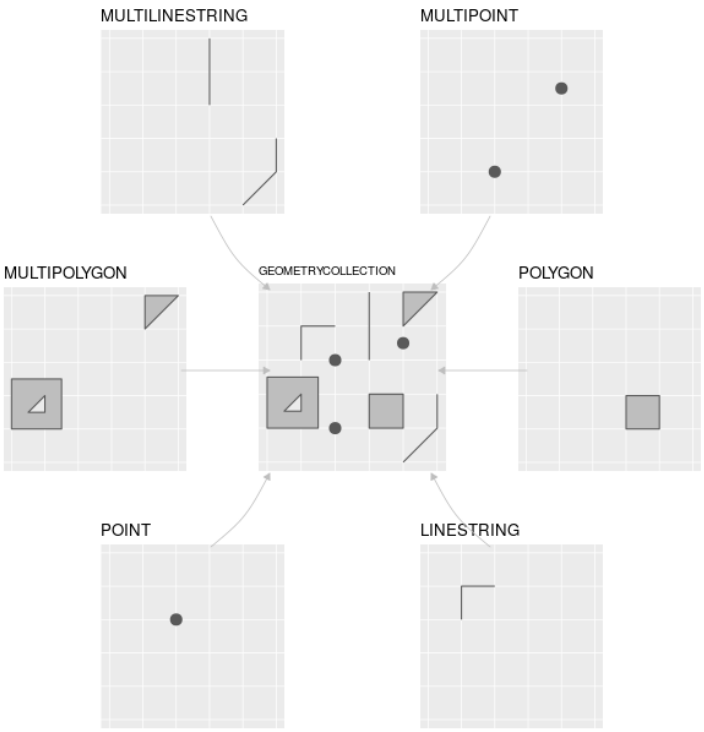

Fig. 4.2: The most commonly used simple feature types.

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 [LNM19].

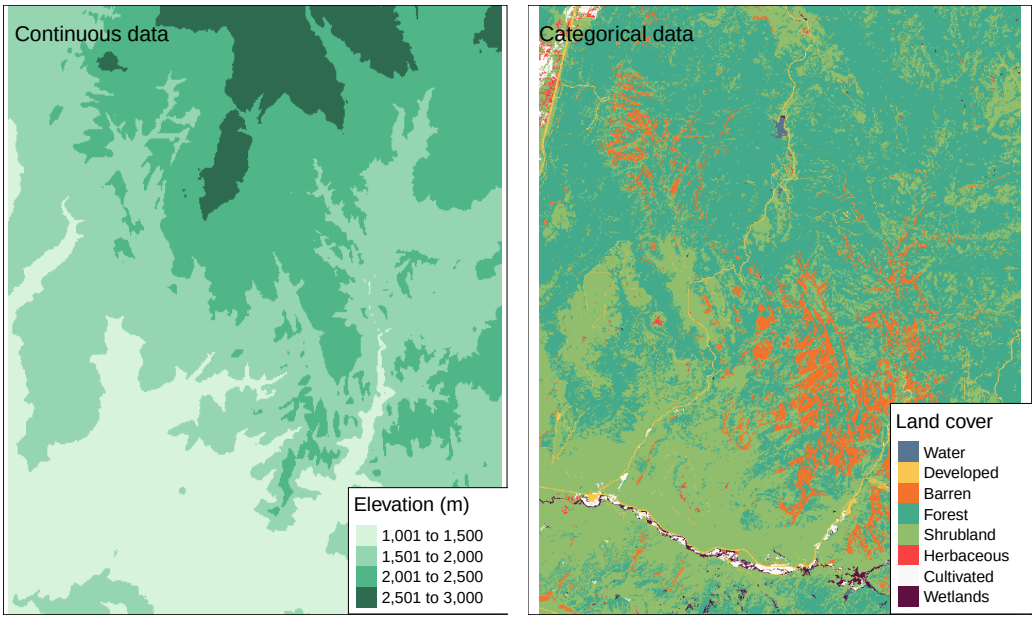

Fig. 4.3: An example of continuous and categorical raster data

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 $Z(\omega)$.
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 $\mu\ inN$ can be written in this particular form.
A point process η on $\mathbb{X}$ is called *proper point process* if random elements $X_1, X_2, ...\ existin\mathbb{X}$ 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 an *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].

4.2.3 Random Measures and Cox Processes

A Poisson process with a random intensity measure, and thus the result of a *doubly stochastic* process, is called a *Cox process*. A random measure is a natural and important generalisation of a point process and it is the determining factor of the distribution of a Cox process.
Since a Cox process η can be interpreted as the result of a doubly stochastic process, a random measure ξ is generated first, followed by a Poisson process with intensity measure ξ.
These processes are often used to simulate spike trains or in financial mathematics for modeling the prices of financial instruments where credit risk is a major factor.

Random Measures

Let $(\mathbb{X}, \mathcal{X})$ be a measurable space and let $M\left(\mathbb{X}\right) \equiv M$ denote the set of all s-finite measures μ on $\mathbb{X}$. Let $\mathcal{M}\left(\mathbb{X}\right) \equiv \mathcal{M}$ be the σ field generated by all sets of the form

$$\left\{\mu \in M : \mu(B) \leq t\right\}, \qquad B \in \mathcal{X}, t \in \mathbb{R}_+.$$

This is equal to the smallest σ-field of subsets of M such that $\mu \mapsto \mu(B)$ is a measurable mapping for all $B \in \mathbb{X}$.
A random measure on $\mathbb{X}$ is defined as a random element ξ of the space $(M, \mathcal{M})$, i.e. a measurable mapping $\xi : \Omega \mapsto M$.
If ξ is a random measure and $B \in \mathcal{X}$, then $\xi(B)$ denotes the random variable $\omega \mapsto \xi\left(\omega, B\right) := \xi\left(\omega\right)\left(B\right)$. This mapping represents a kernel from Ω to $\mathbb{X}$ with the additional property that the measure $\xi\left(\omega, \cdot\right)$ is s-finite for each $\omega \in \Omega$.
A random measure ξ on $\mathbb{X}$ follows the distribution of the probability measure $\mathbb{P}_\xi$ on $(M, \mathcal{M})$ given by $A \mapsto \mathbb{P}\left(\xi \in A\right)$. This distribution is again determined by the family of random vectors $\left(\xi\left(B_1\right), ..., \xi\left(B_m\right)\right)$ for pairwise disjoint $B_1, ..., B_m \in \mathcal{X}$ and $m \in \mathbb{N}$ [LP17, pp. 127–128].

Cox Processes

Let Π_λ denote the distribution of a Poisson process with intensity measure $\lambda \, in M\left(\mathbb{X}\right)$ and let ξ be a random measure on $\mathbb{X}$. A point process η on $\mathbb{X}$ is called a Cox process directed by ξ if

$$\mathbb{P}\left(\eta \in A | \xi\right) = \Pi_\xi(A), \qquad \mathbb{P}\text{-almost surely}, A \in \mathcal{N}. \tag{4.6}$$

$\mathcal{N}\left(\mathbb{X}\right) \equiv \mathcal{N}$ denotes the σ field formed by the set of all subsets of N of the form

$$\left\{\mu \in N : \mu(B) = k\right\}, \qquad B \in \mathcal{X}, k \in \mathbb{N}_0.$$

Thus $\mathcal{N}$ denotes the smallest σ field on N such that $\mu \mapsto \mu(B)$ is measurable for all $B \in \mathcal{X}$ [LP17, p. 129].

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].

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

$$w_{ij} = \begin{cases} 1 & \text{if regions } i \text{ and } j \text{ share a border} \\ 0 & \text{else} \end{cases} \tag{4.7}$$

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

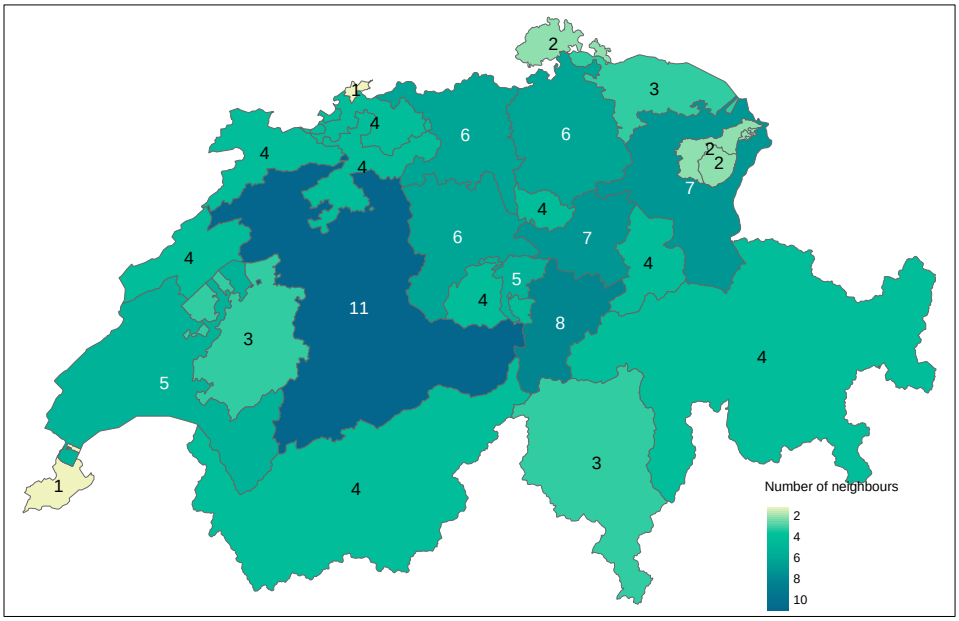

Fig. 4.4: The number of shared borders of cantons in Switzerland

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

$$\text{SIR}_i = \frac{Y_i}{E_i}. \tag{4.8}$$

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

$$E_i = \sum_{j=1}^{m} r_j^{(s)} n_j^{(i)}, \tag{4.9}$$

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].

