



Department of Mathematics and Statistics

## Modelling of Viral Disease Spread

Using the Example of SARS-CoV-2

Nico Hahn

Master's thesis in Statistics - STA-3900 - June 2021



# Abstract

Covid-19 has had a significant impact on daily life since the initial outbreak of the global pandemic in late 2019. Countries have been affected to varying degrees, depending on government actions and country characteristics such as infrastructure and demographics. Using Norway and Germany as a case study, this thesis aims to determine which factors influence the risk of infection in each country, using Bayesian modelling and a non-Bayesian machine learning approach. Specifically, the relationship between infection rates and demographic and infrastructural characteristics in a municipality at a fixed point in time is investigated and the effectiveness of a Bayesian model in this context is compared with a machine learning algorithm. In addition, temporal modelling is used to assess the usefulness of government interventions, the impact of changes in mobility behaviour and the prevalence of different strains of Covid-19 in relation to infection numbers. The results show that a spatial model is more useful than a machine learning model in this context. For Germany, it is found that the logarithmic trade tax in a municipality, the share of the vote for the right-wing AfD party and the population density have a positive influence on the infection figures. For Norway, the number of immigrants in a municipality, the number of unemployed immigrants in a municipality and population density are found to have a positive association with infection rates, while the proportion of women in a municipality is negatively associated with infection rates. The temporal models identify higher workplace mobility as a factor significantly influencing the risk of infection in Germany and Norway.

**Keywords:** Spatial modelling, Bayesian modelling, Disease mapping, Machine learning

# Contents

<b>Symbols</b>	<b>1</b>
<b>1 Introduction</b>	<b>2</b>
1.1 Background . . . . .	2
1.2 Motivation . . . . .	4
1.3 Aim and Objective . . . . .	5
1.4 Corona Virus . . . . .	7
1.5 Related Work and Contribution . . . . .	9
1.6 Thesis Outline . . . . .	14
<b>2 Introduction to Bayesian Inference</b>	<b>15</b>
2.1 Preliminaries . . . . .	16
2.1.1 Matrices and Vectors . . . . .	16
2.1.2 Lattice and Torus . . . . .	17
2.1.3 General Notation and Abbreviations . . . . .	18
2.1.4 Symmetric Positive Definite Matrices . . . . .	18
2.2 Basic Concepts of Bayesian Theory . . . . .	20
2.2.1 Bayes' Theorem . . . . .	20
2.2.2 Conditional Independence . . . . .	21
2.2.3 Undirected Graphs . . . . .	21
2.2.4 Calculation of Summary Statistics . . . . .	22
2.3 Prior Selection . . . . .	24
2.3.1 Conjugate Priors . . . . .	24
2.3.2 Penalized Complexity Priors . . . . .	25
2.4 Markov-Chain-Monte-Carlo-Methods . . . . .	28
2.4.1 Monte Carlo Integration . . . . .	28
2.4.2 Markov Chains . . . . .	29
2.4.3 The Metropolis-Hastings Algorithm . . . . .	32
2.5 Latent Gaussian Models and INLA . . . . .	34
2.5.1 Notation and Basic Properties . . . . .	34
2.5.2 Applications for Latent Gaussian Models . . . . .	35

2.5.3	Gaussian Random Fields . . . . .	37
2.5.4	Gaussian Markov Random Fields . . . . .	39
2.5.5	Integrated Nested Laplace Approximation . . . . .	44
2.6	Bayesian Spatial Models . . . . .	47
2.6.1	Besag Spatial Models . . . . .	47
2.6.2	The Besag-York-Mollié Model . . . . .	49
2.6.3	The Leroux Model . . . . .	49
2.6.4	The BYM2 Model . . . . .	50
2.7	Goodness-of-Fit indicators . . . . .	52
2.7.1	The Akaike Information Criterion . . . . .	52
2.7.2	The Deviance Information Criterion . . . . .	52
2.7.3	The Watanabe-Akaike Information Criterion . . . . .	53
2.7.4	The Conditional Predictive Ordinate . . . . .	54
2.8	Model Issues . . . . .	55
2.9	The Variance Inflation Factor . . . . .	56
<b>3</b>	<b>Analysis of Geospatial Health Data</b>	<b>57</b>
3.1	Geographic Data . . . . .	58
3.1.1	Vector Data . . . . .	58
3.1.2	Raster Data . . . . .	60
3.2	Modeling and Visualising Health Data . . . . .	64
3.2.1	Areal Data . . . . .	64
<b>4</b>	<b>Short Introduction to Machine Learning</b>	<b>70</b>
4.1	Common Machine Learning Algorithms . . . . .	70
4.1.1	K-Nearest Neighbours . . . . .	70
4.1.2	Neural Networks . . . . .	71
4.1.3	Classification and Regression Trees . . . . .	72
4.1.4	Gradient Boosting . . . . .	73
4.1.5	Random Forests . . . . .	74
4.2	Machine Learning Methodology . . . . .	76
4.2.1	Tuning of Machine Learning Models . . . . .	76
4.2.2	Interpretation of Machine Learning Models . . . . .	78
<b>5</b>	<b>Dataset Collection</b>	<b>84</b>
5.1	Covid-19 Data . . . . .	85
5.1.1	Covid-19 Data for Norway . . . . .	85
5.1.2	Covid-19 Data for Germany . . . . .	85
5.2	Vaccination Data . . . . .	86
5.3	Demographic Data . . . . .	87

5.3.1	Demographic Data for Norway . . . . .	87
5.3.2	Demographic Data for Germany . . . . .	87
5.4	Shapefiles . . . . .	89
5.4.1	Shapefiles for Norway . . . . .	89
5.4.2	Shapefiles for Germany . . . . .	89
5.5	OpenStreetMap Data . . . . .	90
5.6	Government Response and Mobility Data . . . . .	91
5.7	Covid-19 Variants Data . . . . .	92
5.8	Data Wrangling . . . . .	93
5.8.1	Data Wrangling for Norway . . . . .	93
5.8.2	Data Wrangling for Germany . . . . .	96
5.8.3	Data Wrangling for the Temporal Models . . . . .	98
<b>6</b>	<b>Data Analysis</b>	<b>100</b>
6.1	Standardised Incidence Ratio (SIR) . . . . .	101
6.1.1	SIR for Germany . . . . .	101
6.1.2	SIR for Norway . . . . .	102
6.2	Data Modelling . . . . .	104
6.2.1	Choice of Likelihood . . . . .	104
6.3	Models without a Spatial Component . . . . .	110
6.3.1	Models without a Spatial Component for Germany . . . . .	111
6.3.2	Models without a Spatial Component for Norway . . . . .	112
6.4	Spatial Models . . . . .	113
6.4.1	Spatial Models for Germany . . . . .	113
6.4.2	Spatial Models for Norway . . . . .	115
6.5	Choice of Hyperpriors . . . . .	118
6.6	Predictive models . . . . .	124
6.6.1	Predictive models for Germany . . . . .	124
6.6.2	Predictive models for Norway . . . . .	129
6.7	Temporal models . . . . .	134
6.7.1	Choice of Likelihood . . . . .	134
6.7.2	Temporal models for Germany . . . . .	138
6.7.3	Temporal models for Norway . . . . .	142
<b>7</b>	<b>Further Analysis using R-Shiny</b>	<b>147</b>
<b>8</b>	<b>Discussion</b>	<b>153</b>
8.1	Discussion of the (Non)-Temporal Models . . . . .	153
8.1.1	Discussion of the (Non)-Temporal Models for Norway . . . . .	153
8.1.2	Discussion of the (Non)-Temporal Models for Germany . . . . .	157

8.1.3 Comparison Between the Spatial Models and the Predictive Models . . . . .	160
8.2 Discussion of the Temporal Models . . . . .	162
<b>9 Conclusion</b>	<b>164</b>
<b>10 Appendix</b>	<b>166</b>
10.1 Probability Distributions and the Exponential Family . . . . .	166
10.1.1 The Exponential Family . . . . .	166
10.1.2 The Normal Distribution . . . . .	167
10.1.3 The Multivariate Normal Distribution . . . . .	167
10.1.4 The Poisson Distribution . . . . .	168
10.1.5 The Negative Binomial Distribution . . . . .	168
10.2 Moments . . . . .	170
10.2.1 Skewness . . . . .	170
10.2.2 Kurtosis . . . . .	170
10.3 Distribution Fits . . . . .	171
10.3.1 Distribution Fits for Germany . . . . .	171
10.3.2 Distribution Fits for Norway . . . . .	172
10.4 Choice of Hyperpriors for Germany . . . . .	176
10.5 OpenStreetMap Key-Value Pairs . . . . .	178
10.6 Software Used . . . . .	181
<b>Bibliography</b>	<b>183</b>
<b>Declaration</b>	<b>201</b>

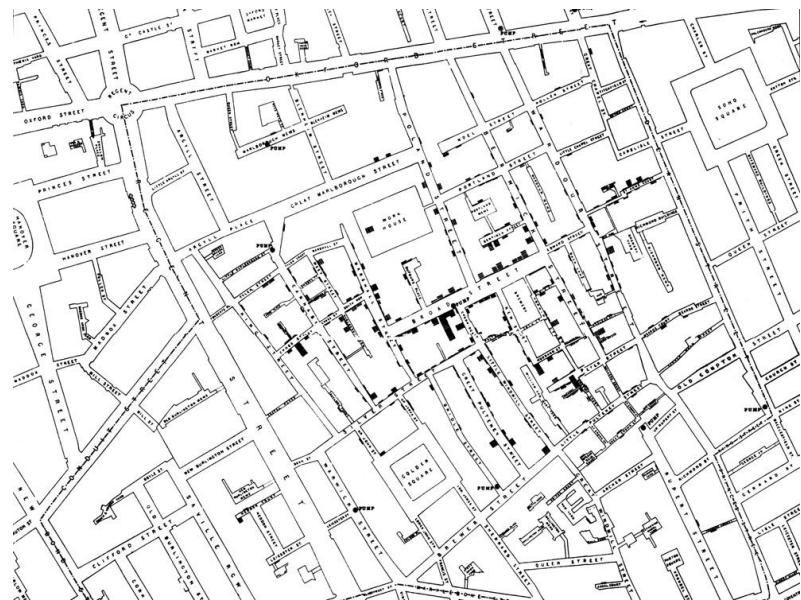
# Symbols

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

# Introduction

## 1.1 Background

Controlling or even trying to prevent an infectious disease is a challenging task, and therefore it is crucial to find ways to combat this type of disease through new and creative ways. The number of infections can vary greatly between different countries or regions, and it is therefore of great interest for local governments or health institutes to find the underlying factors for these differences. This may lead to the identification of previously unrecognized environmental factors that could be the cause of the different risk of disease in different areas. One of the earliest examples of this type of analysis was carried out in relation to a cholera outbreak in south London in 1854 by John Snow. By creating the map shown in Figure 1.1, he was able to show that cholera cases occurred mainly around a water pump in Broad Street. These findings were crucial to understanding that cholera spread through contaminated water supplies, and thus led to the modernization of water supply and sanitation systems in London and the rest of the world (Snow, 1857).



**Fig. 1.1:** The original map of cholera cases in southern London, created by John Snow in 1854.

The location itself (that is, a set of geographic coordinates) is generally unlikely to influence the risk of a certain disease; as there is no reason why one set of coordinates would inherently be at higher risk than another. Instead, the geographic location is a proxy measure, for differences in the attributes of the areas. These differences may relate to physical geography (e.g. temperature, sunlight, precipitation), environmental factors (e.g. air pollution, water quality) or population attributes(e.g. age, income, migration background). The identification of disparities in disease risk across a geographic region can lead to further investigation of the underlying reasons for the differences, which can lead to health breakthroughs such as those noted by Snow. Furthermore, by identifying areas of high risk, health authorities can focus additional resources on these areas in an attempt to influence the behaviours of the population that contribute to an increased risk of disease.

Most approaches to disease mapping are based on dividing the geographic region into spatial units, with disease risks estimated for each of these units. The reason for this is that individual-level data would violate patient confidentiality and governments are more interested in risk levels for the entire population. Each spatial unit has different demographics, so comparisons between spatial units are generally based on the standardized incidence ratio (SIR), defined as the number of observed cases in a given area divided by the number of cases expected for that area based on its population demographics. Most modern methodology for estimating disease risk relies on conditional autoregressive (CAR) models (Besag et al., 1991), which assume the existence of spatial autocorrelation between neighbouring areas, based on the notion that nearby areas are more likely to have more in common than areas that are further apart. This is due to the fact that adjacent areas are more likely to have similar socio-economic characteristics in terms of deprivation and population behaviour. It is assumed by these models that this level of spatial autocorrelation is constant across the spatial region.

## 1.2 Motivation

Covid-19 has had a significant impact on the lives of probably almost everyone on earth. Whether people had to work from home, children suddenly had online classes or people just stayed at home more, everyone was affected. Everyone had to adjust to this new reality where suddenly you could not meet for a coffee or go to the cinema together because those establishments were either closed or people had no desire to risk contracting Covid-19. And that, of course, says nothing about the impact it had on the lives of people who became infected with Covid-19 or had relatives who became infected. The point is that everyone was affected by the impact of the pandemic, and still is, albeit at different levels.

Over time, different countries introduced different strategies to combat Covid-19, such as hard lockdowns where people were only allowed to leave the house if they had a legitimate reason to do so, for example to go to work or to buy groceries, while other countries did not introduce any lockdown measures. Other measures included wearing face masks in public places or limiting how many people can meet in public and private spaces. But even when the same measures were implemented in one country, there were big differences between the proportion of infected people in different parts of the country.

Finally, attitudes towards these measures have become a matter of political identity in various countries, as political parties from different spectrums have different views on how this pandemic should be handled, what measures should be implemented, or whether this pandemic even exists or if it is just much ado about nothing. This has led to the formation of new political movements that regularly protest against the measures taken by the government and demand a return to pre-pandemic conditions.

Understanding the reason why the number of infections varies in different parts of the country can be crucial in helping local governments decide which measures to implement to limit the risk of infection and contribute to the common goal of ending the pandemic as soon as possible.

Identifying areas where people are at higher risk of becoming infected can help governments decide on vaccination strategies, as it may make more sense to vaccinate people in high-risk areas first. By limiting the number of infections in these areas, the likelihood of the virus spreading from one of these areas to one or more neighbouring areas decreases, which in turn slows the spread of Covid-19.

## 1.3 Aim and Objective

The main goal of this work is to analyse what factors drive Covid-19 infection numbers and thus increase the risk of people becoming infected and getting sick from the virus. To identify these factors, it is possible to either look at current infection numbers across different areas to try to find patterns in the data, or to look at infection numbers across a spectrum of time and see if the likelihood of becoming infected changes when new factors, such as vaccination, are introduced or factors, such as government policies, change. Typically, Bayesian spatial models are used for disease mapping, where the neighbourhood structure between the areas of interest plays a crucial role. However, it is possible to neglect this structure and use a non-Bayesian approach to extract key factors. It is not possible to say that one class of models is superior to the other, as this depends mainly on the given data. Nevertheless, one of the aims of this work is to analyse infection counts from a non-temporal point of view and to compare the usefulness of a Bayesian spatial model with that of a non-Bayesian machine learning model. The other main objective of this work is to model the relationship between different features and infection numbers over a period of time using a temporal Bayesian model to see if any pattern emerges or identify which key factors have led to increased or decreased numbers of infections.

To answer these questions, a basic concept of Bayesian theory and Bayesian spatial models is developed and an introduction to geospatial data and the analysis of this particular type of data is given. In addition, a brief introduction to common machine learning methods is given.

For the analysis, different types of data need to be collected from different sources. For non-temporal modelling, these data include:

- Data related to the number of infections in a given municipality.
- Data related to the number of vaccinations in a given municipality.
- Demographic data related to a specific municipality.
- Data related to spatial points of interest in a given municipality.
- Shapefiles for the geographic areas of interest.

For the temporal models, the following data is needed:

- Data related to the number of infections in a given country
- Data related to the number of vaccinations in a given country

- Demographic data related to a given country
- Mobility trends in a given country
- Data keeping track of government measures in a given country
- Data on the relative frequency of different strains of Covid-19 in a given country.

In the analysis, these data are collected for two countries, Germany and Norway. These two countries are not equally affected by the pandemic and the population is distributed differently in the two countries, which makes for an interesting comparison of what factors influence the infection numbers and what kind of models work well in each country.

In order to achieve the main objectives of this thesis, there are the following important goals:

1. Build a basic concept of Bayesian theory, the analysis of geospatial health data as well as non-Bayesian machine learning models.
2. Collect all the data needed for the analysis.
  - 2.1 Collect data related to Covid-19 from the National Institutes of Health.
  - 2.2 Collect demographic data and shapefiles from other official sources.
  - 2.3 Collect infrastructure data by querying OpenStreetMap.
  - 2.4 Collect data for mobility trends and government measures from Our World in Data (OWID).
  - 2.5 Collect data on the frequency of Covid-19 strains from the open-source project CoVariants.
3. Merge all data from different sources to create clean datasets for the analysis.
4. Develop and compare different types of Bayesian spatial models.
5. Train non-Bayesian machine learning models and compare them to the Bayesian models.
6. Develop and compare different types of Bayesian temporal models.
7. Critically evaluate the models and extract factors that significantly influence the risk of infection.

## 1.4 Corona Virus

Viral diseases continue to pose a serious public health threat. Several viral epidemics have occurred in the last 20 years, including the Severe Acute Respiratory Syndrome (SARS) pandemic in 2002/3, H1N1 influenza in 2009, and more recently the Middle East Respiratory Syndrome Coronavirus (MERS-CoV), which was first detected in Saudi Arabia in 2012.

In late 2019, the first few cases of lower respiratory infections were detected in Wuhan, China. In February 2020, this viral disease was officially named "Covid-19", an acronym for "Coronavirus Disease 2019".

Due to the rapid spread of the virus, a Public Health Emergency of International Concern was declared at the end of January 2020, with 18 countries reporting cases and four countries reporting human-to-human transmission.

At the end of February 2020, the World Health Organization (WHO) raised the risk of a Covid-19 epidemic to "very high" before declaring it a pandemic on 11 March. At that time, more than 118,000 cases in 114 countries and 4000 deaths had already been registered.

The first cases of the disease were linked to direct exposure at the Huanan Seafood Wholesale Market in Wuhan, with animal-to-human transmission suspected as the main mechanism. After subsequent cases could not be linked to this mechanism, human-to-human transmission was presumed to be the main transmission mechanism. Furthermore, symptomatic individuals are thought to be the most common source of Covid-19 spread. However, asymptomatic individuals can transmit the virus, therefore isolation is the best way to contain this epidemic.

Similar to other respiratory diseases, e.g. influenza, transmission is thought to occur through respiratory droplets (particles  $> 5 - 10\mu\text{m}$  in diameter) when coughing and sneezing. In closed rooms, transmission by aerosol is possible.

Based on the data from the first cases in Wuhan, the incubation period is generally between 3 and 7 days, with a median of 5.1 days. According to the data, the number of infections doubled about every seven days and the basic reproductive number  $R$  is 2.2, which means that on average each infected individual infects another 2.2 individuals.

According to a report by the Chinese Centre for Disease Control, which studied 72,314 cases, the overall mortality rate of confirmed cases was 2.3%, with most of the fatal cases affecting people over 70 years of age.

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

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

Subsequent reports indicate that the disease is asymptomatic or with very mild symptoms in 70% of patients, while the remaining 30% develop a respiratory syndrome with high fever, cough and even severe respiratory failure, which may require admission to the intensive care unit.

Most countries use some kind of clinical and epidemiological information to determine who should be tested. A molecular test, for example a PCR test, can be used to detect the disease.

The WHO recommends the collection of samples from both the upper and lower respiratory tract. In the laboratory, the genetic material extracted from the saliva or mucus sample is amplified by reverse polymerase chain reaction (RT-PCR), which synthesizes a double-stranded DNA molecule from an RNA form. Once the genetic material is sufficient, the parts of the genetic code of the CoV that are conserved are searched for. The probes used are based on the original gene sequence published by the Shanghai Public Health Clinical Center & School of Public Health, Fudan University, Shanghai, China on Virological.org and subsequent confirmatory evaluation by other laboratories (Casella et al., 2021).

## 1.5 Related Work and Contribution

Since the start of the pandemic in 2019, numerous scientific papers have been written on Covid-19, covering a wide range of topics, such as medicine, social sciences and statistics. Naturally, there have been papers written on the relationship between geographic regions and Covid-19. Incorporating a spatial dimension into the research process can help to better understand different phenomena and make them potentially mappable. The papers considered here can be divided into two categories. The first group consists of disease mapping, spatial analysis and spatio-temporal analysis and refers to studies that analyse the spatial and spatio-temporal patterns of Covid-19. The other group contains research that focuses on other factors that influence the dynamics of the Covid-19 pandemic, but may include spatial and spatio-temporal analysis.

### Disease Mapping, Spatial Analysis and Spatio-Temporal Analysis

Guan et al. (2020) studied cases in mainland China up to 25 February 2020 to determine the defining clinical features and severity of the disease. Among other things, they found that Covid-19 spread rapidly through the country and that the severity of the disease varied. Furthermore, they found that the most common symptoms experienced by patients were cough and fever. They reported a median incubation period of 4 days (Guan et al., 2020).

Chen et al. (2020) analysed how people who emigrated from Wuhan contributed to the early stages of the pandemic in China at the beginning of 2020. They found a strong correlation between the number of confirmed cases of Covid-19 in a given province and emigration from Wuhan. They found that the lockdown of several cities in Hubei province and the implementation of nationwide control measures were effective in preventing the exponential growth of the number of cases (Z.-L. Chen et al., 2020).

Similar to this study, Gross et al. (2020) compare the infection rate in different cities in China and provinces in Italy during the early phases of the pandemic and conclude that the spread of the disease is defined by a two-stage process. The first stage, the authors say, is defined by a constant rate of infection due to a lack of means to detect infected individuals before symptoms appear. In the second stage, they observed an approximately exponential decline due to quarantine. While they found differences between China and Italy, most notably that it took longer for

outbreaks of the disease to be controlled in the Italian provinces, they found similar behaviour in terms of infection rate (Gross et al., 2020).

Chen et al. (2021) analysed the spatio-temporal distribution characteristics and influencing factors of the virus in mainland China using statistical methods, correlation analyses and geographic information system (GIS) mapping. They concluded that the outbreak in non-Hubei provinces can be divided into five phases. The initial outbreak phase, the peak phase where the highest number of new infections is observed, the containment phase where the number of new infections decreases, the rebound phase and a final phase where the number of new infections flattens out. They observed that cities with large population flows from Wuhan were more affected by Covid-19 (Y. Chen et al., 2021).

Saha et al. (2020) provide an overview of how GIS, e.g. mapping dashboards and applications, can be used to monitor the pandemic and related activities. They conclude that the pandemic requires massive data generation and GIS to enable rapid response and analysis to help prevent and guide decisions and movements (Saha et al., 2020).

Gianquintieri et al. (2020) used geo-referenced calls to the emergency number relevant to respiratory problems and subsequent emergency medical service interventions to derive an unbiased representation of Covid-19 diffusion. This study was conducted for the Lombardy region of Italy, which was particularly hard hit by the pandemic in early 2020. The authors reported a strong correlation between Covid-19-related deaths at the provincial level and emergency calls and age- and sex-weighted ambulance dispatches (Gianquintieri et al., 2020).

Lastly, Petrov et al. (2020) examined the spatio-temporal dynamics of the pandemic in the Arctic up to July 2020. They found that the number of infections and morbidity were highly variable, but generally below national levels. They classified the Arctic regions into four groups: Iceland, the Faroe Islands, northern Norway and northern Finland, which were characterized by increased early infection rates but containment of the pandemic through quarantine and other measures; Northern Sweden and Alaska, where the first wave of infection persisted despite weak (Sweden) or variable (Alaska) quarantine measures; northern Russia, where a late start led to a steep rise in infections, deaths and several outbreaks; and northern Canada and Greenland where there was no significant spread of the pandemic (Petrov et al., 2020).

## Other Factors Influencing the Pandemic

Xiong et al. (2020) carried out a correlation analysis for the number of cases in the Hubei province between 30 January 2020 and 18 February 2020. They found a

significant correlation between population, regional GDP, retail sales of consumer goods and the number of confirmed cases of Covid-19 in Hubei province, among others (Xiong et al., 2020).

Ahmadi et al. (2020) analysed the influence of climatic factors on the spread of Covid-19 in Iran and found that areas with low wind speed, humidity and solar radiation support the survival of the virus. The same study found a direct correlation between population density and movement within provinces (Ahmadi et al., 2020). Mehmood et al. (2021) analysed the relationship between air pollution, climate, socioeconomic factors and infection rates in Pakistan. They reported a significant positive correlation between particulate matter ( $PM_{2.5}$ ), an air pollutant, and the number of infections. In contrast to Ahmadi et al. the correlation between the factors humidity and wind speed and the number of infections was positive in some regions and negative in others. They found a small negative relationship between population density and Covid-19 cases, suggesting that areas with higher population density reported proportionally fewer cases (Mehmood et al., 2021).

Pedrosa (2020) analysed the relationship between the number of cases in the US and weather, demographic variables and the infection timeline. He found that only population density and a time series variable, defined as the number of days between the first and the 100th case, showed statistical significance, while the climate in the USA had no influence on infection numbers (Pedrosa, 2020)

As the United States is one of the countries most affected by the pandemic, many studies have attempted to determine what factors are driving up the number of infections in the country. Mollalo et al. (2020) analysed the spatial variability of Covid-19 in the United States up to the 9th of April 2020. Out of 35 environmental, socio-economic, topographical and demographic variables, the four variables found to be most significant were: income inequality, median household income, the proportion of black females and the proportion of nurse practitioners at the county level (Mollalo et al., 2020).

Maiti et al. (2021) analysed infection counts up to 13 May 2020 in the USA. They observed a higher risk of Covid-19 clusters in metropolitan areas compared to rural counties, counties near central airports, more populous counties and counties with the highest proportion of racial and ethnic minorities (Maiti et al., 2021).

Wang et al. (2021) analysed the numbers up to 29 January 2021 in the US and find that factors of ethnicity, crime and income have positive correlations with the number of Covid-19 cases and explain most of the variance in the modelling estimate (Wang et al., 2021).

Allcott et al. (2020) examine partisan differences in Americans' responses to the Covid-19 pandemic, specifically how Republicans and Democrats socially distance themselves and make other efforts to reduce transmission of the disease. They model

not the risk of being infected, but how a person's political beliefs affect their beliefs about the Covid-19 pandemic. They find significant individual-level differences between Republicans and Democrats in self-reported social distancing, beliefs about their personal risk of being infected, and beliefs about the future severity of the pandemic. According to the study, Democrats find it significantly more important to stay inside to prevent the spread of the virus than to go outside to help the economy, compared to Republicans (Allcott et al., 2020).

Bermudi et al. (2021) modelled mortality in the country using latent Gaussian-Bayesian spatial models and found significant relationships between Covid-19 mortality and socioeconomic conditions, as higher socioeconomic levels, as measured by a socioeconomic index, were shown to lead to a lower risk of mortality due to Covid-19. In addition, they showed that men and older persons had the highest risk of mortality due to Covid-19 (Bermudi et al., 2021). Castro et al. (2020), on the other hand, could not find a single narrative that explains the spread of the virus across the states of Brazil, but rather found that layers of complex scenarios intertwine, resulting in a different and simultaneous Covid-19 epidemic across Brazil (Castro et al., 2021).

The situation in India was analysed in a paper by Nandy et al. (2021) and the authors found that higher investment in health and education reduced the likelihood of the spread of Covid-19. In addition, a higher cure rate was found in states with sustained investment in health and education, with mortality rates lower in states that invest more in education (Nandy et al., 2021).

Sannigrahi et al. (2020) found a significant correlation between selected demographic and socio-economic components, including total population, poverty and income, and the number of deaths from Covid-19 in Europe, without controlling for other factors such as environmental variables, socio-ecological status or climate extremes (Sannigrahi et al., 2020).

Studies have been conducted analysing the impact of interdiction measures on the spread of Covid-19. Kasilingam et al. (2020) attempted to predict early containment of Covid-19 using machine learning models based on infrastructural and environmental variables, as well as government-implemented policies and infection-related independent variables for 42 countries. Using logistic regression, a significant positive association was found between healthcare infrastructure and lockdown policies and signs of early containment (Kasilingam et al., 2020). Orea and Álvarez (2020) found a significant positive relationship between interdiction in Spain and its usefulness in preventing the spread of Covid-19 between different provinces in Spain. Furthermore, the same type of relationship was found for the spread of Covid-19 within the same province (Orea & Álvarez, 2020).

## Contribution

The information contained in all the papers mentioned earlier shows just how many factors may or may not be associated with the way that Covid-19 spreads in different countries. Finding a perfect model that explains why numbers are higher in one geographical region than in another is utopian, as there are still too many unknowns even more than a year into the pandemic. Achieving a scientific breakthrough is therefore beyond the scope of this work, the aim is rather to consider a wide range of factors, including infrastructural factors, demographic and socio-economic variables, when discussing the reason for different infection figures within a country and between two different countries. The countries selected for this work, Norway and Germany, are not equally affected by Covid-19, so looking for factors that influence infection numbers in both countries may be indicative of a variable that is driving infection numbers up or down, independent of the country. Of particular interest to this thesis is the link between the political views of people within a municipality and the infection rates in the municipality. Since Germany has been experiencing a lot of anti-hygiene demonstrations and a lot of criticism comes from the right side of the political spectrum, the decision was made to take a close look at whether there is a correlation between the share of votes that certain political parties receive in a given area and the number of infections in that area.

## 1.6 Thesis Outline

The structure of the thesis is as follows. First, an introduction to Bayesian inference is given in Chapter 2. This part includes basic concepts of Bayesian theory, e.g. Bayes' theorem, which are essential for the methodology used in this thesis. Furthermore, different types of priors are introduced, as they form an integral part in Bayesian modelling. In addition, Markov-chain-Monte-Carlo-methods (MCMC methods) and latent Gaussian models are introduced, the latter of which can overcome the shortcomings of MCMC methods and form the basis for Bayesian spatial models. The last part of this chapter includes the introduction of goodness-of-fit indicators used to evaluate model performance and addresses some problems of Bayesian spatial models.

In Chapter 3 a brief introduction to the analysis of geospatial health data is given. First, different types of geospatial data, namely vector data and raster data, are introduced before discussing different methodologies used in modelling this type of data. These methodologies include the standardized incidence ratio (SIR) and the estimation of disease risk in spatial areas.

A short introduction to machine learning is given in Chapter 4. Several commonly used machine learning algorithms are introduced in Section 4.1 before providing a short review of machine learning methodology and introducing a recent field in machine learning, interpretable machine learning, in Section 4.2.

Chapter 5 gives a brief overview of the different types of data collected in this thesis and how the different data sources were combined into a coherent dataset. Chapter 6 focuses on the analysis of this data. First the SIR for the countries is examined, followed by the modelling of the relationship between variables of interest and the infection numbers in Norway and Germany. First, a Bayesian approach to this problem is shown, consisting of models that do not take the neighbourhood structure in the respective countries into account and models that do take such a structure into account. Section 6.5 analyses how these models change, when the prior distribution that is used in the modelling process is changed. Next, non-Bayesian models that are built using the methodology introduced in Section 4.1 are discussed and compared to the Bayesian models. Finally, Bayesian temporal models are evaluated in Section 6.7. As part of this thesis, a dashboard that gives an overview over the used data and allows the modelling of spatial and temporal relationship was developed. A summary of the functionality of this dashboard is found in Chapter 7. The relevant findings of the models calculated during the analysis are discussed in Chapter 8 before the research is wrapped up in Chapter 9, summing up the most important insights of this thesis.

# Introduction to Bayesian Inference

Bayesian inference is a branch of statistics that uses the Bayesian concept of probability and Bayes' theorem to investigate questions of stochastics.

Characteristic for Bayesian statistics is the consistent use of probability distributions or marginal distributions, whose form conveys the accuracy of the procedures or reliability of the data and the procedure.

The Bayesian concept of probability does not presuppose infinitely repeatable random experiments, so that Bayesian methods can be used even with small data sets. A small amount of data leads to a broad probability distribution, which is not strongly localized.

In the Bayesian approach, the parameters of interest are treated as random variables that are governed by their parameters, for instance the mean and standard deviation, and distributions.

Bayesian inference is an essential technique in mathematical statistics and the polar opposite of the frequentist approach, in which a hypothesis is tested without being assigned a probability.

In the Bayesian approach a *prior* distribution  $\pi(\theta)$  is introduced as part of the model. This distribution is intended to express a state of knowledge or ignorance about  $\theta$  prior to obtaining the data. Using the prior distribution, the likelihood function  $\pi(y|\theta)$ , and the observed data  $y$ , it is possible to calculate the probability distribution  $\pi(\theta|y)$  of  $\theta$  given the data  $y$ . This distribution is called the *posterior* distribution of  $\theta$  and is used to make inferences about the parameters (Box & Tiao, 2011, p. 6).

## 2.1 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 (Rue & Held, 2005, pp. 14–19).

### 2.1.1 Matrices and Vectors

Vectors and matrices are indicated by bold notation, such as  $\mathbf{x}$  and  $\mathbf{A}$ . The transpose of  $\mathbf{A}$  is denoted by  $\mathbf{A}^T$ . The element in the  $i$ th row and  $j$ th column of  $\mathbf{A}$  is referenced by  $A_{ij}$ . This notation is used for vectors and  $x_i$  denotes the  $i$ th element of a vector. The vector  $(x_i, x_{i+1}, \dots, x_j)^T$  is abbreviated to  $\mathbf{x}_{i:j}$ . If the columns  $\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_m$  of a  $n \times m$  matrix  $\mathbf{A}$  are stacked on top of each other, this is denoted by  $\text{vec}(\mathbf{A}) = (\mathbf{A}_1^T, \mathbf{A}_2^T, \dots, \mathbf{A}_m^T)$ . Deleting rows and/or columns from  $\mathbf{A}$  creates a *submatrix*. If a submatrix of a  $n \times n$  matrix  $\mathbf{A}$  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  $\mathbf{A}$ .

A diagonal  $n \times n$  matrix  $\mathbf{A}$  is denoted by  $\text{diag}(\mathbf{A})$  and has the following structure:

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

The identity matrix is denoted by  $\mathbf{I}$ .

If  $A_{ij} = 0$  for  $i < j$  or  $A_{ij} = 0$  where  $i > j$ , then  $\mathbf{A}$  is called *upper triangular* and *lower triangular* respectively. The *bandwidth* of a matrix  $\mathbf{A}$  is defined as  $\max\{|i - j| : A_{ij} \neq 0\}$ . The *lower bandwidth* is given by  $\max\{|i - j| : A_{ij} \neq 0 \text{ and } i > j\}$ .  $|\mathbf{A}|$  denotes the *determinant* of a  $n \times n$  matrix  $\mathbf{A}$  and is equal to the product of the eigenvalues of  $\mathbf{A}$ . The *rank* of  $\mathbf{A}$ , referenced by  $\text{rank}(\mathbf{A})$ , is the number of linearly independent rows or columns of  $\mathbf{A}$ . The sum of the diagonal elements is called *trace* of  $\mathbf{A}$ ,  $\text{trace}(\mathbf{A}) = \sum_i A_{ii}$ .

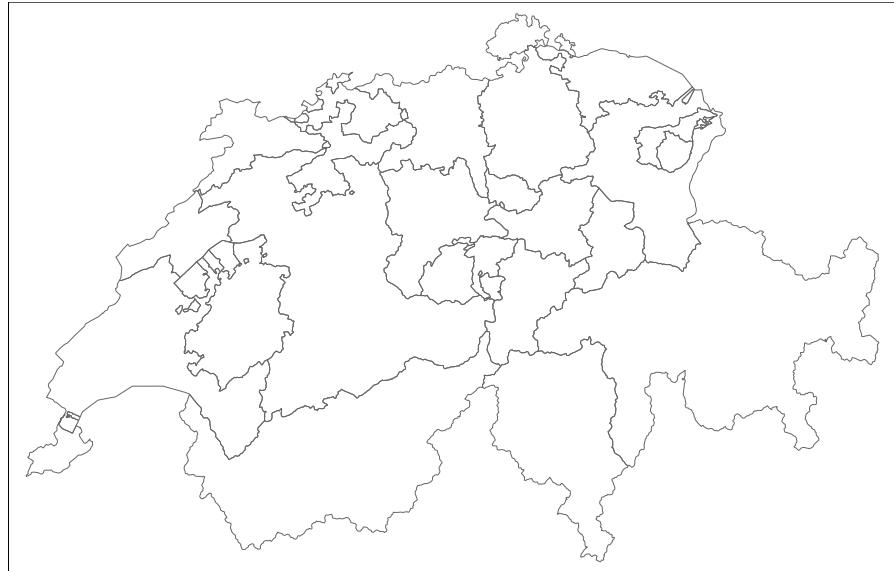
Finally, '⊗' denotes the element-wise multiplication of two matrices of size  $n \times m$ , '⊘' denotes the element-wise division and raising each element of a matrix  $\mathbf{A}$  to a scalar power uses the symbol '⊗' (Rue & Held, 2005, pp. 14–15).

## 2.1.2 Lattice and Torus

$\mathcal{I}_n$  denotes a (regular) **lattice** (or grid) of size  $n = (n_1, n_2)$  (in the two-dimensional case).  $\mathbf{y}$  can take values on  $\mathcal{I}_n$  and  $y_{i,j}$  denotes the value of  $\mathbf{y}$  at location  $ij$ , for  $i = 1, \dots, n_1$  and  $j = 1, \dots, n_2$ . For easier reading this is shortened to  $y_{ij}$ . On an *infinite lattice*  $\mathcal{I}_\infty$ ,  $ij$  are numbered as  $i = 0, \pm 1, \pm 2, \dots$ , and  $j = 0, \pm 1, \pm 2, \dots$ .

A lattice with cyclic or toroidal boundary conditions is referred to as *torus* and is denoted by  $\mathcal{I}_\infty$ . The dimension is  $n = (n_1, n_2)$  (in the two-dimensional case) and all indices are modulus  $n$  and run from 0 to  $n_1 - 1$  or  $n_2 - 1$ . If a GMRF  $\mathbf{y}$  is defined on  $\mathcal{I}_n$ , the toroidal boundary conditions imply that  $y_{-2,n_2}$  is equal to  $y_{n_1-2,0}$  since  $-2 \bmod n_1$  is equal to  $n_1 - 2$  and  $n_2 \bmod n_2$  is equal to 0.

An *irregular lattice* refers to a spatial configuration of regions  $i = 1, \dots, n$  where the regions (mostly) have common boundaries, for instance the states of a nation (Rue & Held, 2005, pp. 15–16). An example of an irregular lattice is shown in Figure 2.1.



**Fig. 2.1:** The cantons of Switzerland, an example of an irregular lattice.

### 2.1.3 General Notation and Abbreviations

For  $C \in \mathcal{I} = \{1, \dots, n\}$  let  $\mathbf{y}_C = \{y_i : i \in C\}$ .  $-C$  denotes the set  $\mathcal{I} - C$  such that  $\mathbf{y}_{-C} = \{y_i : i \notin C\}$ . For two sets  $A$  and  $B$ ,  $A \setminus B = \{i : i \in A \text{ and } i \notin B\}$ .

$\pi(\cdot)$  denotes the density of its arguments, for example  $\pi(\mathbf{y})$  for the density of  $\mathbf{y}$  and  $\pi(\mathbf{y}_A | \mathbf{y}_{-A})$  for the conditional density of  $\mathbf{y}_A$ , given  $\mathbf{y}_{-A}$ . ' $\sim$ ' is used when a variable is 'distributed' according to the law  $l$  (Rue & Held, 2005, p. 16).

### 2.1.4 Symmetric Positive Definite Matrices

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

$$\mathbf{x}^T \mathbf{A} \mathbf{x} > 0, \quad \forall \mathbf{x} \neq \mathbf{0}.$$

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

An SPD matrix  $\mathbf{A}$  has some of the following properties.

1.  $\text{rank}(\mathbf{A}) = n$ .
2.  $|\mathbf{A}| > 0$ .
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.  $\mathbf{A}^{-1}$  is SPD.
8. All principal submatrices of  $\mathbf{A}$  are SPD.

If  $\mathbf{A}$  and  $\mathbf{B}$  are SPD,  $\mathbf{A} + \mathbf{B}$  is SPD, but the reverse is generally not true. Additionally, if  $\mathbf{AB} = \mathbf{BA}$ ,  $\mathbf{AB}$  is SPD.

The following conditions are all sufficient and necessary for a symmetric matrix  $\mathbf{A}$  to be SPD:

1. All eigenvalues  $\lambda_1, \dots, \lambda_n$  of  $\mathbf{A}$  are strictly positive.
2. There exists such a matrix  $\mathbf{C}$  that  $\mathbf{A} = \mathbf{CC}^T$ . If  $\mathbf{C}$  is lower triangular, it is called the *Cholesky triangle* of  $\mathbf{A}$ .

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*:

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

A  $n \times n$  matrix  $\mathbf{A}$  is called a *symmetric positive semidefinite* (SPSD) matrix. An SPSD matrix  $\mathbf{A}$  is sometimes denoted ' $\mathbf{A} \geq 0$ ' (Rue & Held, 2005, pp. 18–19).

## 2.2 Basic Concepts of Bayesian Theory

To understand Bayesian theory, it is helpful to first introduce a few basic concepts, first and foremost Bayes' theorem, which is introduced in Section 2.2.1, one of the most famous concepts in all of statistics. Other notions that are integral to the rest of this thesis are the concept of conditional independence, undirected graphs and the computation of summary statistics, the latter of which is an essential part of the analysis section of this thesis.

### 2.2.1 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  $\mathbf{y}^T = (y_1, \dots, y_n)$  be a vector of  $n$  observations whose probability distribution  $\pi(\mathbf{y}|\boldsymbol{\theta})$  depends on the values of  $k$  parameters  $\boldsymbol{\theta}^T = (\theta_1, \dots, \theta_k)$ . Let  $\pi(\boldsymbol{\theta})$  be the probability distribution of  $\boldsymbol{\theta}$ . Then

$$\pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) = \pi(\mathbf{y}, \boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|\mathbf{y})\pi(\mathbf{y}). \quad (2.1)$$

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

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\mathbf{y})}. \quad (2.2)$$

This last statement is known as Bayes' theorem (Bayes, 1763). The *prior* distribution  $\pi(\boldsymbol{\theta})$  contains knowledge about  $\boldsymbol{\theta}$  without knowledge of the data.  $\pi(\boldsymbol{\theta}|\mathbf{y})$  contains what is known about  $\boldsymbol{\theta}$  given knowledge of the data and is the *posterior* distribution of  $\boldsymbol{\theta}$  given  $\mathbf{y}$ .

If  $\pi(\mathbf{y}|\boldsymbol{\theta})$  is considered as a function of  $\boldsymbol{\theta}$  instead of  $\mathbf{y}$ , it is called the *likelihood function* of  $\boldsymbol{\theta}$  given  $\mathbf{y}$  and can be written as  $l(\boldsymbol{\theta}|\mathbf{y})$ . Thus, Bayes' theorem can be written as

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto l(\boldsymbol{\theta}|\mathbf{y})\pi(\boldsymbol{\theta}). \quad (2.3)$$

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

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

The data  $\mathbf{y}$  modifies the prior knowledge of  $\boldsymbol{\theta}$  through the likelihood function, and thus can be regarded as a representation of the information about  $\boldsymbol{\theta}$  derived from the data (Box & Tiao, 2011).

## 2.2.2 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  $\mathbf{x}_A$ , given  $\mathbf{x}_{-A}$ , the following statement is repeatedly used,

$$\pi(\mathbf{x}_A | \mathbf{x}_{-A}) = \frac{\pi(\mathbf{x}_A, \mathbf{x}_{-A})}{\pi(\mathbf{x}_{-A})} \propto \pi(\mathbf{x}). \quad (2.4)$$

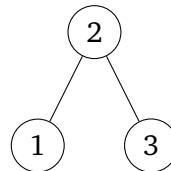
It follows that  $x$  and  $y$  are independent precisely when  $\pi(x, y) = \pi(x)\pi(y)$ , which is expressed by  $x \perp y$ .  $x$  and  $y$  are conditionally independent for a given  $z$  if and only if  $\pi(x, y|z) = \pi(x|z)\pi(y|z)$  (Dawid, 1979). The conditional independence can be easily validated with the help of the following *factorization criterion*,

$$x \perp y | z \iff \pi(x, y, z) = f(x, z)g(y, z), \quad (2.5)$$

for some functions  $f$  and  $g$ , and for all  $z$  with  $\pi(z) > 0$  (Rue & Held, 2005, pp. 16–17).

## 2.2.3 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  $\mathcal{V} = \{1, 2, \dots, n\}$  is assumed, which is referred to as *labelled*. A simple example of an undirected graph is shown in Figure 2.2.



**Fig. 2.2:** 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  $i$ ,

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

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  $\mathcal{V}, i_1, i_2, \dots, i_m$ , for which  $(i_j, i_{j+1}) \in \mathcal{E}$  for  $j = 1, \dots, m - 1$ . 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  $A$ .  $\mathcal{G}^A = \{\mathcal{V}^A, \mathcal{E}^A\}$ , 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 2.2 and  $\mathcal{A} = \{2, 3\}$ , then  $\mathcal{V}^A = \{2, 3\}$  and  $\mathcal{E}^A = \{\{2, 3\}\}$  (Rue & Held, 2005, pp. 17–18).

## 2.2.4 Calculation of Summary Statistics

As the posterior mean and the credibility intervals of coefficient are of interest, calculation of these is performed later on. This allows a better interpretation of the results.

To receive the expected value given a marginal function  $\pi(x)$ , the expected value of a function  $f(x)$  is calculated, i.e.

$$\int f(x) \pi(x) dx \tag{2.6}$$

(Aitkin, 1991).

If necessary, e.g. if the target variable follows a (negative) binomial distribution, the value must be transformed to its original scale, as in these cases the log-likelihood is

modelled. Therefore, in these cases, the expected value would have to be exponentiated to allow a clear interpretation.

In practice, to obtain the credibility interval of a variable, the marginal values are first transformed to their original scale, if necessary, and the 2.5% quantile and the 97.5% quantiles are calculated.

## 2.3 Prior Selection

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

### 2.3.1 Conjugate Priors

One property of exponential families is that they have conjugate priors (Diaconis & Ylvisaker, 1979), which is an important property in Bayesian statistics. If the posterior distribution  $\pi(\boldsymbol{\theta}|\mathbf{y})$  and the prior distribution  $\pi(\boldsymbol{\theta})$  belong to the same probability distribution family, the prior and posterior distributions are called *conjugate* distributions. Furthermore, the prior for the likelihood function  $\pi(\mathbf{y}|\boldsymbol{\theta})$  is called the *conjugate prior*. These priors were first discussed and formalized by Raiffa and Schlaifer in 1961 (Raiffa & Schlaifer, 1961).

The construction of a conjugate prior is done by factorizing 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 (Raiffa & Schlaifer, 1961). In cases where the prior and posterior distributions are part of the same family, the prior is said to be closed under sampling. Furthermore, since the data are only incorporated into the posterior distribution through the sufficient statistics, there exist relatively simple formulas for updating the prior into the posterior.

For an example of the construction of a conjugated prior, see Fink 1997 (Fink, 1997). A drawback of conjugated priors is that the a priori known information about  $\mu$  may be insufficient for determining both parameters or may be inconsistent with the structure imposed by conjugacy (C. P. Robert et al., 2010). Moreover, these priors can be too restrictive and not every belief about the prior can be described (Irwin, 2005).

Thus, although conjugate priors are easy to handle both mathematically and com-

putationally (Irwin, 2005), they are not often used in practice because of these drawbacks.

### 2.3.2 Penalized 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 *penalized complexity priors*, which is a methodology that penalizes the complexity of model components in relation to deviation from simple base model formulations.