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:

$$Y_i \sim \text{Po}\left(E_i\theta_i\right), \quad i = 1, ..., n, \tag{4.10}$$

$$\log\left(\theta_i\right) = \alpha + u_i + v_i. \tag{4.11}$$

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,

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

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 $\boldsymbol{d}_j$ $(j = 1, ..., p)$ 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 smoothes the data

according to a given neighbourhood structure that defines two areas as neighbours if they share a common boundary, specifically,

$$u_i|\boldsymbol{u}_{-i} \sim \mathcal{N}\left(\overline{u}_{\delta_i}, \frac{\sigma_u^2}{n_{\delta_i}}\right), \tag{4.12}$$

where $\overline{u}_{\delta_i}^{-1} = n_{\delta_i}^{-1} \sum_{j \in \delta_i} u_j$, while δ_i and n_{δ_i} represent the set and the amount of neighbours of the area i, respectively. The unstructured component v_i is modeled as an independent and identically distributed (i.i.d.) normal variable with zero mean and variance σ_v^2.
In 2017, Simpson et al. proposed BYM2 [Sim+17], a new parameterisation of the BYM model that yields interpretable parameters and facilitates the assignment of meaningful penalised complexity priors. It uses a scaled, spatially structured component $\boldsymbol{u_*}$ and an unstructured component $\boldsymbol{v_*}$,

$$\boldsymbol{b} = \frac{1}{\sqrt{\tau_b}}\left(\sqrt{1-\phi}\boldsymbol{v_*} + \sqrt{\phi}\boldsymbol{u_*}\right). \tag{4.13}$$

The precision parameter $\tau_b > 0$ controls the marginal variance contribution of the weighted sum of $\boldsymbol{u_*}$ and $\boldsymbol{v_*}$. The mixing parameter $0 \leq \phi \leq 1$ captures the proportion of the marginal variance explained by the structured effect $\boldsymbol{u}_*$. Therefore, the BYM2 model is equal to a pure spatial model for $\phi = 1$ and equal to unstructured spatial noise for $\phi = 0$ [Rie+16]. To define the prior for the marginal accuracy τ_b, the following probability statement is used:

$$\begin{aligned} &\mathbb{P}\left(\frac{1}{\sqrt{\tau_b}} > U\right) = \alpha \\ \iff &\mathbb{P}\left(\phi < U\right) = \alpha \end{aligned} \tag{4.14}$$

[Mor19].

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

$$Y_{ij} \sim \text{Po}\left(E_{ij}, \theta_{ij}\right), \qquad i = 1, ..., I,\ j = 1, ..., J. \tag{4.15}$$

$\log(\theta_{ij})$ 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

$$\log(\theta_{ij}) = \alpha + u_i + v_i + (\beta + \delta_i) \times t_j. \tag{4.16}$$

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 $\delta_i \neq is0$, 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:

$$\log(\theta_{ij}) = \alpha + u_i + v_i + \gamma_j + \phi_j + \delta_{ij}. \tag{4.17}$$

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)

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

or second order random walk in time (RW2)

$$\gamma_j | \gamma_{j-1}, \gamma_{j-2} \sim \mathcal{N}\left(2\gamma_{j-1} - \gamma_{j-2}, \sigma_\gamma^2\right). \tag{4.19}$$

The unstructured temporal effect is given by $\phi_j, \phi_j \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma_\phi^2\right)$. 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 $\delta_{ij} \sim \mathcal{N}\left(0, \sigma_{\delta}^{2}\right)$ [Mor19].

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

$$\left\{ Z(s) : s \in D \subset \mathbb{R}^2 \right\}, \tag{4.20}$$

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.

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]:

$$\left(\kappa^2 - \Delta\right)^{\alpha/2} (\tau x(\boldsymbol{s})) = \mathcal{W}(\boldsymbol{s}). \tag{4.21}$$

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 otherwise. 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

$$x(\boldsymbol{s}) = \sum_{k=1}^{n} \psi_k(\boldsymbol{s}) x_k, \tag{4.22}$$

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 [Mor19].

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.

58

5.1 Covid-19 Data

Covid-19 Data for Norway

The Covid-19 data for Norway comes from a dataset made available to the public via the website Github.com. 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

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.

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.

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, as shown in Table 5.3.

Tab. 5.3: An excerpt from the age data for Norway.

region	sex	age	Persons 2020
K-1103 Stavanger	Males	0 years	842
K-1103 Stavanger	Females	0 years	760
K-1103 Stavanger	Males	1 year	816
K-1103 Stavanger	Females	1 year	807

Next, unemployment data were collected for a given municipality, as shown in Table 5.4.

Tab. 5.4: An excerpt from the unemployment data for Norway.

region	country background	unemployment_perc
1103 Stavanger	Total	3.1
1103 Stavanger	Immigrants	6.0
1507 Ålesund	Total	2.3
1507 Ålesund	Immigrants	5.2

Other data collected include data related to the number of workers in a particular industry and immigration data, as shown in Table 5.5 and Table 5.6, respectively.

Tab. 5.5: An excerpt from the worker industry data for Norway. Does not contain all variables. In the actual dataset, a distinction is made between employees by place of residence and employees by place of work. Here, the number of employees by place of residence are displayed.

region	age	industry	working hours	employees
1103 Stavanger	15-74 years	00-99 All industries	Part-time	17960
1103 Stavanger	20-66 years	00-99 All industries	Full-time	51231
1103 Stavanger	15-74 years	41-43 Construction	Part-time	352
1103 Stavanger	20-66 years	41-43 Construction	Full-time	2772

Tab. 5.6: An excerpt from the immigration data for Norway.

region	category	background	Persons 2020	%population
1103 Stavanger	Immigrants	All	27075	18.9
1103 Stavanger	Norwegian with immigrant parents	All	5972	4.16
1507 Ålesund	Immigrants	All	9066	13.7
1507 Ålesund	Norwegian with immigrant parents	All	1434	2.16

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, as shown in Table 5.7.

Tab. 5.7: An excerpt from the unemployment data for Germany. Does not contain all variables.

District	City	unemployed_total	unemployed_foreigners
2000	Hamburg, City	64774	21994
8221	Heidelberg, City	3055	964
9162	München, City	30557	13487
9187	Rosenheim, District	3155	679

Next, the 2019 European election data were collected, as shown in Table5.8.

Tab. 5.8: An excerpt from the 2019 European election data for Germany. Does not contain all variables.

District	City	voter turnout	Union	SPD	AfD
2000	Hamburg, City	61.9	140966	157840	51649
8221	Heidelberg, City	70.1	13007	10378	4303
9162	München, City	65.4	163350	69403	36282
9187	Rosenheim, District	63.1	53763	8768	10275

Data were also gathered in relation to protection-seekers, as shown in Table 5.9, social welfare, as seen in Table 5.10, and in relation to asylum-seeker benefits, as shown in Table 5.11.

Tab. 5.9: An excerpt from the protection-seekers data for Germany.

District	City	protection_seekers
2000	Hamburg, City	52730
8221	Heidelberg, City	2570
9162	München, City	35105
9187	Rosenheim, District	2730

Tab. 5.10: An excerpt from the social welfare data for Germany. Does not contain all variables.

District	City	welfare_recipients
2000	Hamburg, City	10076
8221	Heidelberg, City	292
9162	München, City	5230
9187	Rosenheim, District	1338

Tab. 5.11: An excerpt from the asylum-seeker benefits data for Germany. Does not contain all variables.

District	City	asylum_seeker_benefits
2000	Hamburg, City	9665
8221	Heidelberg, City	1326
9162	München, City	4887
9187	Rosenheim, District	1132

Finally, trade tax, income tax, and payroll tax data were collected, as shown in Table 5.12.

Tab. 5.12: An excerpt from the tax data for Germany. Does not contain all variables. Income total is the sum of the income tax and the payroll tax for a given city/district.

District	City	trade_tax	income_total
2000	Hamburg, City	472,517,476	42,524,847
8221	Heidelberg, City	23,285,772	3,465,746
9162	München, City	489,476,177	45,844,191
9187	Rosenheim, District	33,758,111	5,891,415

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.

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.

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.

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 5.13.

Tab. 5.13: 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, seen in Table 5.3, 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. This was done for men, women, and both groups combined, as seen in Table 5.14.

Tab. 5.14: An excerpt from the median age data for Norway. Does not contain all variables.

region	median_age_f	median_age_m	median_age
K-1103 Stavanger	37	38	37
K-1507 Ålesund	38	40	39
K-4601 Bergen	36	38	37
K-5001 Trondheim	35	37	36

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. A small extract is shown in Table 5.15.

Tab. 5.15: An excerpt from the merged data. Does not contain all variables.

kommune_no	kommune_name	median_age	unemp_immg	immg_total
1103	Stavanger	37	6.0	18.86
1507	Ålesund	39	5.2	13.68
4601	Bergen	37	7.5	15.18
5001	Trondheim	36	4.8	13.64

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 municipality, this data was then merged with the data shown in Table 5.15 and can be seen in Table 5.16.

Tab. 5.16: An excerpt from the merged data. Does not contain all variables.

kommune_no	kommune_name	median_age	schools	restaurants
1103	Stavanger	37	85	78
1507	Ålesund	39	0	0
4601	Bergen	37	152	225
5001	Trondheim	36	104	153

For each variable representing an absolute number, e.g. the number of schools or the number of employees, this number was calculated per 1000 inhabitants of the respective area.
If there were missing values in the covariates, these values were imputed using the median of the respective variable.
Finally, four new variables were created:

1. higher_educ, which counts the number of universities and colleges in a given area.
2. sex, which gives the proportion of females living in a given area.
3. pop_dens, i.e. the number of people per square kilometer in a given area.
3. urb_dens, i.e. the number of residential buildings per square kilometer in a given area.