PC priors provide a systematic and unified approach to calculating prior distributions for parameters of model components by using an inherited nested structure. This structure contains two models, the base model and a flexible version of the model. The first of the two is generally characterized by a fixed value of the relevant parameter, while the second version is considered a function of the random parameter. By penalizing the deviation from the flexible model to the fixed base model, the PC prior is calculated (Martins et al., 2014).

#### 2.3.2.1 The Principles Behind PC Priors

Four main principles should be followed to calculate priors consistently and to understand their properties.

##### Support to Occam's Razor

Let  $\pi(x|\xi)$  denote the density of a model component  $x$  and  $\xi$  the parameter to which a prior distribution is to be assigned. The base model is characterized by a density  $\pi(x|\xi = \xi_0)$ , where  $\xi_0$  is a fixed value. The prior for  $\xi$  should be such that proper shrinkage is given to  $\xi_0$ . The simplicity of the model is therefore prioritized over the complexity of the model, preventing overfitting (Martins et al., 2014).

## 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 characterized using the Kullback-Leibler divergence (Kullback & Leibler, 1951) to calculate a measure of complexity between the two models,

$$\text{KLD}(f_1||f_0) = \int f_1(x) \log \left( \frac{f_1(x)}{f_0(x)} \right) dx. \quad (2.7)$$

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

$$\text{KLD}(f_1||f_0) = \frac{1}{2} \left( \text{trace}(\Sigma_0^{-1} \Sigma_1) - n - \log \left( \frac{|\Sigma_1|}{|\Sigma_0|} \right) \right), \quad (2.8)$$

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

$$d(\xi) = d(f_1||f_0) = \sqrt{2\text{KLD}(f_1||f_0)} \quad (2.9)$$

which can be interpreted as a measure of distance from  $f_1$  to  $f_0$  (Martins et al., 2014).

## Constant Rate Penalisation

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

$$\frac{\pi_d(d(\xi) + \delta)}{\pi_d(d(\xi))} = r^\delta, \quad d(\xi), \delta \geq 0. \quad (2.10)$$

$r \in (0, 1)$  represents the constant decay rate and thus implies that the relative change in the prior 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  $\xi$  is given (Martins et al., 2014).

## User-Defined Scaling

Since  $\lambda$  characterizes 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  $\lambda$  by including a probability statement of tail events, for example

$$\mathbb{P}(Q(\xi) > U) = \alpha, \quad (2.11)$$

where  $U$  represents an assumed upper bound for an interpretable transformation  $Q(\xi)$  and  $\alpha$  denotes a small probability (Martins et al., 2014).

### 2.3.2.2 Example: PC Prior for the Precision

A PC prior can be used to adjust the smoothness of a spatial field in an intuitive way by specifying such a prior for the precision  $\tau$ . This makes it possible to adjust the smoothness of the spatial field in an intuitive way. In this case, the penalized complexity prior is defined by the parameter  $\sigma_0$ . Equation 2.11 therefore looks like this,

$$\mathbb{P}(\sigma > \sigma_0) = \alpha. \quad (2.12)$$

The actual expression of the prior is given by

$$\pi(\tau) = \frac{\lambda}{2} \tau^{-3/2} \exp\left(-\lambda \tau^{-1/2}\right), \quad \tau > 0 \quad (2.13)$$

and is a type-2-Gumbel distribution.

The prior for  $\tau$  corresponds to an exponential distribution with rate  $\lambda$  for the standard deviation.

$\lambda$  quantifies the size of the penalty for deviation from the base model and is increased with higher values for it.

Here,  $\lambda = \frac{-\log(\alpha)}{\sigma_0}$  (Martins et al., 2014).

## 2.4 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 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.

### 2.4.1 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

$$\mathbb{E}_f [h(X)] = \int_X h(x)f(x)dx. \quad (2.14)$$

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

$$\bar{h}_m = \frac{1}{m} \sum_{j=1}^m h(x_j). \quad (2.15)$$

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

$$v_m = \frac{1}{m^2} \sum_{j=1}^m [h(x_j) - \bar{h}_m]^2. \quad (2.16)$$

For  $m$  large,

$$\frac{\bar{h}_m - \mathbb{E}_f [h(X)]}{\sqrt{v_m}} \quad (2.17)$$

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  $\mathbb{E}_f[h(X)]$  (C. Robert & Casella, 2013, 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" (Metropolis & Ulam, 1949).

## 2.4.2 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 measurable.

In the discrete case, the transition kernel is a matrix  $K$  with elements

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

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

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

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

$$\begin{aligned} \mathbb{P}(X_{t+1} \in A | x_0, x_1, x_2, \dots, x_t) &= \mathbb{P}(X_{t+1} \in A | x_t) \\ &= \int_A K(x_t, dx). \end{aligned} \tag{2.18}$$

These chains were first introduced by Markov in 1906 (Markov, 1906). Markov chains can have certain properties that affect their long-term behaviour and are of particular importance for MCMC algorithms. In Sections 2.4.2.1–2.4.2.5, some of them are introduced.

### 2.4.2.1 Irreducibility

Irreducibility is critical to the construction of Markov chain Monte Carlo algorithms, as it ensures the convergence of such an algorithm. A Markov chain is *irreducible* if all states communicate, that is, for all states  $i$  and  $j$  the probability of going from  $i$  to  $j$  in finite time is true positive.

Formally speaking, given a measure  $\varphi$ , a Markov chain  $(X_n)$  with transition kernel  $K(x, y)$  is  $\varphi$ -*irreducible*, if, for every  $A \in \mathcal{B}(\mathcal{X})$  with  $\varphi(A) > 0$ , there exists  $n$  such that  $K^n(x, A) \forall x \in \mathcal{X}$ . The chain is *strongly  $\varphi$ -irreducible* if  $n = 1 \forall$  measurable  $A$  (C. Robert & Casella, 2013, pp. 213–214).

### 2.4.2.2 Periodicity

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

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

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

$$\mathbf{P} = \begin{pmatrix} 0 & \mathbf{D}_1 & 0 & \dots & 0 \\ 0 & 0 & \mathbf{D}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & & \\ \mathbf{D}_d & 0 & 0 & & 0 \end{pmatrix}, \quad (2.19)$$

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* (C. Robert & Casella, 2013, pp. 217–218).

### 2.4.2.3 Transience and Recurrence

To guarantee an acceptable approximation of a simulated model, a Markov chain needs to have good stability properties. Irreducibility is not strong enough to ensure that the trajectory of  $(X_n)$  enters  $A$  often enough. This leads to the formalization 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 (C. Robert & Casella, 2013, pp. 218–219).

#### 2.4.2.4 Ergodicity

When looking at a Markov chain  $(X_n)$  from a temporal point of view, it is essential to establish to what the chain is converging. A natural candidate for the limiting distribution is the stationary distribution  $\pi$  which leads to the need to define sufficient conditions on  $(X_n)$  for  $X_n$  to be asymptotically distributed according to  $\pi$ . There are several conditions that can be imposed on the convergence of  $P^n$ , the distribution of  $X_n$  to  $\pi$ . The most fundamental and important is that of *ergodicity*, that is, independence of initial conditions.

If a Markov chain  $(X_n)$  is both aperiodic and positive recurrent, it is called an *ergodic* Markov chain (C. Robert & Casella, 2013, pp. 231–234).

#### 2.4.2.5 Stationary distribution

A chain  $(X_n)$  is more stable if the marginal distribution of  $X_n$  is independent of  $n$ . This is a requirement for the existence of a probability distribution  $\pi$  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  $\sigma$ -finite measure  $\pi$  is *invariant* for the transition kernel  $K(\cdot, \cdot)$  if

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

This distribution is referred to as *stationary* if  $\pi$  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 given by

$$\pi_x = (\mathbb{E}_x [\tau_x])^{-1}, \quad x \in \mathcal{X}, \tag{2.20}$$

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  $X_0$ , 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 (C. Robert & Casella, 2013, pp. 223–224).

### 2.4.3 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 (Metropolis et al., 1953).

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  $q(\mathbf{y}|\mathbf{x})$  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  $q(\mathbf{x}|\mathbf{y}) = q(\mathbf{y}|\mathbf{x})$ .

The Metropolis-Hastings algorithm of a target density  $f$  and proposal density  $q$  produces a Markov chain  $(X^{(t)})$  by the following transition.

---

#### Algorithm 1 The Metropolis-Hastings Algorithm

---

Given  $f(\mathbf{x})$  and  $q(\mathbf{y}|\mathbf{x})$

- 1: Initialization: Choose arbitrary  $x_t$  as the first sample
- 2: **for** each iteration  $t$  **do**
- 3:     Generate  $Y_t \sim q(\mathbf{y}|x^{(t)})$
- 4:     Take

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

where

$$\mathbb{P}(x, y) = \min \left\{ \frac{f(\mathbf{y})}{f(\mathbf{x})} \frac{q(\mathbf{x}|\mathbf{y})}{q(\mathbf{y}|\mathbf{x})}, 1 \right\}. \quad (2.21)$$


---

$\mathbb{P}(x, y)$  is the *Metropolis-Hastings acceptance probability*.

The algorithm always accepts values  $y_t$  that lead to an increase in the ratio  $f(y_t)/q(y_t|x^{(t)})$  compared to the previous value  $f(x^{(t)})/q(x^{(t)}|y_t)$ . In the symmetric case, the acceptance probability simplifies to

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

(Hastings, 1970).

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

values of  $y$  such that  $f(y_t) = 0$  are all rejected by the algorithm. As the number of iterations  $t$  increases, the distribution of saved states  $x_0, \dots, x_t$  converges towards the target density  $f(\mathbf{x})$  (C. Robert & Casella, 2013, pp. 270–275).

## 2.5 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 modelling to handle large and complex datasets. 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  $\boldsymbol{x}$  (Gamerman, 1997), while the second problem requires, at least in part, a joint update of  $\boldsymbol{\theta}$  and  $\boldsymbol{x}$ . There are several proposals to solve these shortcomings, but MCMC sampling continues to show poor computational speed (Rue et al., 2009, p. 322).

Bayesian hierarchical models have proven to be effective in capturing complex stochastic structures in spatial processes. A large proportion of these models are based on latent Gaussian models, a subclass of structured additive regression models. These models include Integrated Nested Laplace Approximations (INLA), which are a class of models used to approximate the posterior marginals of a latent Gaussian field. These approximations work by reformulating the regression model as a three-part hierarchical model. The parts are as follows:

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

$\boldsymbol{\theta}$  is a vector of hyperparameters. The hyperparameters  $\boldsymbol{\theta}$  can be, for example, the variance in the Gaussian likelihood or the shape parameter in the likelihood of the gamma distribution. In the case of latent fields, they can be, for instance, dispersion parameters or spatial correlation parameters (Rue et al., 2009).

### 2.5.1 Notation and Basic Properties

For structured additive regression models, the distribution of the response variable  $y_i$  is assumed to be a member of the exponential family, with the mean  $\mu_i$  linked to

a structured additive predictor  $\eta_i$  by a link function  $g(\cdot)$  such that  $g(\mu_i) = \eta_i$ . The predictor  $\eta_i$  takes into account the effect of multiple covariates in an additive way,

$$\eta_i = \alpha + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}) + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \epsilon_i, \quad i = 1, \dots, n \quad (2.22)$$

(Stone, 1985).

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

The conditional density  $\pi(\mathbf{x}|\theta_1)$  is Gaussian with an assumed zero mean and precision matrix  $\mathbf{Q}(\theta_1)$ . The Gaussian density  $\mathcal{N}(\mu, \Sigma)$  with mean  $\mu$  and covariance  $\Sigma$  at configuration  $\mathbf{x}$  is denoted by  $\mathcal{N}(\mathbf{x}; \mu, \Sigma)$ . For simplicity,  $\{\eta_i\}$  is included instead of  $\{\epsilon_i\}$ .

The distribution for the  $n_d$  observational variables  $y = \{y_i : i \in \mathcal{I}\}$  is denoted by  $\pi(\mathbf{y}|\mathbf{x}, \theta_2)$  and is assumed conditionally independent given  $\mathbf{x}$  and  $\theta_2$ . Let  $\boldsymbol{\theta} = (\theta_1^T, \theta_2^T)^T$  with  $\dim(\boldsymbol{\theta}) = m$ . For non-singular  $\mathbf{Q}(\boldsymbol{\theta})$  the posterior is given by

$$\begin{aligned} \pi(\mathbf{x}, \boldsymbol{\theta} | \mathbf{y}) &\propto \pi(\boldsymbol{\theta}) \pi(\mathbf{x} | \boldsymbol{\theta}) \prod_{i \in \mathcal{I}} \pi(y_i | x_i, \boldsymbol{\theta}) \\ &\propto \pi(\boldsymbol{\theta}) |\mathbf{Q}(\boldsymbol{\theta})|^{1/2} \exp \left[ -\frac{1}{2} \mathbf{x}^T \mathbf{Q}(\boldsymbol{\theta}) \mathbf{x} + \sum_{i \in \mathcal{I}} \log \{\pi(y_i | x_i, \boldsymbol{\theta})\} \right]. \end{aligned} \quad (2.23)$$

Most latent Gaussian models satisfy two basic properties:

1. The latent field  $\mathbf{x}$  is of large dimension,  $n \approx 10^2 - 10^5$ . Therefore, the latent field is a Gaussian Markov random field with sparse precision matrix  $\mathbf{Q}(\boldsymbol{\theta})$ .
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 are assumed to be true for the remainder of this work (Rue et al., 2009).

### 2.5.2 Applications for Latent Gaussian Models

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

### 2.5.2.1 Regression Models

Bayesian generalized linear models correspond to the linear relationship  $\eta_i = \alpha + \sum_{k=1}^{n_\beta} \beta_k z_{ki}$  (Dey et al., 2000). Either the linear relationship of the covariates can be relaxed through the  $f(\cdot)$  terms (Fahrmeir & Tutz, 2013), random effects can be introduced through them or both. Smooth covariate effects are frequently modelled using penalized spline models (Lang & Brezger, 2004) or random walk models (Fahrmeir & Tutz, 2013), continuous indexed spline models (Rue & Held, 2005) or Gaussian processes (Chu et al., 2005). 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 (Fahrmeir & Lang, 2001).

### 2.5.2.2 Dynamic Models

Temporal dependence can be introduced by using  $i$  in (2.22) as temporal index  $t$  and defining  $f(\cdot)$  and  $\mathbf{u}$  such that  $f(u_t) = f_t$ . Both a discrete-time and a continuous-time autoregressive model can be modelled by  $\{f_t\}$ . Furthermore, a seasonal effect or the latent process of a structured time series model can be modelled (Kitagawa & Gersch, 1996). Alternatively, a smooth temporal function in the same sense as for regression models can be represented by  $\{f_t\}$ .

### 2.5.2.3 Spatial and Spatio-Temporal Models

Similar to the previous type of model, spatial dependence can be modelled by a spatial covariate  $\mathbf{u}$  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 realizations of some sort. Popular models of this type include the Besag-York-Mollié (Besag et al., 1991) model with extensions for regional data, continuous indexed Gaussian models (Banerjee et al., 2014) and texture models (Marroquin et al., 2001). 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 (Kammann & Wand, 2003).

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 (2.22), sometimes linear or sum-to-zero constraints can be imposed (Rue et al., 2009, pp. 319–321).

### 2.5.3 Gaussian Random Fields

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

$$\left\{ Z(s) : s \in D \subset \mathbb{R}^2 \right\} \quad (2.24)$$

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  $\mathbf{h} \in \mathbb{R}^2$  the distribution of  $Z(\mathbf{s}) = (Z(s_1), \dots, Z(s_n))$  is equal to that of  $Z(\mathbf{s} + \mathbf{h}) = (Z(s_1 + h), \dots, Z(s_n + h))$ . A less constraining requirement is given by second-order stationarity. Under this condition, the process has a constant mean value

$$\mathbb{E}[Z(\mathbf{s})] = \mu, \quad \forall s \in D, \quad (2.25)$$

and the covariances depend only on the differences between locations

$$\text{Cov}(Z(\mathbf{s}), Z(\mathbf{s} + \mathbf{h})) = C(\mathbf{h}), \quad \forall \mathbf{s} \in D, \forall \mathbf{h} \in \mathbb{R}^2. \quad (2.26)$$

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

$$\text{Var}(Z(s_i) - Z(s_j)) = 2\gamma(s_i - s_j), \quad \forall s_i, s_j. \quad (2.27)$$

$2\gamma(\cdot)$  is the variogram and  $\gamma(\cdot)$  is called the semivariogram (Cressie, 2015). Under the assumption of intrinsic stationarity, the constant-mean assumption implies

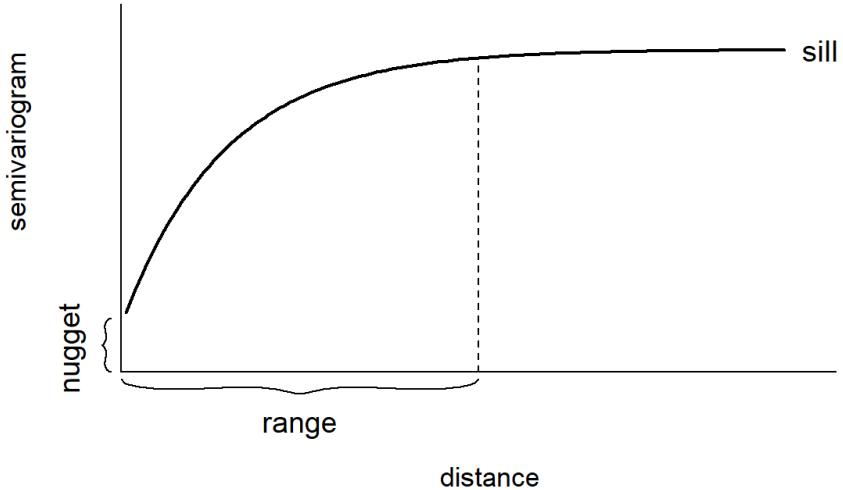
$$2\gamma(\mathbf{h}) = \text{Var}(Z(\mathbf{s} + \mathbf{h}) - Z(\mathbf{s})) = \mathbb{E}[(Z(\mathbf{s} + \mathbf{h}) - Z(\mathbf{s}))^2],$$

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

$$2\hat{\gamma}(\mathbf{h}) = \frac{1}{|N(\mathbf{h})|} \sum_{N(\mathbf{h})} (Z(s_i) - Z(s_j))^2, \quad (2.28)$$

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

Plotting the empirical semivariogram against the separation distance conveys essential information regarding the continuity and spatial variability of the process. Given



**Fig. 2.3:** A typical semivariogram

relatively short distances, the semivariogram tends to be small but increases with distance, indicating the similarity of observations in 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 be indicative of a spatially discontinuous process. The empirical semivariogram is an exploratory tool useful for assessing whether data exhibit spatial correlation. Furthermore, it can be compared to a Monte Carlo envelope of empirical semivariograms calculated from random permutations of the data while keeping the locations fixed (Diggle et al., 2003). If the empirical semivariogram lies outside the Monte Carlo envelope with increasing distance, this is an indication of spatial correlation.

An example of a semivariogram is shown in Figure 2.3. This graph is taken from "Geospatial Health Data: Modelling and Visualization with R-INLA and Shiny" by Paula Moraga (Moraga, 2019).

## 2.5.4 Gaussian Markov Random Fields

### 2.5.4.1 Definition of GMRFs

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

Since  $\boldsymbol{\mu}$  does not affect the pairwise conditional independence properties of  $\mathbf{x}$ , 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  $\mathbf{Q}$  determines  $\mathcal{G}$ , i.e. whether  $x_i$  and  $x_j$  are conditionally independent, and can be derived from  $\mathbf{Q}$ . If  $\mathbf{Q}$  is a fully dense matrix,  $\mathcal{G}$  is fully connected, implying that any normal distribution with SPD covariance matrix is a GMRF and vice versa.

The elements of  $\mathbf{Q}$  are used for conditional interpretations. For any GMRF with respect to  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  with mean  $\boldsymbol{\mu}$  and precision matrix  $\mathbf{Q} > 0$ ,

$$\mathbb{E}[x_i | \mathbf{x}_{-i}] = \mu_i - \frac{1}{Q_{ii}} \sum_{j:j \sim i} Q_{ij} (x_j - \mu_j), \quad (2.29)$$

$$\text{Prec}(x_i | \mathbf{x}_{-i}) = Q_{ii} \quad \text{and} \quad (2.30)$$

$$\text{Corr}(x_i, x_j | \mathbf{x}_{ij}) = -\frac{Q_{ij}}{\sqrt{Q_{ii}Q_{jj}}}, \quad i \neq j \quad (2.31)$$

(Rue & Held, 2005, p. 21).

On the main diagonal of  $\mathbf{Q}$  are the conditional precisions of  $x_i$  given  $\mathbf{x}_{-i}$  are placed, while the other elements, when scaled appropriately, provide information about the conditional correlation between  $x_i$  and  $x_j$  given  $\mathbf{x}_{ij}$ . Since  $\text{Var}(x_i) = \Sigma_{ii}$  and  $\text{Corr}(x_i, x_j) = \Sigma_{ij}/\sqrt{\Sigma_{ii}\Sigma_{jj}}$ , 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 a  $n$ -dimensional distribution to a one- or two-dimensional distribution.  $\mathbf{Q}$  is difficult to interpret marginally because either  $\mathbf{x}_{-i}$  or  $\mathbf{x}_{ij}$  would have to be integrated out of the joint distribution parameterized with respect to  $\mathbf{Q}$ .  $\mathbf{Q}^{-1} = \boldsymbol{\Sigma}$  by definition, and in general  $\Sigma_{ii}$  depends on each element in  $\mathbf{Q}$  and vice versa (Rue & Held, 2005, pp. 20–23).

### 2.5.4.2 Markov Properties of GMRFs

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

The *pairwise Markov property*:

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

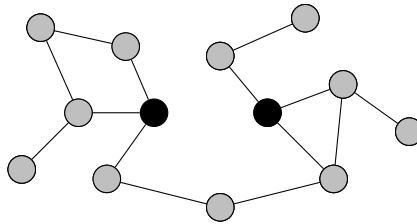
The *local Markov property*:

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

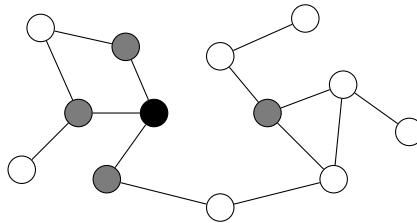
The *global Markov property*:

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

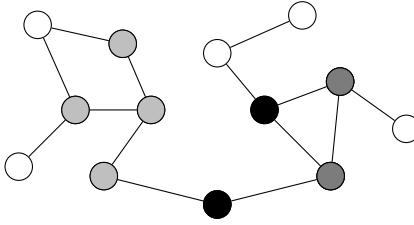
for all disjoint sets  $A$ ,  $B$  and  $C$  where  $A$  and  $B$  are non-empty and separated by  $C$ . Illustrations for these properties are shown in Figure 2.4, Figure 2.5 and Figure 2.6. These illustrations are taken from Rue and Held (Rue & Held, 2005, pp. 23–24).



**Fig. 2.4:** The pairwise Markov property; the black nodes are conditionally independent given the light grey nodes.



**Fig. 2.5:** The local Markov property; the black nodes and white nodes are conditionally independent given the dark grey nodes.



**Fig. 2.6:** The global Markov property; the dark grey and light grey nodes are globally independent given the black nodes.

#### 2.5.4.3 Conditional Properties of GMRFs

An essential result of GMRFs is the conditional distribution for a subset  $\mathbf{x}_A$  given  $\mathbf{x}_{-A}$ . Here the canonical parameterization 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  $-A$ ,

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_A \\ \mathbf{x}_B \end{pmatrix}. \quad (2.32)$$

The mean and the precision are divided accordingly,

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_A \\ \boldsymbol{\mu}_B \end{pmatrix}, \quad \text{and} \quad \mathbf{Q} = \begin{pmatrix} \mathbf{Q}_{AA} & \mathbf{Q}_{AB} \\ \mathbf{Q}_{BA} & \mathbf{Q}_{BB} \end{pmatrix}. \quad (2.33)$$

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

$$\boldsymbol{\mu}_{A|B} = \boldsymbol{\mu}_A - \mathbf{Q}_{AA}^{-1} \mathbf{Q}_{AB} (\mathbf{x}_B - \boldsymbol{\mu}_B) \quad (2.34)$$

and

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

Thus, the explicit knowledge of  $\mathbf{Q}_{A|B}$  is available through  $\mathbf{Q}_{AA}$ , 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  $\mathbf{Q}$  in  $A \cup \text{ne}(A)$ , since  $Q_{ij} = 0 \forall j \notin \text{ne}(i)$ . For successive conditioning, the canonical parameterization for GMRF is useful. A GMRF  $\mathbf{x}$  with respect to  $\mathcal{G}$  and canonical parameters  $\mathbf{b}$  and  $\mathbf{Q} > 0$  has the density

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

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

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

Furthermore,

$$\mathcal{N}(\boldsymbol{\mu}, \mathbf{Q}^{-1}) \iff \mathcal{N}_C(\mathbf{Q}\boldsymbol{\mu}, \mathbf{Q}).$$

If the indices are partitioned into two non-empty sets A and B and  $\mathbf{x}$ ,  $\mathbf{b}$  and  $\mathbf{Q}$  are partitioned as in (2.32) and (2.33), then

$$\mathbf{x}_A | \mathbf{x}_B \sim \mathcal{N}_C(\mathbf{b}_A - \mathbf{Q}_{AB}\mathbf{x}_B, \mathbf{Q}_{AA}). \quad (2.35)$$

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

$$\mathbf{x} | \mathbf{y} \sim \mathcal{N}_C(\mathbf{b} + \mathbf{P}\mathbf{y}, \mathbf{Q} + \mathbf{P}). \quad (2.36)$$

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 parameterization can be repeatedly updated without explicitly calculating the mean until it is needed. The computation of the mean requires the solution of  $\mathbf{Q}\boldsymbol{\mu} = \mathbf{b}$ , but only matrix-vector products are needed for updating the canonical parameterization (Rue & Held, 2005, pp. 25–27).

#### 2.5.4.4 Specification Through Full Conditionals

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

$$\mathbb{E}[x_i | \mathbf{x}_{-i}] = \mu_i - \sum_{j:j \sim i} \beta_{ij} (x_j - \mu_j) \quad \text{and} \quad (2.37)$$

$$\text{Prec}(x_i | \mathbf{x}_{-i}) = \kappa_i > 0 \quad (2.38)$$

for  $i = 1, \dots, n$ , 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  $\pi(\mathbf{x})$  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  $\mathbf{Q}$  must be symmetrical, i.e.,

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

then  $\mathbf{x}$  is a GMRF with respect to a labelled graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  with mean  $\boldsymbol{\mu}$  and precision matrix  $\mathbf{Q} = (Q_{ij})$  (Rue & Held, 2005, p. 27).

#### 2.5.4.5 Multivariate GMRFs

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

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

$x_i$  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  $\mathbf{x}_i$  of dimension  $p$ , resulting in a GMRF of size  $np$ . Such a GMRF is denoted by  $\mathbf{x} = (\mathbf{x}_1^T, \dots, \mathbf{x}_n^T)^T$ . The Markov property with respect to the nodes is preserved, i.e.,

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

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

It follows that

$$\mathbf{x}_i \perp \mathbf{x}_{-i} | \mathbf{x}_{-ij} \iff \tilde{Q}_{ij} = \mathbf{0}.$$

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

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

and

$$\tilde{Q}_{ij} \neq 0 \iff \{i, j\} \in \mathcal{E} \quad \forall i \neq j.$$

A MGMRF<sub>*p*</sub> is equivalent to a GMRF of dimension *np* with identical mean vector and precision matrix. Therefore, all results valid for a GMRF are valid for a MGMRF<sub>*p*</sub>, with modifications, since the graph for a MGMRF<sub>*p*</sub> has size *n* and is defined with respect to  $\{\mathbf{x}_i\}$ , while for a GMRF it has size *np* and is defined with respect to  $\{x_i\}$ . The interpretation of  $\tilde{\mathbf{Q}}_{ii}$  and  $\tilde{\mathbf{Q}}_{ij}$  can be derived from the full conditional  $\pi(\mathbf{x}_i|\mathbf{x}_{-i})$ . The extensions of (2.29) and (2.30) are

$$\mathbb{E}[\mathbf{x}_i|\mathbf{x}_{-i}] = \boldsymbol{\mu}_i - \tilde{\mathbf{Q}}_{ii}^{-1} \sum_{j:j \sim i} \tilde{\mathbf{Q}}_{ij} (\mathbf{x}_j - \boldsymbol{\mu}_j) \quad (2.39)$$

$$\text{Prec}(\mathbf{x}_i|\mathbf{x}_{-i}) = \tilde{\mathbf{Q}}_{ii}. \quad (2.40)$$

In some applications, the full conditionals

$$\mathbb{E}[\mathbf{x}_i|\mathbf{x}_{-i}] = \boldsymbol{\mu}_i - \sum_{j:j \sim i} \boldsymbol{\beta}_{ij} (\mathbf{x}_j - \boldsymbol{\mu}_j) \quad (2.41)$$

$$\text{Prec}(\mathbf{x}_i|\mathbf{x}_{-i}) = \kappa_i > 0, \quad (2.42)$$

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

$$\tilde{\mathbf{Q}}_{ij} = \begin{cases} \kappa_i \boldsymbol{\beta}_{ij} & i \neq j \\ \kappa_i & i = j \end{cases};$$

therefore  $\tilde{\mathbf{Q}} > 0 \iff (\mathbf{I} + (\boldsymbol{\beta}_{ij})) > 0$  (Rue & Held, 2005, pp. 29–30).

## 2.5.5 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 (Rue

et al., 2009).

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

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

As introduced in subsection 2.5.1,  $\mathbf{y}$  are the observed data,  $\mathbf{x}$  is a Gaussian field,  $\boldsymbol{\theta}$  represents the hyperparameters, while  $\mu(\boldsymbol{\theta})$  and  $\mathbf{Q}(\boldsymbol{\theta})$  denote the mean and precision matrix respectively. To ensure fast inference, the dimension of the hyperparameter vector  $\boldsymbol{\theta}$  should be small, since the approximations are computed by numerical integration over the hyperparameter space.

In most cases, the observations  $y_i$  are assumed to belong to the exponential family with mean  $\mu_i = g^{-1}(\eta_i)$ . As shown in equation (2.22),  $\eta_i$  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  $\{f^{(j)}\}$  can take different forms.

Let  $\mathbf{x} = (\alpha, \{\beta_k\} | \theta \sim \mathcal{N}(\mu(\theta), \mathbf{Q}(\theta)^{-1}))$  be the vector of latent Gaussian variables, and let  $\boldsymbol{\theta}$  be the vector of hyperparameters, which are not required to be Gaussian. INLA calculates accurate and fast approximations for the posterior marginals of the components of the latent Gaussian variables

$$\pi(x_i | \mathbf{y}), \quad i = 1, \dots, n,$$

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

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

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

$$\pi(x_i | \mathbf{y}) = \int \pi(x_i | \boldsymbol{\theta}, \mathbf{y}) \pi(\boldsymbol{\theta} | \mathbf{y}) d\boldsymbol{\theta}, \quad (2.43)$$

and the posterior marginal for the hyperparameters can be expressed by

$$\pi(\theta_j | \mathbf{y}) = \int \pi(\boldsymbol{\theta} | \mathbf{y}) d\boldsymbol{\theta}_{-j}. \quad (2.44)$$

$\pi(x_i | \mathbf{y})$  is approximated by combining analytical approximations to the full conditionals  $\pi(x_i | \boldsymbol{\theta}, \mathbf{y})$  and  $\pi(\boldsymbol{\theta} | \mathbf{y})$  and numerical integration routines to integrate out  $\boldsymbol{\theta}$ . Similarly,  $\pi(\theta_j | \mathbf{y})$  is approximated by approximating  $\pi(\boldsymbol{\theta} | \mathbf{y})$  and integrating out  $\boldsymbol{\theta}_{-j}$ . In particular, the posterior density of  $\boldsymbol{\theta}$  is obtained through Gaussian approximation

for the posterior of the latent field,  $\tilde{\pi}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ , evaluated at the posterior mode,  $x^*(\boldsymbol{\theta}) = \arg \max_{\mathbf{x}} \pi_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ ,

$$\tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) \propto \frac{\pi(\mathbf{x}, \boldsymbol{\theta}, \mathbf{y})}{\tilde{\pi}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})} \Big|_{\mathbf{x}=x^*(\boldsymbol{\theta})}. \quad (2.45)$$

Next, the following nested approximations are constructed,

$$\tilde{\pi}(x_i|\mathbf{y}) = \int \tilde{\pi}(x_i|\boldsymbol{\theta}, \mathbf{y}) \tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}, \quad \tilde{\pi}(\theta_j|\mathbf{y}) = \int \tilde{\pi}(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}_{-j}. \quad (2.46)$$

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

$$\tilde{\pi}(x_i|\mathbf{y}) = \sum_k \tilde{\pi}(x_i|\theta_k, \mathbf{y}) \tilde{\pi}(\theta_k|\mathbf{y}) \times \Delta_k, \quad (2.47)$$

$$\tilde{\pi}(\theta_j|\mathbf{y}) = \sum_l \tilde{\pi}(\theta_l^*|\mathbf{y}) \times \Delta_l^*, \quad (2.48)$$

with  $\Delta_k$  and  $\Delta_l^*$  representing the area weights corresponding to  $\theta_k$  and  $\theta_l^*$ .

To obtain the approximations for the posterior marginals for the  $x_i$ 's conditioned on selected values of  $\theta_k$  and  $\tilde{\pi}(x_i|\theta_k, \mathbf{y})$ , a Gaussian, Laplace or simplified Laplace approximation can be used. Using a Gaussian approximation derived from  $\tilde{\pi}_G(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$  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 (Moraga, 2019).

## 2.6 Bayesian Spatial Models

Bayesian spatial models are often used in the field of disease mapping. Bayesian hierarchical models improve estimates of log risk by providing information about neighbouring regions in the spatially structured component as well as regional variation in the unstructured component (Blangiardo & Cameletti, 2015). One of the most well-known spatial models is Besags' spatial model, which is presented in Section 2.6.1. Several models have been developed based on the Besag model, including the Besag-York-Mollié (BYM) model, introduced in Section 2.6.2, the Leroux model, introduced in Section 2.6.3, and more recently the BYM2 model, introduced in Section 2.6.4.

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

### 2.6.1 Besag Spatial Models

#### 2.6.1.1 Besags' Improper Spatial Model

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

$$x_i | \mathbf{x}_{-i}, \tau_x \sim \mathcal{N} \left( \frac{1}{n_{\delta_i}} \sum_{j \in \delta_i} x_j, \frac{1}{n_{\delta_i} \tau_x} \right), \quad (2.49)$$

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

$$\pi(\mathbf{x} | \tau_x) \propto \exp \left( -\frac{\tau_x}{2} \sum_{i \sim j} (x_i - x_j)^2 \right) \propto \exp \left( -\frac{\tau_x}{2} \mathbf{x}^T \mathbf{Q} \mathbf{x} \right). \quad (2.50)$$

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

$$Q_{ij} = \begin{cases} n_{\delta_i} & i = j, \\ -1 & i \sim j, \\ 0 & \text{else.} \end{cases} \quad (2.51)$$

$\mathbf{Q}$  is a singular matrix, i.e. it has a non-empty null space  $\mathbf{V}$ , hence the model is called intrinsic or Besags' improper spatial model.

The Besag model for spatial effects has one hyperparameter, the precision  $\tau_x$ , which is represented as

$$\theta_1 = \log \tau_x. \quad (2.52)$$

The prior is defined on  $\theta_1$  (Besag, 1974; Riebler et al., 2016).

### 2.6.1.2 Besags' Proper Spatial Model

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

$$Q_{ij} = \begin{cases} \tau_x (n_{\delta_i} + d) & i = j, \\ -\tau & \text{else.} \end{cases} \quad (2.53)$$

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

$$x_i | \mathbf{x}_{-i}, \tau_x, d \sim \mathcal{N} \left( \frac{1}{d + n_{\delta_i}} \sum_{i \sim j} x_j \frac{1}{\tau_x (d + n_{\delta_i})} \right). \quad (2.54)$$

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

$$\theta_1 = \log \tau_x \quad (2.55)$$

and the diagonal parameter  $d$ , which is represented as

$$\theta_2 = \log d. \quad (2.56)$$

The priors are defined on  $\theta_1$  and  $\theta_2$ , respectively (Besag, 1974; Riebler et al., 2016).

## 2.6.2 The Besag-York-Mollié Model

The Besag-York-Mollié (BYM) model is a lognormal Poisson model that is a combination of a Besag model  $u$  and an ordinary random effect component  $v$  for non-spatial heterogeneity. It combines the regional spatial effect  $\mathbf{x}$  into the sum of an unstructured and a structured spatial component, so that  $\mathbf{x} = \mathbf{v} + \mathbf{u}$ .

$\mathbf{v} \sim \mathcal{N}(0, \tau_v^{-1} \mathbf{I})$  accounts for pure overdispersion, while  $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \tau_u^{-1} \mathbf{Q}^-)$  is the Besag model.

By using a spatial and a non-spatial error term, the overdispersion that is not modelled by the Poisson variables is taken into account. Thus, if the observed variance is not fully explained by the spatial structure of the data, the error terms explain the rest of the variance.

The resulting covariance matrix of  $\mathbf{x}$  is given by

$$\text{Var}(\mathbf{x} | \tau_u, \tau_v) = \tau_v^{-1} \mathbf{I} + \tau_u^{-1} \mathbf{Q}^-, \quad (2.57)$$

where  $\mathbf{Q}^-$  denotes the generalized inverse of  $\mathbf{Q}$ .

The hyperparameters of the model are the precision  $\tau_u$  of the Besag model  $u$  and the precision  $\tau_v$  of the iid model  $v$ . They are represented as

$$\boldsymbol{\theta} = (\theta_1, \theta_2) = (\log \tau_v, \log \tau_u) \quad (2.58)$$

and the prior is defined on  $\boldsymbol{\theta}$  (Besag et al., 1991; Riebler et al., 2016).

## 2.6.3 The Leroux Model

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

$$\text{Var}(\mathbf{x} | \tau_x, \phi) = \tau_x^{-1} ((1 - \phi) \mathbf{I} + \phi \mathbf{Q})^{-1}, \quad (2.59)$$

with  $\phi \in [0, 1]$  as mixing parameter. For  $\phi = 0$  the model reduces to pure overdispersion and for  $\phi = 1$  to the Besag model. The conditional expected value of  $x_i$  for all other random effects is the weighted mean of the unstructured model with

zero mean and the mean of the Besag model, while the conditional variance is the weighted mean of  $\tau_x^{-1}$  and  $(\tau_x \cdot n_{\delta_i})^{-1}$  (Leroux et al., 2000; Riebler et al., 2016).

## 2.6.4 The BYM2 Model

One problem that all the aforementioned models have is the lack of scaling of the spatially structured component. Scaling facilitates the assignment of hyperpriors and ensures that the interpretation of hyperpriors remains the same across different areas.

Another problem is that the marginal standard deviations of the commonly used IGMRF priors can vary greatly, a fact that should be taken into account by assigning hyperpriors to the precision parameters of these models.

Since the Besag model penalizes a local deviation from its null space, the hyperprior controls this local deviation and thus affects the smoothness of the estimated spatial effects. If the estimate of the field is too smooth, the precision is large and the spatial variation may be blurred. On the other hand, if the precision is too small, the model could overfit due to the large local variability.

The marginal variances  $\tau_x^{-1} [\mathbf{Q}^-]_{ii}$  depend on the structure of the graph, which is reflected in the structure matrix  $\mathbf{Q}$ . A generalized variance can be calculated as the geometric mean of the marginal variance as follows

$$\sigma_{GV}^2(\mathbf{u}) = \exp \left( \frac{1}{n} \sum_{i=1}^n \log \left( \frac{1}{\tau_x} [\mathbf{Q}^-]_{ii} \right) \right) = \frac{1}{\tau_x} \exp \left( \frac{1}{n} \sum_{i=1}^n \log ([\mathbf{Q}^-]_{ii}) \right). \quad (2.60)$$

In order to unify the interpretation of a chosen prior for  $\tau_x$  and make it transferable across domains, the structured effect must be scaled such that  $\sigma_{GV}^2(\mathbf{x}) = \tau_x^{-1}$ . This implies that  $\tau_x$  denotes the accuracy of the (marginal) deviation from a constant level, independent of the underlying graph.

A modification of the BYM model that addresses this scaling problem is the BYM2 model. It uses a scaled structured component  $\mathbf{u}_*$ , where  $\mathbf{Q}_*$  denotes the precision matrix of the Besag model, scaled with the marginal variance  $\sigma_{GV}^2$  as a factor. The random effect is given by

$$\mathbf{x} = \frac{1}{\tau_x} \left( \sqrt{1-\phi} \mathbf{v} + \sqrt{\phi} \mathbf{u}_* \right), \quad (2.61)$$

with covariance matrix

$$\text{Var}(\mathbf{x} | \tau_x, \phi) = \frac{1}{\tau_x} ((1-\phi) \mathbf{I} + \phi \mathbf{Q}_*). \quad (2.62)$$

Equation 2.61 emphasizes the trade-off between pure overdispersion and spatially structured correlation, where  $0 \leq \phi \leq 1$  measures the fraction of the marginal variance explained by the structured effect. For  $\phi = 0$  the model reduces to pure overdispersion, while for  $\phi = 1$  it becomes a Besag model (Martins et al., 2014; Riebler et al., 2016).

## 2.7 Goodness-of-Fit indicators

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

The goodness of fit of a model to available data can be assessed with the help of statistical tests or suitable ratios.

### 2.7.1 The Akaike Information Criterion

The historically oldest criterion was proposed in 1973 by Hirotugu Akaike (1927-2009) as an information criterion and is known today as the Akaike information criterion (AIC). The AIC is one of the most frequently used criteria for model selection in the context of likelihood-based inference.

Let the population contain the distribution of a variable with unknown density function  $p$ . The maximum likelihood estimation assumes a known distribution with an unknown parameter  $\theta$ , hence the density function can be written as  $q(\theta)$ . The Kullback-Leibler divergence is used as a distance measure between  $p$  and  $q(\hat{\theta})$  with  $\hat{\theta}$  the estimated parameter from the maximum likelihood estimation. The better the maximum likelihood model, the small the Kullback-Leibler divergence  $D(P||Q)$ . For a maximum likelihood model with a  $p$ -dimensional parameter vector  $\hat{\theta}$ , the Akaike information criterion is defined as

$$\text{AIC} = -2l(\hat{\theta}_{ML}) + 2p, \quad (2.63)$$

with  $l$  the log-likelihood function (Akaike, 1974).

### 2.7.2 The Deviance Information Criterion

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

information criterion.

The deviance is defined as

$$D(\boldsymbol{\theta}) = -2 \log(l(\mathbf{y}|\boldsymbol{\theta})) + C, \quad (2.64)$$

with  $\mathbf{y}$  the data,  $\boldsymbol{\theta}$  the unknown parameters of the model and  $l$  the likelihood function.  $C$  is a constant that cancels out in all calculations that compare different models and therefore it does not need to be known (Nelder & Wedderburn, 1972).

The DIC is given by

$$\text{DIC} = D(\bar{\boldsymbol{\theta}}) + 2p_D, \quad (2.65)$$

with

$$p_D = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}}), \quad (2.66)$$

where  $\bar{\boldsymbol{\theta}}$  is the expected value of  $\boldsymbol{\theta}$  (Spiegelhalter et al., 2014).

### 2.7.3 The Watanabe-Akaike Information Criterion

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

The WAIC is given by

$$\text{WAIC} = -2\text{LLPD} + 2p_{\text{WAIC}}, \quad (2.67)$$

with the log pointwise predictive density (LLPD) given by

$$\text{LLPD} = \sum_{i=1}^n \log \left( \int \pi(y_i|\boldsymbol{\theta}) \pi_{\text{post}}(\boldsymbol{\theta}) \right) d\boldsymbol{\theta}. \quad (2.68)$$

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

$$p_{\text{WAIC}} = \sum_{i=1}^n \text{Var}_{\text{post}}(\log(\pi(y_i|\boldsymbol{\theta}))), \quad (2.69)$$

where the term represents the variance of the individual terms in the LLPD over all data points (Watanabe & Opper, 2010; Yong, 2018).

## 2.7.4 The Conditional Predictive Ordinate

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

The conditional predictive ordinate is given by

$$\text{CPO} = \pi(y_i | \mathbf{y}_{-i}) \quad (2.70)$$

with  $\mathbf{y}$  the data,  $\mathbf{y}_{-i}$  the data without the  $i$ -th observation, and  $\pi(\cdot | \mathbf{y}_{-i})$  the predictive distribution of a new observation at  $\mathbf{y}_{-i}$ . Low values of CPO are an indication that  $y_i$  is surprising given prior knowledge and the other observations (Cox, 1980; Pettit, 1990).

## 2.8 Model Issues

One problem that plagues these models is that they cannot be directly compared due to their different parameterizations and the fact that the precision in these models is interpreted differently. Since neither a Besag model nor a BYM model nor a Leroux model is scaled, the precision parameter is not representative of the marginal precision but is confounded with the mixing parameter. Therefore, the effect of a prior assigned to the precision parameter is dependent on the graph structure of the application. Thus, a given prior is not transferable between different applications if the underlying graph changes. Furthermore, the goal of the BYM2 model is not to optimize goodness-of-fit indicators, but to provide a meaningful model formulation where all parameters have a clear meaning. By mapping the precision parameter to the marginal standard deviation, the model parameters are flexible and the assignment of meaningful hyperpriors is made easier (Riebler et al., 2016).

Additionally, the goodness-of-fit indicators introduced in Section 2.7 have their own problems. The DIC, for example, produces unreasonable results if the posterior distribution is not well summarized by its mean, while the WAIC is based on a data partition that would create difficulties for structured models, such as for spatial or network data (Gelman et al., 2014).

Finally, the choice of the prior affects the value of these criteria and depending on the values chosen for the PC priors used in this work, overfitting of the models may occur, which is reflected in these criteria, but more on this later.

## 2.9 The Variance Inflation Factor

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

$$\text{VIF}_i = \frac{1}{1 - R^2}, \quad i = 1, \dots, p - 1, \quad (2.71)$$

with  $R^2$  the coefficient of determination. In most literature, a value of at least 5 is suggested as too high and is therefore used as the threshold in this work (Craney & Surles, 2002).

## Analysis of Geospatial Health Data

Healthcare data provides information for detecting public health problems and reacting adequately when they occur. With this information, prevention and control of a multitude of health conditions including infectious diseases, non-communicable diseases, injuries and health-related behaviours can be achieved. To analyse and interpret health data, the process involves a wide variety of system designs, analytical methods, modes of presentation and interpretive uses(Teutsch, Churchill, et al., 2000). Descriptive methods generally form the basis of routine reporting of surveillance data. Rather than focusing on observed patterns in the data, these may attempt to compare the relative occurrence of health outcomes in different subgroups. More specific hypotheses can be explored using inferential methods. The aim of these methods is to draw statistical inferences about patterns or outcomes of health.

The increasing availability of geo-referenced health data, population data, satellite imagery of environmental factors influencing levels of disease activity, and the development of geographic information systems (GIS) and address geocoding software, the rise of studies of spatial and spatio-temporal variation in disease has been facilitated.

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

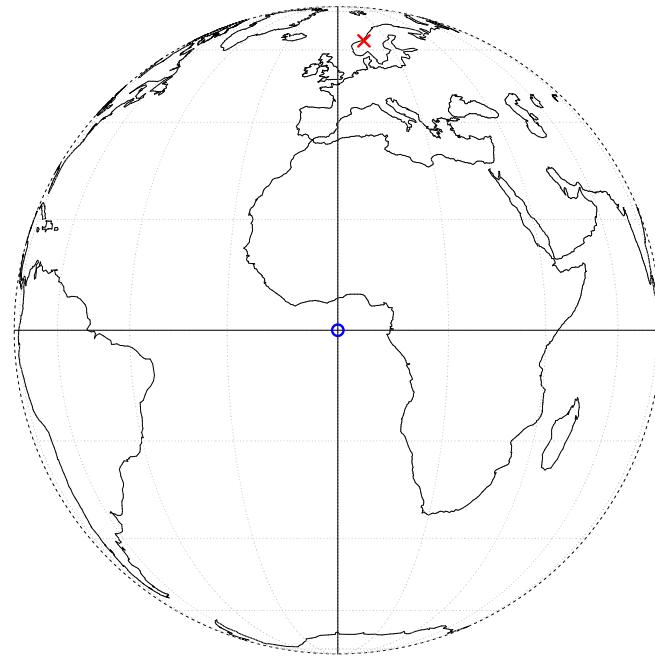
## 3.1 Geographic Data

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 some overlap and both types can be used together or one form can be converted into the other (Lovelace et al., 2019).

### 3.1.1 Vector Data

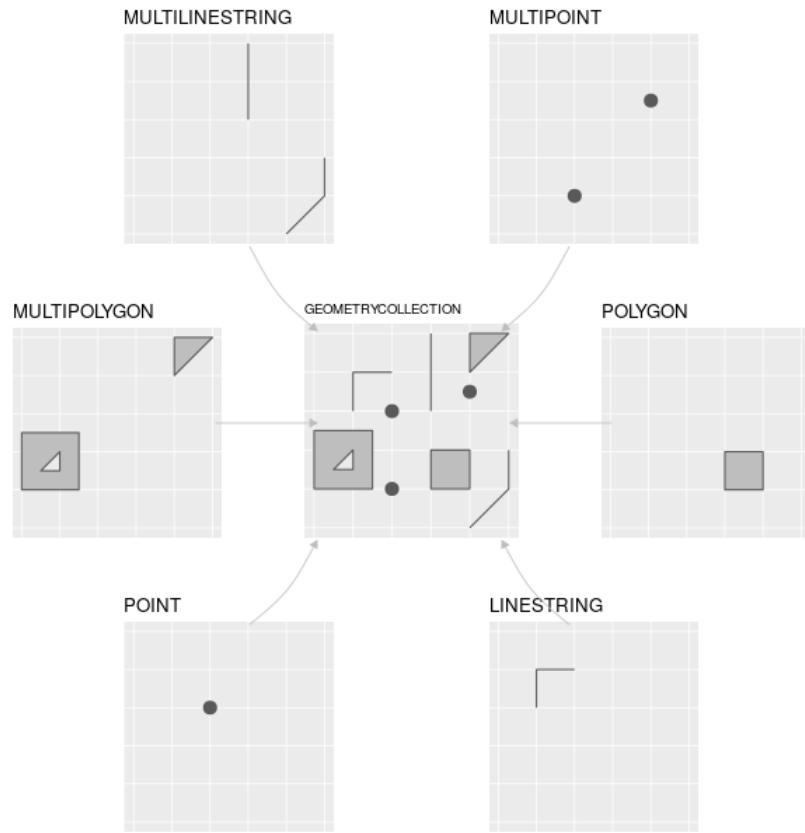
The geographic vector data model is based on points located within a *coordinate reference system* (CRS), in which points either represent self-standing features or form more complex geometric shapes, i.e. lines and polygons. Using this system, Trondheim can be represented by the coordinates (10.4, 63.4), 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 is said about CRS later, but for now it is sufficient to know that it is possible to display coordinates in various ways. An example of a CRS is shown in Figure 3.1.



**Fig. 3.1:** A geographic CRS with an origin at  $0^\circ$  longitude and latitude. The red X denotes the location of Trondheim.

### Different Types of Vector Data

As mentioned earlier, there are different types of vector data. There are 17 different geometry types in the standard *simple features*, but there are seven core types that can be used in most analysis software. These types are visualized in Figure 3.2.



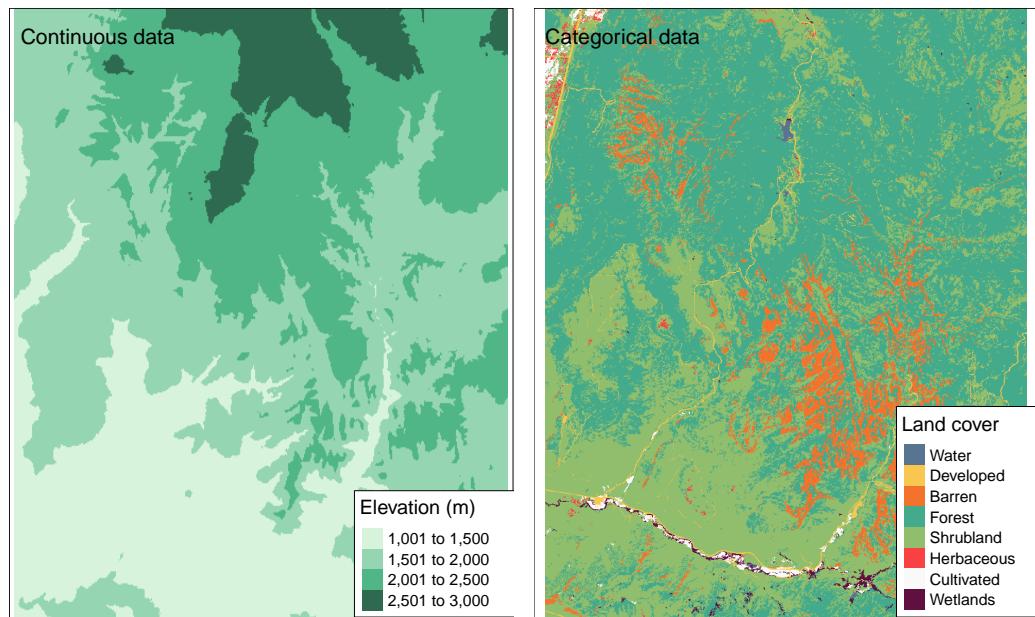
**Fig. 3.2:** The most commonly used simple feature types.

Simple Features was developed by the Open Geospatial Consortium and is an open, standardized, 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 (Lovelace et al., 2019).

### 3.1.2 Raster Data

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

data processing and makes it more efficient, but each raster layer can only contain a single value, which can be either numeric or categorical. Typically, raster maps are used to depict continuous features such as elevation or temperature, but categorical variables, for example soil or land cover, as shown in Figure 3.3 (Lovelace et al., 2019).



**Fig. 3.3:** 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 optimized for any particular location (Lovelace et al., 2019).

## 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 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 minimized 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 (Lovelace et al., 2019).

## 3.2 Modeling and Visualising Health Data

After collecting and cleaning all the data needed to analyse a research question, the next step is the analysis itself. This analysis may involve visualizing the data at hand, for example by visualizing the neighbourhood structure of spatial areas or simply by plotting the locations of all points of interest on a base map. When deciding whether to model spatial dependence, Moran's I can be used to construct a test for spatial correlation between different spatial entities. Another method that can be helpful in deciding whether to use a spatial model is the standardized incidence ratio (SIR), which relates the expected number of cases to the actual number of cases. These two methods are presented in Section 3.2.1.2 and Section 3.2.1.3. If there is a spatial correlation, it is useful to model this spatial effect using models for risk assessment in spatial areas. The way this is done is introduced in Section 3.2.1.4. If data are available over a longer period of time and some variables change their values over time or a temporal trend is present, it may be useful to model the spatial and temporal effect by using spatio-temporal models, which are introduced in Section 3.2.1.5.

### 3.2.1 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 *standardized 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.

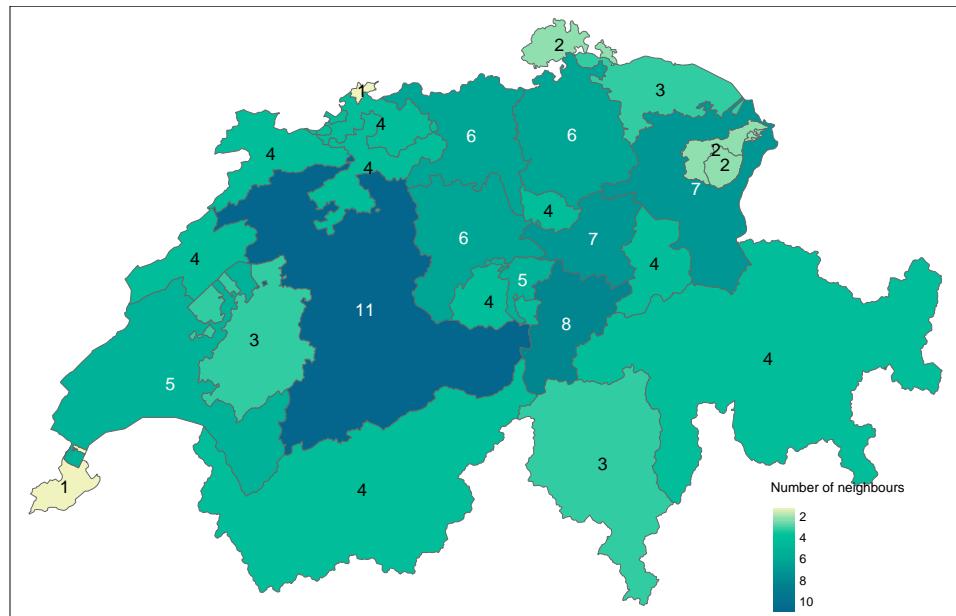
#### 3.2.1.1 Spatial Neighbourhood Matrices

Spatial or proximity matrices are useful for exploratory analysis of area data. Let  $w_{ij}$  denote the  $(i, j)$  element of a *spatial neighbourhood matrix*  $\mathbf{W}$ .  $w_{ij}$  connects the two areas in some spatial way. The neighbourhood structure over the complete study region is defined by  $\mathbf{W}$ , 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

$$w_{ij} = \begin{cases} 1 & \text{if regions } i \text{ and } j \text{ share a border} \\ 0 & \text{else} \end{cases} \quad (3.1)$$

Since a region cannot share a boundary with itself,  $w_{ii} = 0$  (Moraga, 2019). In Figure 3.4, the number of shared borders of each canton in Switzerland are mapped.



**Fig. 3.4:** The number of shared borders of cantons in Switzerland

### 3.2.1.2 Moran's I

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

$$I = \frac{n}{W} \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} (x_i - \bar{x}) (x_j - \bar{x})}{\sum_{i=1}^n}, \quad (3.2)$$

with  $n$  denoting the number of spatial units indexed by  $i$  and  $j$ ,  $x$  the parameter of interest,  $w$  a spatial neighbourhood matrix and  $W$  the sum of all  $w_{ij}$ .

Using Moran's I, a test for spatial autocorrelation can be constructed with the following hypotheses:

$$H_0 : \text{No spatial autocorrelation} \text{ vs. } H_1 : \text{Spatial autocorrelation.} \quad (3.3)$$

Under  $H_0$  the expected value is given by

$$\mathbb{E}[I] = \frac{-1}{n-1}. \quad (3.4)$$

As  $n$  approaches infinity, the expected value therefore approaches 0 (Moran, 1950).

### 3.2.1.3 Standardised Incidence Ratio

A basic measure of disease risk is the *standardized 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

$$\text{SIR}_i = \frac{Y_i}{E_i}. \quad (3.5)$$

$E_i$  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 standardization as

$$E_i = \sum_{j=1}^m r_j^{(s)} n_j^{(i)}, \quad (3.6)$$

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 standardized 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 called the standardized mortality ratio when applied to mortality data (Moraga, 2019).

### 3.2.1.4 Spatial Small Area Disease Risk Estimation

While SIRs may prove useful in some situations, in areas with low population sizes or rare diseases, expected counts may be low, making SIRs insufficiently reliable for

reporting. It is therefore preferable to assess disease risk using models that allow information to be borrowed from neighbouring areas and incorporate information from covariates, thus smoothing or shrinking extreme values due to small sample sizes (Gelfand et al., 2010).

The observed counts  $Y_i$  in area  $i$  are typically modelled with a Poisson distribution with mean  $E_i\theta_i$ , where  $E_i$  is the expected counts and  $\theta_i$  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.  $\theta_i$  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,  $\theta_i = 0.5$ . The general model for spatial data is formulated as follows:

$$Y_i \sim \text{Po}(E_i\theta_i), \quad i = 1, \dots, n, \quad (3.7)$$

$$\log(\theta_i) = \alpha + u_i + v_i. \quad (3.8)$$

The overall risk in the region of study is represented by  $\alpha$ ,  $u_i$  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,  $v_i \sim \mathcal{N}(0, \sigma_v^2)$ . Covariates are often included to measure risk factors and other random effects to deal with different sources of variability. For example,

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

with  $\mathbf{d}_i = (1, d_{i1}, \dots, d_{ip})$  a vector of the intercept and  $p$  covariates corresponding to the area  $i$  and  $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^T$  the vector of coefficients. An increase in  $d_j$  ( $j = 1, \dots, p$ ) by one unit, leads to an increase in the relative risk by a factor of  $\exp(\beta_j)$ , provided that all other covariates remain constant (Moraga, 2019).

### 3.2.1.5 Spatio-Temporal Small Area Disease Risk Estimation

When disease counts are monitored over time, spatio-temporal models are useful as they take into account not only the spatial structure but temporal correlations and spatio-temporal interactions (Martínez-Beneito et al., 2008). Let  $Y_{ij}$  be the counts observed in area  $i$  and at time  $j$ ,  $\theta_{ij}$  be the relative risk,  $E_{ij}$  be the expected number of cases in area  $i$  and at time  $j$ , then

$$Y_{ij} \sim \text{Po}(E_{ij}\theta_{ij}), \quad i = 1, \dots, I, j = 1, \dots, J. \quad (3.9)$$

$\log(\theta_{ij})$  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. (Bernardinelli et al., 1995), for example, propose a spatio-temporal model with parametric time trends that expresses the logarithm of relative risks as

$$\log(\theta_{ij}) = \alpha + u_i + v_i + (\beta + \delta_i) \times t_j. \quad (3.10)$$

The intercept is denoted by  $\alpha$ ,  $u_i + v_i$  is a random area effect,  $\beta$  represents a global linear trend effect and  $\delta_i$  is an interaction between space and time which is the difference between  $\beta$  and the area-specific trend. For modelling  $u_i$  and  $\delta_i$ , a CAR distribution is used and  $v_i$  is iid. 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  $\beta + \delta_i$ .  $\delta_i$  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  $\beta$ . If  $\delta_i \neq 0$ , area  $i$  has a time trend with a slope that is either steeper or less steep than the overall time trend  $\beta$ .

For models that do not demand linearity of the time trend, non-parametric models such as the one proposed by Knorr-Held (Knorr-Held, 2000) can be used. This specific model incorporates spatial effects, temporal random effects and an interaction between space and time as follows:

$$\log(\theta_{ij}) = \alpha + u_i + v_i + \gamma_j + \phi_j + \delta_{ij}. \quad (3.11)$$

The intercept is again denoted by  $\alpha$ ,  $u_i + v_i$  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  $\gamma_j$  follows either a first order random walk in time (RW1)

$$\gamma_j | \gamma_{j-1} \sim \mathcal{N}(\gamma_{j-1}, \sigma_\gamma^2), \quad (3.12)$$

or second order random walk in time (RW2)

$$\gamma_j | \gamma_{j-1}, \gamma_{j-2} \sim \mathcal{N}(2\gamma_{j-1} - \gamma_{j-2}, \sigma_\gamma^2). \quad (3.13)$$

The unstructured temporal effect is given by  $\phi_j \stackrel{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\phi^2)$ . The interaction between space and time,  $\delta_{ij}$ , 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  $(u_i, \gamma_j)$ ,  $(u_i, \phi_j)$ ,  $(v_i, \gamma_j)$  and  $(v_i, \phi_j)$  (Knorr-Held,

2000).

Using the last of these interactions leads to the assumption that there is no spatial or temporal structure on  $\delta_{ij}$ . Thus, the interaction term can be modelled as  $\delta_{ij} \sim \mathcal{N}(0, \sigma_\delta^2)$  (Moraga, 2019).

### 3.2.1.6 Issues With Areal Data

The analysis of spatially aggregated data is subject to the "misaligned data problem" (MIDP), which arises when the data to be analysed is at a different scale from that at which it was collected (Banerjee et al., 2014). 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) (Openshaw, 1984) 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 (Robinson, 2009) 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 (Gotway & Young, 2002; Moraga, 2019).

# Short Introduction to Machine Learning

Machine learning is a generic term for the "artificial" generation of knowledge from experience: an artificial system learns from examples and can generalize these after the learning phase is complete. To do this, machine learning algorithms build a statistical model based on training data. This means that the examples are not simply learned by heart, but patterns and regularities are recognized in the learning data. In this way, the system can also assess unknown data or fail to learn unknown data. This chapter gives a short overview over some commonly used algorithms in machine learning and show how these can be optimized using hyperparameter tuning. A relatively new area of machine learning, interpretable machine learning, and some of the methods used in this area are introduced in the final parts of the chapter.

## 4.1 Common Machine Learning Algorithms

In the following sections, the algorithms that are used to compute the predictive models in Section 6.6 are introduced.

### 4.1.1 K-Nearest Neighbours

The k-nearest neighbour algorithm (knn) is a non-parametric method in which class assignment is performed considering the  $k$  nearest neighbours.

In the simplest case, the classification of an object  $x \in \mathbb{R}^n$  is done by majority vote. The  $k$  nearest already classified objects of  $x$  participate in this voting. To determine which neighbour is closest, many distance measures can be used. Among the most common is the Euclidean distance, defined as

$$d(x, y) = \|x - y\| = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (4.1)$$

(Danielsson, 1980).

Another commonly used distance metric is the Manhattan distance, which defines the distance  $d$  between two points  $x$  and  $y$  as the sum of the absolute difference of the individual coordinates or features,

$$d(x, y) = \sum_{i=1}^n |x_i - y_i| \quad (4.2)$$

(Krause, 1986).

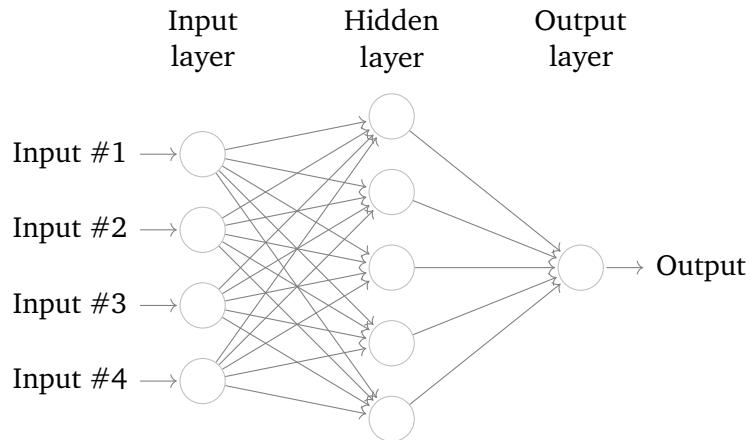
$x$  is assigned to the class that occurs most frequently among the  $k$  neighbours. To avoid ties between two classes, an odd value can be chosen for  $k$ . If the value of  $k$  is small, there is a risk that noise in the training data can lead to worse classification results, while a value that is too large risks including points with a large distance to  $x$  in the decision.

The algorithm can be used for regression problems to estimate continuous variables. While in knn classification the output is the class membership, in knn regression the output is the average value of the  $k$  nearest neighbours (Altman, 1992)

### 4.1.2 Neural Networks

Artificial neural networks, usually referred to simply as neural networks, are computer systems vaguely inspired by the biological neural networks that make up the brains of living things. Neural networks, however, are more about an abstraction (modelling) of information processing, less about replicating biological neural networks and neurons, which is more the subject of computational neuroscience. Motivated by biology, modelling is now so good that many tasks are performed much better than by humans.

In artificial neural networks, topology refers to the structure of the network. This generally means how many artificial neurons are located on how many layers and how they are connected to each other. Artificial neurons can be connected in many ways to form an artificial neural network. In many models, neurons are arranged in layers; a network with only one trainable neuron layer is called a single-layer network. Using a graph, the neurons can be represented as nodes and their connections as edges. The inputs are occasionally represented as nodes. The backmost layer of the network, whose neuron outputs are usually the only ones visible outside the network, is called the output layer. Layers before this are referred to as the hidden layer. Figure 4.1 shows the architecture of a single-layer neural network.

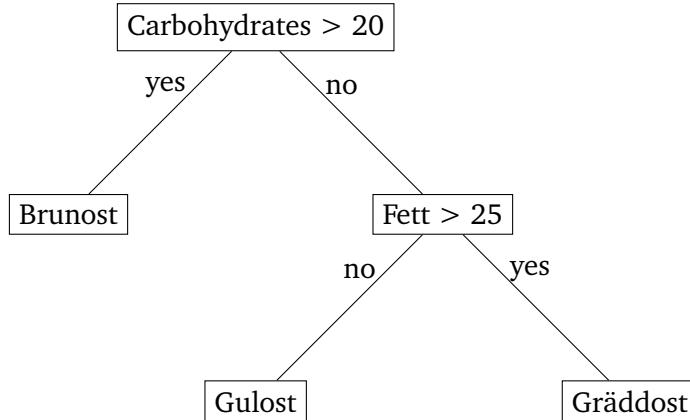


**Fig. 4.1:** A single-layer neural network.

While neural networks are mainly known for their use in deep learning and modelling complex problems such as image recognition, they can be easily adapted for regression problems. In supervised learning, the neural network is given an input pattern and the output produced by the neural network in its current state is compared with the value it is supposed to output. By comparing the target and actual output, it is possible to infer the changes to be made to the network configuration (Ripley, 2007).

#### 4.1.3 Classification and Regression Trees

Classification and regression trees (CART) is an approach to classification or regression problems. A significant feature of the CART algorithm is that only binary trees can be generated, which means that there are always exactly two branches at each node. The central element of this algorithm is therefore finding an optimal binary separation. In the CART algorithm, attribute selection is controlled by maximizing the information content. CARTs are characterized by optimally separating the data in terms of classification. This is achieved with a threshold value that is searched for each attribute. The information content of an attribute is considered high if a classification can be made with a high hit rate by evaluating the attribute characteristics resulting from the division via the threshold values. For the decision trees calculated by the CART algorithm, the following applies: The higher the information content of an attribute in relation to the target variable, the higher up in the tree this attribute is found. Figure 4.2 shows an example of a decision tree.



**Fig. 4.2:** A simple example of a decision tree

In regression analysis, trees are built by a collection of rules based on the available features of the dataset:

- Rules based on the values of the variables are selected to obtain the best split to distinguish the observations based on the dependent variable.
- Once a rule is selected and a node is split into two parts, the same process is applied to each "child" node (i.e. it is a recursive process).
- Splitting stops when the algorithm determines that no further gain can be made or when some preset stop rules are met. (Alternatively, the data is split as much as possible and the tree is pruned later).

Each branch of the tree culminates in a terminal node. Each observation falls into exactly one terminal node, and each terminal node defined by a unique set of rules (Breiman et al., 1984).

#### 4.1.4 Gradient Boosting

CART are considered weak learners, meaning that their predictive performance is only slightly better than chance. Boosting is a method that can be used to convert weak learners into strong learners. Gradient boosting in particular uses CART as a weak learner. Here, each new tree is an adaptation to a modified version of the original tree. After the first tree is grown, the error residuals, defined as the difference between the target value and the predicted target value, are calculated. A new tree is fitted using the error residuals as the target variable while still using the same input variables. The predicted residuals are added to the previous predictions.

This procedure is repeated for the remaining residuals until the loss reaches an acceptable level or no longer improves on an external validation set.

Trees are added one at a time and existing trees in the model are not changed. After calculating the loss, a tree must be added that reduces the loss (i.e. follows the gradient). The mathematical version of the algorithm can be seen in Algorithm 2.

---

**Algorithm 2** The Gradient Boosting Algorithm

---

Given a training set  $\{(x_i, y_i)\}_{i=1}^n$ , a differentiable loss function  $L(y, F(x))$  and the number of iterations  $M$ ,

- 1: Initialization: Compute a model with a constant value:

$$F_0(x) = \arg \min_{\gamma} \sum_{i=1}^n L(y_i, \gamma).$$

- 2: **for** each iteration  $m = 1$  to  $M$  **do**

- 3:     Compute *pseudo-residuals*:

$$r_{im} = - \left[ \frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right]_{F(x)=F_{m-1}(x)}, \quad \text{for } i = 1, \dots, n.$$

- 4:     Fit a weak learner  $h_m(x)$  to the pseudo-residuals, using  $\{(x_i, r_{im})\}_{i=1}^n$  as the training set

- 5:     Compute the multiplier  $\gamma_m$  by solving the following optimization problem:

$$\gamma_m = \arg \min_{\gamma} \sum_{i=1}^n L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)).$$

- 6:     Update the model:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x).$$

- 7: Output:  $F_M(x)$
- 

Gradient boosting of CART produces robust and interpretable models for both regression and classification that achieve high predictive accuracy (Friedman, 2001).

#### 4.1.5 Random Forests

A Random Forest is a classification and regression procedure consisting of several uncorrelated decision trees. All decision trees are grown under a certain type of randomization during the learning process. For a classification, each tree in that forest is allowed to make a decision and the class with the most votes decides the

final classification. Random forests can be used for regression. Random forest uses bagging, a meta-algorithm that can be used to improve the stability and accuracy of machine learning algorithms, for instance CART.

Using  $B$  samples of size  $n$ ,  $B$  models  $F_i(x), i = 1, \dots, B$  are computed. For each  $x$ ,  $B$  predictions  $m_i(x), i = 1, \dots, B$  exist then. The predicted value is given by

$$m^B(x) = \frac{1}{B} \sum_{i=1}^B (m_i(x)). \quad (4.3)$$

This method, as well as random forest, were both developed by Leo Breiman (Breiman, 1996). Random forests differ in only one way from this general scheme: they use a modified tree learning algorithm that selects, at each candidate split in the learning process, a random subset of the features. This process is sometimes called "feature bagging". The reason for doing this is the correlation of the trees in an ordinary bootstrap sample: if one or a few features are strong predictors for the response variable (target output), these features are selected in many of the  $B$  trees, causing them to become correlated. A general overview of the random forest is given in Algorithm 3

---

### Algorithm 3 The Random Forest Algorithm

---

Given a training set  $\{(x_i, y_i)\}_{i=1}^n$ , the number of trees  $B$  and the number of variables  $m$  that should be tried at each split

- 1: Draw  $B$  bootstrap-samples of  $\{(x_i, y_i)\}_{i=1}^n$
  - 2: From the  $M$  features of the training data,  $m \ll M$  features are randomly selected at each node in the tree to be considered as criteria for the cut (split).
  - 3: Each tree is fully grown and not pruned back
- 

To classify an input, it is evaluated in each tree. The class that is chosen most often is the output of the random forest. In the case of regression, the average prediction is used (Breiman, 2001)

## 4.2 Machine Learning Methodology

After introducing some common machine learning algorithms, this section shows how these algorithms can be optimized to achieve the best possible performance with respect to a given performance measure. Even though machine learning models are often referred to as "black boxes", there are methods to make these models more interpretable. These are presented in this section as well.

### 4.2.1 Tuning of Machine Learning Models

In the field of machine learning, hyperparameter optimization, also known as hyperparameter tuning, refers to the search for optimal hyperparameters. A hyperparameter is a parameter that is used to control the training algorithm and whose value, unlike other parameters, must be set before the actual training of the model. There exists a variety of methods when it comes to the algorithms used to explore the hyperparameter space, from simple methods like a grid search to Bayesian optimization or more advanced methods like iterated F-racing.

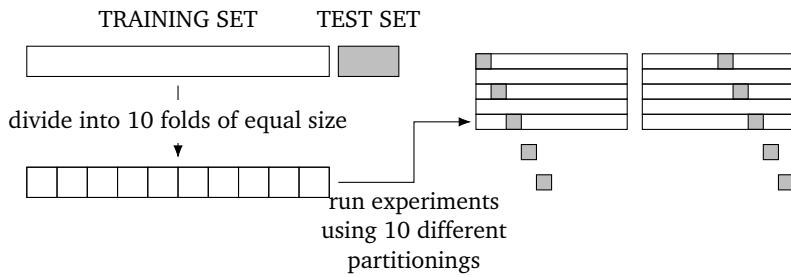
#### Cross-Validation

Cross-validation is a procedure for evaluating the performance of an algorithm in machine learning. Using new datasets that were not used during the training phase, the goodness of the prediction is examined. This is done by partitioning the known dataset into subsets for training and testing the algorithm and the remaining data. Each run of cross-validation involves randomly partitioning the original dataset into a training set and a test set. The training data set is used to train a supervised learning algorithm and the test data set is used to evaluate its performance. This process is repeated several times and the mean cross-validation error is used as a performance indicator. When training a model, it is important not to overfit it with complex algorithms or underfit it with simple algorithms. The choice of training and testing set is critical to reducing this risk. However, it is difficult to split the dataset in a way that maximizes learning and the validity of the test results. This is where cross-validation comes in. To find the best algorithm for the model, cross-validation offers different techniques that split the data differently.

A commonly used cross-validation procedure is k-fold cross-validation. In this method, after splitting the original dataset into a training and a test dataset, the training set is split into  $k$  subsets called folds. Cross-validation iterates through each

fold, using one of the  $k$  folds as the validation set at each iteration, while all the remaining folds are used as the training set. This process is repeated until every fold has been used as a validation set.

Cross-validation helps to select the best performing model by calculating the error using the test set that is not used for training. The test set is used to calculate model accuracy and show how it generalizes with future data (Fushiki, 2011). Figure 4.3 shows an example of 10-fold cross validation.



**Fig. 4.3:** An example of 10-fold cross validation

### Grid and Random Search

The two most commonly used methods for hyperparameter tuning are grid search and random search. Grid search performs an exhaustive search on a manually defined subset of the learning algorithm's hyperparameter space. A grid search must be guided by a performance metric, typically computed by cross-validation on training data or validation data that is not considered during training. For example, for a knn algorithm, a grid search may try any value for  $k$  between 1 and 20 and return the best value for  $k$  with respect to a performance measure. One of the major disadvantages of grid search is that it suffers in terms of dimensionality when the number of hyperparameters grows exponentially. With only four parameters, this problem can become impractical as the number of evaluations required for this strategy increases exponentially with each additional parameter due to the curse of dimensionality.

In the random search, instead of exhaustively trying all combinations, a random selection of values is made within the given hyperparameter space. Unlike grid search, no discretization of the space is required. Random search can outperform grid search in speed and performance, especially when only a few hyperparameters affect the quality of the learning algorithm. This is because grid search tries only a few values for each parameter (but multiple times), while randomly selected values are much better distributed in the search space (Bergstra & Bengio, 2012).

### **F-Race and Iterated F-Racing**

F-Race is a racing algorithm that is used to select the best configuration of parameterized algorithms based on statistical approaches. The main idea is to iteratively evaluate a given finite set of candidates on a stream of instances. After each iteration, some candidate configurations that perform significantly worse than others under the Friedman test with post-hoc analysis for pairs are eliminated and only the remaining ones are evaluated for subsequent iterations. The Friedman test is a statistical test for examining three or more paired samples for equality of the location parameter.

As this process continues, this method focuses more and more on the most promising candidate configurations. An essential part in F-Race is defining the set of candidate configurations of the first step. One way is iterated F-Race, which creates a probability model for a candidate solution. A set of candidates is evaluated at each iteration to update the probability model and steer the next sample towards the better candidate solutions until a termination criterion is met (Birattari et al., 2010).

## **4.2.2 Interpretation of Machine Learning Models**

After creating a model, possibly tuning its hyperparameters and finally training it, the logical next step would be to make predictions for unseen data. The accuracy of the predictions can be evaluated using a performance measure such as the mean absolute error. Beyond that, however, it is often quite difficult to interpret the final model or get an idea of why it predicts a particular value given a set of input variables. This is where interpretable machine learning (IML) comes in, as it aims to shed more light into the black box that is most machine learning algorithms.

### **Feature Importance**

The idea behind feature importance is simple. The importance of a feature is measured by calculating the increase in the model's prediction error after permuting the feature. The more the model error is increased by this permutation, the more important a feature is, as this means that the model relies on the feature for prediction. If the model error does not change, a feature is unimportant because the model ignored the feature for prediction. Algorithm 4 shows how the algorithm works in practice.

---

**Algorithm 4** The Permutation Feature Importance Algorithm

---

Given a trained model  $F$ , a feature matrix  $\mathbf{X}$ , a target vector  $\mathbf{y}$  and a loss function  $L(\mathbf{y}, F)$

- 1: Estimate the original model error  $\varepsilon_{\text{orig}} = L(\mathbf{y}, F(\mathbf{X}))$
  - 2: **for** each feature  $j = 1, \dots, p$  **do**
  - 3:     Generate permuted feature matrix  $\mathbf{X}_{\text{perm}}$  by permuting feature  $j$  in the data  $\mathbf{X}$ , thereby breaking the association between  $j$  and the outcome  $\mathbf{y}$ .
  - 4:     Calculate the permutation error  $\varepsilon_{\text{perm}} = L(\mathbf{y}, f(\mathbf{X}_{\text{perm}}))$  based on the predictions of the permuted data.
  - 5:     Calculate the permutation feature importance  $FI_j = \frac{\varepsilon_{\text{perm}}}{\varepsilon_{\text{orig}}}$ .
  - 6: **Sort** the features by descending  $FI$ .
- 

The advantages of feature importance include that it is easy to interpret, it is comparable across different problems, it accounts for all interactions, and it does not require retraining of the model. On the other hand, there is no clear guideline whether it should be used for training or testing data, the true outcome must be known, and the measurement may be biased if the features are correlated (Fisher et al., 2018; Molnar, 2020).

### Partial Dependence Plots

The partial dependence plot (PDP) shows the marginal effect that one or two features have on the predicted outcome of a machine learning model. A partial dependence plot can reveal whether the relationship between the dependent value and a feature is linear, monotonic or more complex. When applied to a linear regression model, for example, partial dependence plots show a linear relationship every time.

For regression, the partial dependence function is given by

$$\hat{f}_{\mathbf{x}_S}(\mathbf{x}_S) = \mathbb{E}_{\mathbf{x}_C} [\hat{f}(\mathbf{x}_S, \mathbf{x}_C)] = \int \hat{f}(\mathbf{x}_S, \mathbf{x}_C) d\mathbb{P}(\mathbf{x}_C). \quad (4.4)$$

$\mathbf{x}_S$  denotes the features for which the partial dependence function is to be plotted and  $\mathbf{x}_C$  denotes the rest of the features used in the model  $\hat{f}$ . In general, the set  $S$  contains only one or two features and these are the ones for which the effect on the prediction is to be evaluated. The total feature space  $\mathbf{x}$  is composed of the feature vectors  $\mathbf{x}_S$  and  $\mathbf{x}_C$ . By marginalizing the output of  $\hat{f}$  over the distribution of the features in set  $C$ , the function shows the relationship between the features in  $S$  and the predicted outcome. This process produces a function that depends only on the

features in  $S$  and includes interactions with other features.

$\hat{f}_{\mathbf{x}_S}$  is estimated by averaging the training data,

$$\hat{f}_{\mathbf{x}_S}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \hat{f}(\mathbf{x}_S, x_{C,i}). \quad (4.5)$$

For given value(s) of  $S$ , the function returns the average marginal effect on the prediction.  $x_{C,i}$  denotes the feature values from the dataset for the features not of interest and  $n$  denotes the number of observations in the dataset. For a PDP, it is assumed that the features in  $C$  and the features in  $S$  are not correlated. Violation of this assumption leads to improbable or impossible data points for the PDP.

The advantages of PDPs include clear interpretation and intuitiveness of the method, as the partial dependence function at a given feature value represents the average prediction when all data points are forced to assume that feature value.

Disadvantages include the aforementioned assumption of independence and that a maximum of two features can realistically be used, as more than three dimensions are inconceivable to humans (Friedman, 2001; Molnar, 2020).

### Individual Conditional Expectation

The PDP for the average impact of a feature is a global method as it does not focus on specific instances but on a total average. The equivalent of a PDP for individual instances of data is called an individual conditional expectation (ICE) plot. An ICE plot visualizes the dependence of the prediction on a feature for each instance separately, leading to one line per instance, compared to the one line in PDPs. Hence, a PDP represents the average of the lines in an ICE plot. The values relating to a line (and one instance) are calculated by keeping the rest of the features the same, producing variants of that instance by replacing the value of the feature with values from a grid, and making predictions using the model for these newly created instances. This results in a set of points for an instance with the feature value from the grid and the respective predictions.

Since a PDP only displays the average relationship between a feature and the prediction, heterogeneous relationships arising from interactions can be hidden. If the interactions between the features for which the PDP is calculated and the other features are weak, this is not a problem. However, if these interactions are not weak, an ICE plot provides deeper insight.

Formally, for each instance in  $\{(x_{S,i}, x_{C,i})\}_{i=1}^N$ ,  $\hat{f}_{S,i}$  is plotted against  $x_{S,i}$ , where  $x_{C,i}$  remains fixed.

ICE curves are even more intuitive than a PDP, with each line representing the

predictions for an instance when the feature of interest is varied. In addition, they can reveal heterogeneous relationships.

Drawbacks include the fact that only one display can be meaningfully plotted, as plotting multiple lines would require drawing multiple overlapping surfaces that would make it difficult to see anything in the plot. It is difficult to see the average and there may be crowding in the plot if too many lines are drawn (Goldstein et al., 2015; Molnar, 2020)

## Shapley Values

A prediction can be explained with the assumption that each feature value of the instance is a "player" in a game where the prediction is the payout. A fair distribution of the "payout" among the individual features can be obtained using Shapley values - a methodology from coalitional game theory. The Shapley value, is a method that assigns payouts to players depending on how much they contributed to the total payout. Players cooperate in a coalition and as a result receive a certain profit from this coalition.

In machine learning, the "game" is a prediction task for a single observation of the dataset. The difference between the actual prediction of this observation and the average prediction of all instances is the "gain". Feature values of the observation represent the "players" that cooperate to obtain the gain, i.e. to predict a certain value. The Shapley value is thus the average marginal contribution of the value of a feature across all possible coalitions.

The Shapley value is given by a value function of the players in  $S$ . For a single feature value, the Shapley value is its contribution to the payoff, weighted and summed over all possible feature value combinations,

$$\phi_j(\text{val}) = \sum_{S \subseteq \{x_1, \dots, x_p\} \setminus \{x_j\}} \frac{|S|!(p-|S|-1)!}{p!} (\text{val}(S \cup \{x_j\}) - \text{val}(S)). \quad (4.6)$$

$S$  denotes a subset of the features used in the model,  $\mathbf{x}$  is the vector of feature values of the instance to be explained and  $p$  stands for the number of features.  $\text{val}_{\mathbf{x}}(S)$  denotes the prediction for the feature values in  $S$  marginalized over features excluded from  $S$ ,

$$\text{val}_{\mathbf{x}}(S) = \int \hat{f}(x_1, \dots, x_p) d\mathbb{P}_{\mathbf{x} \notin S} - \mathbb{E}_{\mathbf{X}} [\hat{f}(\mathbf{X})]. \quad (4.7)$$

In practice, multiple integrations are performed for each feature that is not included in  $S$ .

The definition of a fair allocation can be viewed as an allocation that has four properties: Efficiency, Symmetry, Dummy and Additivity. The only allocation method that satisfies these four properties is the Shapley value.

**Efficiency:** The contributions of the individual features must add up to the difference between the prediction for  $\mathbf{x}$  and the average prediction,

$$\sum_{j=1}^p \phi_j = \hat{f}(\mathbf{x}) - \mathbb{E}_{\mathbf{X}} [\hat{f}(\mathbf{X})].$$

**Symmetry:** Two characteristic values  $j$  and  $k$  should have the same contribution if their contribution to all possible coalitions is the same. Consequently, if

$$\text{val}(S \cup \{x_j\}) = \text{val}(S \cup \{x_k\}) \quad \forall S \subseteq \{x_1, \dots, x_p\} \setminus \{x_j, x_k\},$$

then

$$\phi_j = \phi_k.$$

**Dummy:** If a feature  $j$  has no influence on the predicted value - regardless of which coalition of feature values it is added to - it should have a Shapley value of 0. Thus if

$$\text{val}(S \cup \{x_j\}) = \text{val}(S) \quad \forall S \subseteq \{x_1, \dots, x_p\},$$

then

$$\phi_j = 0.$$

**Additivity:** For a game that has combined payouts  $\text{val} + \text{val}^+$ , the corresponding Shapley values are

$$\phi_j + \phi_j^+.$$

Suppose a Random Forest has been trained, i.e. the prediction is an average of many decision trees. The additivity property guarantees that for a feature value the Shapley value can be calculated for each tree separately, averaged and thus the Shapley value for the feature value for the Random Forest is obtained.

To calculate the exact Shapley value, all possible coalitions of feature values with and without the  $j$ -th feature must be evaluated. The more features there are in a dataset, the more problematic this calculation becomes, as the number of possible coalitions increases exponentially. An approximation can be achieved by Monte Carlo sampling,

$$\hat{\phi}_j = \frac{1}{M} \sum_{m=1}^M \left( \hat{f}(\mathbf{x}_{+j}^m) - \hat{f}(\mathbf{x}_{-j}^m) \right), \quad (4.8)$$

where  $\hat{f}(\mathbf{x}_{+j}^m)$  denotes the prediction for  $\mathbf{x}$ , but where a random number of feature values are replaced by feature values from a randomly drawn data point  $\mathbf{z}$ , except for the respective value of feature  $j$ .  $\mathbf{x}_{-j}^m$  is almost identical to  $\mathbf{x}_{+j}^m$ , except that the value  $x_j^m$  is also taken from the sample  $\mathbf{z}$ . The calculation of the approximate Shapley value for a single feature value is shown in Algorithm 5

---

**Algorithm 5** The Estimation of Shapley values for a single feature value

---

Given the number of iterations  $M$ , instance of interest  $\mathbf{x}$ , feature index  $j$ , data matrix  $\mathbf{X}$  and machine learning model  $f$

- 1: **for** each feature  $m = 1, \dots, M$  **do**
  - 2:     Draw a random instance  $\mathbf{z}$  from  $\mathbf{X}$ .
  - 3:     Choose random permutation  $o$  of the feature values.
  - 4:     Order  $\mathbf{x} : \mathbf{x}_o = (x_1, \dots, x_j, \dots, x_p)$ .
  - 5:     Order  $\mathbf{z} : \mathbf{z}_o = (z_1, \dots, z_j, \dots, z_p)$ .
  - 6:     Construct two new instances
    - a: With feature  $j : \mathbf{x}_{+j} = (x_1, \dots, x_{j-1}, x_j, z_{j+1}, \dots, z_p)$ .
    - b: Without feature  $j : \mathbf{x}_{-j} = (x_1, \dots, x_{j-1}, z_j, z_{j+1}, \dots, z_p)$ .
  - 7:     Compute the marginal contribution:  $\phi_j^m = \hat{f}(\mathbf{x}_{+j}) - \hat{f}(\mathbf{x}_{-j})$
  - 8:     Compute the Shapley value through averaging:  $\phi_j(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^M \phi_j^m$ .
  - 9:     Output: The Shapley value for the value of the  $j$ -th feature.
- 

This process must be repeated for each of the features to obtain each Shapley value. One of the advantages of Shapley values is that the difference between the prediction and the average prediction is evenly distributed among the feature values of the instance, thus fulfilling the property of efficiency. The Shapley value allows for contrastive explanations. Instead of comparing a prediction to the average prediction of the entire dataset, it can be compared to a subset or even a single data point. Among the disadvantages is that the calculation of Shapley values is computationally intensive, as almost always only an approximation is possible. The Shapley value can also be misinterpreted as the difference in predicted value after removing the feature from model training (Molnar, 2020; Shapley, 1997).

## 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 is the case in this thesis and in the following chapter a brief overview of the data sources used, their pre-processing and how they are combined is given. All the data that is used in this work is openly available, either through government sources, GitHub or the use of R packages.

## 5.1 Covid-19 Data

### 5.1.1 Covid-19 Data for Norway

The Covid-19 data for Norway comes from a dataset made available to the public via a repository on the website GitHub.com, created by the user thohan88. The repository contains a daily updated dataset that is the result of combining several data sources, which include the Institute of Public Health and the Norwegian Directorate of Health. According to the author of the repository, the project is "an open-source effort to make data about the Covid-19 situation in Norway available to the public in a timely and coherent manner" (Hansen, 2020).

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. The number of infections are the cumulative number of infections.

<b>Id</b>	<b>Municipality</b>	<b>Population</b>	<b>Inf. 2020-03-26</b>	<b>Inf. 2020-03-27</b>
1103	Stavanger	143574	87	88
1507	Ålesund	66258	20	20
4601	Bergen	283929	231	248
5001	Trondheim	205163	113	136

### 5.1.2 Covid-19 Data for Germany

In Germany, the Robert Koch Institute publishes daily situation reports in which the number of new cases is published at NUTS 3 level. These reports are available as PDF files via the Institute's website. They can be downloaded and grouped via the R package `covid19germany`(Schmid et al., 2021), as is 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.

<b>Municipality</b>	<b>Date</b>	<b>Cumulative number of infections</b>	<b>Population</b>
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 Vaccination Data

At the end of 2020, the first countries began vaccinating against Covid-19. Vaccination leads to a milder course of the disease, but it is not yet known whether vaccinated people can continue to transmit Covid-19, as of early May 2021. By including the proportion of people vaccinated, an indication as to how much vaccination helps prevent the spread of Covid-19 can potentially be given. These data are available for Norway at the municipality level via the statistics database of the National Institute of Health, FHI (Folkehelseinstituttet, 2020). For Germany, these data are only available at the state level and are therefore not used.

## 5.3 Demographic Data

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

### 5.3.1 Demographic Data for Norway

The demographic data that is collected for Norway comes from Statistisk Sentralbyrå and is made available to the public through their online database, StatBank(Sentralbyrå, 2016).

The first characteristic collected is the age of the population in a given municipality. For each age, starting at 0 and ending at 105, the number of people of that age is known.

Next, unemployment data are collected for a given municipality. For each municipality, the percentage of all people out of work is known, as well as the percentage of all immigrants out of work.

Other data that is collected includes data related to the number of workers in a particular industry, as well as immigration data. Since there is discussion about whether workers from certain industries, in this case construction, contribute to the spread of Covid-19, the decision is made to collect this type of data. For each community, the number of workers across all industries is known, as well as the number of workers in the construction industry. Workers, in this case, are individuals employed in a given municipality who are between the ages of 20 and 66. It is known how many people work full-time and how many people work part-time.

Finally, for immigration data, it is known how many immigrants live in a given municipality and how many Norwegians are born to immigrant parents. These figures are known in terms of the percentage of the population in 2020.

### 5.3.2 Demographic Data for Germany

The demographic data that is collected for Germany comes from the federal and state statistical offices and is made available to the public through their online database, Regionaldatenbank Deutschland (des Bundes und der Länder, 2020).

The first characteristic that is collected is unemployment data at the NUTS 3 level. For each municipality, the number of unemployed people as well as the number of

unemployed foreigners is collected.

Next, data related to the European elections in 2019 are collected. In each municipality, it is known how many people voted in total, how many people voted for the six largest parties, and how many votes the remaining parties received combined. Data is collected in relation to people seeking protection, welfare recipients and in relation to asylum seeker benefits. It is known how many people seek protection in Germany, how many receive social welfare and how many receive asylum seeker benefits.

Finally, trade tax, income tax, and payroll tax data are collected for each municipality.