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 numbers per 1000 inhabitants were calculated for the variables containing absolute numbers. 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 four new variables were created. An example of this final dataset is shown in Table 5.17.

Tab. 5.17: An excerpt from the merged data. Does not contain all variables.

District	City	trade_tax	AfD	shops
2000	Hamburg, City	256638.5	0.0648	0.544
8221	Heidelberg, City	145213.9	0.0616	0.599
9162	München, City	332635.8	0.0598	0.582
9187	Rosenheim, District	129349.8	0.0834	0.682

6 Data Analysis

In this chapter the models calculated 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 regression models are discussed.

6.1 Standardised Incidence Ratio

This section takes a brief look at the standardised incidence ratio for the countries of interest.

6.1.1 Standardised Incidence Ratio for Germany

When looking at the standardised incidence ratio for Germany, 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 standardised incidence ratio compared to the rest of Germany, excluding Saxony, see Figure 6.1

69

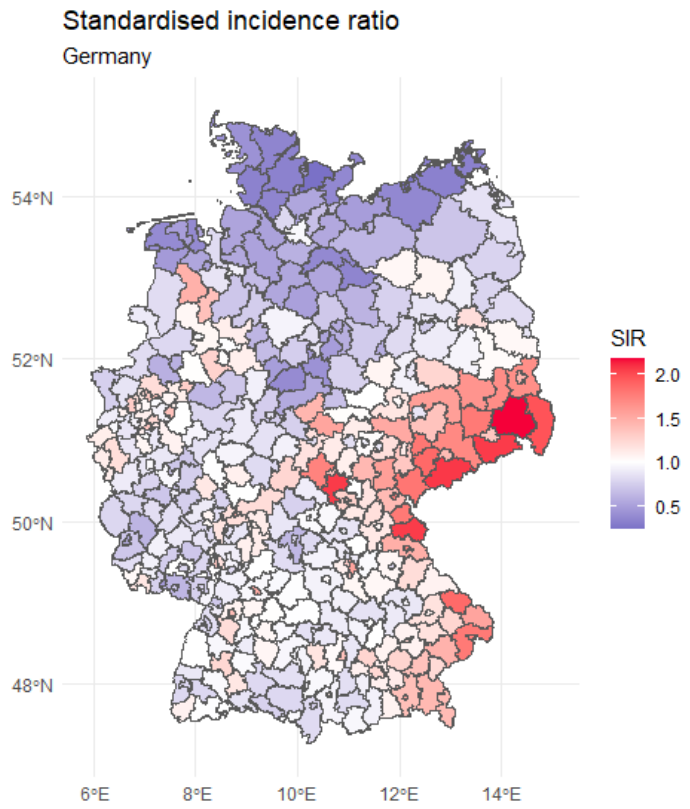

Fig. 6.1: The standardised incidence ratio for Germany based on the data of the 5th of March 2021

6.1.2 Standardised Incidence Ratio for Norway

Looking at the standardised incidence rate for Norway, 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. The SIR in Hyllestad is around 5, following an outbreak in a shipyard in autumn 2020 [Kor20], while Ulvik has a ratio of around 9, 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]. See Figure 6.2 for more information.

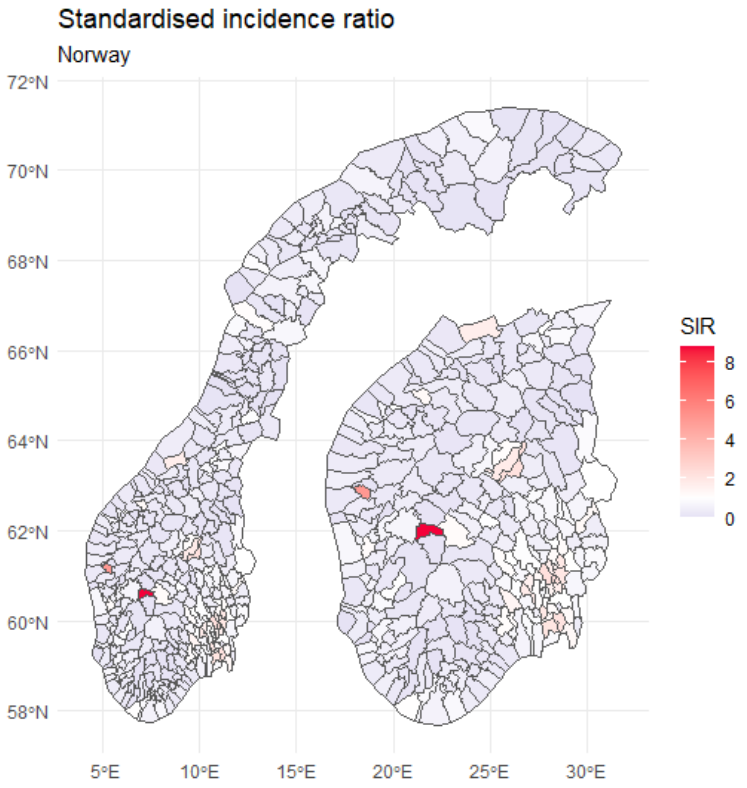

Fig. 6.2: The standardised incidence ratio for Norway based on the data of the 5th of March 2021

6.2 Spatial Models

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 other populations. Three different types of models are used for each country:

1. The Besarg-Yollie-Mollie Model
2. Besags Proper Spatial Model
3. The Leroux-Model

All of these models were computed using the INLA [BGR15] R package.
Before the models are computed, however, the distribution that fits the number of cases must first be found. For this, the function `descdist()` from the `fitdistrplus` R package is used. With this function, descriptive parameters of an empirical distribution can be calculated and a skewness-kurtosis plot is provided. The plots for Germany and Norway can be seen in Figure 6.3 and Figure 6.4.

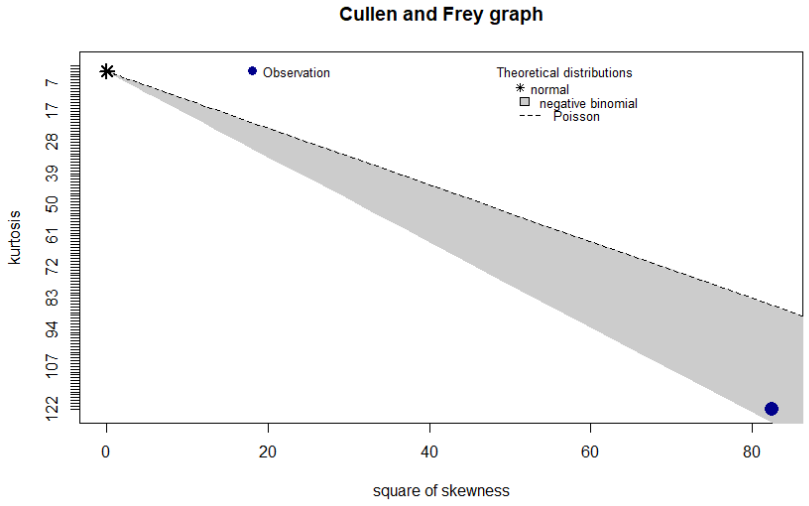

Fig. 6.3: The Cullen and Frey graph for Germany

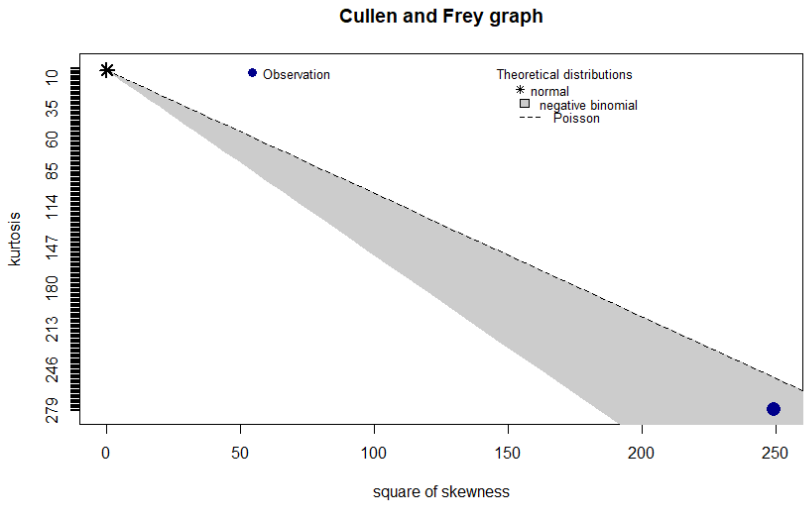

Fig. 6.4: The Cullen and Frey graph for Norway

Following these plots, a negative binomial distribution is fitted to the data using the maximum likelihood method. The function `fitdist()` is used for this. The fits can be seen in Figure 6.5 and Figure 6.6.

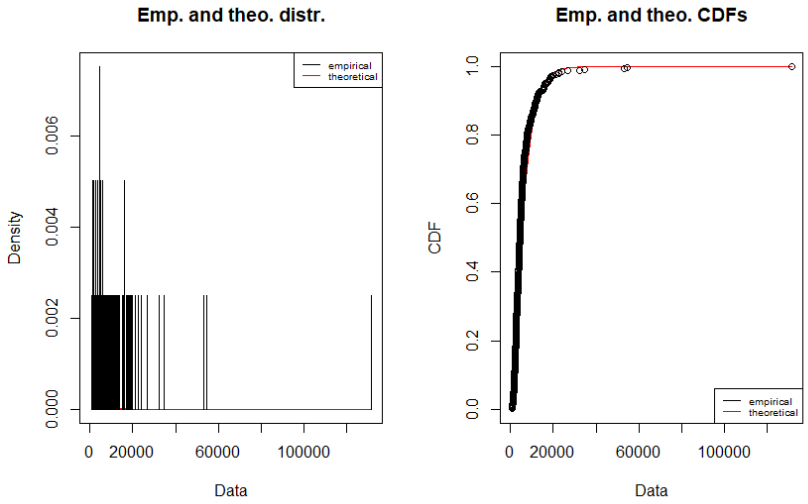

Fig. 6.5: A negative binomial fit to the number of cases in German municipalities

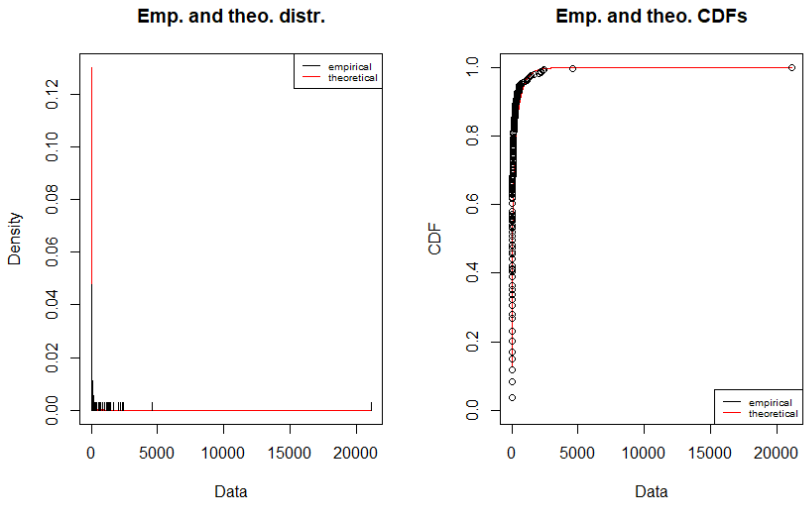

Fig. 6.6: A negative binomial fit to the number of cases in Norwegian municipalities

Lastly, the AIC was 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 was 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	8364
Germany	Poisson	2065103
Germany	Negative Binomial	7726
Norway	Normal	6052
Norway	Poisson	314100
Norway	Negative Binomial	4018

To specify each type of model, the following code shown in Listing 7.1 can be used. Three measures are used to compare the models, the DIC, the WAIC and the CPO. For all countries, the models were computed with

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

    3.1 Without variable selection

3.2 With variable selection

For each model type, different values for the penalised prior component were tried. As this resulted in a large number of models, only the model with the best performance for each model class is examined in more detail. To decide which models perform best, the DIC, WAIC and CPO are used. To select the best model, all calculated models were first ranked according to their DIC, their WAIC and their CPO before the model with the lowest average rank was selected as the best model. Finally, due to the amount of covariates, forwards and backwards stepwise variable selection was performed with the intention of obtaining a model that fits the data well and at the same time is relatively easy to interpret. This can be done with the R package `INLAutils` [Red+17], as shown in Listing 7.2.
A list of all calculated models along with their performance measures is provided in the appendix.

6.2.1 Spatial Models for Germany

First, a look is taken at the spatial models calculated for Germany. These models are based on data from 10 March 2021, when 2534502 people in Germany were confirmed infected with Covid-19. The five municipalities with the most infections are shown in Table 6.2.

Tab. 6.2: The municipalities with the most infections as of March 10th 2021.

Municipality	Population	Number of infections
SK Berlin	3644826	133158
SK Munich	1471508	55263
SK Hamburg	1841179	54233
SK Cologne	1085664	35034
Region Hannover	1157624	33344

Demographic Models

The model with the best performance based on the demographic variables was a Leroux model calculated with the following formula:

```
formula_9_leroux <- CumNumberTestedIll ~
  trade_tax + income_total +
  f(
    idarea_1, model = "generic1",
    Cmatrix = C, hyper = prior_1
```

```
    )
```

Listing 6.1: The formula for the best Leroux model based on the demographic variables

The performance measures of this model and the best-performing BYM2 and Besag proper models are shown in Table 6.3. A total of 66 models were calculated. The Leroux model had the best DIC and the fourth best CPO. Of all calculated models, the 20 best models in terms of the CPO were all Leroux models. The BYM2 model showed good performance in terms of the DIC and WAIC while Besags proper spatial model only showed good performance in terms of the WAIC.

Tab. 6.3: The performance measures for the best performing demographic model of each type. The rank in relation to each performance measure is shown in the brackets.

Model	DIC	WAIC	CPO
Besags Proper	5696 (15)	5603 (2)	-3358 (33)
BYM2	5639 (4)	5578 (1)	-3375 (28)
Leroux	5611 (1)	5682 (18)	-3872 (2)

The summary of the fixed effects is shown in Table 6.4. To compute the posterior mean of the coefficients, the following code shown in Listing 6.2

```
inla.emarginal(exp, model_9_leroux$marginals.fixed$trade_tax)
# calculate the increase in risk if the
# trade tax increase by 25000
inla.emarginal(
  exp,
  model_9_leroux$marginals.fixed$trade_tax
  ) ^ 25000
```

Listing 6.2: Calculating the posterior mean of a coefficent.

To obtain a credibility interval of the fixed effects on the transformed scale, the code in Listing 6.3 can be used.

```
inla.qmarginal(
  c(0.025,0.975),
  inla.tmarginal(
    exp,
    model_9_leroux$marginals.fixed$trade_tax
  )
)
```

Listing 6.3: Extracting the credibility interval for a coefficient

Tab. 6.4: The fixed effects for model 10. Values are rounded.

Variable	Mean	exp(mean$_p$)	exp(q0025$_p$)	exp(q0975$_p$)
(Intercept)	0.1483	1.1658	0.9604	1.4012
trade_tax	0.0000007026	1.000	1.000	1.000
income_total	-0.00001633	1.000	1.000	1.000

The posterior mean of the exponentiated intercept implies a 16.58% risk rate across Germany with a credibility interval ranging from -3.96% to 40.12%.
From this summary, it may seem difficult to interpret the impact of the trade tax and total income on the risk of contracting Covid-19, but it must be borne in mind that the difference in these variables is usually far greater than 1.
The range of the trade tax per 1000 inhabitants is from around 45000€ to 775000€. An increase in the trade tax by 25000€ per 1000 inhabitants leads to an increase in the risk of contracting Covid-19 by 1.77%, given that the total income remains the same.
The range of total income, defined as the sum of income tax and payroll tax, per 1000 inhabitants ranges from 12500€ to about 36000€. An increase of 1000€ per 1000 inhabitants, with no change in trade tax, reduces the risk of contagion by about 1.62%.
Since trade tax must be paid by all companies operating in Germany, this tax is highest in cities, as many companies have their headquarters there. This could therefore be an indication that the risk of infection is higher in cities.
On the other hand, the risk of infection decreases in regions where there is a higher total income based on payroll and income tax, i.e. where people earn a higher salary. Traditionally, people in western Germany have a higher income compared to eastern Germany, which could explain why eastern Germany, especially Saxony, is more affected by Covid-19 than some of the western parts of Germany.

Infrastructure Models

When it comes to the models based on the infrastructural variables, a BYM2 model performed the best. It was computed using the following formula:

```
formula_26_bym2 <- CumNumberTestedIll ~
  marketplace + entertainment + sport + clinic +
  hairdresser + shops + place_of_worship + retail + nursing_home +
  restaurant + aerodrome + office + platform + schools +
  higher_education + kindergarten + bakeries +
  f(
```

```
    idarea_1, model = "bym2", graph = g,
    scale.model = TRUE, hyper = prior_2
  )
```

Listing 6.4: The formula for the best BYM2 model based on the infrastructural variables

The performance measures of the computed models are shown in Table 6.5. In total, 24 models were calculated.

Tab. 6.5: The performance measures for the best performing demographic model of each type. The rank in relation to each performance measure is shown in the brackets.

Model	DIC	WAIC	CPO
Besags Proper	5737 (12)	5653 (3)	-3387 (13)
BYM2	5680 (1)	5647 (2)	-3365 (20)
Leroux	5699 (7)	5681 (10)	-3568 (8)

Again, the models with the best CPO were all Leroux models.
The effect of the covariates are shown in Table 6.6.

Tab. 6.6: The fixed effects for the BYM2 model. Values are rounded.

Variable	Mean	exp(mean$_p$)	exp(q0025$_p$)	exp(q0975$_p$)
(Intercept)	0.09038	1.100	0.9023	1.327
marketplace	0.9468	3.293	0.6496	10.09
entertainment	0.3053	1.454	0.6566	2.798
hairdresser	0.1362	1.154	0.9034	1.453
shops	0.04251	1.058	0.7512	1.448
bakeries	0.01805	1.031	0.7482	1.385
place_of_worship	-0.007848	0.9925	0.9445	1.042
platform	-0.01340	0.9867	0.9768	0.9966
nursing_home	-0.02752	1.057	0.4365	2.160
schools	-0.02930	0.9764	0.7926	1.190
kindergarten	-0.03017	0.9772	0.7689	1.224
aerodrome	-0.03255	1.2162	0.2559	3.615
retail	-0.03583	0.9661	0.8700	1.070
restaurant	-0.04270	0.9587	0.8968	1.024
higher_education	-0.07287	0.9363	0.7372	1.172
office	-0.08885	0.9172	0.7993	1.047
clinic	-0.09167	0.9135	0.8288	1.004
sport	-0.1211	0.8883	0.7693	1.020

Some findings from these results are that the average risk ratio in Germany is 10%. Marketplaces increase the risk by over 200%, however, there is a maximum

of only 0.1265 marketplaces per 1000 inhabitants, resulting in such a high value. If the number of marketplaces increases by 0.1 per 1000 inhabitants, the risk of infection increases by 12.7%. More common sites that increase the risk of infection include bakeries (between 0.1169 and 0.9643 per 1000 inhabitants), hairdressers (between 0.09655 and 1.191 per 1000 inhabitants) and shops (between 0.3710 and 1.129 per 1000 inhabitants). An increase of 0.1 per 1000 for each of these types of establishments results in a risk increase of 0.3%, 1.4% and 0.6%, respectively. It should not be surprising, in particular, that shops increase the risk of infection, as people tend to congregate there and people living in areas where a supermarket is nearby may prefer to go shopping more often in favour of fresh produce, rather than just once a week when a supermarket is further away.
There are also effects that have a posterior mean of less than 1, suggesting that they reduce the risk of infection. However, this is not necessarily the case, as it is probably safer for a person to stay at home rather than go to a restaurant, for example.