## 5.4 Shapefiles

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

### 5.4.1 Shapefiles for Norway

The data for the Norwegian shapefiles comes from Geonorge (Geonorge, 2021) and is downloaded from a GitHub repository, as the data there is in a cleaner state (Smistad, 2020). In addition to the geographic shape, the dataset includes a variable that contains the ID of each municipality.

### 5.4.2 Shapefiles for Germany

The data for the German shapefiles comes from Esri Germany (Deutschland, 2020).

## 5.5 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 licence, the Open Database Licence. The core of the project is therefore an openly accessible database of all contributed geoinformation (OpenStreetMap contributors, 2017).

In R, the OpenStreetMap API can be queried using the R package `osmdata` (Padgham et al., 2017). 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 the "biergarten" value is used together with the "amenity" key, the locations of all beer gardens in a given geographic region is downloaded.

OpenStreetMap's 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 are used for this work. If a location is mapped as either POLYGON or MULTIPOLYGON, the centroid of the location is calculated.

A complete list of all key-value pairs used for this work can be found in Section 10.5 in the Appendix.

## 5.6 Government Response and Mobility Data

Our World in Data (OWID) is an online publication that provides information on the historical development of human living conditions. It looks at demographic, developmental economic, geographical and cultural aspects, among others. Our World in Data often takes a historical perspective and provides information on the historical development of humanity's living conditions.

OWID is structured according to problem areas. Each article discusses a global problem - from health problems, to hunger, poverty, war, education, to environmental issues. Depending on the completeness of the entry, these present the historical development of an aspect, the causes and consequences of this development, and the quality of the underlying data. All topics are also presented graphically.

On their website, they provide statistics on governments' policy responses to the coronavirus pandemic. These statistics come from the Oxford Coronavirus Government Response Tracker (OxCGRT), which contains data from public sources collected by a team of over a hundred Oxford University students and staff from around the world. The tracker contains 17 indicators ranging from containment and closure policies, e.g. school closures, to economic policies such as income support and health system policies, e.g. testing regimes. 9 of these indicators are used to calculate a Government Stringency Index, which scales from 0 to 100, with 100 denoting the most stringent government policies (Hale et al., 2020; Ritchie et al., 2020).

Besides government responses, another type of data available through OWID, are the mobility reports by Google. The mobility reports are designed to provide information on what has changed as a result of the regulations to address the Corona crisis. The reports present movement trends broken down by geographic regions and place categories - for example, retail and recreation, grocery shops, parks, public transport stations and stops, places of work and places of residence. To do this, Google measures the number of visitors to these locations each day and compares them to a pre-pandemic base day. A base day represents a normal value for that day of the week, given as a median value over the five-week period from 3 January 2020 to 6 February 2020 (Google, 2020; Ritchie et al., 2020)

## 5.7 Covid-19 Variants Data

The last data source is the open source project CoVariants. The project provides an overview of SARS-CoV-2 variants and mutations of interest. It tracks for different countries the proportion of the total number of sequences (not cases) over time that fall into defined variant groups, such as the B.1.1.7. variant, better known as the UK variant of Covid-19. In addition to the prevalence of the different variants, the project also provides data on the common mutations between the different strains, but this is not of interest in this work (Hodcroft, 2021).

## 5.8 Data Wrangling

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

### 5.8.1 Data Wrangling for Norway

The initial step in creating the final dataset is to convert the data from a wide format, as seen in Table 5.1, to a long format. This is done using the function `melt()` from the R package `reshape2` (Wickham, 2007). The long version of the dataset is shown in Table 5.3.

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

<b>Id</b>	<b>Municipality</b>	<b>Population</b>	<b>Date</b>	<b>Infections</b>
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 is loaded and processed. Since the age data contains the number of people of a certain age, the median age is calculated for each region based on how many people of each age group live in each region.

The other demographic variables are left unchanged. To combine the demographic data with the Covid-19 data, the municipality IDs are extracted using the `str_extract()` function from the `stringr` (Wickham, 2019) R package using the regular expression `[0-9]{4}`. Next, all demographic datasets and the Covid-19 dataset are merged using the `merge()` function.

Using the `st_intersects()` function from the `sf` (Pebesma, 2018) R package, the number of points of interest downloaded via OpenStreetMap is calculated for each municipality. Since the shapefiles contain the ID for each community, these data are merged with the data containing the demographic and Covid-19 data.

For each numeric variable, e.g. the number of schools or the number of employees, this number is scaled.

If there are missing values in the covariates, these values are imputed using the median of the respective variable.

Next, the vaccination data for Norway is loaded. As the daily number of vaccinated persons for each municipality is included in the data, these numbers only need to

be cumulated before being merged with the rest of the data based on municipality name and date.

Finally, seven new variables are created:

1. Expected count, which is the expected number of cases in each municipality.
2. SIR, which is the standardized incidence ratio in each municipality.
3. An area ID, which is a unique ID given to each municipality.
4. Higher education, which counts the number of universities and colleges in a given area.
5. Sex, which gives the proportion of females living in a given area.
6. Population density, i.e. the number of people per square kilometre in a given area.
7. Urban density, i.e. the number of residential buildings per square kilometre in a given area.

The final dataset contains the variables shown in Table 5.4.

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

Variable Name	Explanation	Scale
Id	The municipality ID	None
Municipality	The municipality name	None
Population	Population in a municipality	None
Date	The date of the data used	None
Infections	The number of infected people	None
Median age	The median age	scaled
Total unemployment	The proportion of unemployed people	scaled
Unemployed immigrants	The proportion of unemployed immigrants	scaled
Full-time workers	The number of full-time workers	scaled
Part-time workers	The number of part-time workers	scaled
Full-time construction	The number of full-time construction workers	scaled
Part-time construction	The number of part-time construction workers	scaled
Total immigrants	The proportion of immigrants	scaled
Marketplace	The number of marketplaces	scaled
Entertainment	The number of entertainment venues	scaled
Sport	The number of sports amenities	scaled
Clinic	The number of clinics	scaled
Hairdresser	The number of hairdresser	scaled
Shops	The number of shops	scaled
Place of worship	The number of places of worship	scaled
Retail	The number of retail stores	scaled
Nursing home	The number of nursing homes	scaled
Restaurant	The number of restaurants	scaled
Aerodrome	The number of aerodromes	scaled
Office	The number of offices	scaled
Platform	The number of public transport platforms	scaled
Kindergarten	The number of kindergartens	scaled
Schools	The number of schools	scaled
Bakeries	The number of bakeries	scaled
Residential	The number of residential buildings	None
Higher education	The number of colleges and universities	scaled
Expected count	The expected number of infections	None
SIR	The standardized incidence ratio	None
Area Id	A unique ID	None
Area	The area in km <sup>2</sup>	None
Population density	People per km <sup>2</sup>	scaled
Urban density	Residential buildings per km <sup>2</sup>	scaled
Sex	The proportion of females	scaled
Vaccinations	The proportion of vaccinated	scaled

## 5.8.2 Data Wrangling for Germany

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

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

Variable Name	Explanation	Scale
Id	The municipality ID	None
Municipality	The municipality name	None
Population	Population in a municipality	None
Date	The date of the data used	None
Infections	The number of infected people	None
Logarithmic trade tax	The trade tax in Euros	scaled
Logarithmic income tax	The income tax in Euros	scaled
Logarithmic total income	The income and payroll tax in Euros	scaled
Asyl benefits	The number of people receiving asylum seeker benefits	scaled
Welfare recipients	The number of welfare recipients	scaled
Unemployed people	The number of unemployed people	scaled
Unemployed foreigners	The number of unemployed foreigners	scaled
Protection seekers	The number of protection seekers	scaled
Die Union	Percentage of vote for Union	scaled
SPD	Percentage of vote for SPD	scaled
Greens	Percentage of vote for the Greens	scaled
FDP	Percentage of vote for FDP	scaled
The left	Percentage of vote for the left	scaled
AfD	Percentage of vote for AfD	scaled
Marketplace	The number of marketplaces	scaled
Entertainment	The number of entertainment venues	scaled
Sport	The number of sports amenities	scaled
Clinic	The number of clinics	scaled
Hairdresser	The number of hairdresser	scaled
Shops	The number of shops	scaled
Place of worship	The number of places of worship	scaled
Retail	The number of retail stores	scaled
Nursing home	The number of nursing homes	scaled
Restaurant	The number of restaurants	scaled
Aerodrome	The number of aerodromes	scaled
Office	The number of offices	scaled
Platform	The number of public transport platforms	scaled
Kindergarten	The number of kindergartens	scaled
Schools	The number of schools	scaled
Bakeries	The number of bakeries	scaled
Residential	The number of residential buildings	None
Higher education	The number of colleges and universities	scaled
Expected count	The expected number of infections	None
SIR	The standardized incidence ratio	None
Area Id	A unique ID	None
Area	The area in $\text{km}^2$	None
Population density	People per $\text{km}^2$	scaled
Urban density	Residential buildings per $\text{km}^2$	scaled
Sex	The proportion of females	scaled

### 5.8.3 Data Wrangling for the Temporal Models

Obtaining data for the temporal models is simple. OWID provides a dataset that includes case numbers for over 200 countries, along with vaccination numbers for that country, among other variables. Next, all data related to government response and mobility data are loaded. As these data are not available over the same time period for each country, some assumptions are made to minimize missing data. These assumptions are

- If vaccination data before the administration of the first vaccine dose in a country are missing, no people were vaccinated at these points in time.
- If vaccination data are missing for time points after the last available data, the vaccination rate has remained the same.
- If government policy data are missing before the first recorded government response to a particular policy, then the policy is "No response".
- If government policy data are missing after the last tracked government response to a particular policy, then the policy has remained the same since that time.
- If mobility data is missing between points in time, then a constant decline / slope is assumed for this data.

After imputing all missing values, the next step is to merge these data. This can be done simply by using the unique combinations between the date and the country variable. In the last step, instead of using the infection figures provided by OWID, which mostly come from Johns Hopkins University, the infection figures that are also used for the spatial models are used, as these come from official government sources.

The final dataset contains the variables shown in Table 5.6.

**Tab. 5.6:** The variables contained in the final dataset.

Variable Name	Explanation	Scale
Country Code	The iso2 code of a country	None
Country	The name of a country	None
Population	Population in a municipality	None
Date	The date of the data used	None
Infections	The number of infected people	None
Mobility retail & recreation	The change in mobility in retail & recreation	scaled
Mobility grocery & pharmacies	The change in mobility in groceries & pharmacies	scaled
Mobility residential	The change in mobility in residential areas	scaled
Mobility transport stations	The change in mobility at public transport areas	scaled
Mobility parks	The change in mobility in parks	scaled
Mobility workplaces	The change in mobility in workplaces	scaled
Testing policy	Policies implemented related to testing	factor
Contact tracing	Policies implemented related to contact tracing	factor
Vaccination policy	Policies implemented related to vaccination	factor
Facial coverings policy	Policies implemented related to facial coverings	factor
Income support	Policies implemented related to income support	factor
Restrictions on internal movement	Policies implemented related to the restriction on internal movement	factor
International travel controls	Policies implemented related to international travel controls	factor
Public information campaigns	Public information campaigns on Covid-19	factor
Cancellation of public events	Policies implemented related to the cancellation of public events	factor
Restriction of gatherings	Policies implemented related to the restriction of gatherings	factor
Closing of public transport	Policies implemented related to the closing of public transport	factor
Closing of schools	Policies implemented related to the closing of schools	factor
Closing of workplaces	Policies implemented related to the closing of workplaces	factor
Stay home requirements	Policies implemented related to stay home requirements	factor
Stringency index	The government stringency index	scaled
People vaccinated per hundred	The proportion of people who have received at least 1 dose of a vaccine	scaled
People fully vaccinated per hundred	The proportion of fully vaccinated people	scaled
Variant 20E	The proportion of the total number of sequences of the 20E variant	scaled
Variant 20L	The proportion of the total number of sequences of the 20L variant	scaled
Other variants	The proportion of the total number of sequences that are not tracked	scaled
Season	The season of the date	factor
Expected count	The expected number of infections	None

# Data Analysis

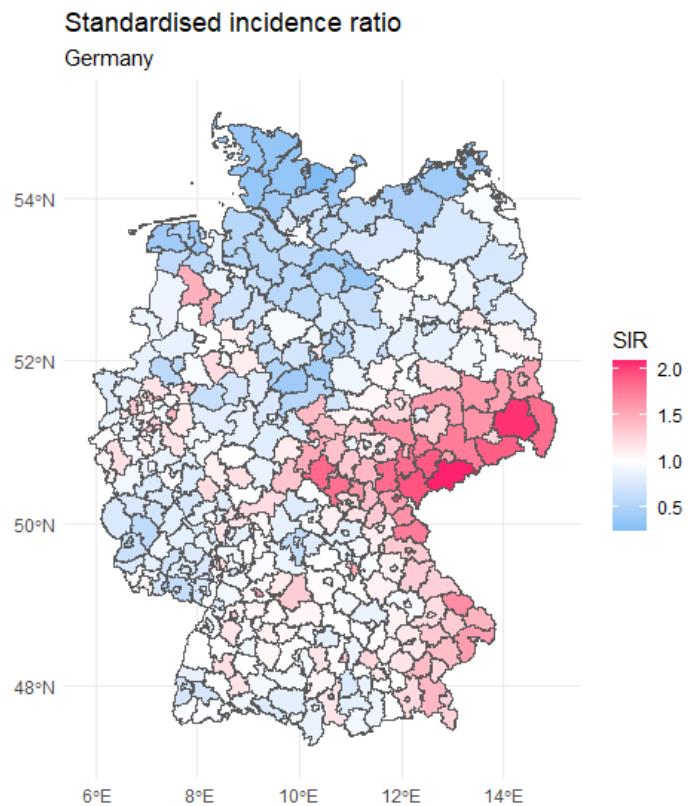
This chapter focuses on the analysis of the research questions. First, a look is taken at the standardized incidence ratio in each country. Next, the relationship between the number of infections and several factors of interest is analysed. For this analysis, a probability must be found that fits the data reasonably well. One way to do this is presented in Section 6.2.1. Thereafter, two different types of models are calculated; models without a spatial component and models with a spatial component. Section 6.3 contains the results of the models without a spatial component and Section 6.4 contains the results of the models with the spatial component. Next, it is analysed how the performance of the spatial models changes when changing the value for the standard deviation  $\sigma_0$  in Equation 2.12. Next, predictive models calculated using the methodology described in chapter 4 are shown and compared with the spatial models. Finally, a look is taken at temporal models in Section 6.7.

## 6.1 Standardised Incidence Ratio (SIR)

This section takes a brief look at the SIR for the countries of interest. Recall from Equation 3.5, that the SIR is defined as the ratio of observed counts to expected counts.

### 6.1.1 SIR for Germany

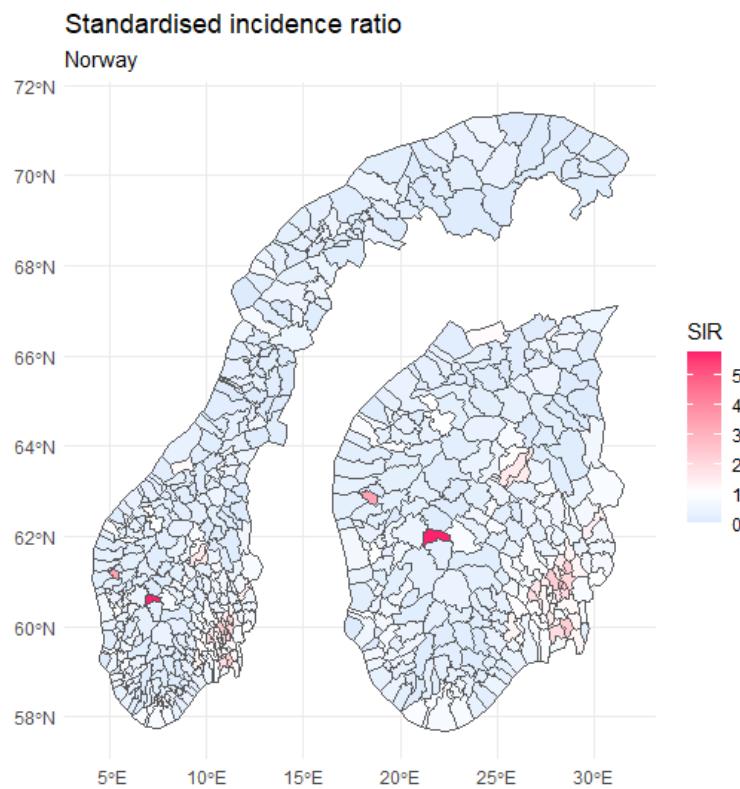
When looking at the SIR for Germany in Figure 6.1, it is noticeable that the actual number of infections in the eastern parts of Germany, especially in Saxony, is considerably higher than the expected number of infections. Furthermore, parts of Bavaria have an increased SIR compared to the rest of Germany,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



**Fig. 6.1:** The SIR for Germany based on the data of the 2nd of May 2021

## 6.1.2 SIR for Norway

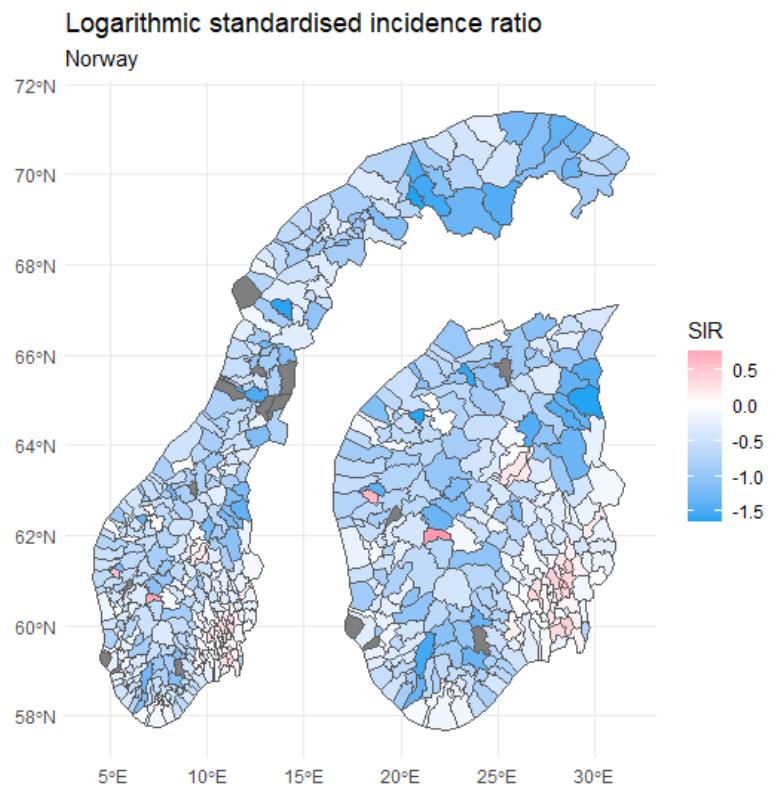
Looking at the standardized incidence ratio for Norway in Figure 6.2, a standardized incidence ratio of less than 1 can be seen for most municipalities north of Trondheim. In the southern parts of Norway there are several municipalities with a ratio above 1, for example the standardized incidence ratio around the capital Oslo is around 2. However, the two small municipalities, Hyllestad and Ulvik, have the highest standardized incidence ratio in Norway. In Hyllestad, 95 of 1328 people are infected with Covid-19 so far, while in Ulvik, 134 of 1080 people are infected so far. The SIR in Hyllestad is around 3.4, following an outbreak in a shipyard in autumn 2020 (Korsvoll, 2020), while Ulvik has a ratio of around 5.7, 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 (NTB, 2021).



**Fig. 6.2:** The SIR for Norway based on the data of the 2nd of May 2021

Because the high numbers from two small municipalities complicate the interpretation of Figure 6.2, Figure 6.3 shows the SIR on a log10 scale. On this scale, a value of 0 means that the risk of infection in a given municipality is neither lower

nor higher. Values below 0 mean that the risk of infection in a municipality is lower than average, while values above 1 mean that the risk of infection in a municipality is higher than average. It is now clearer that the standardized incidence ratio is below 1 in most parts of Norway, but that there is a higher risk in the region around Oslo.



**Fig. 6.3:** The  $\log_{10}$  SIR for Norway based on the data of the 2nd of May 2021

## 6.2 Data Modelling

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

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

All of these models are computed using the INLA (Bivand et al., 2015) R package. The measures introduced in Section 2.7, namely the DIC, the WAIC, the CPO and the mean absolute error (MAE), are used to compare the models.

In addition to specifying what type of spatial model to use, if any, there is the option of specifying a prior.

As can be seen in Section 2.3.2, a pc prior can be specified for the precision parameter  $\tau$ , which is what is done here.

For the parameters  $\sigma_0$  and  $\alpha$  in Equation 2.13 the values 1 and 0.01 are chosen.

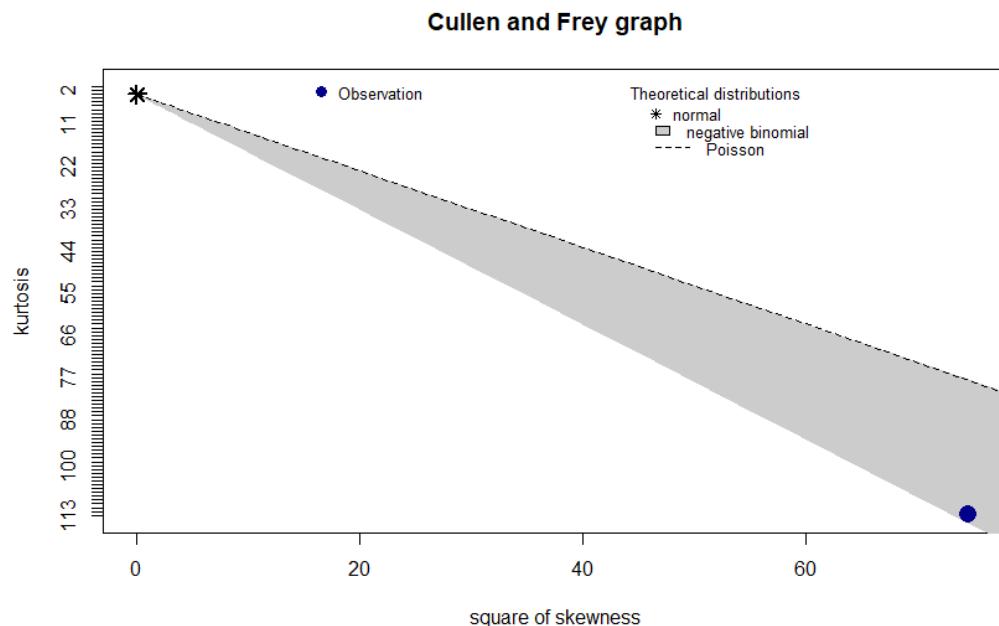
The models are compared using the mean absolute error. For this, 20% of the observations are removed from the training set and used for testing instead. The predicted number of infections for these municipalities is compared to the actual numbers.

### 6.2.1 Choice of Likelihood

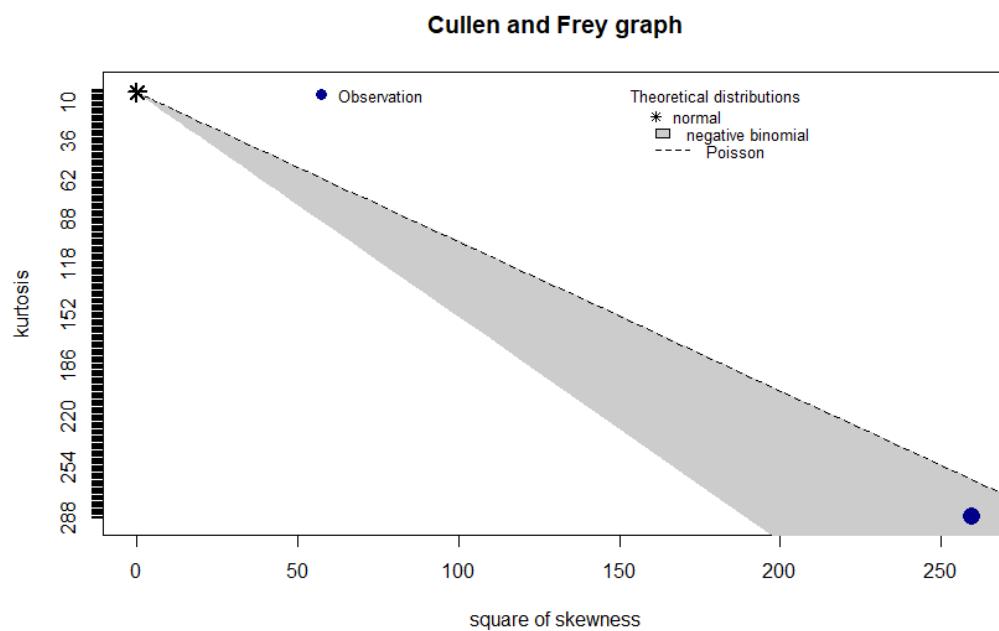
Before the models are computed, however, the distribution that fits the number of cases must first be found. One way to do this, the function `descdist()` from the `fitdistrplus` R package is used. The Cullen and Frey graph illustrates how "close" a sample is to a theoretical distribution based on the kurtosis and the square of the skewness, defined in Equation 10.18 and Equation 10.17. It can be used to get a preliminary idea of which distributions fit the data, in this case the number of infections, reasonably well.

The plots for Germany and Norway can be seen in Figure 6.4 and Figure 6.5. The blue dot represents the data, the star a theoretical normal distribution, the dashed line a theoretical Poisson distribution and the grey area a theoretical negative binomial distribution. In both cases, the blue dot is relatively far from the star and

lies in the region of a negative binomial distribution. For the Norwegian sample shown in Figure 6.5, the sample is closer to a Poisson distribution than is the case for the German sample in Figure 6.4.



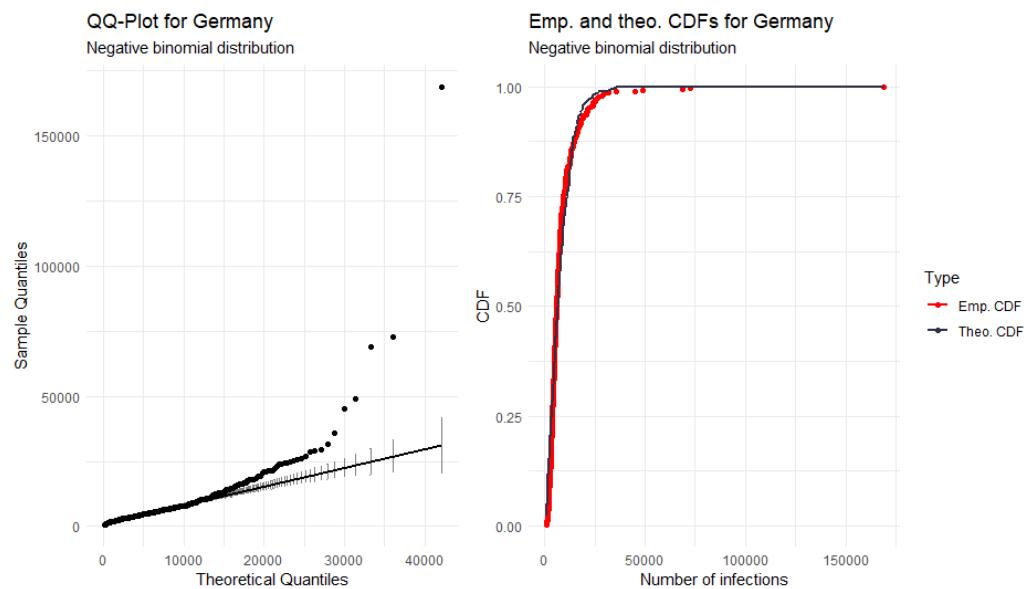
**Fig. 6.4:** The Cullen and Frey graph for Germany



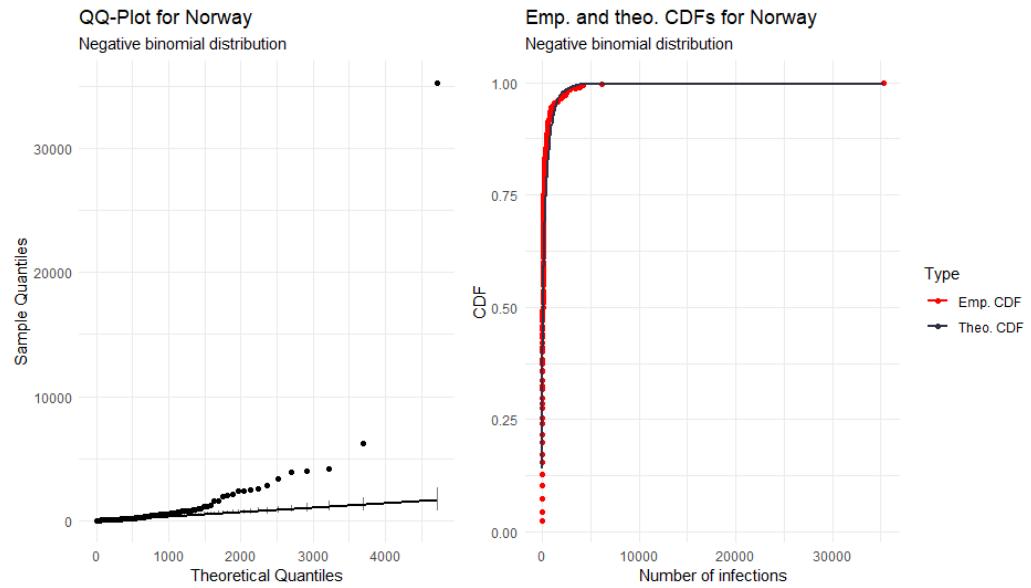
**Fig. 6.5:** The Cullen and Frey graph for Norway

Next, a negative binomial distribution, a normal distribution, and a Poisson distribution are fitted to the data using the maximum likelihood method. The negative binomial fits for both countries can be seen in Figure 6.6 and Figure 6.7. The fits for the normal and Poisson distribution for both countries, are shown in the Appendix in Figure 10.1, Figure 10.2, Figure 10.4 and Figure 10.5.

The QQ-plot for Germany and Norway looks quite similar, as there appears to be a linear relationship between the theoretical quantile and the sample quantiles, up to a certain point where the sample quantiles have a higher value than the theoretical quantiles, indicating that the distribution is right skewed. Since there are many municipalities with relatively few cases and few municipalities with many cases, this is to be expected. It can be seen that the empirical cumulative density function closely follows the theoretical cumulative density function.



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



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

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

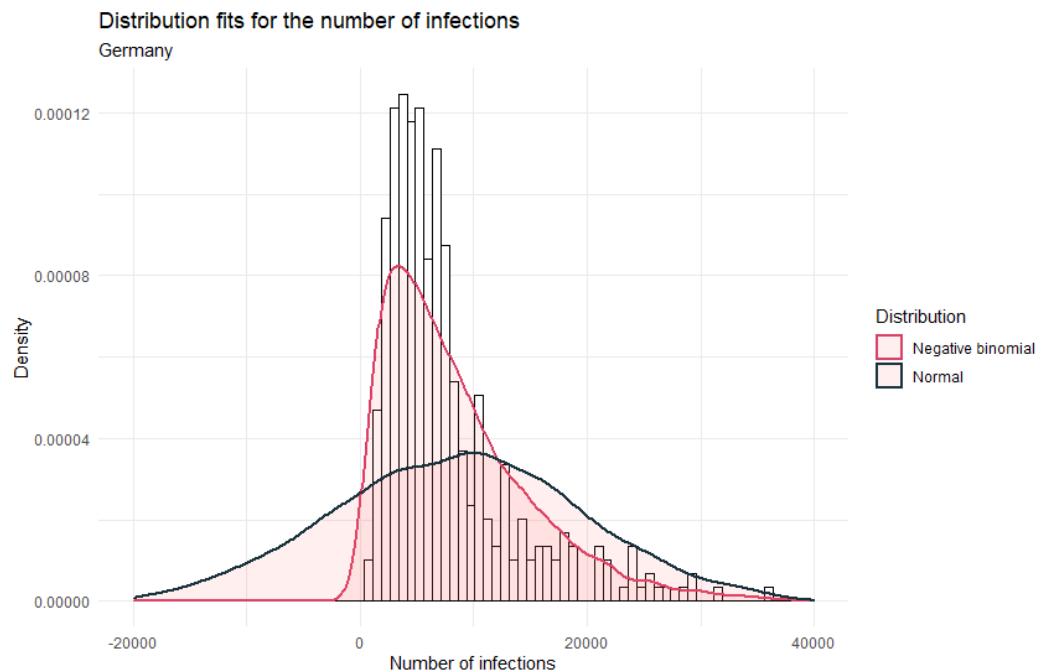
**Tab. 6.1:** The AIC for different distributions for Germany and Norway

Country	Distribution	AIC
Germany	Normal	8619
Germany	Poisson	2699357
Germany	Negative Binomial	8010
Norway	Normal	6400
Norway	Poisson	509139
Norway	Negative Binomial	4293

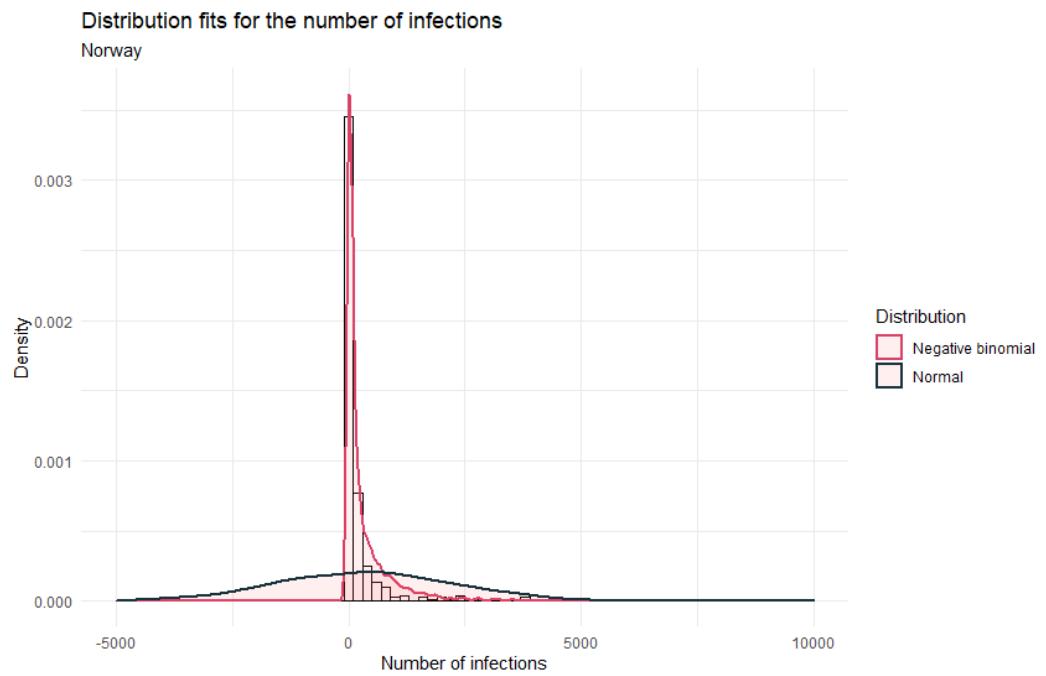
The poor fit for the Poisson distribution can be explained by looking at the range of the number of confirmed cases in a given municipality. For Germany, this number ranges from 697 to 169021 (as of May 2, 2021), while for Norway, the number ranges from 0 to 34654 (as of May 2, 2021). This results in a mean and standard deviation for Germany of 8550 and 11204, respectively. For Norway, the values for these metrics are 326 and 1930. This is problematic because, as shown in Equation 10.12 and Equation 10.13, for a Poisson distribution the expected value

and the variance should be equal.

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



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



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

## 6.3 Models without a Spatial Component

To establish a baseline, a look is first taken at models that do not include a spatial effect. This way, it can be observed how the means and credibility intervals of the covariates change when a spatial effect is added to a model and how the performance of the model changes with respect to the goodness-of-fit indicators introduced in Section 2.7.

Before computing the models, using the VIF introduced in Section 2.9, predictors are removed if their VIF is above 5. To do this, a GLM is first run on all variables, followed by the calculation of VIF. Then the variable with the highest value is removed before running the model again with all remaining variables. This process is repeated until only variables with a VIF of less than 5 remain.

In sections 6.3.1 and 6.3.2 the results of these models are presented. Their goodness-of-fit indicators as well as the coefficients together with their credibility intervals calculated as in Section 2.2.4 are reported.

These models are based on data from 2 May 2021, when 3,428,487 people are infected with Covid-19 in Germany, while 87,537 people are infected in Norway. The five municipalities with the most infections in Germany are shown in Table 6.2 and in Table 6.3 for Norway.

**Tab. 6.2:** The German municipalities with the most infections as of May 2nd 2021.

Municipality	Population	Number of infections
SK Berlin	3644826	169021
SK Hamburg	1841179	72595
SK Munich	1471508	68762
SK Cologne	1085664	48831
Region Hannover	1157624	45043

**Tab. 6.3:** The Norwegian municipalities with the most infections as of May 2nd 2021.

Municipality	Population	Number of infections	% first vaccine shot
Oslo	693494	34654	32.7%
Bergen	283929	6039	30.0%
Drammen	101386	4097	34.1%
Bærum	127731	3860	31.7%
Lillestrøm	85983	3819	30.2%

### 6.3.1 Models without a Spatial Component for Germany

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

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

DIC	WAIC	CPO	MAE <sub>train</sub>	MAE <sub>test</sub>
5593	5596	-2814	1434	1284

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

Variable	mean <sub>p</sub>	exp(mean <sub>p</sub> )	exp(q0025p)	exp(q0975p)	sig.
(Intercept)	-0.041	0.960	0.937	0.985	*
AfD	0.172	1.190	1.057	1.335	*
Population density	0.164	1.178	1.127	1.232	*
Logarithmic trade tax	0.082	1.086	1.039	1.134	*
Platform	0.037	1.038	0.988	1.089	
Die Union	0.026	1.029	0.893	1.180	
Higher Education	0.018	1.018	0.981	1.058	
Sex	0.017	1.018	0.986	1.050	
Urban density	0.009	1.009	0.973	1.047	
FDP	0.005	1.005	0.971	1.041	
Place of worship	-0.010	0.991	0.951	1.031	
Clinic	-0.011	0.990	0.942	1.041	
Nursing home	-0.013	0.987	0.958	1.018	
Office	-0.031	0.970	0.929	1.013	
Marketplace	-0.044	0.957	0.905	1.012	
SPD	-0.091	0.914	0.847	0.984	*
The left	-0.119	0.888	0.822	0.958	*
Greens	-0.156	0.857	0.753	0.970	*

### 6.3.2 Models without a Spatial Component for Norway

Table 6.6 contains the performance measures for the baseline model for Germany, while Table 6.7 contains the posterior mean, the exponentiated posterior mean and the credibility intervals of the coefficients. It can be seen that the intercept as well as four of the coefficients are significant.

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

DIC	WAIC	CPO	MAE <sub>train</sub>	MAE <sub>test</sub>
2859	2864	-703	229	92

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

Variable	mean <sub>p</sub>	exp(mean <sub>p</sub> )	exp(q0025 <sub>p</sub> )	exp(q0975 <sub>p</sub> )	sig.
(Intercept)	-0.809	0.446	0.409	0.486	*
Total immigrants	0.237	1.270	1.112	1.446	*
Unemployed immigrants	0.233	1.266	1.084	1.474	*
Urban density	0.181	1.202	1.037	1.408	*
Vaccinations	0.055	1.058	0.947	1.179	
Marketplace	0.043	1.045	0.954	1.155	
Total unemployment	0.037	1.043	0.852	1.269	
Platform	0.035	1.038	0.903	1.196	
Nursing home	0.006	1.007	0.930	1.106	
Higher education	0.004	1.005	0.928	1.105	
Median age	-0.028	0.974	0.864	1.093	
Place of worship	-0.035	0.968	0.840	1.118	
Office	-0.146	0.867	0.741	1.013	
Sex	-0.175	0.840	0.748	0.941	*

## 6.4 Spatial Models

Looking at the SIR value for Germany in Figure 6.1 and the SIR value for Norway in Figure 6.2 and Figure 6.3, it clearly looks like there is a correlation between the SIR and the spatial units. For Germany, the SIR is higher in Eastern Germany than in the rest of the country and in Norway the risk seems to be mainly concentrated around the Oslo region.

To check whether there is spatial autocorrelation, Moran's I, introduced in Section 3.2.1.2, can be calculated and the Moran test can be used to tell whether there is spatial autocorrelation. Under the null hypothesis of no spatial autocorrelation, a p-value greater than 0.05 would be expected. The results of the test are presented in Table 6.8. Looking at the p-value for both countries, it can be seen that the number of infections in a municipality and the spatial units are correlated.

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

Country	Moran's I	$\mathbb{E}[I]$	p-Value
Germany	0.110	-0.003	< 0.01
Norway	0.110	-0.003	< 0.01

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

### 6.4.1 Spatial Models for Germany

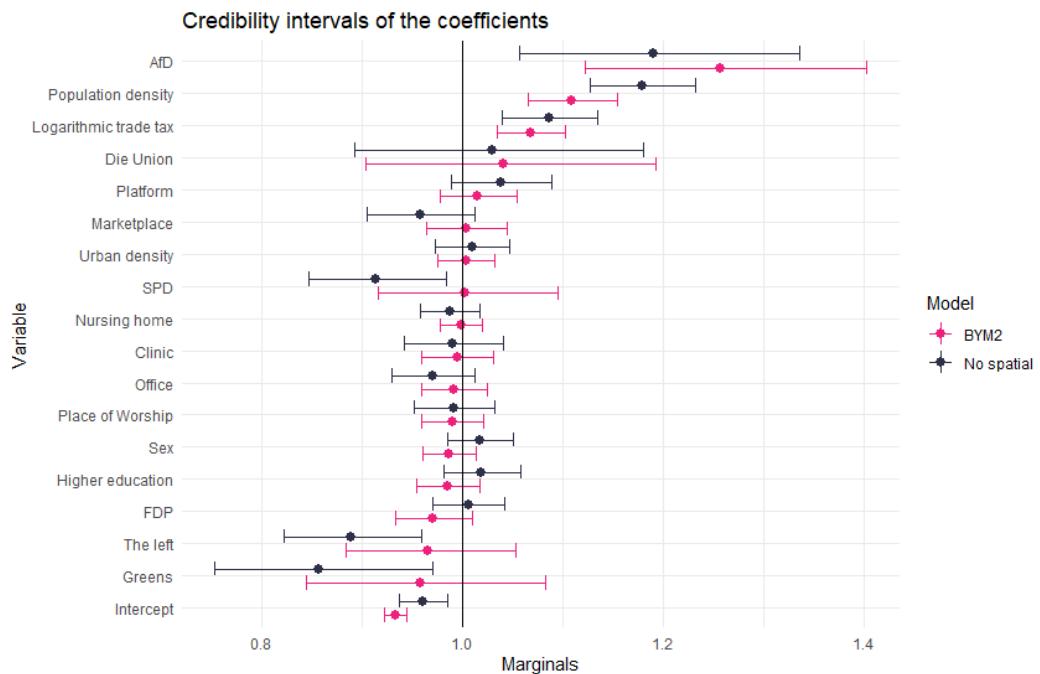
Looking at the performance of the spatial models and the model with the spatial component shown in Table 6.9, it can be seen that the spatial models perform better in terms of the DIC, WAIC and MAE, while they perform equally well or better in terms of the CPO.

The best performance of all models, in terms of MAE, is observed for the Besag model, just ahead of the BYM2 model.

**Tab. 6.9:** The performance measures for the best performing model of each type.

Model	DIC	WAIC	CPO	MAE <sub>train</sub>	MAE <sub>test</sub>
No spatial	5593	5596	-2814	1434	1284
Besag	4842	4897	-2784	143	1027
BYM2	4747	4806	-2781	103	1043
Leroux	4830	4843	-2966	88	1264

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



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

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

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

Variable	mean <sub>p</sub>	exp(mean <sub>p</sub> )	exp(q0025 <sub>p</sub> )	exp(q0975 <sub>p</sub> )	sig.
(Intercept)	-0.070	0.932	0.922	0.944	*
AfD	0.227	1.257	1.122	1.402	*
Population density	0.103	1.109	1.065	1.154	*
Logarithmic trade tax	0.066	1.068	1.035	1.102	*
Die Union	0.037	1.041	0.904	1.192	
Platform	0.015	1.015	0.977	1.054	
Marketplace	0.004	1.004	0.964	1.045	
Urban density	0.003	1.003	0.974	1.032	
SPD	0.001	1.002	0.915	1.095	
Nursing Home	-0.002	0.998	0.977	1.020	
Clinic	-0.006	0.994	0.959	1.031	
Office	-0.008	0.992	0.960	1.025	
Place of Worship	-0.011	0.989	0.959	1.021	
Sex	-0.014	0.987	0.960	1.014	
Higher Education	-0.015	0.985	0.954	1.017	
FDP	-0.030	0.971	0.933	1.009	
The left	-0.036	0.966	0.884	1.053	
Greens	-0.045	0.958	0.844	1.082	

For the hyperparameters, a value of 19.75 is reported for the precision and a value of 0.929 for  $\phi$ . Hence, 92.9% of the marginal variance is explained by the structured effect. Therefore, this model is far from reducing to pure overdispersion and comes close to a Besag model, which is reflected in the similar values of the goodness-of-fit indicators in Table 6.9.

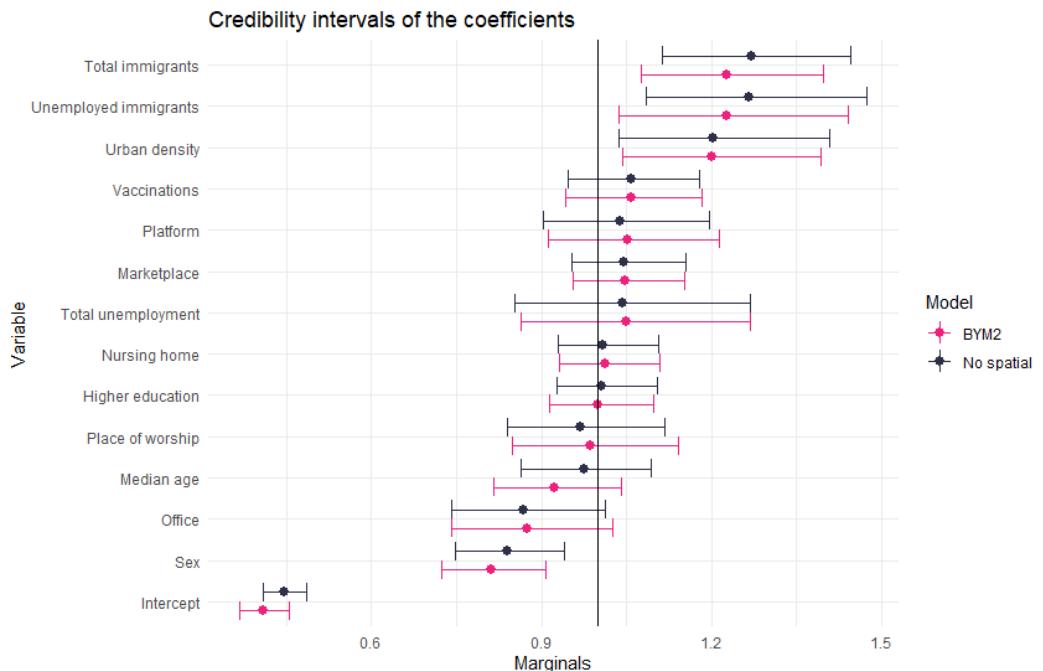
## 6.4.2 Spatial Models for Norway

Comparing the performance of the models in Table 6.11, the spatial models again showed better performance in terms of DIC and WAIC and this time significantly better performance in terms of CPO. The best performance in terms of MAE is observed for the Leroux model, but the MAE is quite close for all models.