Tab. 6.7: How different variables affect the relative risk. Does not contain all variables.

Variable	Scale	Increase	Risk increase
marketplace	per 1000 inhabitants	0.1	12.7%
entertainment	per 1000 inhabitants	0.1	3.8%
hairdresser	per 1000 inhabitants	0.1	1.4%
shops	per 1000 inhabitants	0.1	0.6%
bakeries	per 1000 inhabitants	0.1	0.3%
nursing_home	per 1000 inhabitants	0.1	0.5%
aerodrome	per 1000 inhabitants	0.1	2.0%

Combined Models

Finally, models are considered that include both the infrastructural and demographic covariates. Due to the amount of variables, all models run are based on variable selection. Since forwards and backwards variable selection was performed, a total of 12 models were run. The models with the best performance were a Leroux model that used forwards variable selection and a BYM2 model that used backwards variable selection. The difference between these types of variable selection can be seen in the Listing 6.5, as the Leroux model contains fewer variables.

```
formula_31_bym2 <- CumNumberTestedIll ~
  asylum_seeker_benefits + trade_tax +
  total_income + income_tax + Union + SPD +
  FDP + die_linke + AfD + other_parties + protection_seekers +
  welfare_recipients +  unemployed_total + unemployed_foreigners +
  entertainment +  sport + clinic + hairdresser + shops +
```

```
  nursing_home + restaurant + aerodrome + platform + kindergarten +
  schools + bakeries +  pop_dens + urb_dens + sex +
  f(
    idarea_1, model = "bym2", graph = g,
    scale.model = TRUE, hyper = prior_1
  )
formula_34_leroux <- CumNumberTestedIll ~
  pop_dens + other_parties + SPD +
  AfD + entertainment + unemployed_foreigners + unemployed_total +
  welfare_recipients + schools + clinic +
  f(
    idarea_1, model = "generic1",
    Cmatrix = C, hyper = prior_2
  )
```

Listing 6.5: The formulas for the best models containing all variables.

The performance measures of the computes models are shown in Table 6.8.

Tab. 6.8: The performance measures for the best performing demographic model of each type. The rank in relation to each performance measure is shown in the brackets.

Model	DIC	WAIC	CPO
Besags Proper	6083 (9)	6078 (9)	-3300 (11)
BYM2	5798 (3)	5774 (3)	-3347 (5)
Leroux	5815 (4)	5827 (5)	-3551 (2)

A look is only taken at the covariates of the Leroux model, as it has a lower CPO and is therefore better at predicting. The covariates of the BYM2 model are shown in Table 7.1 in the Appendix. For the Leroux model, the covariates are shown in Table 6.9.

Tab. 6.9: The fixed effects for model 34. Values are rounded.

Variable	Mean	exp(mean$_p$)	exp(q0025$_p$)	exp(q0975$_p$)
(Intercept)	-0.244	0.7857	0.6712	0.9137
other_parties	29.01	1.176e+51	-2.453e+25	-4.135e+20
AfD	4.115	65.03	31.16	120.2
schools	0.1743	1.198	0.9504	1.491
unemployed_ foreigners	0.0611	1.063	1.049	1.077
pop_dens	0.00003913	1.000	1.000	1.000
clinic	-0.002106	0.9990	0.9118	1.092
welfare_ recipients	-0.01726	0.9829	0.9676	0.9984
unemployed_ total	-0.01744	0.9827	0.9775	0.9880
entertainment	-0.6267	0.5732	0.2569	1.110
SPD	-1.666	0.1981	0.1042	0.3426

What is immediately noticeable here is the extremely high coefficient for *other_parties*. However, one should bear in mind that in the 2019 European election the highest combined percentage of all other political parties was 0.4%. Therefore, an increase of 0.005% only increases the risk by 0.59%. One of the larger political parties in Germany, the AfD, is known for its opposition to the corona rules in Germany. If the proportion of votes for this party increases by 1% in a given area, the risk of getting Covid-19 increases by 4.26%. Since the people who voted for them tend not to wear masks or keep a safe distance from others, it is not surprising that the infection figures are also higher in Bavaria and Saxony, both federal states where the AfD receives a higher proportion of the vote. Another positive effect worth mentioning is the number of schools per 1000 inhabitants. If these increase by 0.1, then the risk of infection increases by 1.8%. If the number of unemployed foreigners per 1000 inhabitants increases by 1, then the risk of infection increases by 6.3%. Finally, if the number of people per square kilometre increases by a value of 250, the risk of infection increases by 1.0%.

Finally, a look is taken at the posterior means $\boldsymbol{\zeta} = \exp(\xi)$ of the area-specific risks. Since the uncertainty associated with them can also be mapped, the posterior probability $\mathbb{P}(\zeta_i > 1|y)$ is visualised as well. In Figure 6.7 it can be seen that the relative risk is higher in the southern parts of Bavaria, as well as in some parts of Saxony, Lower Saxony, North Rhine-Westphalia and Brandenburg. The posterior probability is highest in large parts of Bavaria, Saxony, North Rhine-Westphalia, Hesse and Lower Saxony. Both risks are low in the northern parts of Germany and in Baden-Württemberg, which is located in the south, west of Bavaria.

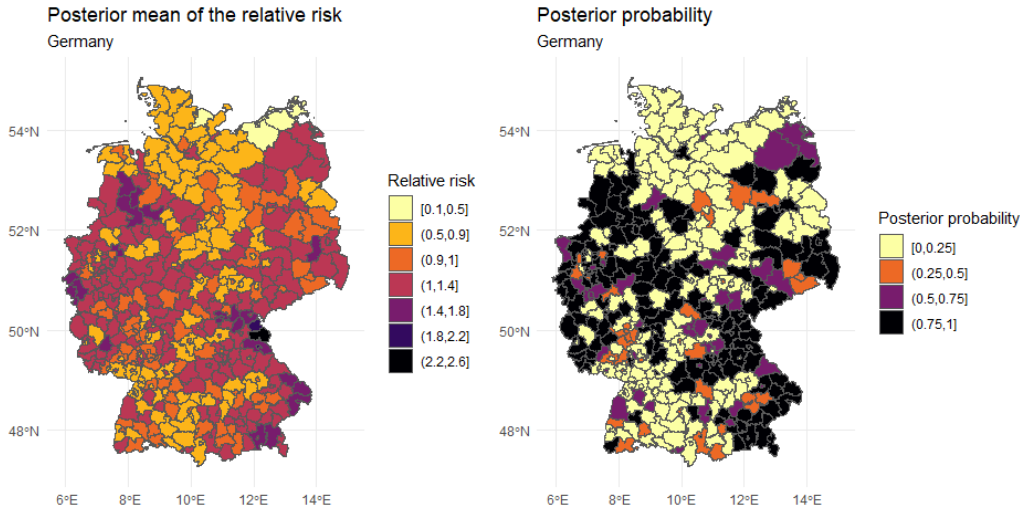

Fig. 6.7: Posterior mean of the area-specific risk and the posterior probabilit.

6.2.2 Spatial Models for Norway

Next, the same types of models are evaluated for Norway. These models are based on data from 12 March 2021, when 76577 people in Norway were confirmed infected with Covid-19. The five municipalities with the most infections are shown in Table 6.10.

Tab. 6.10: The municipalities with the most infections as of March 12th 2021.

Municipality	Population	Number of infections
Oslo	693494	22248
Bergen	283929	4618
Drammen	101386	2549
Lillestrøm	85983	2390
Fredrikstad	82385	2373

Demographic Models

As with Germany, the best demographic model was again a Leroux model. Only this time, the covariates of the model were selected using forwards variable selection. The following formula was used for the model:

```
formula_22_leroux <- value ~
  pop_dens + immigrants_pure + median_age +
  sex +
  f(
    idarea_1, model = "generic1",
    Cmatrix = C, hyper = prior_2
  )
```

Listing 6.6: The formula for the best Leroux model based on the demographic variables

The performance measures for all three types of models are shown in Table 6.11.

Tab. 6.11: The performance measures for the best performing demographic model of each type. The rank in relation to each performance measure is shown in the brackets.

Model	DIC	WAIC	CPO
Besags Proper	2804 (45)	2820 (45)	-10433 (23)
BYM2	2716 (3)	2670 (14)	-9786 (25)
Leroux	2703 (1)	2648 (1)	-11077 (14)

The Leroux model had the best performance in terms of both DIC and WAIC and also performed well in terms of CPO. Just like for Germany, a total of 66 models were calculated. Of the 14 best models in terms of CPO, 13 were Leroux models. The effects of the covariates are shown in Table 6.12.

Tab. 6.12: The fixed effects for model 22. Values are rounded.

Variable	Mean	exp(mean$_p$)	exp(q0025$_p$)	exp(q0975$_p$)
(Intercept)	1.087	222.4	-8.203e+09	-9.140e+04
pop_dens	0.001775	1.002	1.001	1.003
immigrants_pure	-0.03358	0.9671	0.9392	0.9948
median_age	-0.03628	0.9645	0.9396	0.9897
sex	-0.0426	2.902e+06	-6.904e+26	-1.164e+22

There is little meaningful interpretation for the posterior mean of the intercept in this model due to its high value. Of note is that if the population density increases by 100 persons per square kilometre, the risk of infection increases by about 19.4%. Furthermore, if the proportion of women in an area increases by 0.05, the risk of infection increases by around 7.7%. Finally, if the average age in a community increases by 1 year, the risk of infection decreases by roughly 3.5%. Because younger people tend to be more mobile, they come into contact with more people than older

generations, so it makes sense that people living in younger areas have a higher risk of becoming infected with Covid-19.

Infrastructure Models

Looking at the 24 models computed for Norway that are based on the infrastructural variables, once again a Leroux model based on forwards variable selection had the best performance. The formula is shown in Listing 6.7.

```
formula_30_leroux <- value ~
  pop_dens + shops + retail + place_of_worship +
  schools + nursing_home + kindergarten +
  f(
    idarea_1, model = "generic1",
    Cmatrix = C, hyper = prior_2
  )
```

Listing 6.7: The formula for the best Leroux model based on the infrastructural variables

The performance measure for the models are shown in Table 6.13.

Tab. 6.13: The performance measures for the best performing demographic model of each type. The rank in relation to each performance measure is shown in the brackets.

Model	DIC	WAIC	CPO
Besags Proper	2816 (17)	2813 (17)	-10347 (8)
BYM2	2728 (2)	2671 (2)	-9145 (14)
Leroux	2718 (1)	2665 (1)	-9553 (11)

The effect of the covariates are shown in Table 6.14.

Tab. 6.14: The fixed effects for model 30. Values are rounded.

Variable	Mean	exp(mean$_p$)	exp(q0025$_p$)	exp(q0975$_p$)
(Intercept)	-0.7263	0.4858	0.4024	0.5816
nursing_home	0.2017	1.270	0.7174	2.081
retail	0.1541	1.169	1.023	1.329
kindergarten	0.07385	1.082	0.8940	1.295
pop_dens	0.001853	1.002	1.001	1.003
schools	-0.1022	0.9065	0.7564	1.076
place_of_worship	-0.1098	0.8987	0.7710	1.040
shops	-0.1509	0.8633	0.7228	1.022

The posterior mean of the exponential intercept implies a risk rate of -51.42% across Norway, with a credibility interval ranging from -59.76% to -41.84%. However, the risk rate increases if the number of nursing homes per 1000 inhabitants (between 0 and 1.851), retail shops per 1000 inhabitants (between 0 and 10.72), kindergartens per 1000 inhabitants (between 0 and 5.051) or the population density (between 0.1479 and 1437) increases. A 0.1 increase in nursing homes increases the risk by 2.4%, an increase of 0.1 retail shops per 1000 inhabitants increases the risk by 1.6% and an increase of 0.1 kindergartens per 1000 inhabitants increases the risk by 0.8%. Finally, if the population density increases by 50 persons per square kilometre, the risk of infection increases by 20.4%. Again, there are covariates that negatively affect the risk rate, but as stated earlier, it is probably safer to stay at home than to attend these facilities.

Combined Models

Looking at the models with both types of variables for Norway, a BYM2 model that used backwards variable selection performed the best. Its formula can be seen in Listing 6.8

```
formula_31_bym2 <- value ~
  median_age + unemp_tot + unemp_immg + workers_ft_com +
  workers_pt_com + mining_ft_com + mining_pt_com +
  construction_pt_com + immigrants_total + immigrants_norge +
  immigrants_pure + marketplace + entertainment + clinic +
  hairdresser + shops + place_of_worship + retail +
  nursing_home + aerodrome + platform + kindergarten +
  schools + bakeries + higher_education + pop_dens + urb_dens +
  f(
    idarea_1, model = "BYM2", graph = g,
    scale.model = TRUE, hyper = prior_1
  )
```

Listing 6.8: The formula for the best BYM2 model based all variables

In Table 6.15, the performance measure of the respective models can be seen.

Tab. 6.15: The performance measures for the best performing demographic model of each type. The rank in relation to each performance measure is shown in the brackets.

Model	DIC	WAIC	CPO
Besags Proper	2926 (9)	2958 (9)	-5811 (9)
BYM2	2711 (1)	2673 (2)	-9609 (4)
Leroux	2734 (6)	2666 (1)	-12201 (1)

The effects of the covariates are shown in Table-6.16.

Tab. 6.16: The fixed effects for model 31. Values are rounded.

Variable	Mean	exp(mean$_p$)	exp(q0025$_p$)	exp(q0975$_p$)
(Intercept)	0.5638	2.126	0.5236	5.868
marketplace	0.5562	2.944	0.2227	12.83
entertainment	0.4206	1.575	0.9194	2.525
bakeries	0.3080	1.408	0.8087	2.2557
unemployment_ immigrants	0.1610	1.176	1.081	1.285
clinic	0.1375	1.157	0.8903	1.474
retail	0.08740	1.094	0.9575	1.243
nursing_home	0.06441	1.097	0.6728	1.689
kindergarten	0.04593	1.052	0.8618	1.271
place_of_worship	0.02894	1.033	0.8831	1.200
mining_pt_com	0.01075	1.013	0.9013	1.134
immigrants_ norway	0.006865	1.542e+38	-1.480e+13	-2.495e+08
mining_ft_com	0.003582	1.004	0.9934	1.014
workers_pt_com	0.003072	1.003	0.9957	1.011
construction_ pt_com	0.002205	1.003	0.9529	1.054
pop_dens	0.001339	1.001	1.000	1.003
platform	0.001339	1.001	0.9916	1.011
workers_ft_com	-0.0009058	0.9991	0.9968	1.001
aerodrome	-0.003797	0.9966	0.9402	1.045
urb_dens	-0.004528	0.9955	0.9825	1.009
immigrants_total	-0.02282	1.494e+38	-1.435e+13	-2.418e+08
immigrants_pure	-0.02963	1.486e+38	-1.425e+13	-2.403e+08
schools	-0.03689	0.9674	0.8133	1.142
median_age	-0.03716	0.9636	0.9363	0.9914
hairdresser	-0.04620	0.9715	0.6628	1.371
shops	-0.09960	0.9090	0.7566	1.082
unemp_tot	-0.1105	0.9018	0.6975	1.116
higher_education	-1.279	0.3591	0.06780	1.117

How different variables influence the relative risk is shown in Table 6.17.

Tab. 6.17: How different variables affect the relative risk. Does not contain all variables.

Variable	Scale	Increase	Risk increase
marketplace	per 1000 inhabitants	0.1	11.4%
entertainment	per 1000 inhabitants	0.1	4.6%
bakeries	per 1000 inhabitants	0.1	3.5%
clinic	per 1000 inhabitants	0.1	1.4%
retail	per 1000 inhabitants	0.1	0.9%
nursing_home	per 1000 inhabitants	0.1	0.9%
kindergarten	per 1000 inhabitants	0.1	0.5%
place_of_worship	per 1000 inhabitants	0.1	0.3%
platform	per 1000 inhabitants	0.1	0.01%
unemployment_immigrants	per 1000 inhabitants	0.5	8.4%
mining_pt_com	per 1000 inhabitants	0.5	0.6%
mining_ft_com	per 1000 inhabitants	0.5	0.02%
workers_pt_com	per 1000 inhabitants	0.5	0.02%
construction_pt_com	per 1000 inhabitants	0.5	0.01%
pop_dens	people per km^2	100	14.3%
urb_dens	residential buildings per km^2	10	-4.4%

Finally, the posterior mean and posterior probability for Norway can be seen in Figure 6.8. It can be seen that the risk is the highest in the region around the capital Oslo. In addition, there is a greater risk of infection in the Bergen region. The covid outbreak in Tromsø, which occurred in early March 2021, is also visible as the posterior probability for the region is between 0.75 and 1, while the relative risk is between 1.1 and 1.4.

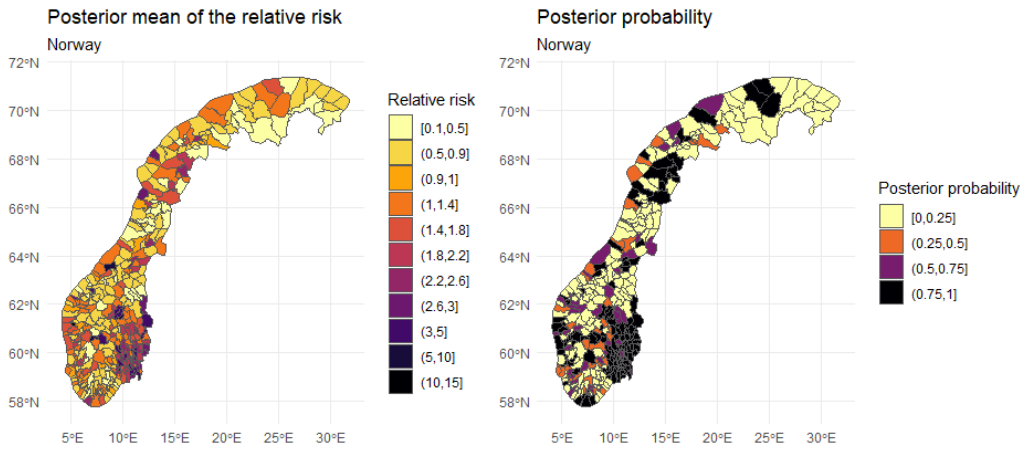

Fig. 6.8: Posterior mean of the area-specific risk and the posterior probability.