**Tab. 6.11:** The performance measures for the best performing model of each type.

Model	DIC	WAIC	CPO	MAE <sub>train</sub>	MAE <sub>test</sub>
No spatial	2859	2864	-703	229	92
Besag	2854	2860	-1856	227	90
BYM2	2825	2832	-3646	160	89
Leroux	2382	2373	-7949	10	81

Looking at the differences between the coefficients and credibility intervals in Figure 6.11, the picture is similar to Figure 6.10. This time, however, all significant effects of the model without the spatial component are still significant. This again shows that the spatial effect is weaker in Norway than in Germany, as no new variables turn out to be significant, and no variables lose their significance when the spatial term is added.



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

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

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

Variable	meanp	exp(meanp)	exp(q0025p)	exp(q0975p)	sig.
(Intercept)	-0.895	0.409	0.368	0.454	*
Total immigrants	0.203	1.228	1.075	1.397	*
Unemployed immigrants	0.200	1.226	1.037	1.441	*
Urban density	0.180	1.201	1.042	1.393	*
Vaccinations	0.055	1.058	0.942	1.182	
Platform	0.049	1.053	0.912	1.213	
Marketplace	0.045	1.047	0.956	1.153	
Total unemployment	0.044	1.050	0.863	1.267	
Nursing home	0.012	1.013	0.931	1.109	
Higher education	-0.003	0.998	0.915	1.097	
Place of worship	-0.018	0.985	0.849	1.142	
Median age	-0.082	0.923	0.816	1.041	
Office	-0.138	0.874	0.742	1.026	
Sex	-0.211	0.811	0.723	0.906	*

For the hyperparameters, a value of 7.668 is reported for the precision and a value of 0.073 for  $\phi$ . Hence, 7.3% of the marginal variance is explained by the structured effect. Therefore, this model is close to pure overdispersion.

## 6.5 Choice of Hyperpriors

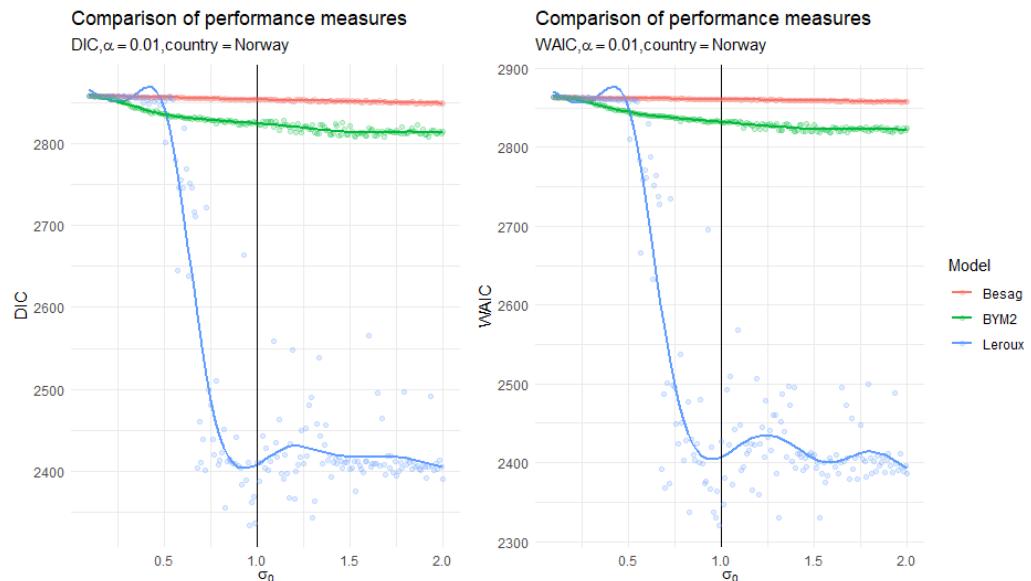
As can be seen in Equation 2.12, there is flexibility when it comes to choosing the values for the standard deviation  $\sigma_0$  as well as the probability  $\alpha$ . Therefore, an upper bound for the standard deviation can be chosen as well as the weight placed on this "tail event", describing how informative the resulting prior is.

Some of the issues that come with the choice of these hyperpriors are already discussed in Section 2.8.

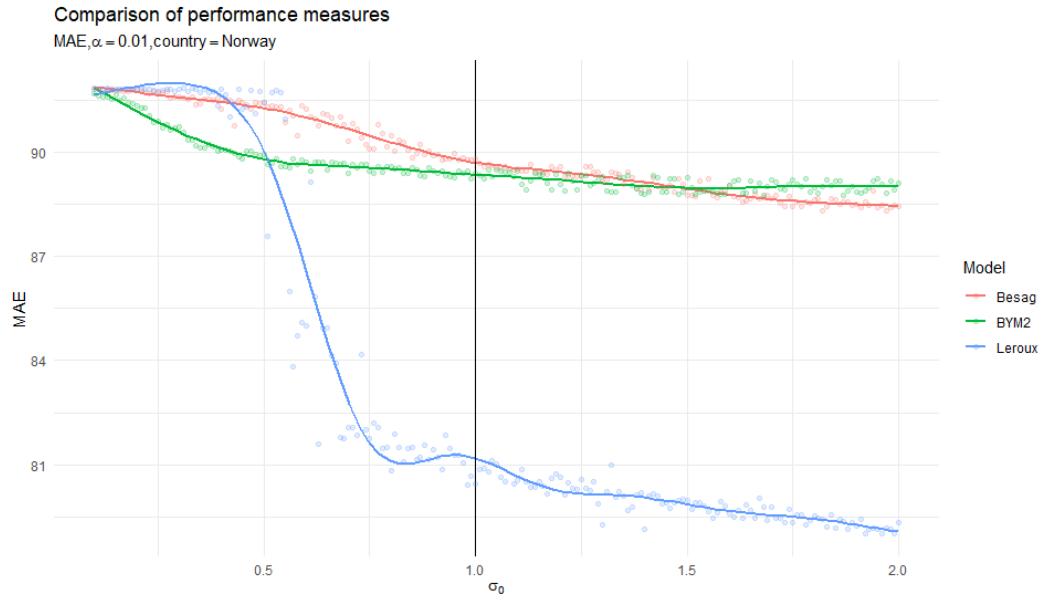
In the following, an assessment is made of how the performance of a Besag model, a BYM2 model and a Leroux model changes when playing around with the value for the standard deviation  $\sigma_0$ . To create these plots, models are calculated with  $\sigma_0$  values of  $\sigma_0 = (0.1, 0.11, 0.12, \dots, 5)$ .

In Figure 6.12 it can be seen that when choosing a higher value for  $\sigma_0$ , the DIC and WAIC is lower in the case of the Besag model and the BYM2 model. For the Leroux model, on the other hand, the WAIC gets lower until about 2 before it rises until around  $\sigma_0 = 2.5$  and then flattens out. It is a positive sign that this is not the case with the BYM2 model, as it is designed to avoid exactly this kind of thing.

For the MAE in Figure 6.13, it can be seen that for all model types, a higher value for  $\sigma_0$  leads to a lower MAE, however not by much.



**Fig. 6.12:** Values of the DIC and the WAIC when changing the value for  $\sigma_0$ . The black line highlights the values for  $\sigma_0 = 1$ .

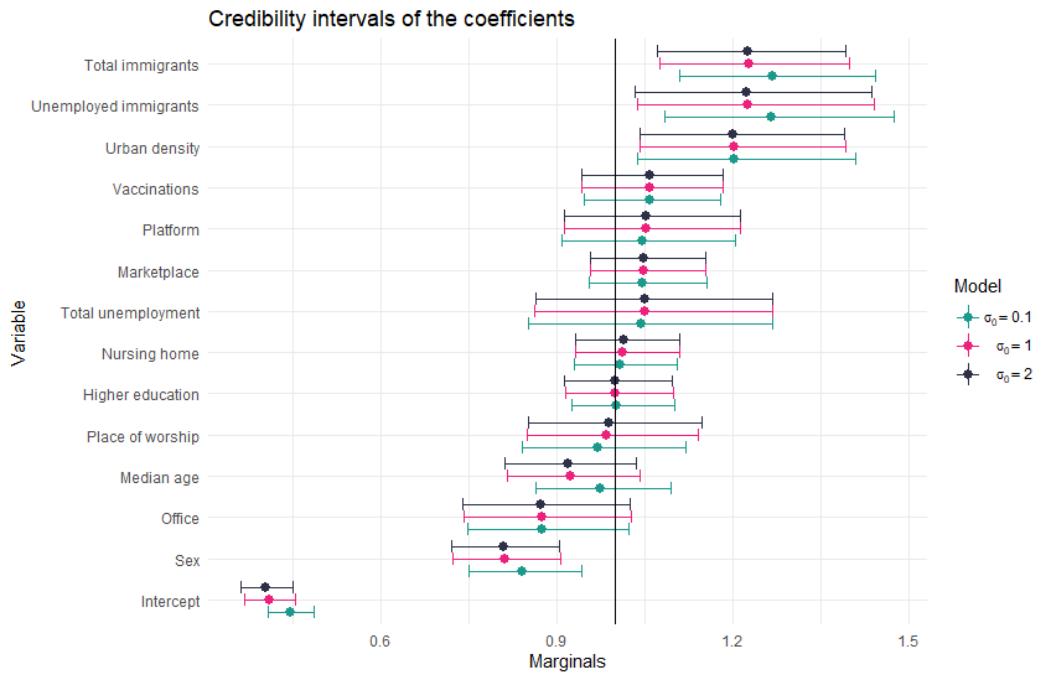


**Fig. 6.13:** Value of the MAE when changing the value for  $\sigma_0$ . The black line highlights the values for  $\sigma_0 = 1$ .

By allowing the precision to be greater, the variance is forced to be smaller. Hence, choosing a lower value for the precision leads to lower values for the WAIC. While this indicates a better fit to the training data, Figure 6.13 shows that the MAE increases when a higher value for  $\sigma_0$  is chosen, as the models overfit on the training data and therefore make worse predictions.

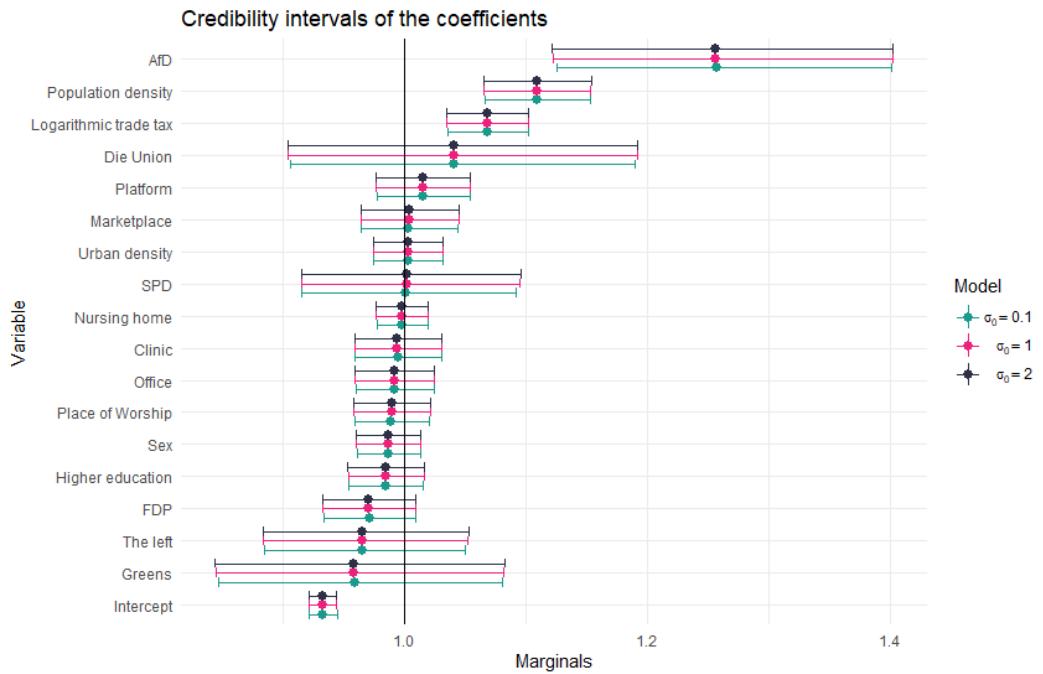
The corresponding figures for Germany are shown in Figure 10.10 and Figure 10.11 in the Appendix.

Figure 6.14 shows how the credibility intervals of the coefficients of a BYM2 model change when the value for  $\sigma_0$  is increased. The values of the coefficients tend to remain relatively similar most of the time, especially when the value of the coefficient is close to 1. However, a few times, for example for the total number of immigrants and the number of unemployed immigrants, the values differ.



**Fig. 6.14:** Comparison of the credibility intervals of a BYM2 model for different values of  $\sigma_0$ .

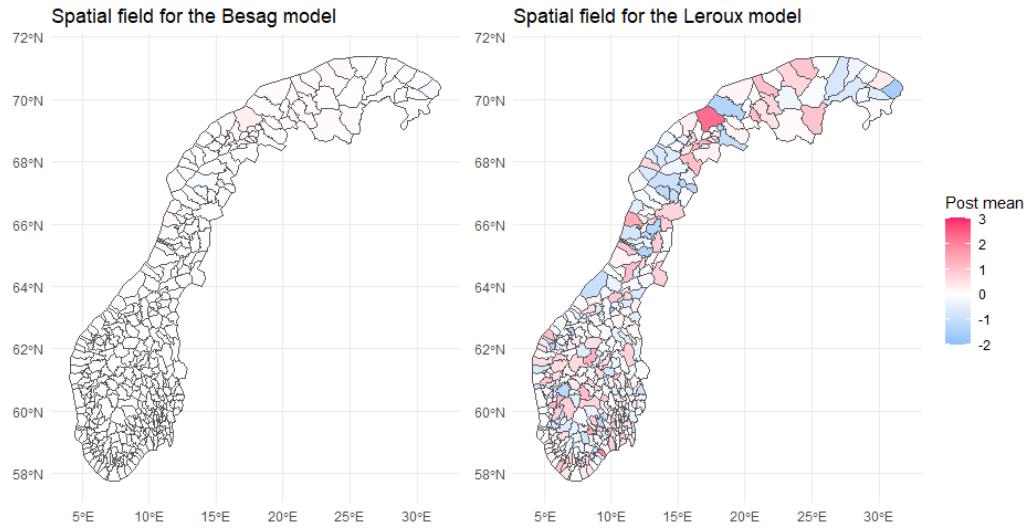
Having credibility intervals and posterior means that are very similar to each other, regardless of the value chosen for  $\sigma_0$ , is a great sign as it means that the calculated model is robust. That is, it means that the fixed covariates are important in their own right because they provide different information than the spatial field, even if the spatial field is allowed to be very smooth or coarse. This is exactly the case for Germany, which can be seen in Figure 6.15.



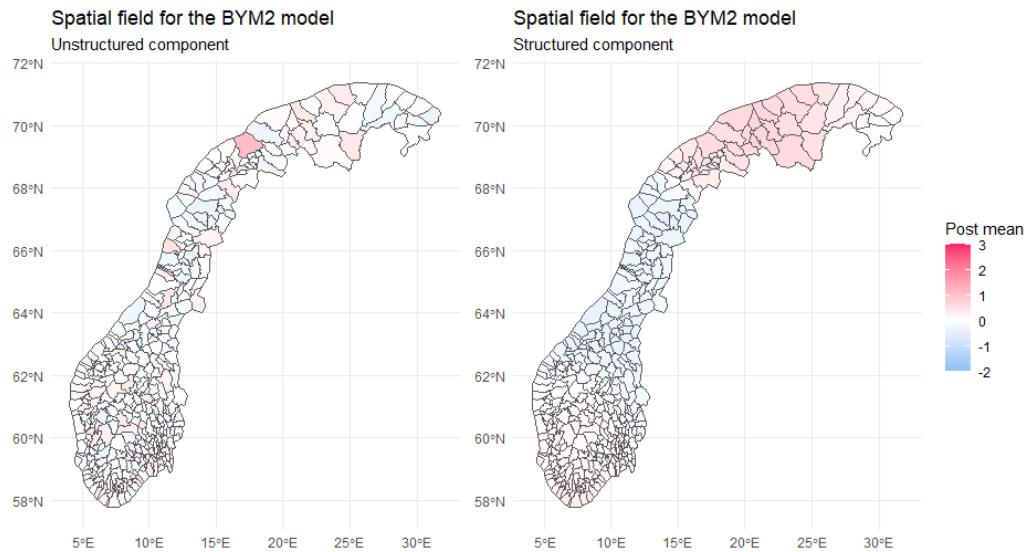
**Fig. 6.15:** Comparison of the credibility intervals of a BYM2 model for different values of  $\sigma_0$

Figure 6.16 and Figure 6.17 underline the problem of the models not being comparable with each other. Figure 6.16 already shows huge differences in the spatial field of the Besag model and the Leroux model, with the spatial field of the Besag model looking smoother than the spatial field of the Leroux model. This is a sign that the Leroux model is quickly overfitting to the data. In the left part of Figure 6.17 the values of Equation 2.61 are plotted, while in the right part the values of  $u_*$  are plotted.

For the spatial field of the unstructured random effect, the values of the posterior mean are similar to the values of the Besag model. For the structured component, however, the absolute values are somewhat higher. Nevertheless, in both cases the values are far from the values for the Leroux model.



**Fig. 6.16:** Spatial field for a Besag model and a Leroux model.

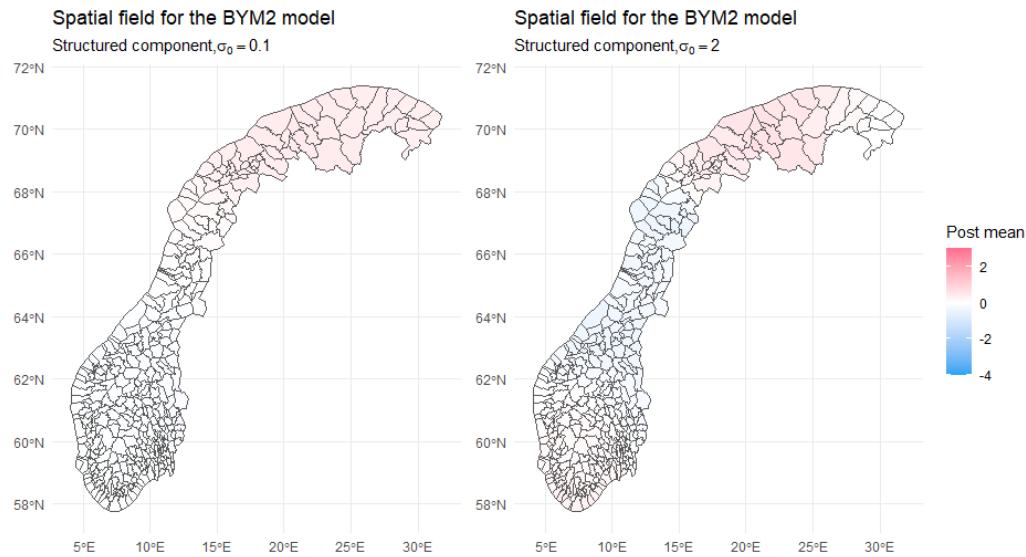


**Fig. 6.17:** Spatial fields for a BYM2 model.

The corresponding figures for Germany are shown in Figure 10.12 and Figure 10.13 in the Appendix.

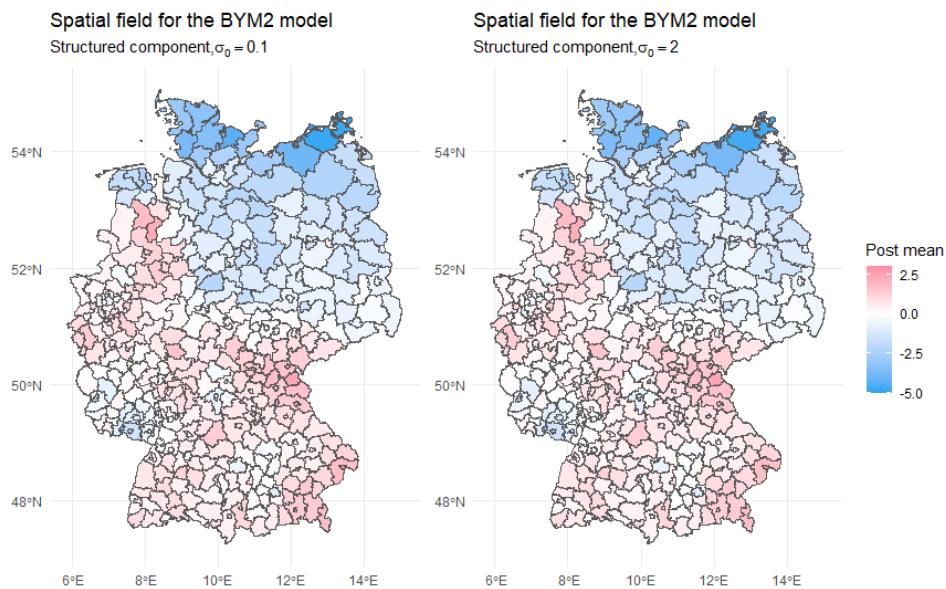
Finally, looking at the spatial field of the structured component when changing the value for  $\sigma_0$ , as seen in Figure 6.18, it can be seen that for a small value like  $\sigma_0 = 0.1$ , the values of the posterior mean are mostly around 0, while for a higher value these

values get slightly higher. The reason for this is simply that higher values for  $\sigma_0$  make the spatial field fit the data more closely, making it less smooth.



**Fig. 6.18:** Spatial fields for the structured component of a BYM2 model when changing the value for  $\sigma_0$ .

Looking at Figure 6.19, there are hardly any differences in the spatial fields for  $\sigma_0 = 0, 1$  and  $\sigma_0 = 2$ , which again shows how robust the model is for Germany.



**Fig. 6.19:** Spatial fields for the structured component of a BYM2 model when changing the value for  $\sigma_0$ .

## 6.6 Predictive models

Another way to determine whether calculating a Bayesian spatial model is the right approach for this research question is to compare the calculated models with a completely different class of models, namely predictive models calculated using machine learning methods. The algorithms used in this chapter do not assume any kind of prior distributions or assumptions about the likelihood of the data.

The approach used to calculate these models is described in detail below.

1. Define the five base learners, which are:
  - A regression tree
  - A k-nearest neighbours (knn) algorithm
  - A neural net
  - A random forest
  - An eXtreme Gradient Boosting (xGBoost) algorithm
2. Tune each algorithm using iterated F-Racing.
3. Train each model using the parameters determined during the tuning process.
4. Make predictions on the data.

The remainder of this chapter presents the results of the models calculated for Norway and Germany and provides some insights into how different factors influence infection rates.

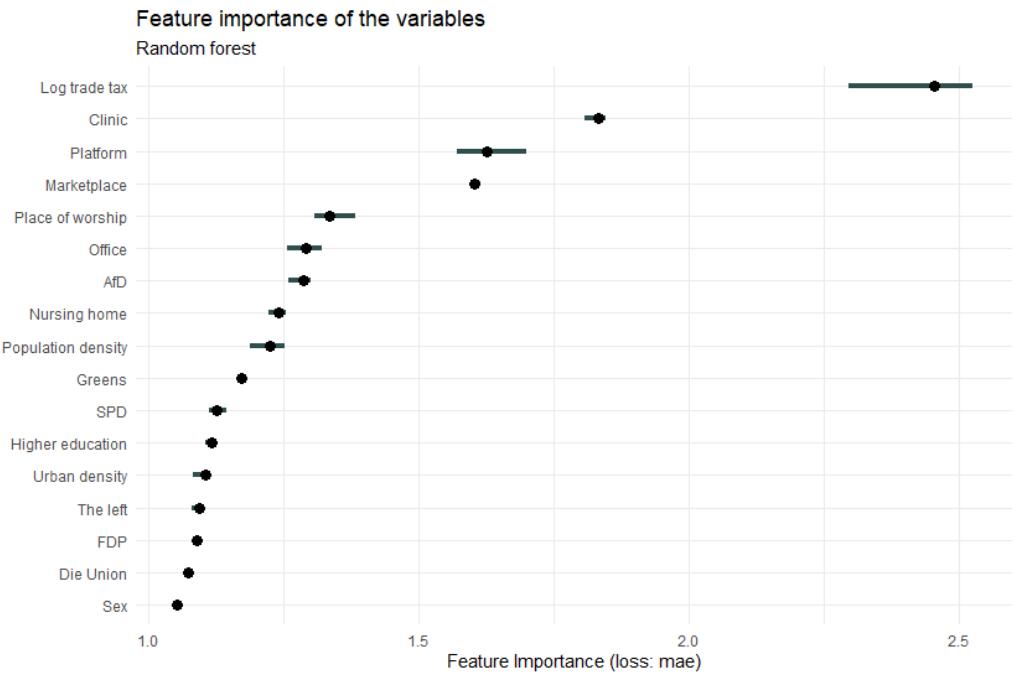
### 6.6.1 Predictive models for Germany

A look at Table 6.13 shows that the random forest had the lowest MAE of all the predictive models for both the training and test data, but still underperforms compared to the BYM2 model from Section 6.4.1. In the remainder of this section, the random forest model is evaluated in more detail to gain an understanding of which variables influence infection rates according to the model.

**Tab. 6.13:** The MAE for the BYM2 model and the predictive models.

Model	MAE <sub>train</sub>	MAE <sub>test</sub>
BYM2	103	1043
Regression tree	2759	3249
K-nearest neighbours	1510	2430
Neural net	3942	4133
Random forest	1109	1895
eXtreme Gradient Boosting	1941	3036

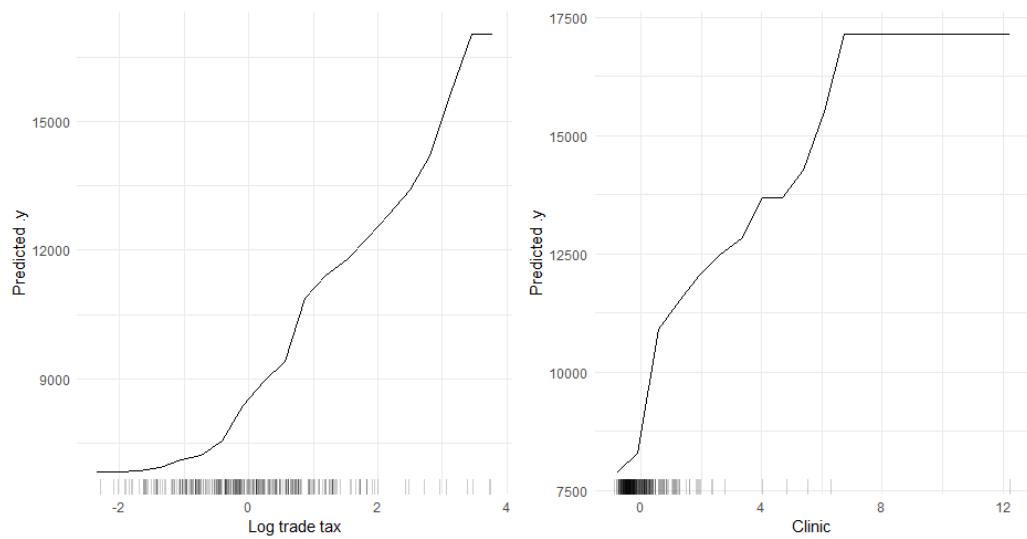
Looking at the feature importance of the variables in Figure 6.20, it can be seen that log trade tax is the most important feature, followed by the number of clinics, platforms and marketplaces in a municipality. The next most important characteristic is the number of places of worship, but there is quite a large difference between the importance of this characteristic and the importance of the number of marketplaces. The graph shows the confidence interval for each feature and in this case each confidence interval is narrow, suggesting that these are robust estimates of importance. Comparing the most important features of the random forest with the coefficients of the BYM2 model in Table 6.10, the logarithmic trade tax appears to be the only feature that is important in both types of models, as it is a significant coefficient in the BYM2 model and the most important feature in the random forest model. The number of clinics, platforms and marketplaces are not significant in the BYM2 model, while the share of the vote for the AfD and population density are significant in the BYM2 model, but are only the 7th and 9th most important features in the random forest, respectively.



**Fig. 6.20:** The variable importance plots for the random forest.

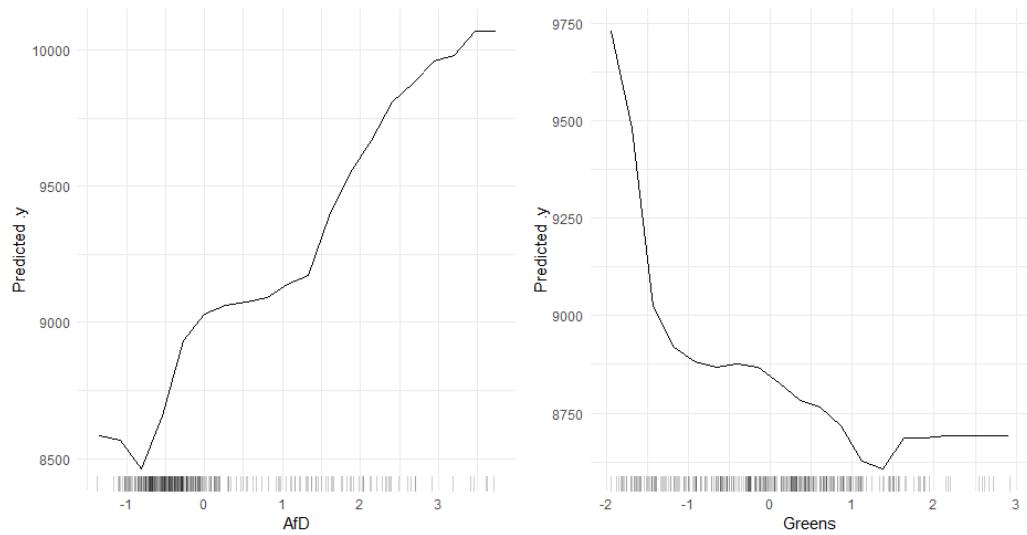
To get a better idea of how the features influence the predicted outcome, the partial dependence plots for the two most important features are shown in Figure 6.21, while Figure 6.22 shows how the two most important parties, the AfD and the Greens, influence the predicted outcome. For the logarithmic trade tax, a slow increase up until 0.5 can be seen before a steeper incline sets in, which eventually turns into a linear slope. For the number of clinic, it starts with a steep slope before changing to a less steep linear relationship and eventually flattening out. In general, however, the higher either value is, the higher the predicted outcome. Looking at the plots for the political parties, it can be seen that the higher the share of the vote for the AfD, the higher the predicted number of infections, while a higher share of the vote for the Greens leads to a lower predicted number of infections. The same relationship is observed in the BYM2 model, as can be seen in Table 6.10, but the effect for the Greens is not significant in this case.

Partial dependence plot of the features  
Random forest



**Fig. 6.21:** The partial dependence plots for the logarithmic trade tax and the number of clinics.

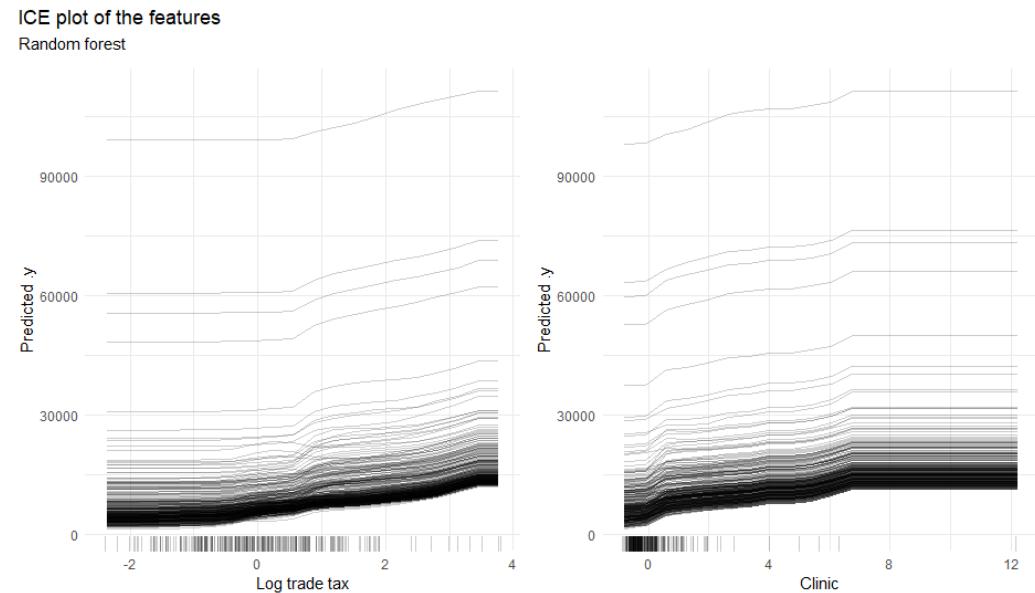
Partial dependence plot of the features  
Random forest



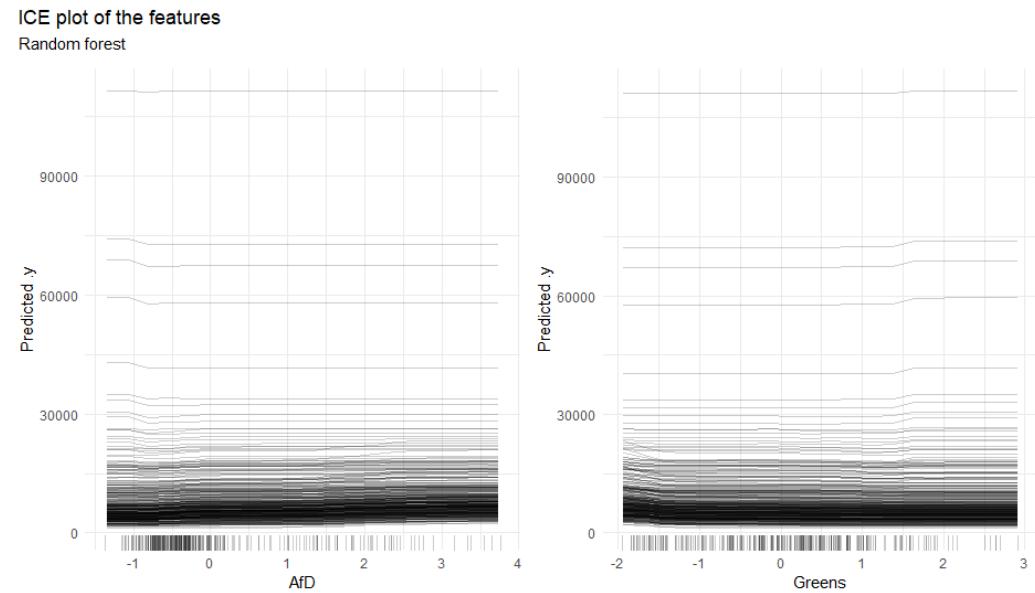
**Fig. 6.22:** The partial dependence plots for the share of the vote the AfD and the Greens get.

To check whether there are any interactions that could distort the relationship between the features shown in Figure 6.21 and Figure 6.22, the ICE plots of these variables in Figure 6.23 and Figure 6.24 can be looked at. All curves seem to follow the same course, so there are no obvious interactions. This means that the PDP

already represents a good summary of the relationships between the displayed features and the predicted number of infections.

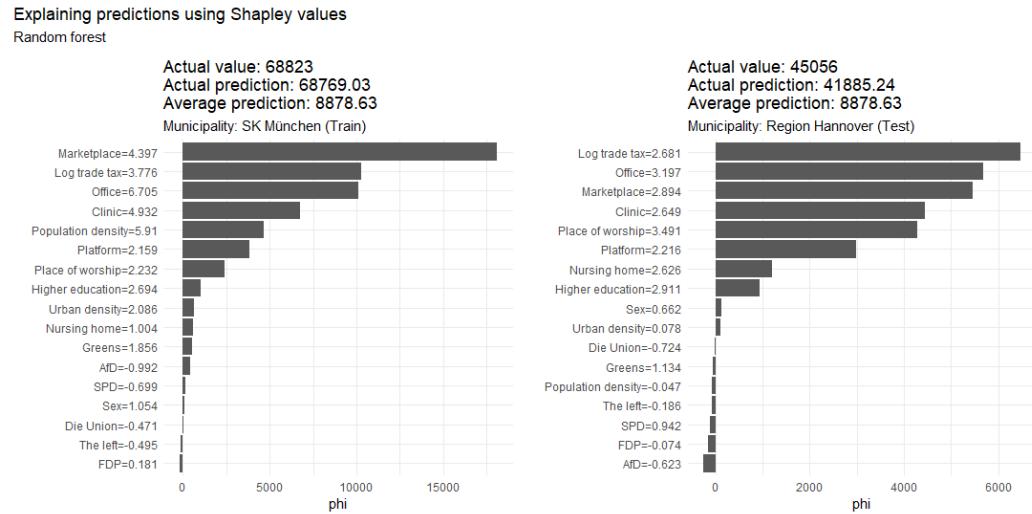


**Fig. 6.23:** The individual conditional expectation for the logarithmic trade tax and the number of clinics.



**Fig. 6.24:** The individual conditional expectation for the share of the vote the AfD and the Greens get.

Lastly, looking at the Shapley values for the region of Hannover in Figure 6.25, the prediction is nearly 50,000 above the average prediction. The number of marketplaces in the region increases the prediction the most, followed by the logarithmic trade tax and the number of offices in the region. For the city of Munich it is the same three features, however this time the logarithmic trade tax increases the prediction the most, followed by the number of offices and the number of marketplaces. For Munich, the actual prediction is around 33,000 above the average prediction.



**Fig. 6.25:** Shapley values for the cities of Munich and Hannover.

### 6.6.2 Predictive models for Norway

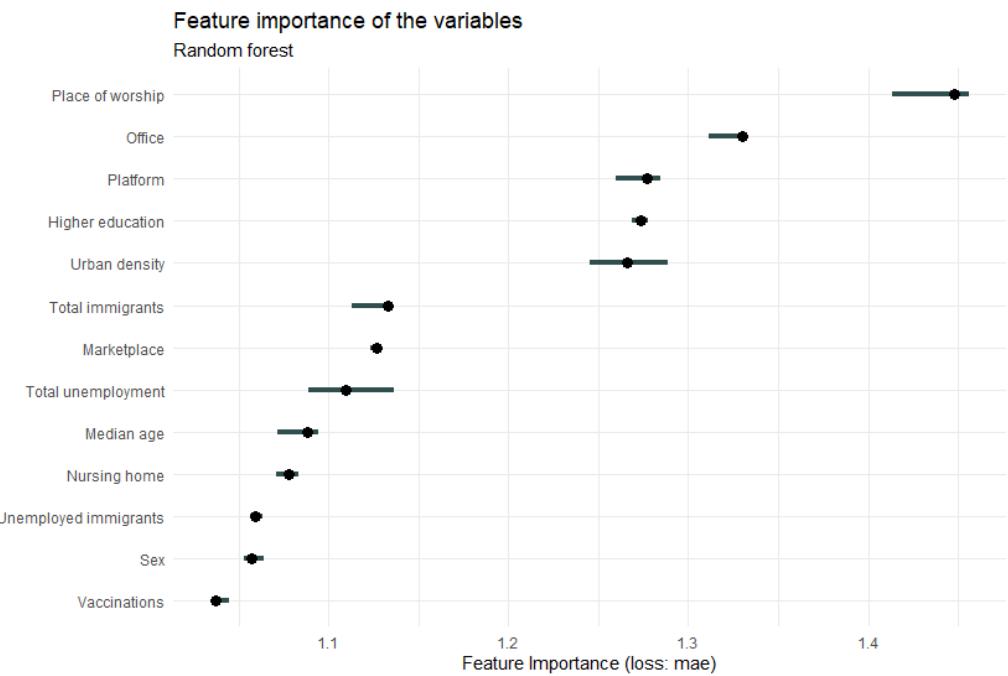
Looking at the performance of the different algorithms in Table 6.14, again, the random forest model had the lowest MAE of all the predictive models for the training and test data, but performs worse than the BYM2 model from Section 6.4.2. Therefore, the random forest model is evaluated in more detail to gain an understanding of which variables influence the infection rates according to the model.

**Tab. 6.14:** The MAE for the BYM2 model and the predictive models.

Model	MAE <sub>train</sub>	MAE <sub>test</sub>
BYM2	160	89
Regression tree	399	279
K-nearest neighbours	184	148
Neural net	392	297
Random forest	161	134
eXtreme Gradient Boosting	251	196

Looking at the feature importance of the variables in Figure 6.26, the most important feature is the number of places of worship in a municipality, followed by the number of offices, the number of public transport platforms, the number of higher education buildings and the urban density. After that, the difference in feature importance clearly increases between urban density and the total number of immigrants. Again, all confidence intervals in the plot are narrow, indicating robustness.

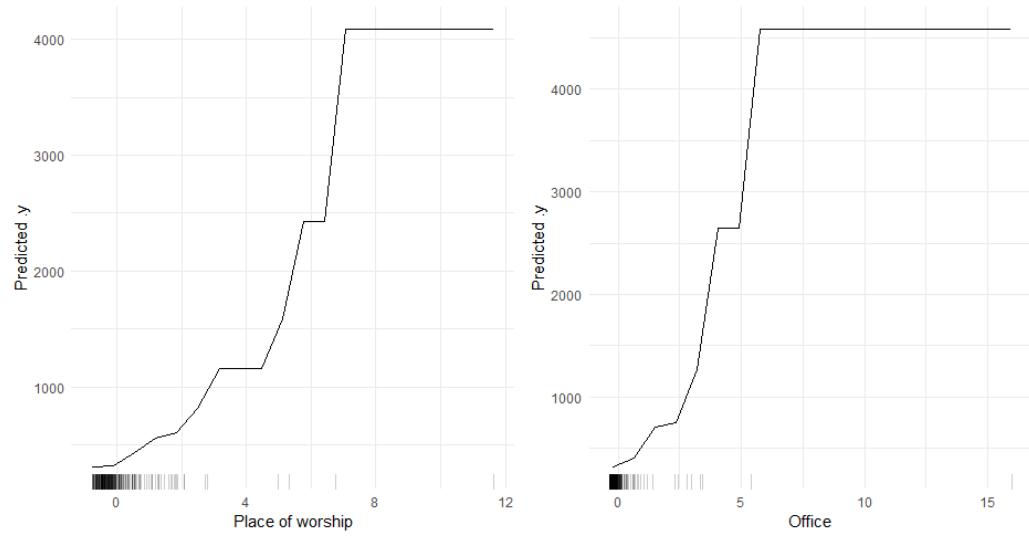
Comparing the important features with the fixed effects of the BYM2 model in Table 6.12, urban density appears to be the only feature that is relevant in both models, being the fifth most important feature in the random forest and a significant effect in the BYM2 model. The four most important features in Figure 6.26 are all non-significant in the BYM2 model, while the other three significant effects in the BYM2 model, the total number of immigrants, the number of unemployed immigrants and the proportion of females are the 11th and 8th and 12th most important features in the random forest model, respectively.



**Fig. 6.26:** The variable importance plots for the random forest.

Figure 6.27 shows how the predicted number of infections changes as the number of places of worship and offices in a community increases. For places of worship, a slow increase is seen until 3, followed by a steep increase that continues until 7 and ends in a plateau. For the number of offices, it is a slow increase until about 2.5, followed by a steep increase until about 6 before ending on a plateau.

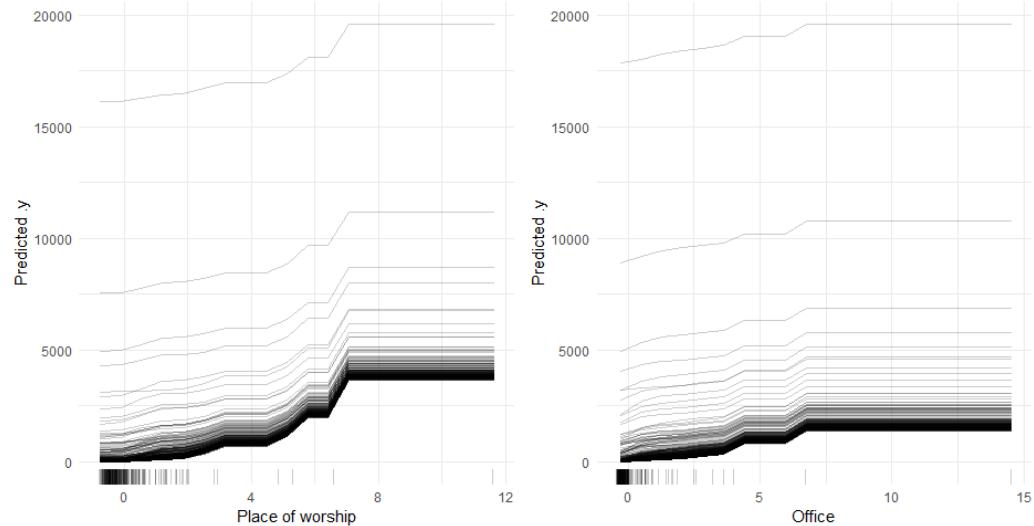
Partial dependence plot of the features  
Random forest



**Fig. 6.27:** The partial dependence plots for the number of places of worship and the number of offices.

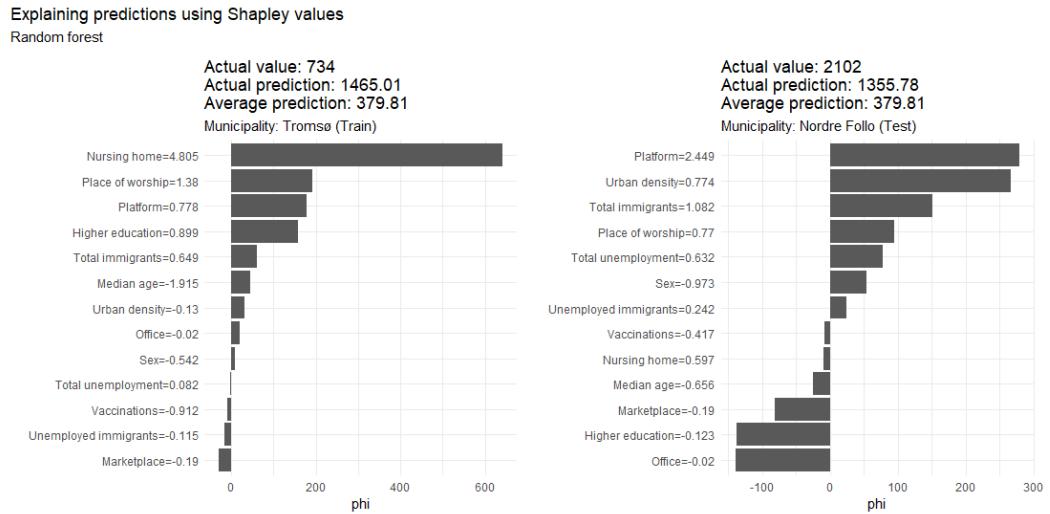
The curves display in the ICE plot in Figure 6.27 all follow the same course, so there are no obvious interactions. Therefore, again, the PDP already represents a good summary of the relationships between the displayed features and the predicted number of infections.

ICE plot of the features  
Random forest



**Fig. 6.28:** The individual conditional expectation for the logarithmic trade tax and the number of clinics.

Finally, looking at the Shapley values for Tromsø municipality in Figure 6.29, the prediction is about 1,100 above the average prediction. The feature that increases the prediction the most is the number of nursing homes in the municipality, while the number of marketplaces is the feature that decreases the prediction the most. For Nordre Follo, the prediction is about 1,000 above the average prediction, with the number of public transport platforms and urban density being the two features that increase the prediction the most, while the number of offices and higher educational buildings are the feature that decrease the prediction the most.



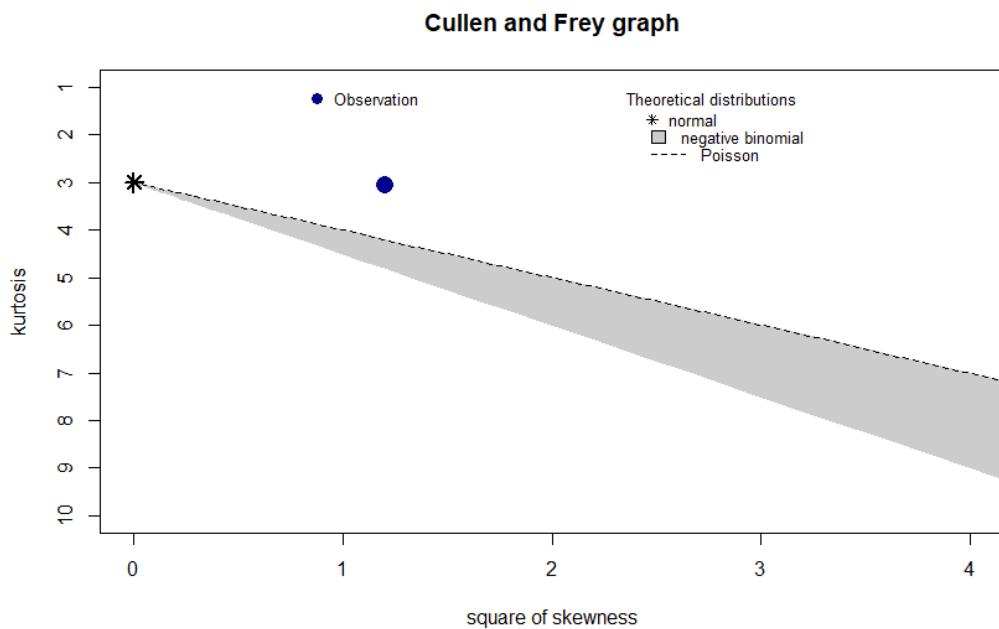
**Fig. 6.29:** Shapley values for the municipalities of Tromsø and Nordre Follo.

## 6.7 Temporal models

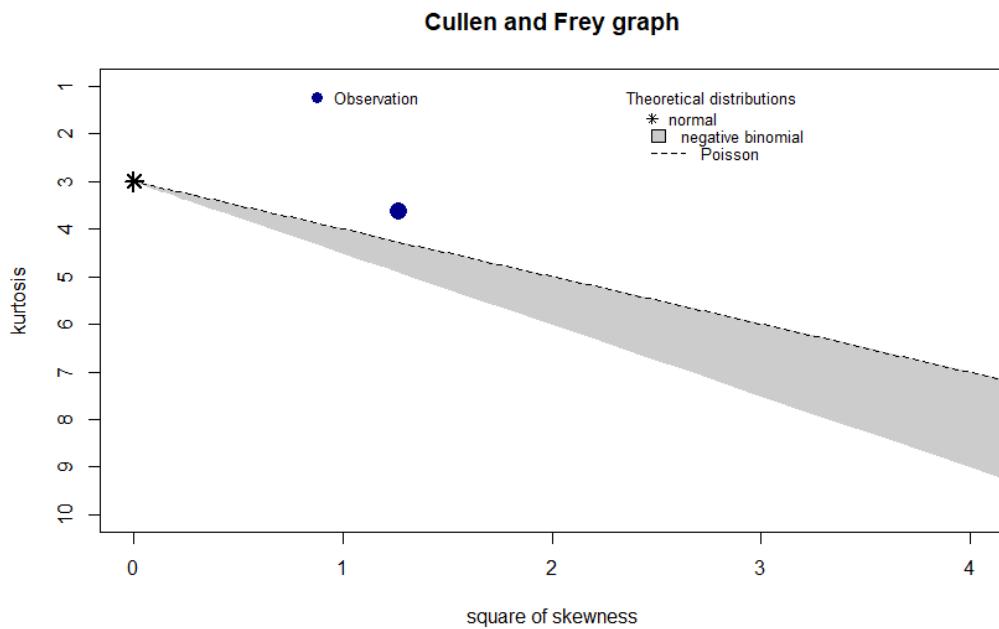
Since the previous sections have focused only on a fixed point in time, this section will address the research question from a temporal perspective. The first step for this is again to find a likelihood that describes the data reasonably well. After identifying such a likelihood, Section 6.7.2 and Section 6.7.3 present the results of temporal modelling of Covid-19 infection numbers in Germany and Norway, respectively.

### 6.7.1 Choice of Likelihood

As with the non-temporal models calculated in Section 6.3 and Section 6.4, a probability must first be found that describes the distribution of the daily number of infections in each country. Returning to the Cullen and Frey graph, Figure 6.30 and Figure 6.31 show that the observations do not actually follow a normal, negative binomial or Poisson distribution.

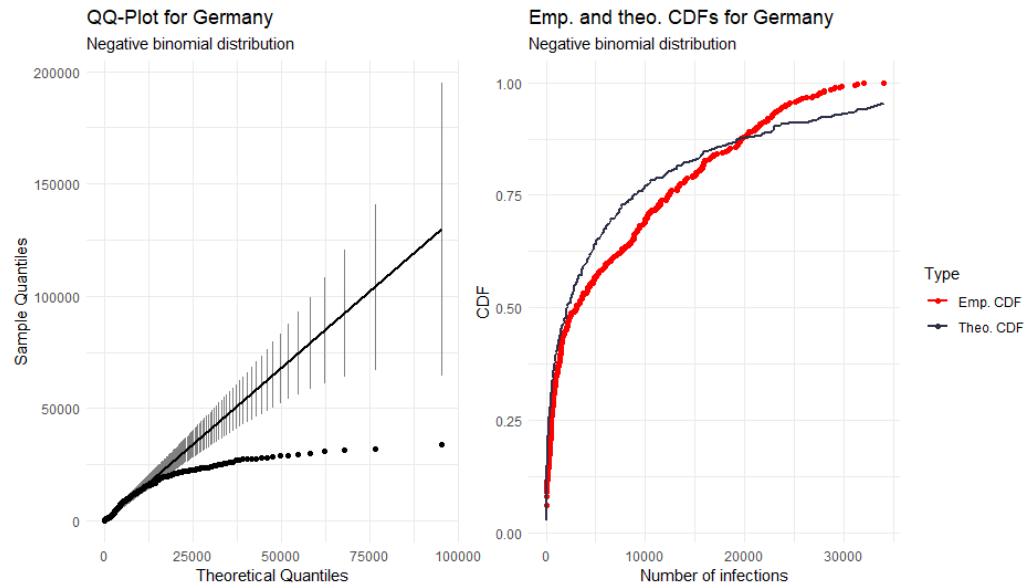


**Fig. 6.30:** The Cullen and Frey graph for Germany

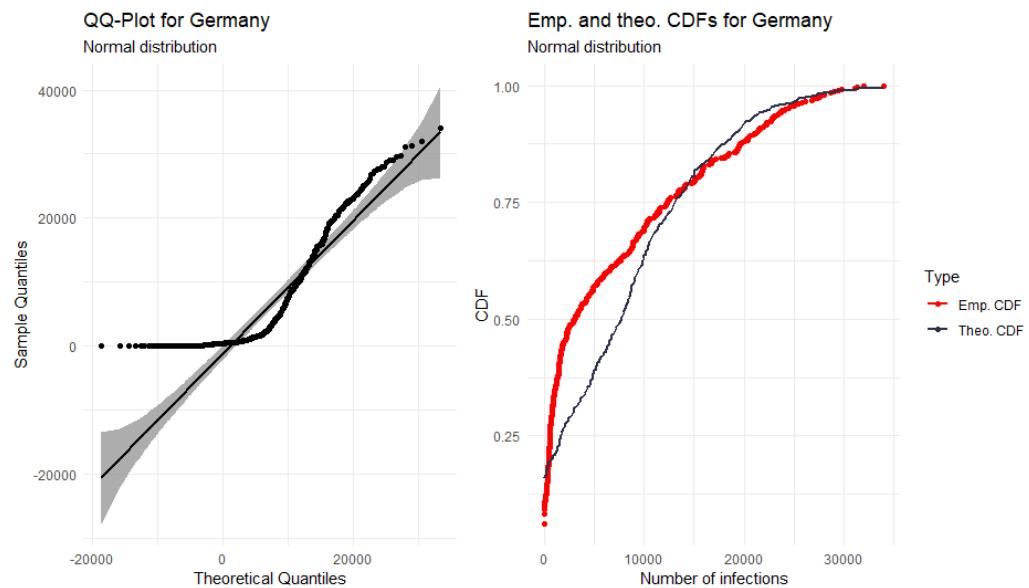


**Fig. 6.31:** The Cullen and Frey graph for Norway

After fitting all three distributions to the data using the maximum likelihood method, QQ-plots and the empirical and theoretical cumulative density function can be used to help decide which likelihood to choose. The QQ-plots in Figure 6.32 and Figure 6.33 clearly show that there is no linear relationship between the theoretical quantiles and the sample quantiles for both distributions, while the empirical CDF roughly follows the theoretical CDF in both cases. The plots for the Poisson distribution as well as all plots for the Norwegian municipalities are in Section 10.3.1.2 and Section 10.3.2.1 in the Appendix.



**Fig. 6.32:** A negative binomial fit to the number of cases in German municipalities



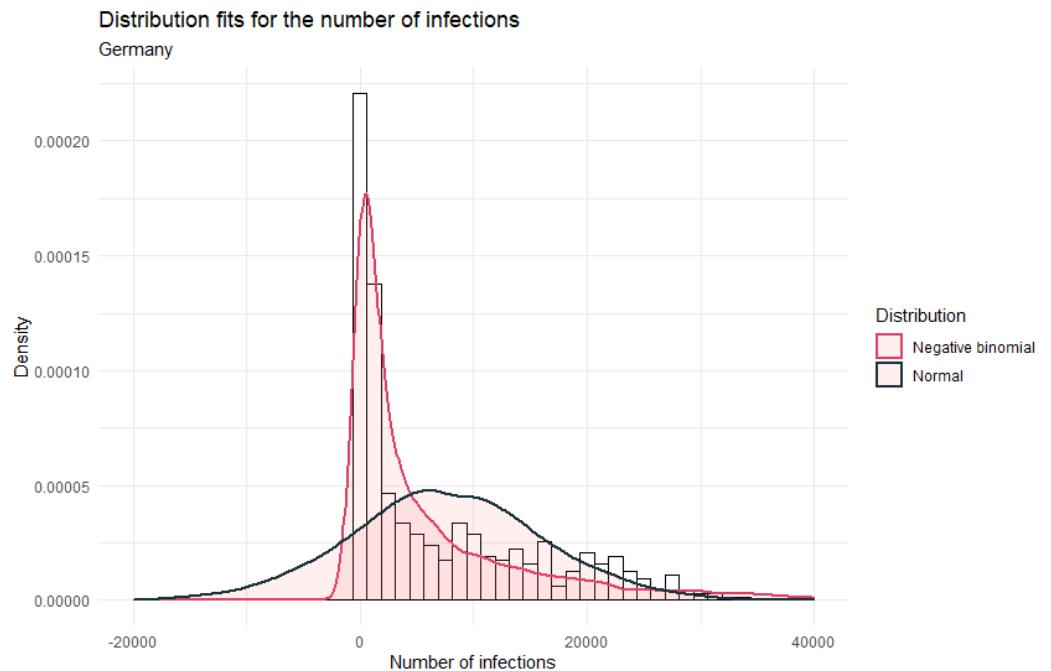
**Fig. 6.33:** A normal fit to the number of cases in German municipalities

Looking at the AIC for all the distribution fits shown in Table 6.15, the negative binomial distribution again seems to be the best fit for the data. For Germany the mean and standard deviation are 7274 and 8429 respectively, while for Norway these moments are 256 and 266. This again explains the poor fit of the Poisson distribution, since an equal mean and variance are assumed for the Poisson distribution.

**Tab. 6.15:** The AIC for different distributions for Germany and Norway

Country	Distribution	AIC
Germany	Normal	10461
Germany	Poisson	4762604
Germany	Negative Binomial	9462
Norway	Normal	6600
Norway	Poisson	131469
Norway	Negative Binomial	6076

Overlaying the number of infections with a negative binomial distribution and a normal distribution using the estimated parameters, as shown in Figure 6.34, reinforces the assumption of the negative binomial distribution, which is used as the assumed probability of the data for the models calculated in this chapter. The graph for Norway can be seen in Figure 10.9.



**Fig. 6.34:** Histogram for the number of cases in German municipalities with a normal and a negative binomial distribution overlayed.

## 6.7.2 Temporal models for Germany

For Germany, 502 data points are available, starting on 7 January 2020 and ending on 22 May 2021. Of these 502 data points, the first 482 are used for model training, while the last 20 are used for the analysis. Due to the temporal effect, a random split would not make sense as this may lead to a look-ahead bias as the model would use information that is not yet known or available during the analysed period. After removing variables using the VIF method already used for the non-temporal models, the trained model uses the following variables:

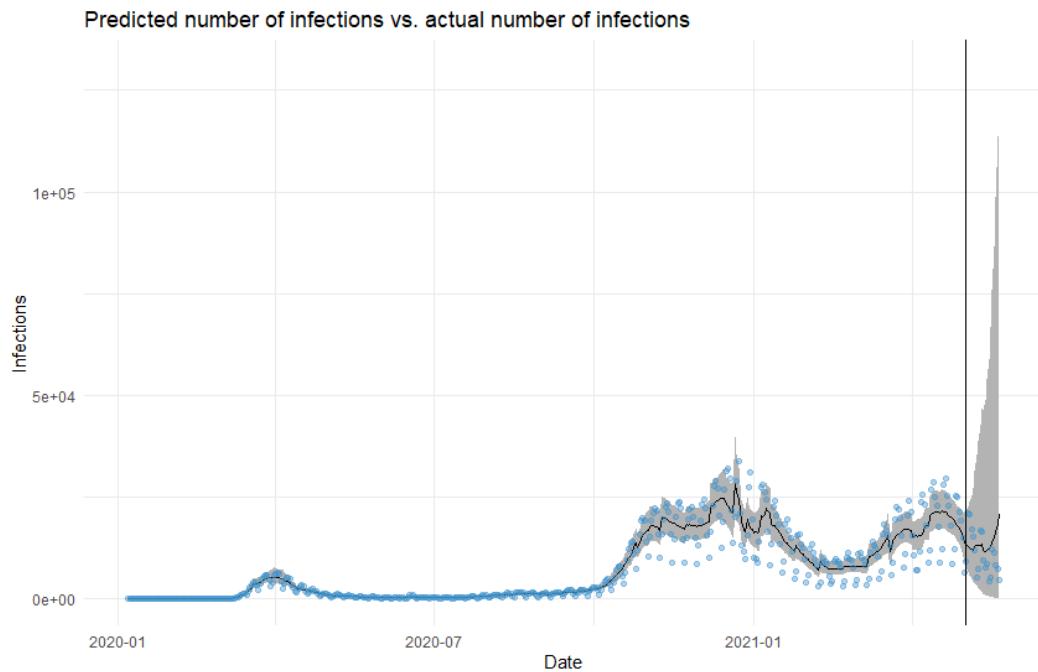
- Measures taken in relation to contact tracing.
- Measures taken in relation to restrictions on internal movement.
- Measures taken in relation to the use of face masks.
- Measures taken in relation to the requirement to remain at home.
- The mobility at workplaces.
- The season of the year.
- The relative frequency of variant 20E, the most prevalent in Western Europe in the summer of 2020
- The relative frequency of variant 20L, also known as B.1.1.7
- The relative frequency of the other strains, that are not tracked
  - The relative frequencies of the variant variables do not add up to 1 because the CoVariants tracks numerous variants and combines all untracked variants into one group. This group is predominant in each country at the beginning of the pandemic.

A second-order random walk is chosen for the temporal effect. The performance measures of this model and the temporal baseline models are shown in Table 6.16. While the AR(1) process has by far the lowest MAE in training, it has the second-highest MAE in testing. The temporal model, besides the AR(1), has the lowest MAE during training and the overall lowest MAE during testing, indicating that the model fits the data best.

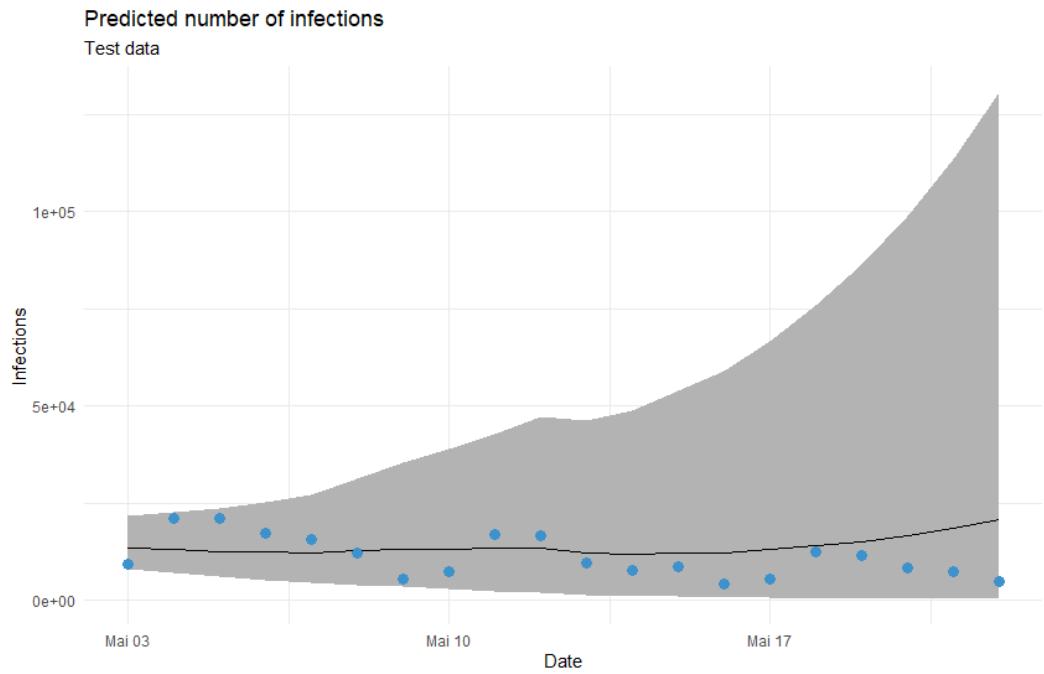
**Tab. 6.16:** The performance measures for different types of temporal models for Germany.

Model	DIC	WAIC	CPO	MAE <sub>train</sub>	MAE <sub>test</sub>
Only date as covariate	8815	8816	-4408	4915	27450
Random walk of second order	7522	7519	-3760	2009	6180
AR(1) process	5950	5924	-3877	18	7524
Temporal model	7524	7521	-3777	1931	5821

Looking at the predicted number of infections from the temporal model in Figure 6.35, it can be seen that the predictions for the training data follow roughly the same path as the actual numbers, as three waves can be clearly seen. For the test data depicted more clearly in Figure 6.36, the predicted number of cases shows a very slow increase, and it can be seen that the "weekend effect", which means that fewer infections are reported on the weekend, has not been captured. Furthermore, the more recent a data point is in the test set, the greater the uncertainty associated with the prediction.

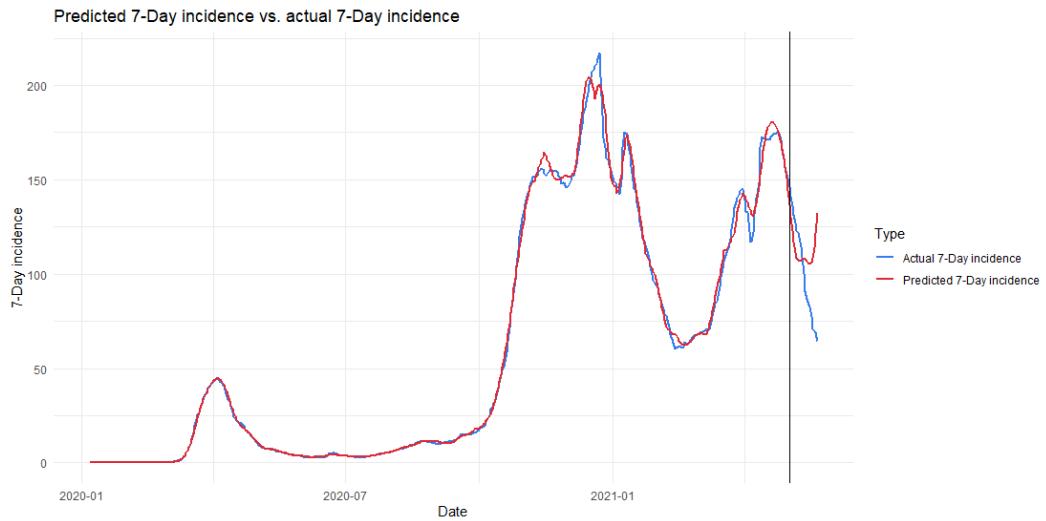


**Fig. 6.35:** The predicted number of infections in Germany according to the temporal model. The vertical line indicates where the test data begins.



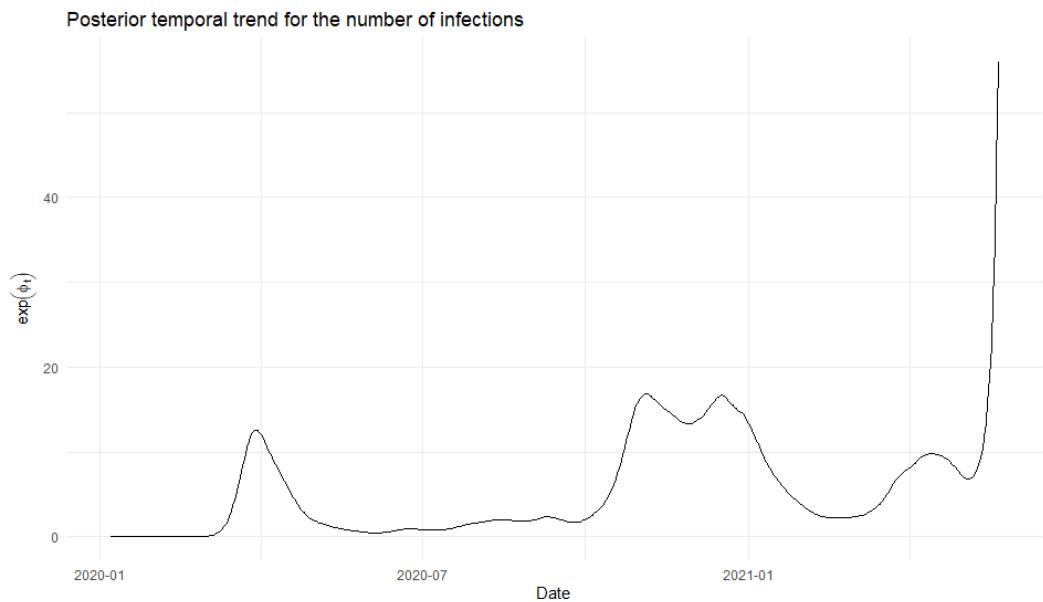
**Fig. 6.36:** The predicted number of infections in Germany according to the temporal model.

Another way to look at the predictions is to compare the 7-day incidence of the actual number of infections and the predicted number of infections, as daily variations such as the "weekend effect" should be smoothed out. Figure 6.37 shows that the predicted 7-day incidence again follows the actual data very well, even a little for the test data. However, this is because for the first few days of test data, the incidence depends on the predicted number of infections for a few days of training data. Once it is only test data, the 7-day decline slowly stops before a steep slope starts, in contrast to the actual 7-day incidence, which is still falling.



**Fig. 6.37:** The 7-day incidence of the actual number of infections and the predicted number of infections. The vertical line indicates where the test data begins.

Looking at the posterior temporal trend for Germany in Figure 6.38, the first two waves are clearly visible, with the second wave having two peaks, the first in early November 2020 and the second in mid-December 2020. The peak of the first and third waves is in late March 2020 and mid-April 2021, respectively. After the third peak, a steep increase sets in, as with the 7-day incidence in Figure 6.37, which is clearly not the case when looking at the actual data.



**Fig. 6.38:** The posterior temporal trend for the number of infections.

The covariates of the model are presented in Table 6.17. Apart from the intercept, the only significant effect is the mobility at workplaces.

**Tab. 6.17:** The fixed effects for the model. Values are rounded. A \* denotes a significant effect.

Variable	meanp	exp(meanp)	exp(q0025p)	exp(q0975p)	sig.
(Intercept)	-1.592	0.228	0.079	0.518	*
Variant 20L	0.658	2.117	0.836	4.476	
Variant 20E	0.320	1.407	0.926	2.059	
Restrictions internal movement	0.235	1.298	0.813	1.970	
Movement restricted					
Workplace mobility	0.168	1.186	1.029	1.359	*
Season winter	0.112	1.161	0.653	1.918	
Season spring	0.034	1.082	0.574	1.858	
Stay home requirements	0.022	1.044	0.687	1.517	
Recommended					
Facial coverings	0.018	1.070	0.551	1.884	
Recommended					
Facial coverings					
Required in some public spaces	-0.008	1.109	0.398	2.506	
Stay home requirements	-0.014	1.009	0.650	1.492	
Required (exc. essent.)					
Restrictions internal movement	-0.021	1.004	0.632	1.522	
Recommended					
Contact tracing	-0.047	0.976	0.630	1.454	
Limited tracing					
Other variants	-0.134	0.920	0.470	1.622	
Season summer	-0.341	0.739	0.412	1.216	
Contact tracing	-2.095	0.228	0.010	1.035	
No tracing					

### 6.7.3 Temporal models for Norway

For Norway, 473 data points are available, starting on 5 February 2020 and ending on 22 May 2021. Of these 473 data points, the first 453 are used for model training, while the last 20 are used for the analysis. After the removal of some variables, the following features remained:

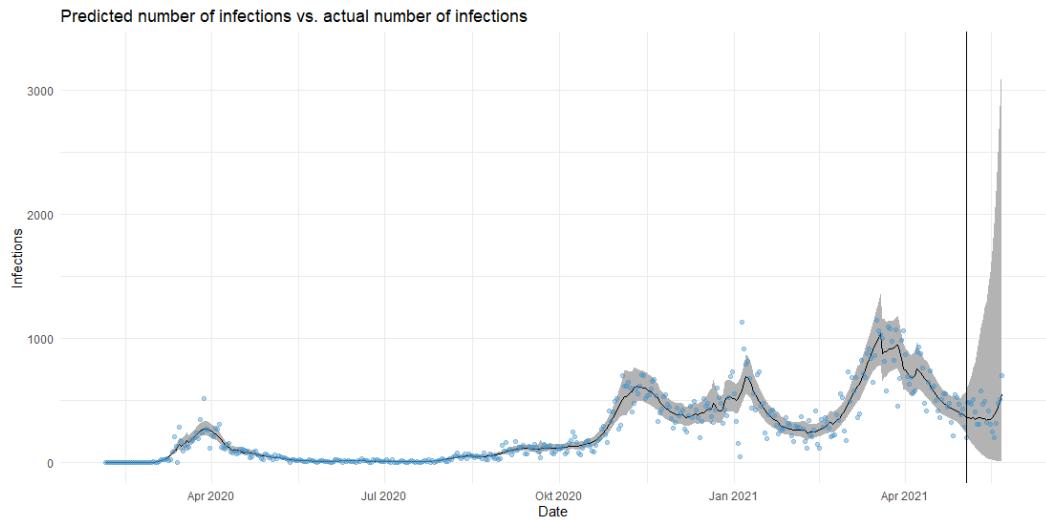
- Measures taken in relation to restrictions on internal movement.
- Measures taken in relation to the requirement to remain at home.
- The mobility at workplaces.
- The mobility at groceries and pharmacies.
- The season of the year.
- The relative frequency of variant 20E, the most prevalent in Western Europe in the summer of 2020
- The relative frequency of variant 20L, also known as B.1.1.7
- The relative frequency of the other strains, that are not tracked
  - The relative frequencies of the variant variables do not add up to 1 because the CoVariants tracks numerous variants and combines all untracked variants into one group. This group is predominant in each country at the beginning of the pandemic.

Again, a second-order random walk is chosen for the temporal effect. The performance measures of this model and the temporal baseline models are shown in Table 6.18. The lowest MAE during training is again observed for the AR(1) process, followed by the temporal model and the second-order random walk. The temporal model shows the best performance in terms of MAE for the test set, followed by the second-order random walk and the AR(1) process.

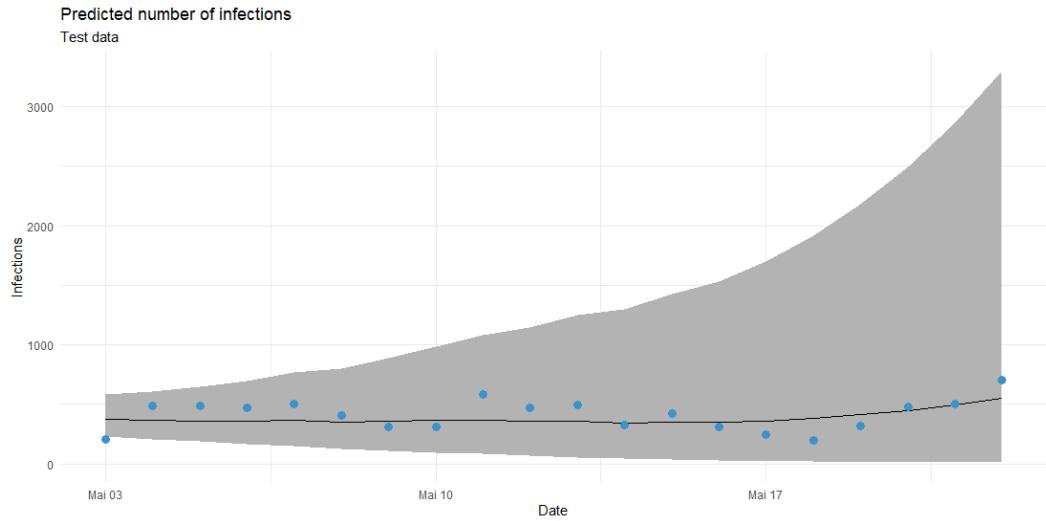
**Tab. 6.18:** The performance measures for different types of temporal models for Norway.

Model	DIC	WAIC	CPO	MAE <sub>train</sub>	MAE <sub>test</sub>
Only date as covariate	5567	5568	-2784	130	446
Random walk of second order	4603	4603	-2387	55	119
AR(1) process	4602	4599	-2315	38	110
Temporal model	4613	4613	-2432	53	100

For the predicted number of infections shown in Figure 6.39, 4 peaks are seen for the training data, just like for the actual data. The rest of the predicted infection numbers follow the actual infection numbers very closely, which is not too surprising given the low MAE shown in Table 6.18. For the test data shown in Figure 6.40, it is very similar to the graph for Germany in Figure 6.36, as the "weekend effect" is not captured, a slow increase in the predicted infection numbers can be seen, as well as an increasing uncertainty in the predictions.

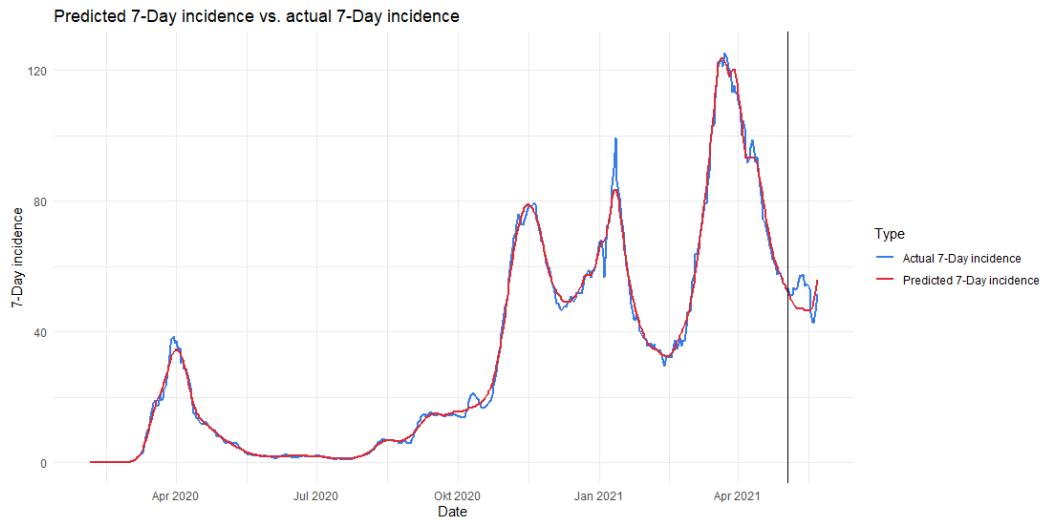


**Fig. 6.39:** The predicted number of infections in Norway according to the temporal model. The vertical line indicates where the test data begins.



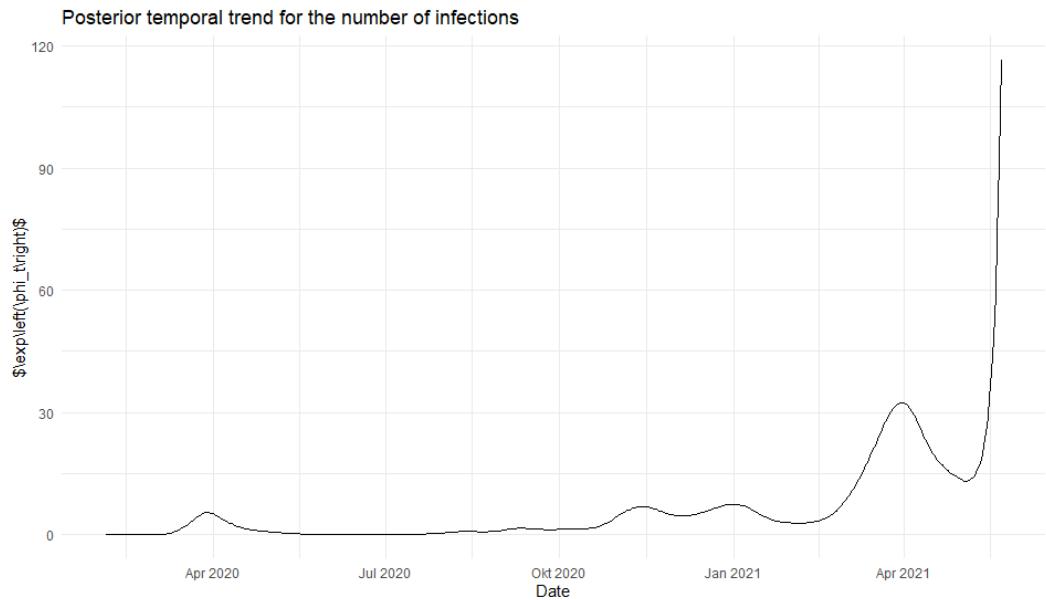
**Fig. 6.40:** The predicted number of infections in Norway according to the temporal model.

Comparing the 7-day incidence of the predicted data and the actual data in Figure 6.41, the incidences of the training data and the actual data follow the same pattern. However, the predicted 7-day incidence for the test data initially falls before rising, while the actual 7-day incidence initially rises and then falls before rising again. Although still not ideal, the predicted incidence for Norway looks slightly better than that for Germany, which can be seen in Figure 6.37.



**Fig. 6.41:** The 7-day incidence of the actual number of infections and the predicted number of infections. The vertical line indicates where the test data begins.

The posterior temporal trend for Norway in Figure 6.42 clearly shows all three waves, including the double peak of the second wave. It can be seen that the third wave is clearly the worst in terms of infection numbers. On the other hand, however, the first and second waves appear to be equally bad in this graphic, while in reality the second wave in Norway was worse than the first. Just as with Germany in Figure 6.38, a steep increase sets in after the third wave.



**Fig. 6.42:** The posterior temporal trend for the number of infections.

The covariates of the model are presented in Table 6.19. Apart from the intercept, the only significant effect is the mobility at workplaces.

**Tab. 6.19:** The fixed effects for the model. Values are rounded. A \* denotes a significant effect.

Variable	meanp	exp(meanp)	exp(q0025p)	exp(q0975p)	sig.
(Intercept)	-0.926	0.408	0.249	0.631	*
Workplace mobility	0.179	1.201	1.011	1.417	*
Season winter	0.109	1.159	0.645	1.924	
Other variants	0.056	1.073	0.760	1.482	
Variant 20E	0.041	1.043	0.942	1.149	
Restrictions internal movement	0.012	1.063	0.545	1.854	
Recommended Stay home requirements	-0.058	1.033	0.413	2.163	
Recommended Mobility grocery and pharmacy	-0.081	0.923	0.831	1.022	
Variant 20L	-0.086	0.963	0.499	1.684	
Season spring	-0.089	0.962	0.489	1.706	
season summer	-0.178	0.886	0.436	1.624	
Restrictions internal movement	-0.209	0.842	0.477	1.384	
Movement restricted					

## Further Analysis using R-Shiny

R-Shiny is a framework that enables the creation of web applications and dashboards to visualize data interactively, to make statistics accessible to people without programming skills or a mathematical background, or to enable further analysis on research questions, to name just a few use cases. Dashboards have also been used by governments during the pandemic to effectively communicate infection data related to Covid-19, e.g. the Robert Koch Institute's Covid-19 Dashboard, which displays daily infection figures for each municipality in Germany that can be grouped by age group and gender, displays key figures such as the 7-day incidence and contains a choropleth map, among other features. The dashboard can be accessed at

<https://experience.arcgis.com/experience/478220a4c454480e823b17327b2bf1d4>

A similar dashboard is available for Norway and specifically for the municipality of Oslo:

<https://experience.arcgis.com/experience/742a281a0fa74ab79147a76e6b52833b>

As part of this thesis, a dashboard was developed using the R-Shiny framework, with the intention to give the reader a bit more insight into the data. The dashboard comes with four main features:

1. The Data Explorer
2. The SIR for Norway and Germany
3. Spatial Modelling
4. Temporal Modelling

## The Data Explorer

The main idea behind the data explorer is to display the data used to create all the models calculated in this work. The data can be visualized for Norway, Germany or Europe. Each tab contains a map and several drop-down menus where the user can select different things. For Norway and Germany, in addition to the variable to be displayed on the map, there is also the option to choose between three map types, hexagon maps, heat maps and choropleth maps. Hexagon maps can be seen as a mixture of a heat map and a histogram. Each bar represents the number of times a feature is present within a certain radius. The higher a bar is, the more often the feature is present. Figure 7.1 shows the number of bakeries within a radius of 2.5 km. It is clear that most bakeries are found in Berlin and Munich.

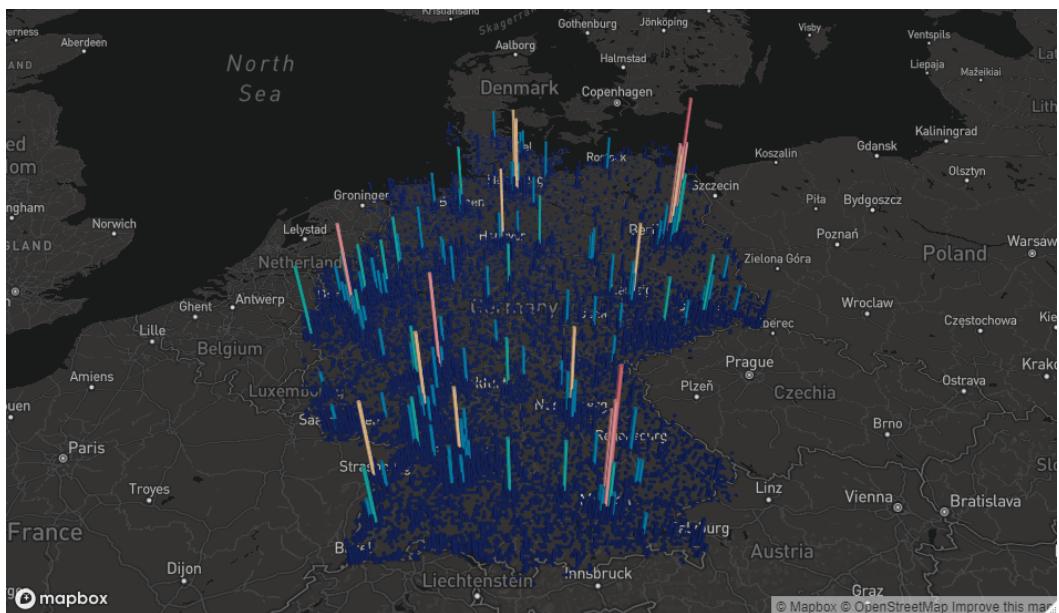
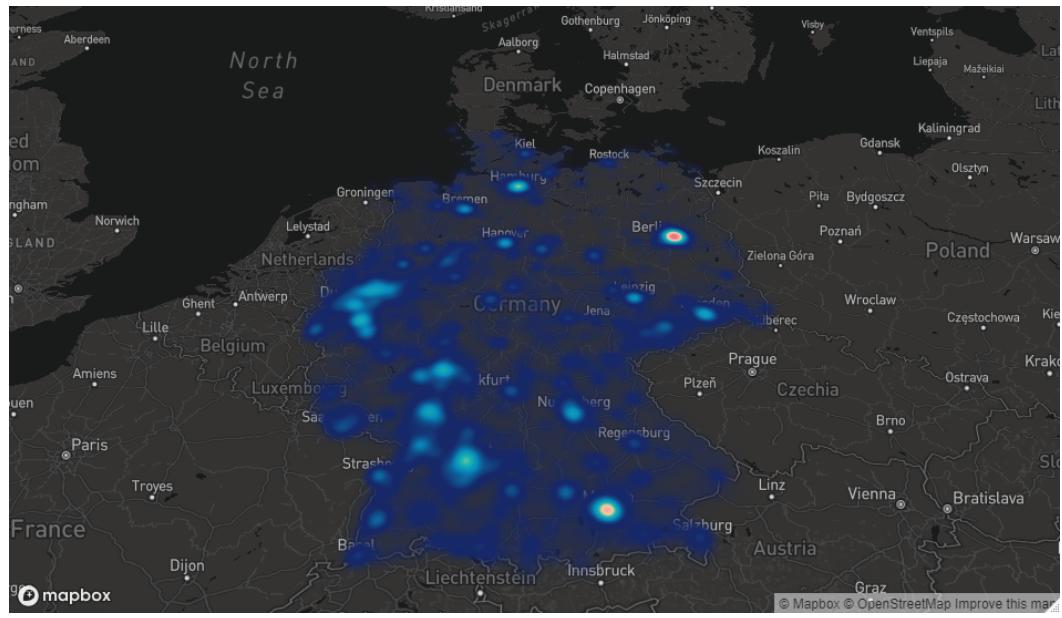


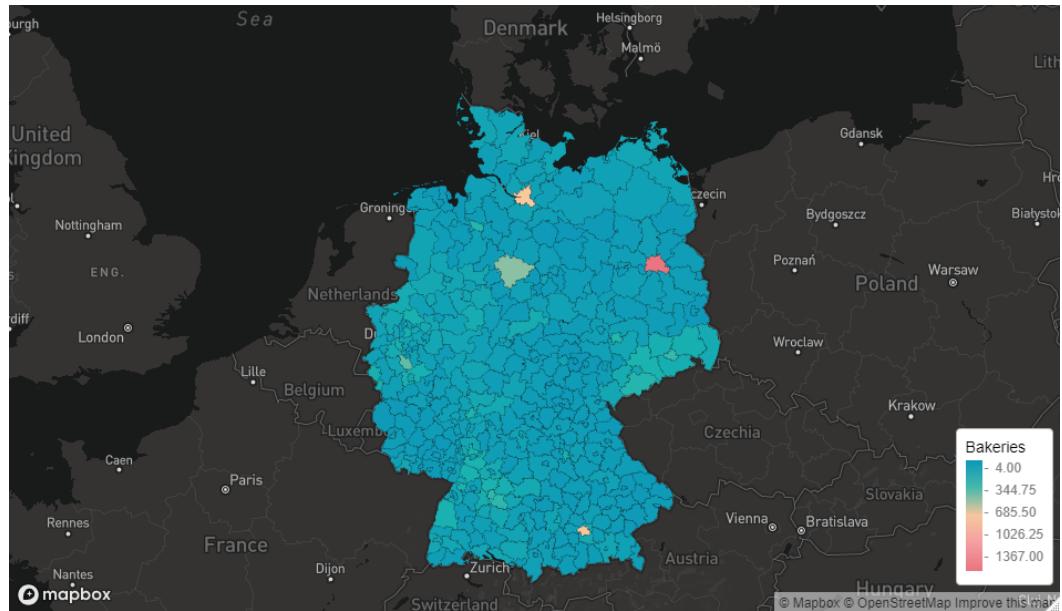
Fig. 7.1: A hexagon map of all bakeries in Germany.

The same information can also be conveyed using a heat map, as shown in Figure 7.2.



**Fig. 7.2:** A heat map of all bakeries in Germany.

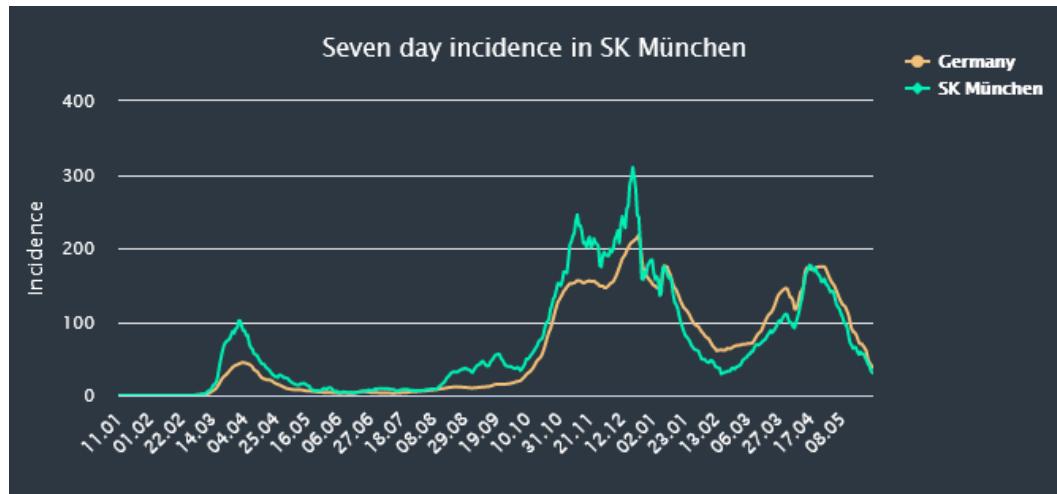
Through a choropleth map the number of bakeries in each municipality can be visualized, as seen in Figure 7.3.



**Fig. 7.3:** A choropleth map of all bakeries in Germany.

The app also displays a histogram of the variable selected by the user. Finally, the user has the option to compare a municipality with the rest of the country in terms of indicators of Covid-19 severity. These indicators are the daily number of infections,

the total number of infections and the seven-day incidence. This can be used to create charts like the one in Figure 7.4.



**Fig. 7.4:** The seven-day incidence in Munich compared to Germany.

For Europe, the user can view the features used in the calculation of the temporal models. The map can show mobility in different European countries, government measures and general key figures. The user can select a date to see, for example, which measurements were taken at Christmas 2020. Only a choropleth map is available here. Again, the user can view indicators of Covid-19 severity, this time comparing a country with the rest of Europe.

### The SIR for Norway and Germany

This is self-explanatory. The user can display the standardized incidence ratio either in Norway or in Germany.