6.3 Spatio-Temporal Models

6.3.1 Spatio-Temporal Models for Germany

6.3.2 Spatio-Temporal Models for Norway

6.4 Regression Models

6.4.1 Regression Models for Germany

6.4.2 Regression Models for Norway

7 Appendix

7.1 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

$$f(k,r,p) = \mathcal{P}(X = k) = \binom{k+r-1}{r-1} (1-p)^k p^r, \tag{7.1}$$

with r the number of successes, k the number of failures, and p the probability of success [Hal41].

7.2 Code examples

Specifying the Different Types of Models

```
# Specify a penalized prior
prior_1 <- list(
  prec = list(
    prior = "pc.prec",
    param = c(1, 0.01)
  )
)
# Create a neighbourhood list
nb <- poly2nb(newest_numbers_germany)
# save the neighbourhood
nb2INLA("maps/map_2.adj", nb)
# load it
g <- inla.read.graph(filename = "maps/map_2.adj")
# specify the model formula for the bym2 model
formula_bym2 <- CumNumberTestedIll ~
  # add the demographic vars and pop density
  pop_dens + urb_dens + sex +
  # specify the model with neighbourhood matrix
```


91

```
  f(
    idarea_1, model = "bym2", graph = g,
    scale.model = TRUE, hyper = prior_1
  )
# compute the model
res_bym2 <- inla(
  formula_bym2,
  family = "nbinomial",
  data = newest_numbers,
  E = expected_count,
  control.predictor = list(
    compute = TRUE
  ),
  control.compute = list(dic = TRUE, waic = TRUE, cpo = TRUE)
)
# specify the model formula for besags proper spatial model
formula_bp <- CumNumberTestedIll ~
  # add the demographic vars and pop density
  pop_dens + urb_dens + sex +
  # specify the model with neighbourhood matrix
  f(
    idarea_1, model = "besagproper",
    graph = g, hyper = prior_1
  )
# compute the model
res_bp <- inla(
  formula_bp,
  family = "nbinomial",
  data = newest_numbers,
  E = expected_count,
  control.predictor = list(
    compute = TRUE
  ),
  control.compute = list(dic = TRUE, waic = TRUE, cpo = TRUE)
)
# compute the Q matrix
Q <- Diagonal(x = sapply(nb, length))
for(i in 2:nrow(newest_numbers)) {
  Q[i - 1, i] <- -1
  Q[i, i - 1] <- -1
}
# compute the C matrix
C <- Diagonal(x = 1, n = nrow(newest_numbers)) - Q
# specify the model formula for the leroux model
formula_leroux <- CumNumberTestedIll ~
  # add the demographic vars and pop density
  pop_dens + urb_dens + sex +
  # specify the model with neighbourhood matrix
```

```
  f(
    idarea_1, model = "generic1",
    Cmatrix = C, hyper = prior_1
  )
# compute the model
res_leroux <- inla(
  formula_leroux,
  family = "nbinomial",
  data = newest_numbers,
  E = expected_count,
  control.predictor = list(
    compute = TRUE
  ),
  control.compute = list(dic = TRUE, waic = TRUE, cpo = TRUE)
)
```

Listing 7.1: Specifying different models in INLA.

Variable Selection using INLA

```
# define the stack
stack_all <- inla.stack(
  data = list(
  CumNumberTestedIll = newest_numbers$CumNumberTestedIll
    ),
  A = list(1),
  effects = list(
    data.frame(
      Intercept = 1,
      newest_numbers[, c(2:5, 8:18, 26:36, 38:40, 43:48, 56:59)]
    )
  )
)
# run the selection
result_backwards <- INLAstep(
  fam1 = "nbinomial",
  newest_numbers,
  in_stack = stack_all,
  invariant = "Intercept",
  direction = "backwards",
  include = c(2:5, 8:18, 26:36, 38:40, 43:48, 56:59),
  y = "CumNumberTestedIll",
  y2 = "CumNumberTestedIll",
  power1 = 1,
  inter = 1,
  thresh = 2
)
```

```
# run the selection
result_forwards <- INLAstep(
  fam1 = "nbinomial",
  newest_numbers,
  in_stack = stack_all,
  invariant = "Intercept",
  direction = "forwards",
  include = c(2:5, 8:18, 26:36, 38:40, 43:48, 56:59),
  y = "CumNumberTestedIll",
  y2 = "CumNumberTestedIll",
  powerl = 1,
  inter = 1,
  thresh = 2
)
```

Listing 7.2: The code for variable selection in INLA.

Infrastructure Models for Germany

```
prior_2 <- list(
  prec = list(
    prior = "pc.prec",
    param = c(0.5 / 0.31, 0.01)
  )
)
formula_22_bym2 <- CumNumberTestedIll ~
  pop_dens + urb_dens + marketplace + entertainment + sport +
  clinic + toilet + hairdresser + shops + place_of_worship +
  retail + nursing_home + restaurant + aerodrome + office +
  platform + schools + higher_education + banks + kindergarten +
  bakeries + gas + atm +
  f(
    idarea_1, model = "bym2",
    graph = g, scale.model = TRUE,
    hyper = prior_2
  )
formula_23_bym2 <- CumNumberTestedIll ~
  marketplace + entertainment + sport + clinic +
  toilet + hairdresser + shops + place_of_worship + retail +
  nursing_home + restaurant + aerodrome + office + platform +
  schools + higher_education + banks + kindergarten + bakeries +
  gas + atm +
  f(
    idarea_1, model = "bym2", graph = g,
    scale.model = TRUE, hyper = prior_1
  )
formula_26_leroux <- CumNumberTestedIll ~
```

```
  marketplace + entertainment + clinic + toilet + hairdresser +
  place_of_worship + retail + nursing_home + restaurant +
  terminal + platform + kindergarten + schools + bakeries +
  gas + banks + atm + pop_dens + higher_education +
  f(
    idarea_1, model = "generic1",
    Cmatrix = C, hyper = prior_2
  )
```

Listing 7.3: The code for the demographic models.

Coefficients of BYM2 Model with all Variables for Germany

Tab. 7.1: The fixed effects for the BYM2 model. Values are rounded.

Variable	Mean	exp(mean$_p$)	exp(q0025$_p$)	exp(q0975$_p$)
(Intercept)	-2.752	0.1308	0.005269	0.6603
AfD	3.930	57.70	19.04	135.1
sex	3.726	436.4	-528.5	2156
Union	0.8027	2.406	1.044	4.758
aerodrome	0.3879	1.806	0.4220	5.111
kindergarten	0.1613	1.182	0.9437	1.463
entertainment	0.1275	1.211	0.5643	2.282
shops	0.06951	1.084	0.7961	1.442
nursing_home	0.06812	1.144	0.5251	2.175
bakeries	0.05613	1.068	0.8035	1.392
sport	0.04839	1.052	0.9203	1.120
unemployed_ foreigners	0.04093	1.042	1.026	1.057
asylum_seeker_ benefits	0.004844	1.005	0.9950	1.015
pop_dens	0.00007097	1.000	1.000	1.000
total_income	0.00001266	1.000	1.000	1.000
urb_dens	0.000009849	1.0000	0.9993	1.001
trade_tax	0.0000001765	1.0000	1.000	1.000
income_tax	-0.00006125	0.9999	0.9998	1.000
welfare_ recipients	-0.001508	0.9985	0.9852	1.012
protection_ seekers	-0.003726	0.9963	0.9926	1.000
unemployed_ total	-0.008041	0.9920	0.9857	0.9984
platform	-0.008767	0.9913	0.9821	1.000
schools	-0.01115	0.9935	0.8194	1.193
FDP	-0.01648	2.490	0.04743	14.02
restaurant	-0.03560	0.9654	0.9113	1.022
clinic	-0.03697	0.9646	0.8853	1.049
hairdresser	-0.05900	0.9483	0.7617	1.1166
SPD	-0.1807	0.9439	0.3160	2.197
die_linke	-0.2930	1.043	0.1483	3.682
other_parties	-12.22	1.469e+31	-8.710e+8	1.468e+04

Bibliography

[BCG14] Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical modeling and analysis for spatial data*. CRC press, 2014 (cit. on pp. 28, 55).

[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. 8).

[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. 54).

[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. 28, 50, 52).

[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. 72).

[BT11] George EP Box and George C Tiao. *Bayesian inference in statistical analysis*. Vol. 40. John Wiley & Sons, 2011 (cit. on pp. 3, 8, 14).

[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. 31).

[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. 27).

[Cre15] Noel Cressie. *Statistics for spatial data*. John Wiley & Sons, 2015 (cit. on p. 29).

[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. 9).

[DGM00] Dipak K Dey, Sujit K Ghosh, and Bani K Mallick. *Generalized linear models: A Bayesian perspective*. CRC Press, 2000 (cit. on p. 27).

[DY79] Persi Diaconis and Donald Ylvisaker. "Conjugate priors for exponential families". In: *The Annals of statistics* (1979), pp. 269–281 (cit. on p. 13).

[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. 30).

[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. 11, 27).

[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. 13).

[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. 28).

[Gel+10] Alan E Gelfand, Peter Diggle, Peter Guttorp, and Montserrat Fuentes. *Handbook of spatial statistics*. CRC press, 2010 (cit. on pp. 30, 52).

[GG84] Stuart Geman and Donald Geman. “Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images”. In: *IEEE Transactions on pattern analysis and machine intelligence* 6 (1984), pp. 721–741 (cit. on p. 24).

[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. 55).

[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. 91).