### Spatial Modelling

Here the user can calculate his own BYM2 model, either for Norway or for Germany. Different variables can be selected to be added to the model, as well as values for  $\sigma_0$  and  $\alpha$ , which are used for the PC prior. After pressing the button to calculate the model, which takes a few seconds, the map is updated and the user can now choose to display the relative risk, the posterior mean of the random effects, the exceedance probability or the spatial field. Two tables are also displayed. One containing all relevant performance measures and one containing all significant effects. Each time

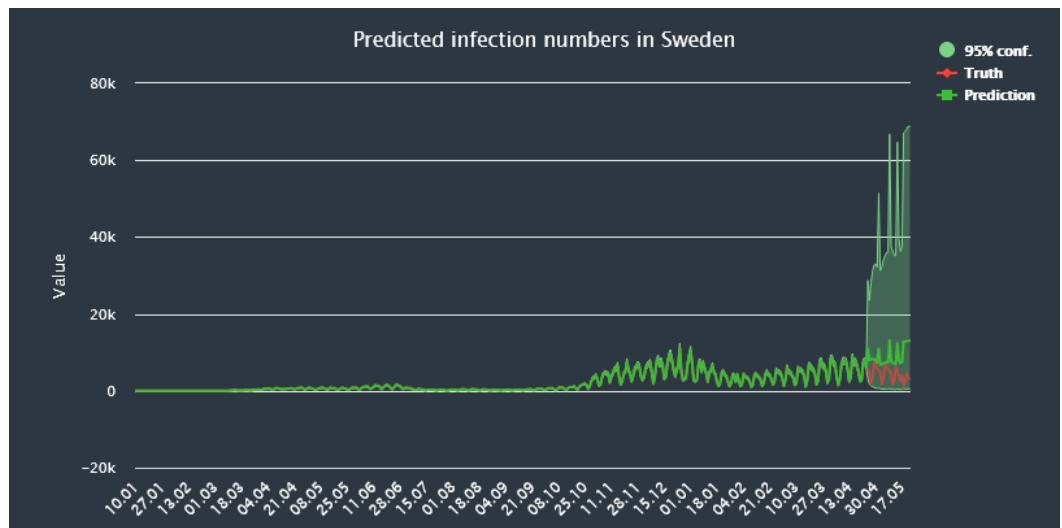
the user calculates a new model, new rows are added to the table. Via the ID column, the user can keep track of his models and see how the performance of a model changes or which effects turn out to be significant in which model.

## Temporal Modelling

This is the same as spatial modelling, but this time the user has the option to specify the type of temporal term to use, the country to use for modelling and the test size. The user can choose between five types of temporal terms:

1. An iid term
2. A random walk of length 1
3. A random walk of length 2
4. An autoregressive process of order 1
5. An Ornstein–Uhlenbeck process

33 countries can be selected for modelling and any number can be set for the test size. After the models have been calculated, two graphs are displayed. One for the predicted infection numbers in a country and one for the posterior temporal trend in that country. The graph for the predicted numbers includes a confidence interval for these predictions as well as the actual infection numbers in the country, an example of which can be seen in Figure 7.5.



**Fig. 7.5:** Predicted numbers in Sweden using an ar1 model with a test size of 28.

The same tables used in the Spatial Modelling tab are also used here and contain the same functionality.

The dashboard can be accessed via the following link:

<https://dashboard.f-hahn.de/>

# Discussion

After the calculation of the models, the next step is to critically examine and evaluate the results. The evaluation of the spatial models follows a simple scheme. First, a brief look is taken at the models without the spatial component, before the spatial models are reviewed and compared with the non-spatial models. Next, it is examined which factors significantly influence the risk of infection and why these factors might have an impact. For this, the coefficient of the BYM2 model are analysed. Finally, a look at area-specific risk is taken to see which regions in a given country are most at risk.

## 8.1 Discussion of the (Non)-Temporal Models

### 8.1.1 Discussion of the (Non)-Temporal Models for Norway

The non-spatial model for Norway identifies five significant effects:

- The number of unemployed immigrants
- The total number of immigrants
- The urban density
- The proportion of females

Moreover, the intercept is significant.

Comparing the spatial models with the non-spatial models using Table 6.11, it can be seen that BYM2, Besag and the non-spatial model perform almost equally well in terms of the DIC and WAIC, while the Leroux model performs best in terms of these two metrics. In terms of CPO, all spatial models perform better than the non-spatial model, with the Leroux model again performing best. However, the MAE shows that the non-spatial model has the best predictive performance, ahead of the Besag model, the BYM2 model and the Leroux model. This shows that the Leroux model overfits the training data more than the other models, a characteristic that can be seen in Figure 6.12. The fact that the non-spatial model performs better than the

spatial models is an indication that the spatial effect in Norway is not strong and that a different class of models may be better suited to identify critical factors affecting infection rates. Table 6.12 shows that adding the spatial effect did not suddenly cause any variables to become significant, while all variables, including the intercept, that are significant in the non-spatial model remain significant. Nevertheless, significant effects are found and have to be discussed.

The exponentiated intercept implies a risk rate of -59.1% across Norway. Factors related to immigration play a crucial role in the relative risk of developing Covid-19. A one standard deviation increase in the number of unemployed immigrants leads to a 22.6% increase in risk and a one standard deviation increase in the total number of immigrants leads to a 22.8% increase in risk. Unfortunately, there are no studies analysing the relationship between unemployed immigrants and the risk of developing Covid-19, as studies mostly focus on the relationship between unemployment in general and the risk of disease. Unemployment, however, turns out to be non-significant in the BYM2 model. Looking at older studies, Elkeles and Seifert (1996) conducted a longitudinal study that looked at the unemployment status and health of labour migrants in Germany. They found that immigrants are often employed in jobs with higher health risks and stress and therefore had poorer health (Elkeles & Seifert, 1996). Sia et al. (2019) analysed the association between immigration status and unemployment and men's and women's health using data from the Canadian Health Measures Survey. The study provided evidence of biological associations between unemployment and the likelihood of common chronic diseases, inflammation and possible malnutrition, with unemployed immigrants, and particularly unemployed immigrant women, being more prone to chronic diseases (Sia et al., 2019). Thus, there is precedent for an association between unemployment and immigrants and a higher risk of disease, however, further research would need to be conducted to analyse the association between unemployed immigrants and the risk of Covid-19 infection.

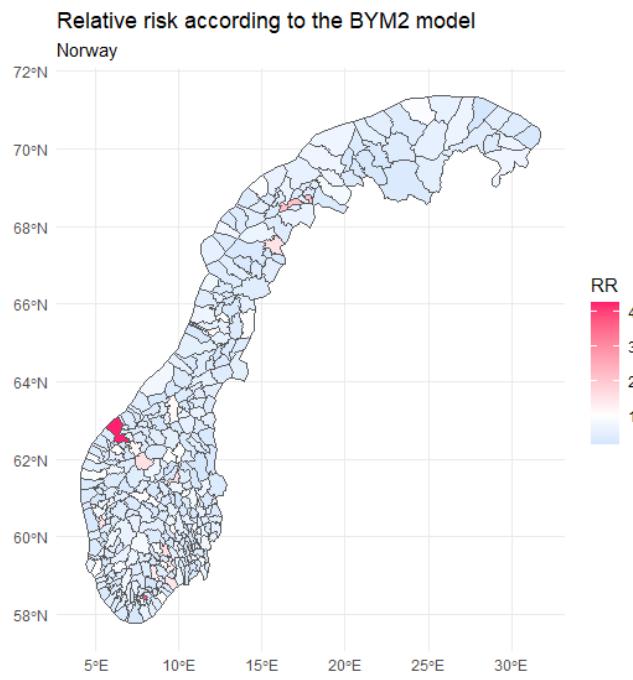
The last factor that influences the relative risk is urban density, where an increase of 1 standard deviation leads to an increase in the relative risk of 20.1%. The relationship between a higher number of residential buildings in a given area and the number of infections is probably due to the fact that with an increasing number of residential buildings comes an increasing number of inhabitants. A look at the Bravais-Pearson correlation coefficient confirms this assumption with  $\rho = 0.7070$ . Furthermore, the correlation between urban density and population density is 0.8037. This is convenient since studies analysing the relationship between urban density and Covid-19 usually define urban density as the number of inhabitants per square kilometre. Because of the positive relationship between urban density and population density, urban density acts as a proxy for population density in these models, so the relation-

ship between higher population density and higher case numbers should account for the higher infection rates in areas with higher urban density. According to Jamshidi et al. (2020) and Whittle & Diaz-Artiles (2020), higher urban/population density and higher infection numbers are due to a decrease in proximity between people and an increase in the likelihood of interpersonal contact (Jamshidi et al., 2020; Whittle & Diaz-Artiles, 2020). Sigler et al. (2020) found that dense urban environments provide more opportunities for the virus to spread, but that higher density had a stronger effect earlier and decreased in strength and importance over time(Sigler et al., 2020).

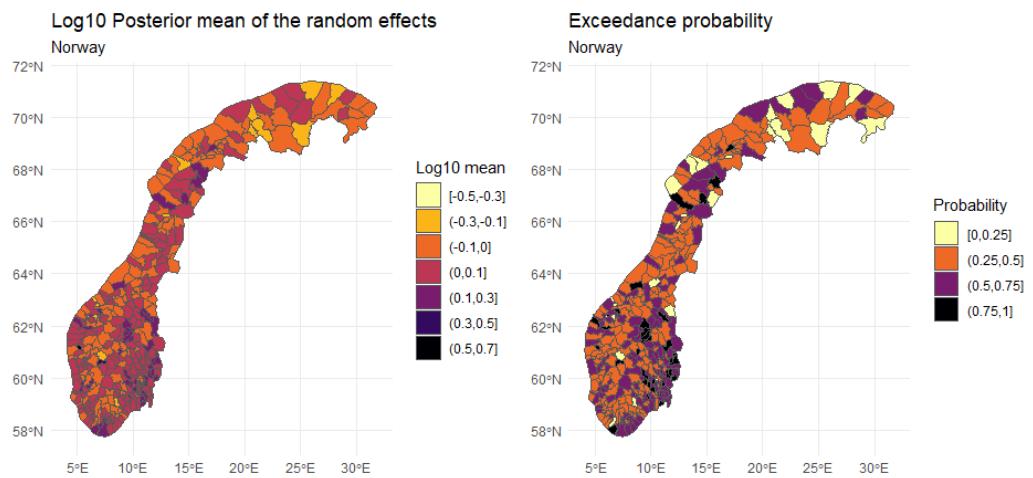
A higher risk among immigrants in Norway is found in a study by Indseth et al. (2020). The reasons given are barriers to adequate information due to low health literacy in certain groups and misconceptions about Covid-19 or test criteria, as well as other socio-economic and environmental factors (Indseth et al., 2020). Therefore, the results of the BYM2 model are consistent with this study.

Looking at the factors that reduce the risk of infection for Covid-19, a 1 standard deviation increase in the proportion of women leads to an 18.9% reduction in the risk of infection. This result is not consistent with current research. Many studies have investigated the relationship between gender and Covid-19 and most come to the same conclusion. While the risk of infection is the same in men and women, men tend to experience a higher severity and mortality rate for Covid-19 compared to women (Gausman & Langer, 2020; Kopel et al., 2020; Mukherjee & Pahan, 2021; Spagnolo et al., 2020).

A look at the relative risk of infection in Figure 8.1 shows that the relative risk is below 1 in most of Norway, which is not too surprising considering that Norway has managed the pandemic quite well so far. The two municipalities with the highest relative risk are Iveland and Ålesund. However, a look at the posterior probability in Figure 8.2 shows that the posterior probability of the risk being greater than 1 is around 0.5 for Ålesund (0.503) and less than 0.75 for Iveland (0.640). Looking at the log posterior mean of the random effects, there is an increased risk in the regions around Oslo and in large parts of southern and central Norway, while the risk tends to be lower in the northern parts of the country.



**Fig. 8.1:** Relative risk of contracting Covid-19 in Norway.



**Fig. 8.2:** Posterior mean of the municipality-specific relative risks  $\zeta = \exp(\xi)$  compared with the whole of Norway (left) and posterior probability  $\mathbb{P}(\zeta_i > 1 | \mathbf{y})$

### 8.1.2 Discussion of the (Non)-Temporal Models for Germany

In the non-spatial model for Germany, six coefficients are significant, in addition to the intercept:

- The percentage of the vote for the right-wing populist AfD
- The population density
- The logarithmic trade tax
- The percentage of the vote for the SPD
- The percentage of the vote for the left-wing party "Die Linke"
- The percentage of the vote for the Green party

A look at Table 6.9 shows that the spatial models outperform the non-spatial model in terms of the DIC and the WAIC, while all models perform about equally well in terms of the CPO. The predictive performance is best for the BYM2 model, just ahead of the Besag model, as indicated by the lowest values for the MAE. When adding the spatial term for the BYM2 model, several variables lose their significance, namely the percentage of the vote for the SPD, "Die Linke" and the Greens, while the intercept, the percentage of the vote for the AfD, population density and the logarithmic trade tax remain significant.

The exponentiated intercept implies a risk rate of -6.8% across Germany. The greatest influence on the relative risk is determined to be the percentage of the vote for the far-right AfD. An increase in this variable by 1 standard deviation leads to an increase in the relative risk by 25.7%. The AfD openly criticizes the measures taken by the government in Germany to prevent the spread of Covid-19, which leads to a large proportion of the party's voters not taking the measures seriously and refusing to keep a safe distance or wear a mask in public spaces. Several studies have taken a look at the response of right-wing parties to the pandemic and how people who vote for these types of parties have reacted. Wondreys and Mudde (2020) point out that these parties were quick to warn about the virus, but once cases spiked, they criticized the measures taken to contain the spread of the virus. They noted that right-wing parties often rejected the measures proposed by the leading parties because they themselves are part of the opposition (Wondreys & Mudde, 2020), as is the case with the AfD in Germany. Vieten (2020) shows how the far-right mobilizes people for anti-hygiene or anti-lockdown protests and how this is used to normalize the global far-right (Vieten, 2020). Eberl et al. (2020) analysed data from the Austrian Corona Panel Project to test whether populism

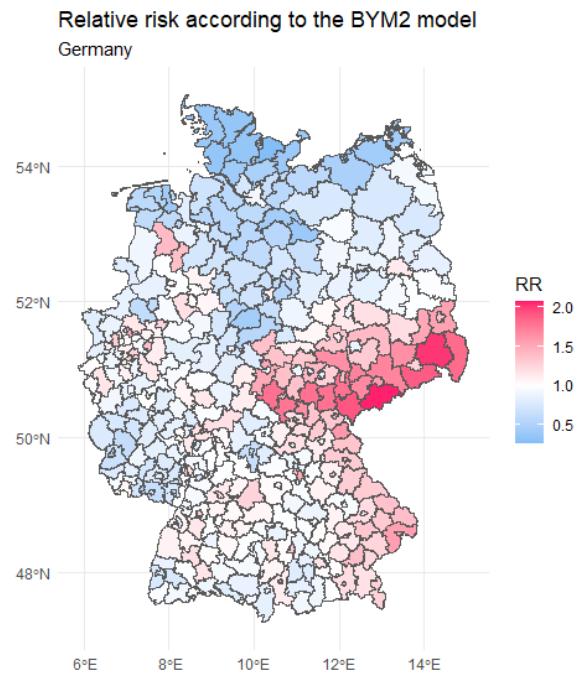
attitudes and belief in conspiracy theories related to Covid-19 are correlated. Using structural equation modelling, they show that populism indirectly influences Covid-19 conspiracy beliefs through trust in political and scientific institutions. They find that populist attitudes have a negative correlation with trust in the government and the parliament. Furthermore, they find that higher populist attitudes are negatively correlated with trust in science, a factor that reduces belief in Covid-19 conspiracy theories (Eberl et al., 2020). Finally, Farias and Pilati (2020) conducted a study in Brazil to predict social distancing violation intention and past non-compliance during the Covid-19 pandemic, controlling for the effects of interolerance of insecurity and socio-demographic variables. Their results included that individuals who support right-wing parties are more likely to violate social distancing measures (Farias & Pilati, 2020).

An increase in population density by 1 standard deviation leads to a risk increase of 10.9%. Reasons why a higher population density correlates positively with the number of infections are discussed in Section 8.1.1.

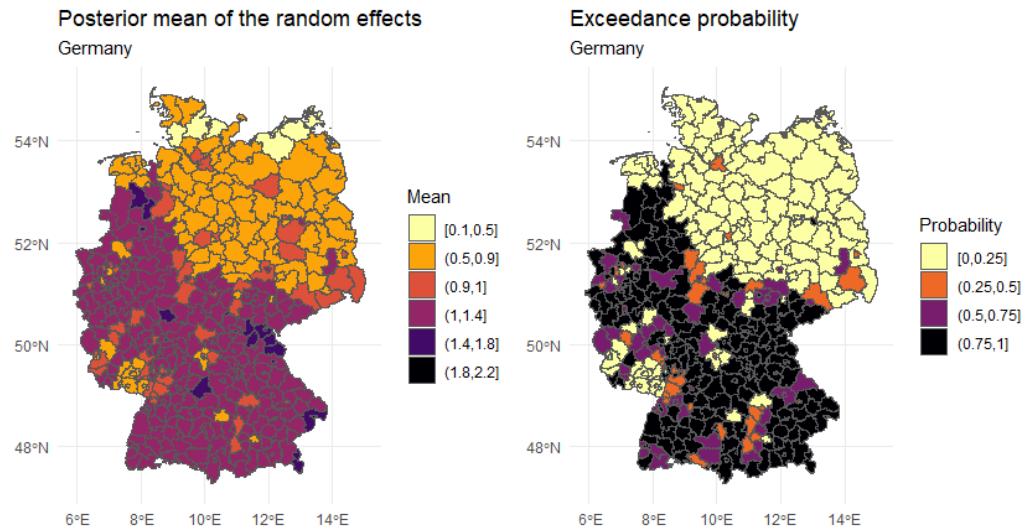
The last significant effect is the logarithmic trade tax with an increase of 1 standard deviation leading to a 6.8% increase in risk. There are no studies that specifically analyse the relationship between the trade tax and the risk of infection, but some studies have analysed infection rates in different spatial areas while controlling for factors such as income. In general, a negative relationship is found between areas with higher income and infection rates, meaning that areas with higher income had fewer cases of Covid-19. Cordes and Castro (2020) found that postcode areas in New York City with a low proportion of positive tests had higher incomes and tested less compared to lower income areas. They found that people in lower income areas are more likely to be without health insurance (Cordes & Castro, 2020). Coven and Gupta (2020) found that New York City residents who come from wealthier neighbourhoods are more likely to flee the city, and that people who live in low-income neighbourhoods are more likely to have frontline occupations and visit retail shops more often, increasing their exposure to Covid-19 (Coven & Gupta, 2020). The situation seems to be the same in Europe, as Baena et al. (2020) found that districts in Barcelona with a lower average income had a higher Covid-19 incidence. They found that the incidence in the district with the lowest income is 2.5 times higher than in the district with the highest income (Baena-Díez et al., 2020). Again, the results of the BYM2 model are not consistent with current research.

The relative risk of infection shown in Figure 8.3 is highest in Eastern Germany, more specifically in the federal state of Saxony. Saxony has established itself as the political stronghold of the AfD in recent years, which has even led to the ruling party, the Union, moving further to the right on the political spectrum. In addition to Saxony, the traditionally highly conservative Bavaria and the populous Ruhr region

have a relative risk of over 1. Figure 8.4 shows that for most of these regions the posterior mean of the random effects is over 1, mostly with a posterior probability of at least 0.75.



**Fig. 8.3:** Relative risk of contracting Covid-19 in Germany.



**Fig. 8.4:** Posterior mean of the municipality-specific relative risks  $\zeta = \exp(\xi)$  compared with the whole of Germany (left) and posterior probability  $\mathbb{P}(\zeta_i > 1 | \mathbf{y})$

### 8.1.3 Comparison Between the Spatial Models and the Predictive Models

Looking at the predictive models again and comparing them to the spatial models, both Table 6.13 and Table 6.14 show that the predictive models perform significantly worse than the spatial models. For Germany, the test MAE of the random forest is about 80% higher than that of the BYM2 model, while in Norway it is about 50% higher. This could indicate that the spatial effect in Norway is not as strong as in Germany, mainly due to the fact that most of the cases in Norway are located in Oslo and the surrounding municipalities. However, the smaller relative difference could be due to the fact that the infection numbers in Norway are significantly lower than in Germany, which naturally leads to a smaller mean absolute error. Nevertheless, the comparison between the two model classes is interesting, as different features turn out to be important for both countries, compared to the BYM2 model for each country. For Germany, it is the number of clinics, the number of public transport platforms and the number of marketplaces, while for Norway it is the number of places of worship, the number of offices, the number of public transport platforms and the number of colleges and universities. For Germany, the logarithmic business tax is the only feature that is important in both models, while for Norway, urban

density is the only feature that is important in the BYM2 model and the random forest.

Most interesting here, however, is that in both countries the number of public transport platforms is an important factor in explaining infection rates in a municipality. The association between the risk of Covid-19 infection and public transport is well studied. Hu et al. (2021) conducted a study in which they analysed the risk of transmission of the disease on high-speed trains in China. To do this, they used data from around 2300 patients and 72,000 close contacts who had a travel time between 0 and 8 hours from 19 December 2019 to 6 March 2020. They analysed the spatial and temporal distribution of Covid-19 transmission among train passengers to identify associations between infection, spatial distance and travel time. They found a high risk of transmission among train passengers, which showed significant differences depending on travel time and seat position. They suggest reducing the risk of transmission by increasing seat pitch, reducing passenger density and using hygiene equipment (Hu et al., 2021). Musselwhite (2020) compares Covid-19 with other infectious diseases that may have similar properties, noting that other types of human coronavirus, e.g. SARS coronavirus or MERS coronavirus, can survive on inanimate surfaces such as metal, glass or plastic for up to nine days and similar surface stability has been observed for SARS-CoV-2. In addition, a link has already been observed between acute respiratory infection (ARI) in winter and the use of buses or trams in the five days before the onset of symptoms. Finally, the greatest risk for infectious diseases on public transport is proximity between people in an enclosed environment. If people do not close their mouths when coughing and sneezing, public transport can become a significant source of microorganisms (Musselwhite et al., 2020).

## 8.2 Discussion of the Temporal Models

The temporal models for Germany and Norway do not perform as well as hoped. One of the problems for the models is that no distribution is found that fits the data reasonably well. In the end, a negative binomial distribution is used because it seems to fit slightly better than a normal distribution. Especially the QQ-plots in Figure 6.32 and Figure 10.6 illustrate this problem. The models themselves do not show an unreasonably high value for the MAE of the test data, but this is mainly due to the fact that the test set is kept relatively small with only 20 observations. Table 8.1 shows what happens when the test size for the temporal models for Germany is increased. The MAE in training remains surprisingly low, but the test MAE increases exponentially as the test size increases.

**Tab. 8.1:** The performance measures for different types of temporal models for Germany.

Model	Test size	MAE <sub>train</sub>	MAE <sub>test</sub>
Only date as covariate	20	4915	27450
Random walk of second order	20	2009	6180
AR(1) process	20	18	7524
Temporal model	20	1931	5821
Only date as covariate	40	4782	22559
Random walk of second order	40	1832	2090247
AR(1) process	40	7	26561
Temporal model	40	1764	10457609
Only date as covariate	60	4597	22525
Random walk of second order	60	1656	5.81e+12
AR(1) process	60	8	99471
Temporal model	60	1584	5.09e+16

The same effect can be observed for Norway. Even with the small test size, which leads to a low mean absolute error of the test, the quality of the prediction in Germany is not as high, since according to the prognosis the number of infections increases, while in reality the number of infections in Germany decreases. This is illustrated by the comparison between the predicted 7-day incidence and the actual 7-day incidence in Figure 6.37, which shows an increase in the predicted 7-day incidence, while in reality the incidence decreases rapidly. For Norway, the comparison between incidences is slightly better, as at least for the last days of the test set, the predicted incidence and the actual incidence both increase as can be seen in Figure 6.41.

Finally, despite trying a variety of models during this work, hardly any coefficient turned out to be significant and interpretable, as often significant coefficients had

values in the tens of thousands. In the models that are ultimately used in this work, the only significant feature besides the intercept turned out to be workplace mobility, with a one standard deviation increase in the mobility leading to a risk increase of 18.6% in Germany and 20.1% in Norway.

Studies on indoor infection risk are numerous and mostly point to the same thing, namely that aerosols from infected people can effectively transmit the virus indoors. Measures such as wearing face masks and ventilating rooms can reduce the risk of infection, but the most effective measure is still not going to the workplace (Bedford et al., 2020; Hall et al., 2020; Lelieveld et al., 2020).

# Conclusion

The aim of this work was to identify factors that have a significant impact on the current Covid-19 infection numbers in Germany and Norway using both a Bayesian approach that takes into account the spatial neighbourhood structure of each country and a non-Bayesian machine learning approach, and to compare these approaches to see which one proves more useful for this type of analysis. Another goal was to determine factors influencing infection numbers over time, namely government actions in response to the pandemic, changes in population mobility patterns, and the prevalence of different strains of Covid-19.

Based on the analysis of the differences between the spatial Bayesian and the non-Bayesian models for which no temporal effect is used, it can be concluded that the use of the Bayesian model with inclusion of a spatial term showed superior performance in terms of the mean absolute error, both in training and in testing. This was observed for Germany and for Norway.

For Germany, the factors that have a significant impact on the prevalence of Covid-19 are population density, the percentage of votes for the right-wing party AfD and the logarithmic trade tax. All three factors had a positive effect on the predicted infection rates, with current research clearly supporting the link between population density and infection rates, as well as the link between right-wing parties and the tendency for people who support these parties to be at higher risk of infection due to not following Covid-19 guidelines. On the other hand, recent research suggests that people living in higher income or richer regions have a lower risk of infection, while the opposite is suggested by the model calculated for Germany.

For Norway, these factors are urban density, the number of unemployed immigrants in a municipality, the total number of immigrants in a municipality and the proportion of women. The proportion of women is the only effect where a negative influence on the predicted infection numbers is observed. However, this association is not supported by current research which suggests that men and women are equally likely to contract Covid-19. No research has yet been conducted to analyse the association between unemployed immigrants and the risk of Covid-19, which makes it difficult to evaluate this effect as there is no research to compare it to. The association between immigrants in Norway and the risk of infection with Covid-19, as well as the association between the risk of infection and urban density,

is consistent with current research which suggests that immigrants in Norway have a higher risk of infection and that higher density in urban areas leads to a higher risk of infection.

Temporal modelling proved more difficult, with no distribution fitting the data reasonably well for Germany or Norway. A small test size limited the interpretation of the results. For both countries, the only significant effect on infection rates found was mobility in workplaces with a higher workplace mobility leading to a higher risk of infection. This is in line with current research.

Based on these conclusions, people living in areas with characteristics such as a higher proportion of people voting for right-wing parties or a higher population or urban density should be cautious in their daily lives and keep a safe distance from other people to limit their risk of contracting Covid-19.

Further research is needed to determine the relationship between areas in Germany that have a higher logarithmic trade tax and the risk of contracting Covid-19 in these areas. Research can never take all factors into account. Therefore, the association between these two variables could possibly be explained by a third variable. The same applies to the association between the proportion of women in a Norwegian municipality and the risk of contracting Covid-19.

Future research could analyse the spatio-temporal relationship between the risk of infection and various factors such as government measures. In addition to nationwide government measures, there are local government measures, both in Germany and Norway. Obtaining this data is a time-consuming task that would require manual collection of this data from official local government sources, as currently no comprehensive dataset exists for either country. Developing a spatio-temporal model based on the demographic and infrastructural variables used for the spatial models in this thesis, as well as the variables used for the temporal models, has the potential to lead to new insights into which factors are driving up infection rates and which measures are successful in preventing new infections.

In summary, this thesis has shown that a Bayesian approach that models the spatial neighbourhood structure in a country is superior to an approach where no spatial neighbourhood structure is modelled. Furthermore, several factors were found to positively influence the risk of Covid-19 infection, both in Germany and Norway, and are supported by current research.

# Appendix

## 10.1 Probability Distributions and the Exponential Family

### 10.1.1 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  $\mathbf{y}$  belongs to the exponential family if the discrete or continuous density with respect to a  $\sigma$ -finite measure of  $\mathbf{y}$  has the form

$$f(\mathbf{y}|\boldsymbol{\theta}, \lambda) = \exp\left(\frac{\mathbf{y}^T \boldsymbol{\theta} - b(\boldsymbol{\theta})}{\lambda} + c(\mathbf{y}, \lambda)\right), \quad (10.1)$$

with  $c(\mathbf{y}, \lambda) \geq 0$ .  $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^q$  is the *natural* or *canonical* parameter of the exponential family, while  $\lambda > 0$  is a *dispersion* or *nuisance* parameter (Holland & Leinhardt, 1981). The natural parameter space  $\Theta$  is the set of all  $\boldsymbol{\theta}$  satisfying

$$0 < \int \exp\left(\left(\mathbf{y}^T \boldsymbol{\theta} - b(\boldsymbol{\theta})\right)/\lambda + c(\mathbf{y}, \lambda)\right) d\mathbf{y} < \infty \quad (10.2)$$

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

$$\mathbb{E}_{\boldsymbol{\theta}}(\mathbf{y}) = \mu(\boldsymbol{\theta}) = \frac{\partial b(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \quad (10.3)$$

$$\text{Cov}_{\boldsymbol{\theta}}(\mathbf{y}) = \Sigma(\boldsymbol{\theta}) = \lambda \frac{\partial^2 b(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}. \quad (10.4)$$

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

$$v(\mu) = \frac{\partial^2 b(\boldsymbol{\theta}(\mu))}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \quad (10.5)$$

is given and the covariance can be written as

$$\text{Cov}_{\boldsymbol{\theta}}(\mathbf{y}) = \lambda v(\mu). \quad (10.6)$$

Important members of the exponential family are the normal, binomial, Poisson, gamma and multivariate normal distribution (Fahrmeir & Tutz, 2013, p. 433).

### 10.1.2 The Normal Distribution

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

The density is given by

$$f(\mathbf{x}|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\mathbf{x}-\mu}{\sigma}\right)^2\right). \quad (10.7)$$

The first two moments of the distribution are given by

$$\mathbb{E}[X] = \mu \quad (10.8)$$

$$\text{Var}[X] = \sigma^2. \quad (10.9)$$

The graph of this density function has a "bell-shaped form" and is symmetrical with  $\mu$  as the centre of symmetry (Fahrmeir et al., 2016, pp. 83–85).

### 10.1.3 The Multivariate Normal Distribution

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

$$\pi(\mathbf{y}) = (2\pi)^{-n/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})\right), \quad \mathbf{y} \in \mathbb{R}^n \quad (10.10)$$

Here,  $\mu_i = \mathbb{E}[y_i]$ ,  $\Sigma_{ij} = \text{Cov}(y_i, y_j)$ ,  $\Sigma_{ii} = \text{Var}(y_i) > 0$  and  $\text{Corr}(y_i, y_j) = \Sigma_{ij}/(\Sigma_{ii}\Sigma_{jj})^{1/2}$ . This is written as  $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . For  $n = 1$ ,  $\mu = 0$  and  $\Sigma_{11} = 1$  the standard normal distribution is obtained.

$\mathbf{y}$  is now split up into  $\mathbf{y} = (\mathbf{y}_A^T, \mathbf{y}_B^T)$  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.  $\mathbf{y}_A \sim \mathcal{N}(\boldsymbol{\mu}_A, \Sigma_{AA})$ .
2.  $\Sigma_{AB} = \mathbf{0}$  precisely when  $\mathbf{y}_A$  and  $\mathbf{y}_B$  are independent.
3. The conditional distribution  $\pi(\mathbf{y}_A | \mathbf{y}_B)$  is  $\mathcal{N}(\boldsymbol{\mu}_{A|B}, \Sigma_{A|B})$ , where

$$\boldsymbol{\mu}_{A|B} = \boldsymbol{\mu}_A + \Sigma_{AB}\Sigma_{BB}^{-1}(\mathbf{y}_B - \boldsymbol{\mu}_B) \text{ and}$$

$$\Sigma_{A|B} = \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}.$$

4. If  $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$  and  $\mathbf{y}' \sim \mathcal{N}(\boldsymbol{\mu}', \Sigma')$  are independent,  
then  $\mathbf{y} + \mathbf{y}' \sim \mathcal{N}(\boldsymbol{\mu} + \boldsymbol{\mu}', \Sigma + \Sigma')$  (Rue & Held, 2005, pp. 19–20).

#### 10.1.4 The Poisson Distribution

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

The density is given by

$$f(k) = \mathbb{P}(X = k) = \begin{cases} \frac{\lambda^k}{k!} \exp(-\lambda) & \text{for } x \in \{0, 1, \dots\} \\ 0 & \text{else} \end{cases} \quad (10.11)$$

with  $\lambda$  representing the expected value of  $X$ .

The first two moments of the distribution are given by

$$\mathbb{E}[X] = \lambda \quad (10.12)$$

$$\text{Var}[X] = \lambda. \quad (10.13)$$

For  $\lambda \geq 10$  the distribution becomes approximately symmetrical and can thus be approximated by a normal distribution (Fahrmeir et al., 2016, p. 243).

#### 10.1.5 The Negative Binomial Distribution

The negative binomial distribution is a univariate probability distribution that belongs to the discrete probability distributions. It models the number of trials required

to achieve a given number of successes in a Bernoulli process.

The density is given by

$$f(k, r, p) = \mathbb{P}(X = k) = \binom{k+r-1}{r-1} (1-p)^k p^r, \quad (10.14)$$

with  $r$  the number of successes,  $k$  the number of failures, and  $p$  the probability of success.

The first two moments of the distribution are given by

$$\mathbb{E}[X] = \frac{pr}{1-p} \quad (10.15)$$

$$\text{Var}[X] = \frac{pr}{(1-p)^2}. \quad (10.16)$$

For large values of  $r$ , the negative binomial distribution can be approximated by a normal distribution (Haldane, 1941).

## 10.2 Moments

### 10.2.1 Skewness

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

The skewness of a random variable  $X$  is the third standardized moment, defined as

$$\text{Skew}[X] = \mathbb{E} \left[ \left( \frac{X - \mu}{\sigma} \right)^3 \right] = \frac{\mathbb{E} [(X - \mu)^3]}{\left( \mathbb{E} [(X - \mu)^{3/2}] \right)^2} = \frac{\mu_3}{\sigma^3}, \quad (10.17)$$

with  $\mu_3$  the third central moment and  $\sigma$  the standard deviation (Doane & Seward, 2011; Wilkins, 1944).

### 10.2.2 Kurtosis

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

The kurtosis of a random variable  $X$  is the fourth standardized moment, defined as

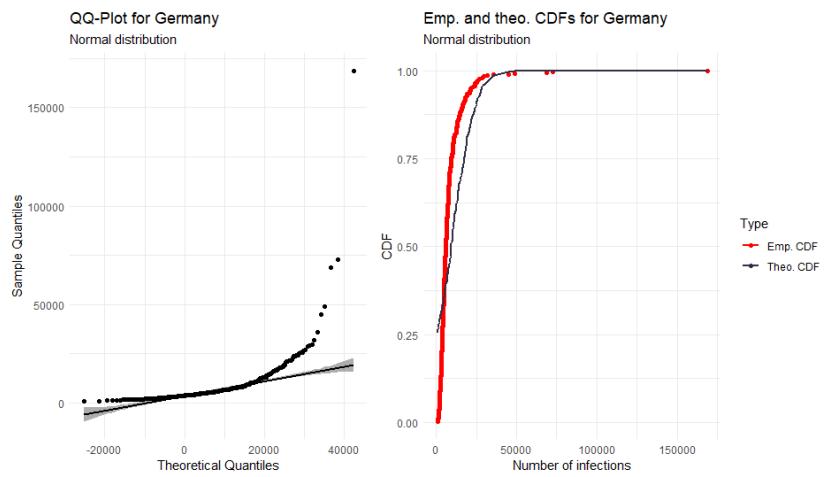
$$\text{Kurt}[X] = \mathbb{E} \left[ \left( \frac{X - \mu}{\sigma} \right)^4 \right] = \frac{\mathbb{E} [(X - \mu)^4]}{\left( \mathbb{E} [(X - \mu)^2] \right)^2} = \frac{\mu_4}{\sigma^4}, \quad (10.18)$$

with  $\mu_4$  the fourth central moment and  $\sigma$  the standard deviation (DeCarlo, 1997; Wilkins, 1944).

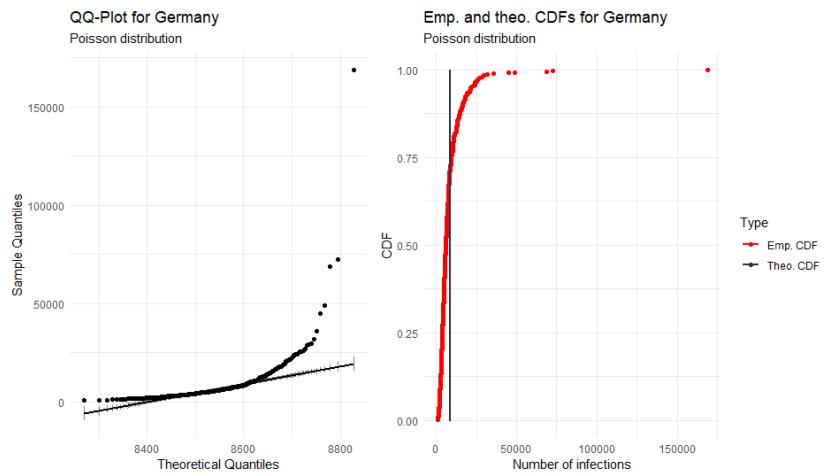
## 10.3 Distribution Fits

### 10.3.1 Distribution Fits for Germany

#### 10.3.1.1 Fits for the Non-Temporal Models



**Fig. 10.1:** A normal fit to the number of cases in German municipalities



**Fig. 10.2:** A Poisson fit to the number of cases in German municipalities

### 10.3.1.2 Fits for the Temporal Models

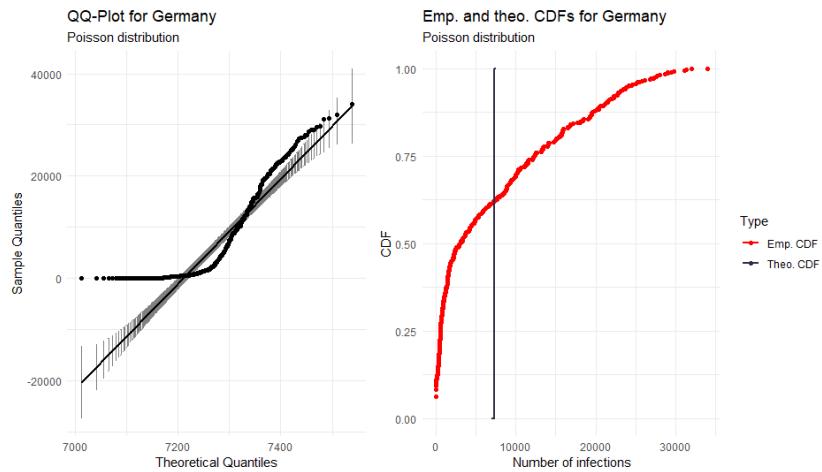


Fig. 10.3: A Poisson fit to the number of cases in German municipalities

## 10.3.2 Distribution Fits for Norway

### 10.3.2.1 Fits for the Non-Temporal Models

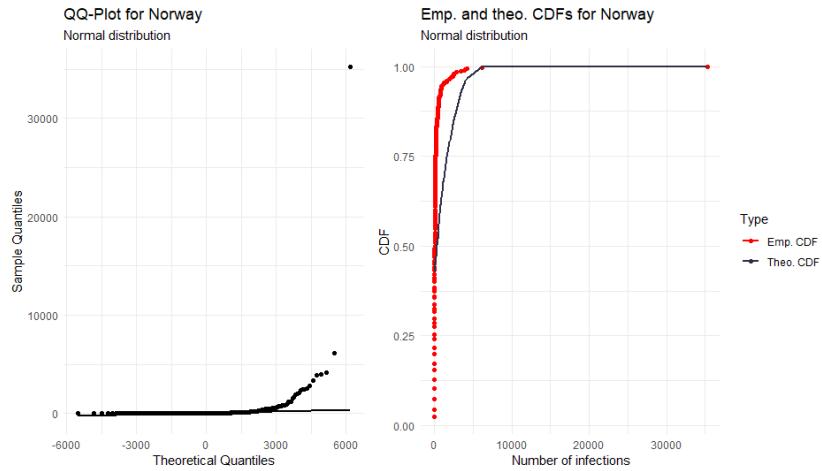
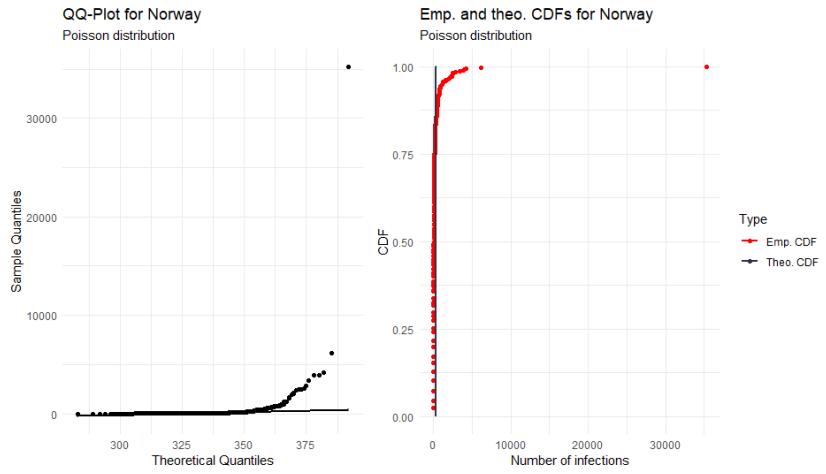
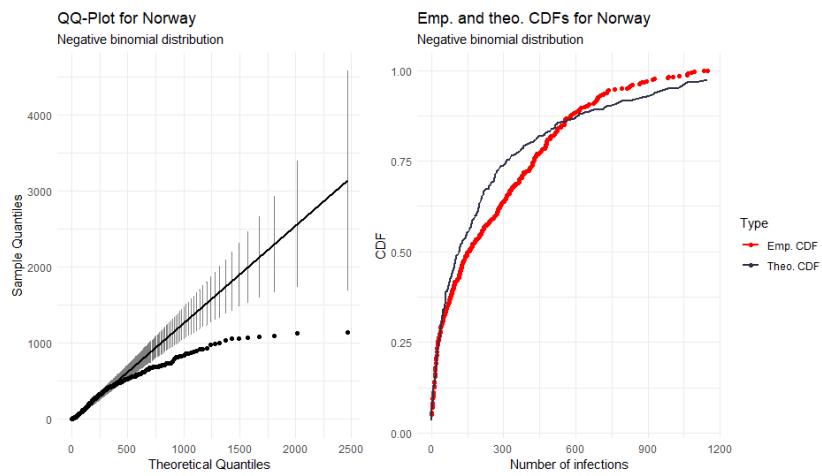


Fig. 10.4: A normal fit to the number of cases in Norwegian municipalities

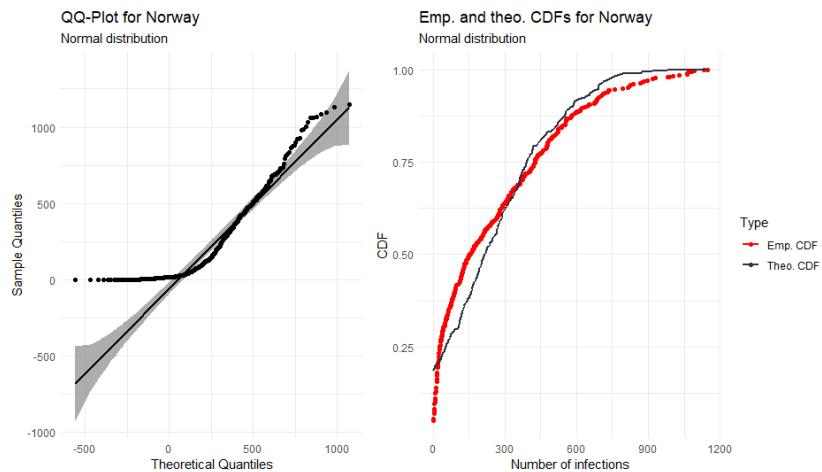


**Fig. 10.5:** A Poisson fit to the number of cases in Norwegian municipalities

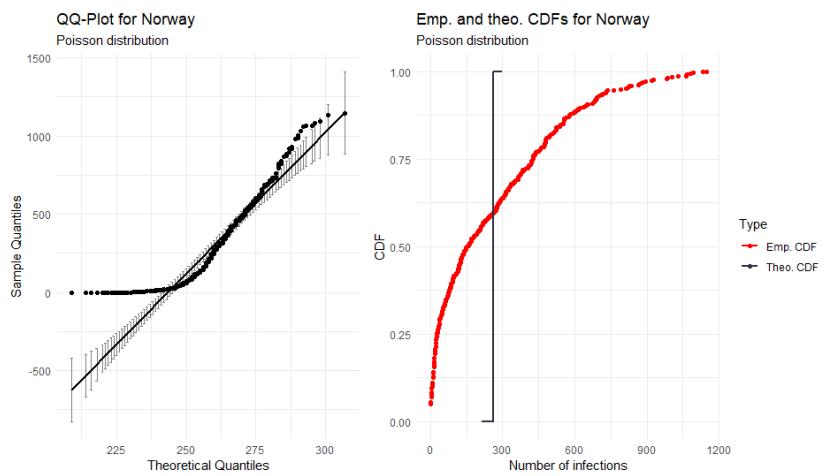
### 10.3.2.2 Fits for the Temporal Models



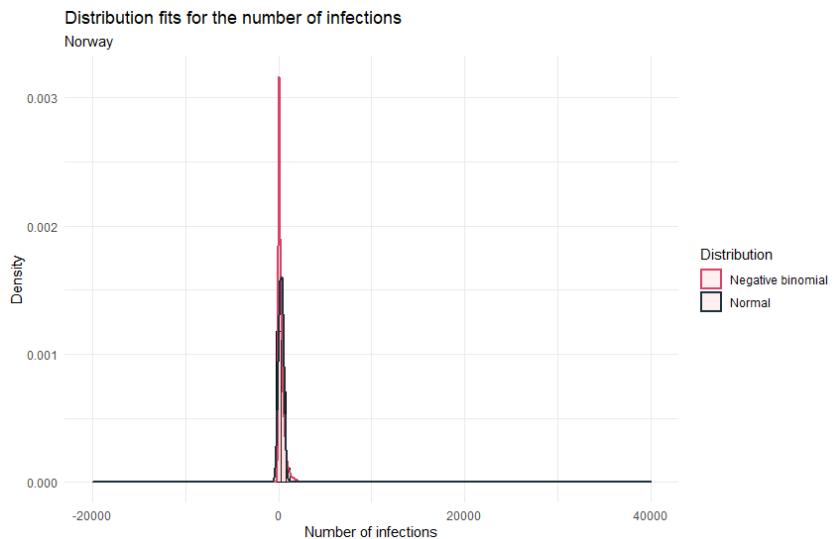
**Fig. 10.6:** A negative binomial fit to the number of cases in Norwegian municipalities



**Fig. 10.7:** A normal fit to the number of cases in Norwegian municipalities

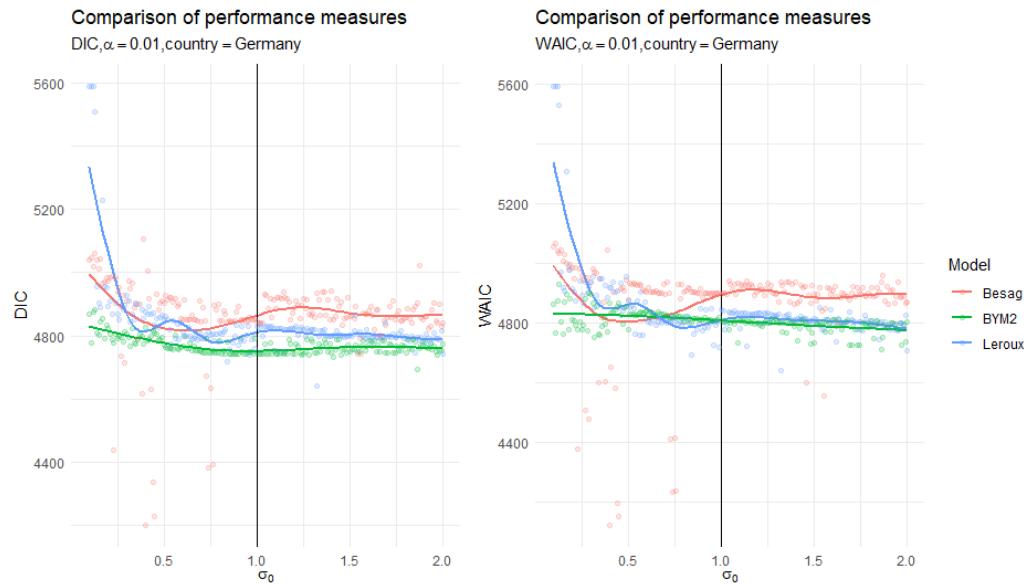


**Fig. 10.8:** A Poisson fit to the number of cases in Norwegian municipalities

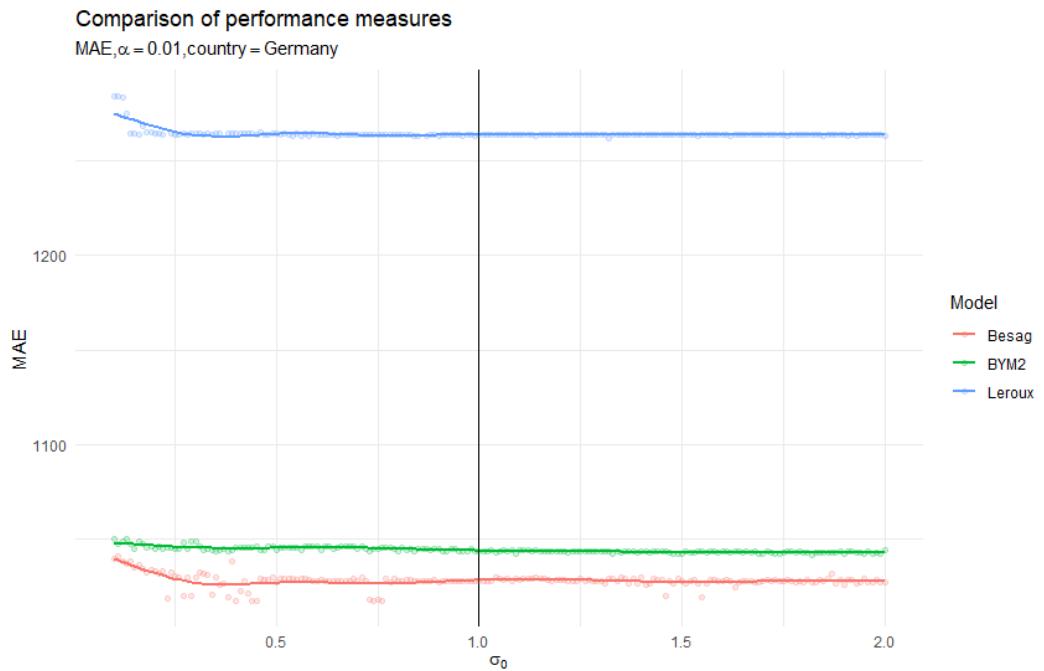


**Fig. 10.9:** Histogram for the number of cases in Norwegian municipalities with a normal and a negative binomial distribution overlayed.