[Han20] Thomas Hansen. *Public COVID-19 Data for Norway (covid19data.no)*. `https://github.com/thohan88/covid19-nor-data`. 2020 (cit. on p. 59).

[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. 10).

[Irw05] ME Irwin. *Prior choice, summarizing the posterior*. 2005 (cit. on p. 13).

[Jef46] Harold Jeffreys. “An invariant form for the prior probability in estimation problems”. In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 186.1007 (1946), pp. 453–461 (cit. on p. 13).

[Jor10] Michael I. Jordan. *Lecture 7: Jeffreys Priors and Reference Priors*. Feb. 2010 (cit. on p. 14).

[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. 28).

[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. 54).

[Kor20] Astrid Iren Solheim Korsvoll. "Nye smittetilfelle i Hyllestad- no har sju personar korona". In: *Firda* (Oct. 1, 2020) (cit. on p. 70).

[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. 15).

[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. 27).

[LP17] Günter Last and Mathew Penrose. *Lectures on the Poisson process*. Vol. 7. Cambridge University Press, 2017 (cit. on pp. 47–49).

[LNM19] Robin Lovelace, Jakub Nowosad, and Jannes Muenchow. *Geocomputation with R*. CRC Press, 2019 (cit. on pp. 41, 43–45).

[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. 28).

[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. 53).

[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. 31, 39, 50, 52, 53, 55, 57).

[NTB21] NTB. “Ordfører: Ulvik-utbruddet spredte seg trolig blant barna”. In: *Haugesunds Avis* (Feb. 8, 2021) (cit. on p. 70).

[Ope84] Stan Openshaw. *The modifiable areal unit problem. Norwich, UK*. 1984 (cit. on p. 55).

[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. 65).

[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. 67).

[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. 13).

[Red+17] Dave W Redding, Tim C D Lucas, Tim Blackburn, and Kate E Jones. “Evaluating Bayesian spatial methods for modelling species distributions with clumped and restricted occurrence data”. In: *PLoS ONE* (2017) (cit. on p. 75).

[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 p. 53).

[RC13] Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013 (cit. on pp. 20–22, 24, 25).

[RMR10] Christian P Robert, Jean-Michel Marin, and Judith Rousseau. “Bayesian inference”. In: *arXiv preprint arXiv:1002.2080* (2010) (cit. on p. 13).

[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. 55).

[RH05] Havard Rue and Leonhard Held. *Gaussian Markov random fields: theory and applications*. CRC press, 2005 (cit. on pp. 4–7, 9, 10, 12, 27, 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. 27, 28).

[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. 59).

[Sim+17] Daniel Simpson, Håvard Rue, Andrea Riebler, Thiago G Martins, and Sigrunn H Sørbye. “Penalising model component complexity: A principled, practical approach to constructing priors”. In: *Statistical science* (2017), pp. 1–28 (cit. on pp. 15, 16, 53).

[SR17] Sigrunn Holbek Sørbye and Håvard Rue. “Penalised complexity priors for stationary autoregressive processes”. In: *Journal of Time Series Analysis* 38.6 (2017), pp. 923–935 (cit. on pp. 16–18).

[WG04] Lance A Waller and Carol A Gotway. *Applied spatial statistics for public health data*. Vol. 368. John Wiley & Sons, 2004 (cit. on p. 56).

[Whi63] Peter Whittle. “Stochastic-processes in several dimensions”. In: *Bulletin of the International Statistical Institute* 40.2 (1963), pp. 974–994 (cit. on p. 56).

[Wic07] Hadley Wickham. “Reshaping Data with the reshape Package”. In: *Journal of Statistical Software* 21.12 (2007), pp. 1–20 (cit. on p. 66).

[Wic19] Hadley Wickham. *stringr: Simple, Consistent Wrappers for Common String Operations*. R package version 1.4.0. 2019 (cit. on p. 66).

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. 61).

[@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. 64).

[@Geo21] Geonorge. *Kartkatalogen*. 2021. URL: `https://www.geonorge.no/` (visited on Feb. 28, 2021) (cit. on p. 64).

[@Ope17] OpenStreetMap contributors. *Planet dump retrieved from https://planet.osm.org*. 2017 (cit. on p. 65).

[@Sen16] Statistisk Sentralbyrå. *StatBank Norway*. 2016. URL: `https://www.ssb.no/en/statbank/` (visited on Feb. 28, 2021) (cit. on p. 60).

[@Smi20] Erik Smistad. *konverter-norgeskart-projeksjon*. 2020 (cit. on p. 64).

List of Figures

3.1 The cantons of Switzerland, an example of an irregular lattice. 5
3.2 An undirected labelled graph with 3 nodes, $\mathcal{V} = \{1, 2, 3\}$ and $\mathcal{E} = \{\{1, 2\}\{2, 3\}\}$. 9
3.3 A typical semivariogram 30
3.4 Covariance functions corresponding to exponential and Matérn models. 31
3.5 The pairwise Markov property; the black nodes are conditionally independent given the light gray nodes. 33
3.6 The local Markov property; the black nodes and white nodes are conditionally independent given the dark gray nodes. 33
3.7 The global Markov property; the dark gray and light gray nodes are globally independent given the black nodes. 34

4.1 A geographic CRS with an origin at 0° longitude an latitude. The red X denotes the location of Trondheim. 42
4.2 The most commonly used simple feature types. 43
4.3 An example of continuous and categorical raster data 44
4.4 The number of shared borders of cantons in Switzerland 51

6.1 The standardised incidence ratio for Germany based on the data of the 5th of March 2021 70
6.2 The standardised incidence ratio for Norway based on the data of the 5th of March 2021 71
6.3 The Cullen and Frey graph for Germany 72
6.4 The Cullen and Frey graph for Norway 73
6.5 A negative binomial fit to the number of cases in German municipalities 73
6.6 A negative binomial fit to the number of cases in Norwegian municipalities 74
6.7 Posterior mean of the area-specific risk and the posterior probabilit. . . 82
6.8 Posterior mean of the area-specific risk and the posterior probability. . 88

List of Tables

5.1 An excerpt from the Covid-19 data for Norway. Does not contain all variables. . . . 59
5.2 An excerpt from the Covid-19 data for Germany. Does not contain all variables. . . . 59
5.3 An excerpt from the age data for Norway. . . . 60
5.4 An excerpt from the unemployment data for Norway. . . . 60
5.5 An excerpt from the worker industry data for Norway. Does not contain all variables. In the actual dataset, a distinction is made between employees by place of residence and employees by place of work. Here, the number of employees by place of residence are displayed. . . . 61
5.6 An excerpt from the immigration data for Norway. . . . 61
5.7 An excerpt from the unemployment data for Germany. Does not contain all variables. . . . 61
5.8 An excerpt from the 2019 European election data for Germany. Does not contain all variables. . . . 62
5.9 An excerpt from the protection-seekers data for Germany. . . . 62
5.10 An excerpt from the social welfare data for Germany. Does not contain all variables. . . . 62
5.11 An excerpt from the asylum-seeker benefits data for Germany. Does not contain all variables. . . . 62
5.12 An excerpt from the tax data for Germany. Does not contain all variables. Income total is the sum of the income tax and the payroll tax for a given city/district. . . . 63
5.13 An excerpt from the long version of the Norwegian Covid-19 data. Does not contain all variables. . . . 66
5.14 An excerpt from the median age data for Norway. Does not contain all variables. . . . 66
5.15 An excerpt from the merged data. Does not contain all variables. . . . 67
5.16 An excerpt from the merged data. Does not contain all variables. . . . 67
5.17 An excerpt from the merged data. Does not contain all variables. . . . 68

6.1 THE AIC for different distributions for Germany and Norway 74

6.2 The municipalities with the most infections as of March 10th 2021. . . 75
6.3 The performance measures for the best performing demographic model of each type. The rank in relation to each performance measure is shown in the brackets. 76
6.4 The fixed effects for model 10. Values are rounded. 77
6.5 The performance measures for the best performing demographic model of each type. The rank in relation to each performance measure is shown in the brackets. 78
6.6 The fixed effects for the BYM2 model. Values are rounded. 78
6.7 How different variables affect the relative risk. Does not contain all variables. 79
6.8 The performance measures for the best performing demographic model of each type. The rank in relation to each performance measure is shown in the brackets. 80
6.9 The fixed effects for model 34. Values are rounded. 81
6.10 The municipalities with the most infections as of March 12th 2021. . . 82
6.11 The performance measures for the best performing demographic model of each type. The rank in relation to each performance measure is shown in the brackets. 83
6.12 The fixed effects for model 22. Values are rounded. 83
6.13 The performance measures for the best performing demographic model of each type. The rank in relation to each performance measure is shown in the brackets. 84
6.14 The fixed effects for model 30. Values are rounded. 84
6.15 The performance measures for the best performing demographic model of each type. The rank in relation to each performance measure is shown in the brackets. 85
6.16 The fixed effects for model 31. Values are rounded. 86
6.17 How different variables affect the relative risk. Does not contain all variables. 87

7.1 The fixed effects for the BYM2 model. Values are rounded. 96

List of Listings

6.1 The formula for the best Leroux model based on the demographic variables . . . 75
6.2 Calculating the posterior mean of a coefficent. . . . 76
6.3 Extracting the credibility interval for a coefficient . . . 76
6.4 The formula for the best BYM2 model based on the infrastructural variables . . . 77
6.5 The formulas for the best models containing all variables. . . . 79
6.6 The formula for the best Leroux model based on the demographic variables . . . 83
6.7 The formula for the best Leroux model based on the infrastructural variables . . . 84
6.8 The formula for the best BYM2 model based all variables . . . 85

7.1 Specifying different models in INLA. . . . 91
7.2 The code for variable selection in INLA. . . . 93
7.3 The code for the demographic models. . . . 94