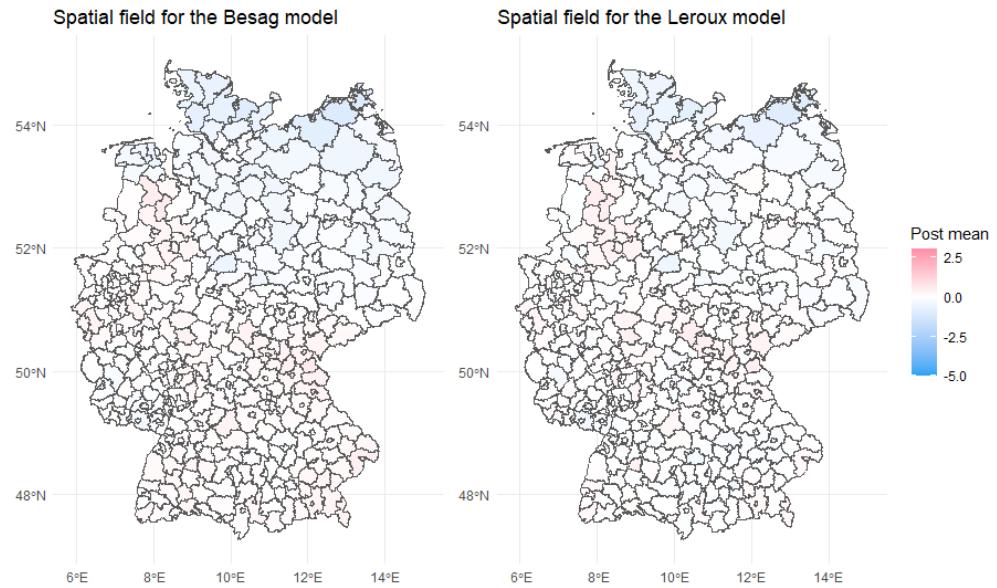
## 10.4 Choice of Hyperpriors for Germany



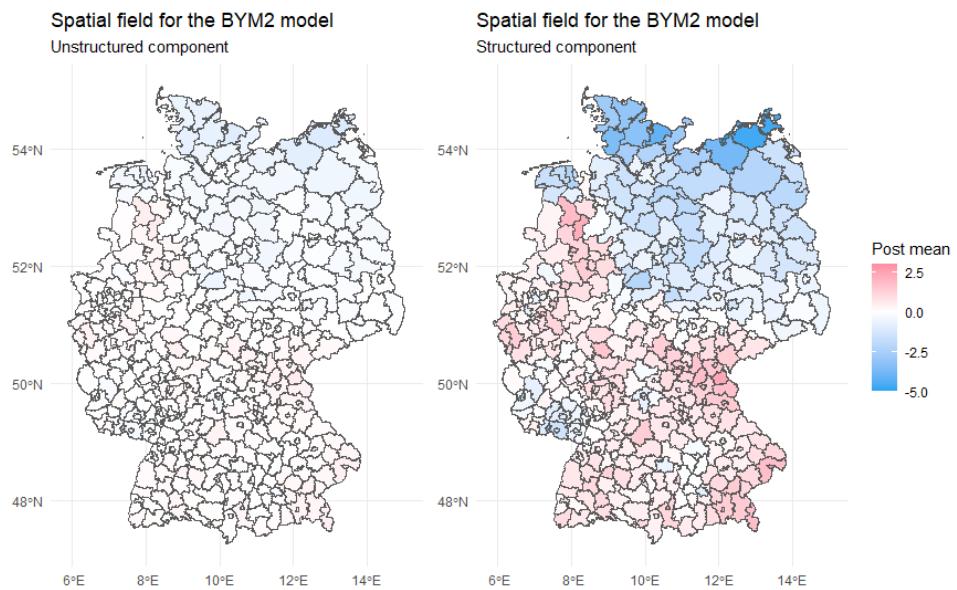
**Fig. 10.10:** Values of the DIC and the WAIC when changing the value for  $\sigma_0$ . The black line highlights the values for  $\sigma_0 = 1$ .



**Fig. 10.11:** Values of the MAE when changing the value for  $\sigma_0$ . The black line highlights the values for  $\sigma_0 = 1$ .



**Fig. 10.12:** Spatial field for a Besag model and a Leroux model.



**Fig. 10.13:** Spatial fields for a BYM2 model.

## 10.5 OpenStreetMap Key-Value Pairs

**Tab. 10.1:** A list of all the key-value pairs used to query OpenStreetMap, except the ones used for residential buildings

Key	Value	Description
amenity	cinema	A place where films are shown
	clinic	A medium-sized medical facility or health centre.
	college	Campus or buildings of an institute of Further Education
	dentist	A dentist practice / surgery
	doctors	A doctor's practice / surgery
	hospital	A hospital providing in-patient medical treatment
	kindergarten	For children too young for a regular school
	marketplace	A marketplace where goods and services are traded daily or weekly
	nightclub	A place to drink and dance
	nursing_home	A home for disabled or elderly persons
	place_of_worship	A church, mosque, or temple, etc.
	restaurant	A restaurant
building	school	School and grounds - primary, middle and secondary schools
	theatre	A theatre or opera house
leisure	university	An university campus: an institute of higher education
	fitness_centre	Fitness centre, health club or gym
public_transport	sports_centre	Facility where sports take place within an enclosed area
	platform	Public transport platforms
shop	bakery	Shop focused on selling bread
	chemist	Shop focused on selling articles of personal hygiene, cosmetics, household products
	convenience	A small local shop carrying a small subset of the items you would find in a supermarket
	hairdresser	Here you can get your hair cut, coloured
	supermarket	A large store with groceries and other items

**Tab. 10.2:** A list of all the key-value pairs used to query OpenStreetMap for residential buildings

Key	Value	Description
building	apartments	A building arranged into individual dwellings, often on separate floors
	bungalow	A single-storey detached small house
	cabin	A small, roughly built house
	detached	A free-standing residential building
	dormitory	A shared building intended for college/university students
	farm	A residential building on a farm
	ger	A permanent or seasonal round yurt or ger used as dwelling
	hotel	A building designed with separate rooms available for overnight accommodation
	house	A dwelling unit inhabited by a single household
	houseboat	A boat used primarily as a home
	residential	A building used primarily for residential purposes
	semidetached_house	A residential house that shares a common wall with another on one side
	static caravan	A mobile home (semi)permanently left on a single site
	terrace	A linear row of residential dwellings

## Combination of Key-Value Pairs

The following table shows which key-value pairs were combined, due to their similarity.

**Tab. 10.3:** A list of all the key-value pairs that were combined to create variables

Key	Value	Variable
amenity	cinema nightclub theatre	Entertainment
amenity	clinic dentist doctors	Clinic
amenity	college university	Higher Education
leisure	fitness_centre sports_centre	Sports
shop	chemist convenience supermarket	Shop

## 10.6 Software Used

All the analysis that was carried out was done so using the 4.1 version of the statistical software R (R Core Team, 2021). The following packages were used as part of the analysis:

- `covid19germany` (Schmid et al., 2021)
- `covidregionaldata` (Abbott et al., 2020)
- `data.table` (Dowle & Srinivasan, 2021)
- `dashboardthemes` (Lilovski, 2020)
- `dplyr` (Wickham et al., 2021)
- `DT` (Xie et al., 2021)
- `eurostat` (Lahti et al., 2017)
- `fitdistrplus` (Delignette-Muller & Dutang, 2015)
- `furrr` (Vaughan & Dancho, 2021)
- `geosphere` (Hijmans, 2019)
- `ggplot2` (Wickham, 2016)
- `here` (Müller, 2020)
- `highcharter` (Kunst, 2020)
- `iml` (Molnar et al., 2018)
- `INLA` (Kourounis et al., 2018)
- `ISOcodes` (Buchta & Hornik, 2021)
- `LaCroixColoR` (Bjork, 2021)
- `lwgeom` (Pebesma, 2021)
- `latex2exp` (Meschiari, 2021)
- `mapdeck` (Cooley, 2020)
- `MASS` (Venables & Ripley, 2002)
- `mlr` (Casalicchio et al., 2017)

- `osmdata` (Padgham et al., 2017)
- `parallelMap` (Bischl et al., 2020)
- `patchwork` (Pedersen, 2020)
- `pbapply` (Solymos & Zawadzki, 2020)
- `qqplotr` (Almeida et al., 2018)
- `readr` (Wickham & Hester, 2020)
- `regclass` (Petrie, 2020)
- `reshape2` (Wickham, 2007)
- `rlist` (Ren, 2016)
- `RSelenium` (Harrison, 2020)
- `sass` (Cheng et al., 2021)
- `sf` (Pebesma, 2018)
- `shiny` (Chang et al., 2021)
- `shinybusy` (Meyer & Perrier, 2021)
- `shinydashboard` (Chang & Borges Ribeiro, 2018)
- `shinyWidgets` (Perrier et al., 2021)
- `SpatialEpi` (Kim & Wakefield, 2018)
- `spdep` (Bivand et al., 2013)
- `stringr` (Wickham, 2019)
- `tibble` (Müller & Wickham, 2021)
- `units` (Pebesma et al., 2016)
- `waiter` (Coene, 2021)

# Bibliography

- Abbott, S., Sherratt, K., Palmer, J., Martin-Nielsen, R., Bevan, J., Gibbs, H., & Funk, S. (2020). Covidregionaldata: Subnational data for the covid-19 outbreak (cit. on p. 181).
- Ahmadi, M., Sharifi, A., Dorosti, S., Ghoushchi, S. J., & Ghanbari, N. (2020). Investigation of effective climatology parameters on covid-19 outbreak in iran. *Science of the Total Environment*, 729, 138705 (cit. on p. 11).
- Aitkin, M. (1991). Posterior bayes factors. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(1), 111–128 (cit. on p. 22).
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6), 716–723 (cit. on p. 52).
- Allcott, H., Boxell, L., Conway, J., Gentzkow, M., Thaler, M., & Yang, D. (2020). Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic. *Journal of Public Economics*, 191, 104254 (cit. on p. 12).
- Almeida, A., Loy, A., & Hofmann, H. (2018). *ggplot2 compatible quantile-quantile plots in r*. (Cit. on p. 182).
- Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3), 175–185 (cit. on p. 71).
- Baena-Diez, J. M., Barroso, M., Cordeiro-Coelho, S. I., Diaz, J. L., & Grau, M. (2020). Impact of covid-19 outbreak by income: Hitting hardest the most deprived. *Journal of Public Health*, 42(4), 698–703 (cit. on p. 158).
- Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. CRC press. (Cit. on pp. 36, 69).
- Bayes, T. (1763). 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. *Philosophical transactions of the Royal Society of London*, (53), 370–418 (cit. on p. 20).
- Bedford, J., Enria, D., Giesecke, J., Heymann, D. L., Ihekweazu, C., Kobinger, G., Lane, H. C., Memish, Z., Oh, M.-d., Schuchat, A., et al. (2020). Covid-19: Towards controlling of a pandemic. *The lancet*, 395(10229), 1015–1018 (cit. on p. 163).
- Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2) (cit. on p. 77).

- Bermudi, P. M. M., Lorenz, C., de Aguiar, B. S., Failla, M. A., Barrozo, L. V., & Chiaravallotti-Neto, F. (2021). Spatiotemporal ecological study of covid-19 mortality in the city of São Paulo, Brazil: Shifting of the high mortality risk from areas with the best to those with the worst socio-economic conditions. *Travel medicine and infectious disease*, 39, 101945 (cit. on p. 12).
- Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M., & Songini, M. (1995). Bayesian analysis of space—time variation in disease risk. *Statistics in medicine*, 14(21-22), 2433–2443 (cit. on p. 68).
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2), 192–225 (cit. on p. 48).
- Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1), 1–20 (cit. on pp. 3, 36, 49).
- Birattari, M., Yuan, Z., Balaprakash, P., & Stützle, T. (2010). F-race and iterated f-race: An overview. *Experimental methods for the analysis of optimization algorithms*, 311–336 (cit. on p. 78).
- Bischl, B., Lang, M., & Schratz, P. (2020). *Parallelmap: Unified interface to parallelization back-ends* [R package version 1.5.0]. (Cit. on p. 182).
- Bivand, R. S., Gómez-Rubio, V., & Rue, H. (2015). Spatial data analysis with R-INLA with some extensions. *Journal of Statistical Software*, 63(20), 1–31 (cit. on p. 104).
- Bivand, R. S., Pebesma, E., & Gomez-Rubio, V. (2013). *Applied spatial data analysis with R, second edition*. Springer, NY. (Cit. on p. 182).
- Bjork, J. (2021). *Lacroixcolor: Lacroix water color palettes* [R package version 0.1.0]. (Cit. on p. 181).
- Blangiardo, M., & Cameletti, M. (2015). *Spatial and spatio-temporal bayesian models with r-inla*. John Wiley & Sons. (Cit. on p. 47).
- Box, G. E., & Tiao, G. C. (2011). *Bayesian inference in statistical analysis* (Vol. 40). John Wiley & Sons. (Cit. on pp. 15, 21).
- Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2), 123–140 (cit. on p. 75).
- Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32 (cit. on p. 75).
- Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press. (Cit. on p. 73).
- Buchta, C., & Hornik, K. (2021). *Isocodes: Selected iso codes* [R package version 2021.02.24]. (Cit. on p. 181).
- Casalicchio, G., Bossek, J., Lang, M., Kirchhoff, D., Kerschke, P., Hofner, B., Seibold, H., Vanschoren, J., & Bischl, B. (2017). Openml: An R package to connect to the machine learning platform openml. *Computational Statistics*, 1–15 (cit. on p. 181).

- Cascella, M., Rajnik, M., Cuomo, A., Dulebohn, S. C., & Di Napoli, R. (2021). Features, evaluation, and treatment of coronavirus (covid-19). *Statpearls [internet]* (cit. on p. 8).
- Castro, M. C., Kim, S., Barberia, L., Ribeiro, A. F., Gurzenda, S., Ribeiro, K. B., Abbott, E., Blossom, J., Rache, B., & Singer, B. H. (2021). Spatiotemporal pattern of covid-19 spread in brazil. *Science* (cit. on p. 12).
- Chang, W., & Borges Ribeiro, B. (2018). *Shinydashboard: Create dashboards with 'shiny'* [R package version 0.7.1]. (Cit. on p. 182).
- Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2021). *Shiny: Web application framework for r* [R package version 1.6.0]. (Cit. on p. 182).
- Chen, Y., Li, Q., Karimian, H., Chen, X., & Li, X. (2021). Spatio-temporal distribution characteristics and influencing factors of covid-19 in china. *Scientific Reports*, 11(1), 1–12 (cit. on p. 10).
- Chen, Z.-L., Zhang, Q., Lu, Y., Guo, Z.-M., Zhang, X., Zhang, W.-J., Guo, C., Liao, C.-H., Li, Q.-L., Han, X.-H., et al. (2020). Distribution of the covid-19 epidemic and correlation with population emigration from wuhan, china. *Chinese medical journal* (cit. on p. 9).
- Cheng, J., Mastny, T., Iannone, R., Schloerke, B., & Sievert, C. (2021). *Sass: Syntactically awesome style sheets ('sass')* [R package version 0.4.0]. (Cit. on p. 182).
- Chu, W., Ghahramani, Z., & Williams, C. K. (2005). Gaussian processes for ordinal regression. *Journal of machine learning research*, 6(7) (cit. on p. 36).
- Coene, J. (2021). *Waiter: Loading screen for 'shiny'* [R package version 0.2.1]. (Cit. on p. 182).
- Cooley, D. (2020). *Mapdeck: Interactive maps using 'mapbox gl js' and 'deck.gl'* [R package version 0.3.4]. (Cit. on p. 181).
- Cordes, J., & Castro, M. C. (2020). Spatial analysis of covid-19 clusters and contextual factors in new york city. *Spatial and Spatio-temporal Epidemiology*, 34, 100355 (cit. on p. 158).
- Coven, J., & Gupta, A. (2020). Disparities in mobility responses to covid-19. *New York University* (cit. on p. 158).
- Cox, D. (1980). Discussion of ‘sampling and bayes’ inference in scientific modelling and robustness. *JR Stat Soc Ser A*, 143, 410 (cit. on p. 54).
- Craney, T. A., & Surles, J. G. (2002). Model-dependent variance inflation factor cutoff values. *Quality Engineering*, 14(3), 391–403 (cit. on p. 56).
- Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons. (Cit. on p. 37).
- Danielsson, P.-E. (1980). Euclidean distance mapping. *Computer Graphics and image processing*, 14(3), 227–248 (cit. on p. 71).

- Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(1), 1–15 (cit. on p. 21).
- DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological methods*, 2(3), 292 (cit. on p. 170).
- Delignette-Muller, M. L., & Dutang, C. (2015). fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software*, 64(4), 1–34 (cit. on p. 181).
- Dey, D. K., Ghosh, S. K., & Mallick, B. K. (2000). *Generalized linear models: A bayesian perspective*. CRC Press. (Cit. on p. 36).
- Diaconis, P., & Ylvisaker, D. (1979). Conjugate priors for exponential families. *The Annals of statistics*, 269–281 (cit. on p. 24).
- Diggle, P. J., Ribeiro, P. J., & Christensen, O. F. (2003). An introduction to model-based geostatistics. *Spatial statistics and computational methods* (pp. 43–86). Springer. (Cit. on p. 38).
- Doane, D. P., & Seward, L. E. (2011). Measuring skewness: A forgotten statistic? *Journal of statistics education*, 19(2) (cit. on p. 170).
- Dowle, M., & Srinivasan, A. (2021). *Data.table: Extension of ‘data.frame’* [R package version 1.14.0]. (Cit. on p. 181).
- Eberl, J.-M., Huber, R. A., & Greussing, E. (2020). From populism to the ‘plandemic’: Why populists believe in covid-19 conspiracies (cit. on p. 158).
- Elkeles, T., & Seifert, W. (1996). Immigrants and health: Unemployment and health-risks of labour migrants in the federal republic of germany, 1984–1992. *Social science & medicine*, 43(7), 1035–1047 (cit. on p. 154).
- Fahrmeir, L., Heumann, C., Künstler, R., Pigeot, I., & Tutz, G. (2016). *Statistik: Der weg zur datenanalyse*. Springer-Verlag. (Cit. on pp. 167, 168).
- Fahrmeir, L., & Lang, S. (2001). Bayesian inference for generalized additive mixed models based on markov random field priors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(2), 201–220 (cit. on p. 36).
- Fahrmeir, L., & Tutz, G. (2013). *Multivariate statistical modelling based on generalized linear models*. Springer Science & Business Media. (Cit. on pp. 36, 167).
- Farias, J. E. M., & Pilati, R. (2020). Violating social distancing amid covid-19 pandemic: Psychological factors to improve compliance (cit. on p. 158).
- Fink, D. (1997). A compendium of conjugate priors. See <http://www.people.cornell.edu/~pages/df36/CONJINTRnew%20TEX.pdf>, 46 (cit. on p. 24).
- Fisher, A., Rudin, C., & Dominici, F. (2018). Model class reliance: Variable importance measures for any machine learning model class, from the “rashomon” perspective. *arXiv preprint arXiv:1801.01489*, 68 (cit. on p. 79).
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics*, 1189–1232 (cit. on pp. 74, 80).

- Fushiki, T. (2011). Estimation of prediction error by using k-fold cross-validation. *Statistics and Computing*, 21(2), 137–146 (cit. on p. 77).
- Gamerman, D. (1997). Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing*, 7(1), 57–68 (cit. on p. 34).
- Gausman, J., & Langer, A. (2020). Sex and gender disparities in the covid-19 pandemic. *Journal of women's health*, 29(4), 465–466 (cit. on p. 155).
- Gelfand, A. E., Diggle, P., Guttorp, P., & Fuentes, M. (2010). *Handbook of spatial statistics*. CRC press. (Cit. on p. 67).
- Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and computing*, 24(6), 997–1016 (cit. on p. 55).
- Gianquintieri, L., Brovelli, M. A., Pagliosa, A., Dassi, G., Brambilla, P. M., Bonora, R., Sechi, G. M., & Caiani, E. G. (2020). Mapping spatiotemporal diffusion of covid-19 in lombardy (italy) on the base of emergency medical services activities. *ISPRS International Journal of Geo-Information*, 9(11), 639 (cit. on p. 10).
- Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1), 44–65 (cit. on p. 81).
- Gotway, C. A., & Young, L. J. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association*, 97(458), 632–648 (cit. on p. 69).
- Gross, B., Zheng, Z., Liu, S., Chen, X., Sela, A., Li, J., Li, D., & Havlin, S. (2020). Spatio-temporal propagation of covid-19 pandemics. *EPL (Europhysics Letters)*, 131(5), 58003 (cit. on p. 10).
- Guan, W.-j., Ni, Z.-y., Hu, Y., Liang, W.-h., Ou, C.-q., He, J.-x., Liu, L., Shan, H., Lei, C.-l., Hui, D. S., et al. (2020). Clinical characteristics of coronavirus disease 2019 in china. *New England journal of medicine*, 382(18), 1708–1720 (cit. on p. 9).
- Haldane, J. B. (1941). The fitting of binomial distributions. *Annals of Eugenics*, 11(1), 179–181 (cit. on p. 169).
- Hale, T., Petherick, A., Phillips, T., & Webster, S. (2020). Variation in government responses to covid-19. *Blavatnik school of government working paper*, 31, 2020–11 (cit. on p. 91).
- Hall, M. T., Bui, H. Q., Rowe, J., & Do, T. A. (2020). Covid-19 case and contact investigation in an office workspace. *Military medicine*, 185(11-12), e2162–e2165 (cit. on p. 163).
- Hansen, T. (2020). Public covid-19 data for norway (covid19data.no). (Cit. on p. 85).
- Harrison, J. (2020). *Rselenium: R bindings for 'selenium webdriver'* [R package version 1.7.7]. (Cit. on p. 182).
- Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications (cit. on p. 32).
- Hijmans, R. J. (2019). *Geosphere: Spherical trigonometry* [R package version 1.5-10]. (Cit. on p. 181).

- Hodcroft, E. (2021). Covariants: Sars-cov-2 mutations and variants of interest. (Cit. on p. 92).
- Holland, P. W., & Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the american Statistical association*, 76(373), 33–50 (cit. on p. 166).
- Hu, M., Lin, H., Wang, J., Xu, C., Tatem, A. J., Meng, B., Zhang, X., Liu, Y., Wang, P., Wu, G., et al. (2021). Risk of coronavirus disease 2019 transmission in train passengers: An epidemiological and modeling study. *Clinical Infectious Diseases*, 72(4), 604–610 (cit. on p. 161).
- Indseth, T., Grøsland, M., Arnesen, T., Skyrud, K., Kløvstad, H., Lamprini, V., Telle, K., & Kjøllesdal, M. (2020). Covid-19 among immigrants in norway, notified infections, related hospitalizations and associated mortality: A register-based study. *Scandinavian Journal of Public Health*, 1403494820984026 (cit. on p. 155).
- Irwin, M. (2005). Prior choice, summarizing the posterior. (Cit. on pp. 24, 25).
- Jamshidi, S., Baniasad, M., & Niyogi, D. (2020). Global to usa county scale analysis of weather, urban density, mobility, homestay, and mask use on covid-19. *International journal of environmental research and public health*, 17(21), 7847 (cit. on p. 155).
- Kammann, E., & Wand, M. P. (2003). Geoadditive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52(1), 1–18 (cit. on p. 36).
- Kasilingam, D., Sathiya Prabhakaran, S. P., Rajendran, D. K., Rajagopal, V., Santhosh Kumar, T., & Soundararaj, A. (2020). Exploring the growth of covid-19 cases using exponential modelling across 42 countries and predicting signs of early containment using machine learning. *Transboundary and Emerging Diseases* (cit. on p. 12).
- Kim, A. Y., & Wakefield, J. (2018). *Spatialepi: Methods and data for spatial epidemiology* [R package version 1.2.3]. (Cit. on p. 182).
- Kitagawa, G., & Gersch, W. (1996). *Smoothness priors analysis of time series* (Vol. 116). Springer Science & Business Media. (Cit. on p. 36).
- Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in medicine*, 19(17-18), 2555–2567 (cit. on p. 68).
- Kopel, J., Perisetti, A., Roghani, A., Aziz, M., Gajendran, M., & Goyal, H. (2020). Racial and gender-based differences in covid-19. *Frontiers in public health*, 8, 418 (cit. on p. 155).
- Kourounis, D., Fuchs, A., & Schenk, O. (2018). Towards the next generation of multiperiod optimal power flow solvers. *IEEE Transactions on Power Systems*, PP(99), 1–10 (cit. on p. 181).
- Krause, E. F. (1986). *Taxicab geometry: An adventure in non-euclidean geometry*. Courier Corporation. (Cit. on p. 71).
- Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1), 79–86 (cit. on p. 26).

- Kunst, J. (2020). *Highcharter: A wrapper for the 'highcharts' library* [R package version 0.8.2]. (Cit. on p. 181).
- Lahti, L., Huovari, J., Kainu, M., & Biecek, P. (2017). Eurostat r package [Version 3.7.5]. (Cit. on p. 181).
- Lang, S., & Brezger, A. (2004). Bayesian p-splines. *Journal of computational and graphical statistics*, 13(1), 183–212 (cit. on p. 36).
- Lelieveld, J., Helleis, F., Borrmann, S., Cheng, Y., Drewnick, F., Haug, G., Klimach, T., Sciare, J., Su, H., & Pöschl, U. (2020). Model calculations of aerosol transmission and infection risk of covid-19 in indoor environments. *International Journal of Environmental Research and Public Health*, 17(21), 8114 (cit. on p. 163).
- Leroux, B. G., Lei, X., & Breslow, N. (2000). Estimation of disease rates in small areas: A new mixed model for spatial dependence. *Statistical models in epidemiology, the environment, and clinical trials* (pp. 179–191). Springer. (Cit. on p. 50).
- Lilovski, N. (2020). *Dashboardthemes: Customise the appearance of 'shinydashboard' applications using themes* [R package version 1.1.3]. (Cit. on p. 181).
- Lovelace, R., Nowosad, J., & Muenchow, J. (2019). *Geocomputation with r*. CRC Press. (Cit. on pp. 58, 60–63).
- Maiti, A., Zhang, Q., Sannigrahi, S., Pramanik, S., Chakraborti, S., Cerdá, A., & Pilla, F. (2021). Exploring spatiotemporal effects of the driving factors on covid-19 incidences in the contiguous united states. *Sustainable cities and society*, 68, 102784 (cit. on p. 11).
- Markov, A. A. (1906). Extension of the law of large numbers to dependent quantities. *Izv. Fiz.-Matem. Obsch. Kazan Univ.(2nd Ser)*, 15, 135–156 (cit. on p. 29).
- Marroquin, J. L., Velasco, F. A., Rivera, M., & Nakamura, M. (2001). Gauss-markov measure field models for low-level vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(4), 337–348 (cit. on p. 36).
- Martínez-Beneito, M. A., López-Quilez, A., & Botella-Rocamora, P. (2008). An autoregressive approach to spatio-temporal disease mapping. *Statistics in medicine*, 27(15), 2874–2889 (cit. on p. 67).
- Martins, T. G., Simpson, D. P., Riebler, A., Fuglstad, G., Rue, H., & Sørbye, S. (2014). Penalising model component complexity: A principled, practical approach to constructing priors. *arXiv preprint arXiv:1403.4630* (cit. on pp. 25–27, 51).
- Mehmood, K., Bao, Y., Abrar, M. M., Petropoulos, G. P., Soban, A., Saud, S., Khan, Z. A., Khan, S. M., Fahad, S., et al. (2021). Spatiotemporal variability of covid-19 pandemic in relation to air pollution, climate and socioeconomic factors in pakistan. *Chemosphere*, 271, 129584 (cit. on p. 11).
- Meschiari, S. (2021). *Latex2exp: Use latex expressions in plots* [R package version 0.5.0]. (Cit. on p. 181).

- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6), 1087–1092 (cit. on p. 32).
- Metropolis, N., & Ulam, S. (1949). The monte carlo method. *Journal of the American statistical association*, 44(247), 335–341 (cit. on p. 29).
- Meyer, F., & Perrier, V. (2021). *Shinybusy: Busy indicator for 'shiny' applications* [R package version 0.2.2.9100]. (Cit. on p. 182).
- Mollalo, A., Vahedi, B., & Rivera, K. M. (2020). Gis-based spatial modeling of covid-19 incidence rate in the continental united states. *Science of the total environment*, 728, 138884 (cit. on p. 11).
- Molnar, C. (2020). *Interpretable machine learning*. Lulu. com. (Cit. on pp. 79–81, 83).
- Molnar, C., Bischl, B., & Casalicchio, G. (2018). Iml: An r package for interpretable machine learning. *JOSS*, 3(26), 786 (cit. on p. 181).
- Moraga, P. (2019). *Geospatial health data: Modeling and visualization with r-inla and shiny*. CRC Press. (Cit. on pp. 38, 46, 65–67, 69).
- Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2), 17–23 (cit. on p. 66).
- Mukherjee, S., & Pahan, K. (2021). Is covid-19 gender-sensitive? *Journal of Neuroimmune Pharmacology*, 1–10 (cit. on p. 155).
- Müller, K. (2020). *Here: A simpler way to find your files* [R package version 1.0.1]. (Cit. on p. 181).
- Müller, K., & Wickham, H. (2021). *Tibble: Simple data frames* [R package version 3.1.2]. (Cit. on p. 182).
- Musselwhite, C., Avineri, E., & Susilo, Y. (2020). Editorial jth 16—the coronavirus disease covid-19 and implications for transport and health. *Journal of Transport & Health*, 16, 100853 (cit. on p. 161).
- Nandy, A., Tiwari, C., & Kundu, S. (2021). Managing the covid-19 pandemic: Does social infrastructure matter? evidence from india. *Transforming Government: People, Process and Policy* (cit. on p. 12).
- Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3), 370–384 (cit. on p. 53).
- Openshaw, S. (1984). The modifiable areal unit problem. norwich, uk. (Cit. on p. 69).
- Orea, L., & Álvarez, I. C. (2020). How effective has the spanish lockdown been to battle covid-19? a spatial analysis of the coronavirus propagation across provinces. *Documento de trabajo*, 3, 1–33 (cit. on p. 12).
- Padgham, M., Rudis, B., Lovelace, R., & Salmon, M. (2017). Osmda. *The Journal of Open Source Software*, 2(14) (cit. on pp. 90, 182).

- Pebesma, E. (2021). *Lwgeom: Bindings to selected 'liblwgeom' functions for simple features* [R package version 0.2-6]. (Cit. on p. 181).
- Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10(1), 439–446 (cit. on pp. 93, 182).
- Pebesma, E., Mailund, T., & Hiebert, J. (2016). Measurement units in R. *R Journal*, 8(2), 486–494 (cit. on p. 182).
- Pedersen, T. L. (2020). *Patchwork: The composer of plots* [R package version 1.1.1]. (Cit. on p. 182).
- Pedrosa, R. H. (2020). The dynamics of covid-19: Weather, demographics and infection timeline. *MedRxiv* (cit. on p. 11).
- Perrier, V., Meyer, F., & Granjon, D. (2021). *Shinywidgets: Custom inputs widgets for shiny* [R package version 0.6.0]. (Cit. on p. 182).
- Petrie, A. (2020). *Regclass: Tools for an introductory class in regression and modeling* [R package version 1.6]. (Cit. on p. 182).
- Petrov, A. N., Welford, M., Golosov, N., DeGroote, J., Degai, T., & Savelyev, A. (2020). Spatiotemporal dynamics of the covid-19 pandemic in the arctic: Early data and emerging trends. *International Journal of Circumpolar Health*, 79(1), 1835251 (cit. on p. 10).
- Pettitt, L. (1990). The conditional predictive ordinate for the normal distribution. *Journal of the Royal Statistical Society: Series B (Methodological)*, 52(1), 175–184 (cit. on p. 54).
- R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. (Cit. on p. 181).
- Raiffa, H., & Schlaifer, R. (1961). Applied statistical decision theory, division of research, graduateschool of business administration, harvard university, boston, 1961. *RaiffaApplied Statistical Decision Theory1961* (cit. on p. 24).
- Ren, K. (2016). *Rlist: A toolbox for non-tabular data manipulation* [R package version 0.4.6.1]. (Cit. on p. 182).
- Riebler, A., Sørbye, S. H., Simpson, D., & Rue, H. (2016). An intuitive bayesian spatial model for disease mapping that accounts for scaling. *Statistical methods in medical research*, 25(4), 1145–1165 (cit. on pp. 48–51, 55).
- Ripley, B. D. (2007). *Pattern recognition and neural networks*. Cambridge university press. (Cit. on p. 72).
- Ritchie, H., Ortiz-Ospina, E., Beltekian, D., Mathieu, E., Hasell, J., Macdonald, B., Giattino, C., Appel, C., Rodés-Guirao, L., & Roser, M. (2020). Coronavirus pandemic (covid-19). *Our World in Data* (cit. on p. 91).
- Robert, C., & Casella, G. (2013). *Monte carlo statistical methods*. Springer Science & Business Media. (Cit. on pp. 29–31, 33).

- Robert, C. P., Marin, J.-M., & Rousseau, J. (2010). Bayesian inference. *arXiv preprint arXiv:1002.2080* (cit. on p. 24).
- Robinson, W. S. (2009). Ecological correlations and the behavior of individuals. *International journal of epidemiology*, 38(2), 337–341 (cit. on p. 69).
- Rue, H., & Held, L. (2005). *Gaussian markov random fields: Theory and applications*. CRC press. (Cit. on pp. 16–19, 21, 22, 36, 39, 40, 42–44, 168).
- Rue, H., Martino, S., & Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2), 319–392 (cit. on pp. 34–36, 44).
- Saha, A., Gupta, K., Patil, M., et al. (2020). Monitoring and epidemiological trends of coronavirus disease (covid-19) around the world. *Matrix Science Medica*, 4(4), 121 (cit. on p. 10).
- Sannigrahi, S., Pilla, F., Basu, B., Basu, A. S., & Molter, A. (2020). Examining the association between socio-demographic composition and covid-19 fatalities in the european region using spatial regression approach. *Sustainable cities and society*, 62, 102418 (cit. on p. 12).
- Schmid, C., Schiffels, S., Boog, J., Lange, M., & Aschoff, M. (2021). *Covid19germany: Load, visualise and analyse daily updated data on the covid-19 outbreak in germany* [R package version 0.1.1]. (Cit. on pp. 85, 181).
- Shapley, L. S. (1997). A value for n-person games. *Classics in game theory*, 69 (cit. on p. 83).
- Sia, D., Miszkurka, M., Batal, M., Delisle, H., & Zunzunegui, M. V. (2019). Chronic disease and malnutrition biomarkers among unemployed immigrants and canadian born adults. *Archives of Public Health*, 77(1), 1–10 (cit. on p. 154).
- Sigler, T., Mahmuda, S., Kimpton, A., Loginova, J., Wohland-Jakhar, P., Charles-Edwards, E., & Corcoran, J. (2020). The socio-spatial determinants of covid-19 diffusion: The impact of globalisation, settlement characteristics and population (cit. on p. 155).
- Snow, J. (1857). Cholera, and the water supply in the south districts of london. *British Medical Journal*, 1(42), 864 (cit. on p. 2).
- Solymos, P., & Zawadzki, Z. (2020). *Pbapply: Adding progress bar to 'apply' functions* [R package version 1.4-3]. (Cit. on p. 182).
- Spagnolo, P. A., Manson, J. E., & Joffe, H. (2020). Sex and gender differences in health: What the covid-19 pandemic can teach us. (Cit. on p. 155).
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 485–493 (cit. on p. 53).
- Stone, C. J. (1985). Additive regression and other nonparametric models. *The annals of Statistics*, 689–705 (cit. on p. 35).

- Teutsch, S. M., Churchill, R. E. et al. (2000). *Principles and practice of public health surveillance*. Oxford University Press, USA. (Cit. on p. 57).
- Vaughan, D., & Dancho, M. (2021). *Furrr: Apply mapping functions in parallel using futures* [R package version 0.2.2]. (Cit. on p. 181).
- Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth) [ISBN 0-387-95457-0]. Springer. (Cit. on p. 181).
- Vietsen, U. M. (2020). The “new normal” and “pandemic populism”: The covid-19 crisis and anti-hygienic mobilisation of the far-right. *Social Sciences*, 9(9), 165 (cit. on p. 157).
- Wang, Y., Liu, Y., Struthers, J., & Lian, M. (2021). Spatiotemporal characteristics of the covid-19 epidemic in the united states. *Clinical infectious diseases*, 72(4), 643–651 (cit. on p. 11).
- Watanabe, S., & Opper, M. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research*, 11(12) (cit. on p. 53).
- Whittle, R. S., & Diaz-Artiles, A. (2020). An ecological study of socioeconomic predictors in detection of covid-19 cases across neighborhoods in new york city. *BMC medicine*, 18(1), 1–17 (cit. on p. 155).
- Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. (Cit. on p. 181).
- Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12), 1–20 (cit. on pp. 93, 182).
- Wickham, H. (2019). *Stringr: Simple, consistent wrappers for common string operations* [R package version 1.4.0]. (Cit. on pp. 93, 182).
- Wickham, H., François, R., Henry, L., & Müller, K. (2021). *Dplyr: A grammar of data manipulation* [R package version 1.0.6]. (Cit. on p. 181).
- Wickham, H., & Hester, J. (2020). *Readr: Read rectangular text data* [R package version 1.4.0]. (Cit. on p. 182).
- Wilkins, J. E. (1944). A note on skewness and kurtosis. *The Annals of Mathematical Statistics*, 15(3), 333–335 (cit. on p. 170).
- Wondreys, J., & Mudde, C. (2020). Victims of the pandemic? european far-right parties and covid-19. *Nationalities Papers*, 1–34 (cit. on p. 157).
- Xie, Y., Cheng, J., & Tan, X. (2021). *Dt: A wrapper of the javascript library 'datatables'* [R package version 0.18]. (Cit. on p. 181).
- Xiong, Y., Wang, Y., Chen, F., & Zhu, M. (2020). Spatial statistics and influencing factors of the novel coronavirus pneumonia 2019 epidemic in hubei province, china (cit. on p. 11).
- Yong, L. (2018). Loo and waic as model selection methods for polytomous items. *arXiv preprint arXiv:1806.09996* (cit. on p. 53).

## Webpages

- des Bundes und der Länder, S. Ä. (2020). *Regionaldatenbank deutschland*. Retrieved February 28, 2021, from <https://www.regionalstatistik.de/genesis/online>. (Cit. on p. 87)
- Deutschland, E. (2020). *Kreisgrenzen 2017*. Retrieved February 28, 2021, from [https://opendata-esri-de.opendata.arcgis.com/datasets/affd8ace4c204981b5d32070f9547eb9\\_0](https://opendata-esri-de.opendata.arcgis.com/datasets/affd8ace4c204981b5d32070f9547eb9_0). (Cit. on p. 89)
- Folkehelseinstituttet. (2020). *Statistikkbankene i fhi*. Retrieved April 5, 2021, from <https://statistikk.fhi.no/>. (Cit. on p. 86)
- Geonorge. (2021). *Kartkatalogen*. Retrieved February 28, 2021, from <https://www.geonorge.no/>. (Cit. on p. 89)
- Google, L. (2020, February 6). *Google covid-19 community mobility reports* [Accessed: 2021-05-28]. (Cit. on p. 91).
- Korsvoll, A. I. S. (2020, October 1). *Nye smittetilfelle i hyllestad- no har sju personar korona*. Retrieved March 5, 2021, from <https://www.firda.no/nye-smittetilfelle-i-hyllestad-no-har-sju-personar-korona/s/5-15-1070915>. (Cit. on p. 102)
- NTB. (2021, February 8). *Ordfører: Ulvik-utbruddet spredte seg trolig blant barna*. Retrieved March 5, 2021, from <https://www.h-avis.no/ordforer-ulvik-utbruddet-spredte-seg-trolig-blant-barna/s/5-62-1138922>. (Cit. on p. 102)
- OpenStreetMap contributors. (2017). *Planet dump retrieved from https://planet.osm.org*. (Cit. on p. 90).
- Sentralbyrå, S. (2016). *Statbank norway*. Retrieved February 28, 2021, from <https://www.ssb.no/en/statbank/>. (Cit. on p. 87)
- Smistad, E. (2020). *Konverter-norgeskart-projeksjon*. (Cit. on p. 89).

# List of Figures

1.1	The original map of cholera cases in southern London, created by John Snow in 1854. . . . .	2
2.1	The cantons of Switzerland, an example of an irregular lattice. . . . .	17
2.2	An undirected labelled graph with 3 nodes, $\mathcal{V} = \{1, 2, 3\}$ and $\mathcal{E} = \{\{1, 2\} \{2, 3\}\}$ . . . . .	21
2.3	A typical semivariogram . . . . .	38
2.4	The pairwise Markov property; the black nodes are conditionally independent given the light grey nodes. . . . .	40
2.5	The local Markov property; the black nodes and white nodes are conditionally independent given the dark grey nodes. . . . .	40
2.6	The global Markov property; the dark grey and light grey nodes are globally independent given the black nodes. . . . .	41
3.1	A geographic CRS with an origin at $0^\circ$ longitude and latitude. The red X denotes the location of Trondheim. . . . .	59
3.2	The most commonly used simple feature types. . . . .	60
3.3	An example of continuous and categorical raster data . . . . .	61
3.4	The number of shared borders of cantons in Switzerland . . . . .	65
4.1	A single-layer neural network. . . . .	72
4.2	A simple example of a decision tree . . . . .	73
4.3	An example of 10-fold cross validation . . . . .	77
6.1	The SIR for Germany based on the data of the 2nd of May 2021 . . . . .	101
6.2	The SIR for Norway based on the data of the 2nd of May 2021 . . . . .	102
6.3	The $\log_{10}$ SIR for Norway based on the data of the 2nd of May 2021 . . . . .	103
6.4	The Cullen and Frey graph for Germany . . . . .	105
6.5	The Cullen and Frey graph for Norway . . . . .	105
6.6	A negative binomial fit to the number of cases in German municipalities	106
6.7	A negative binomial fit to the number of cases in Norwegian municipalities	107
6.8	Histogram for the number of cases in German municipalities with a normal and a negative binomial distribution overlayed. . . . .	108

6.9	Histogram for the number of cases in Norwegian municipalities with a normal and a negative binomial distribution overlayed. . . . .	109
6.10	The posterior mean and credibility intervals of the coefficients . . . . .	114
6.11	The posterior mean and credibility intervals of the coefficients . . . . .	116
6.12	Values of the DIC and the WAIC when changing the value for $\sigma_0$ . The black line highlights the values for $\sigma_0 = 1$ . . . . .	118
6.13	Value of the MAE when changing the value for $\sigma_0$ . The black line highlights the values for $\sigma_0 = 1$ . . . . .	119
6.14	Comparison of the credibility intervals of a BYM2 model for different values of $\sigma_0$ . . . . .	120
6.15	Comparison of the credibility intervals of a BYM2 model for different values of $\sigma_0$ . . . . .	121
6.16	Spatial field for a Besag model and a Leroux model. . . . .	122
6.17	Spatial fields for a BYM2 model. . . . .	122
6.18	Spatial fields for the structured component of a BYM2 model when changing the value for $\sigma_0$ . . . . .	123
6.19	Spatial fields for the structured component of a BYM2 model when changing the value for $\sigma_0$ . . . . .	123
6.20	The variable importance plots for the random forest. . . . .	126
6.21	The partial dependence plots for the logarithmic trade tax and the number of clinics. . . . .	127
6.22	The partial dependence plots for the share of the vote the AfD and the Greens get. . . . .	127
6.23	The individual conditional expectation for the logarithmic trade tax and the number of clinics. . . . .	128
6.24	The individual conditional expectation for the share of the vote the AfD and the Greens get. . . . .	128
6.25	Shapley values for the cities of Munich and Hannover. . . . .	129
6.26	The variable importance plots for the random forest. . . . .	131
6.27	The partial dependence plots for the number of places of worship and the number of offices. . . . .	132
6.28	The individual conditional expectation for the logarithmic trade tax and the number of clinics. . . . .	132
6.29	Shapley values for the municipalities of Tromsø and Nordre Follo. . . . .	133
6.30	The Cullen and Frey graph for Germany . . . . .	134
6.31	The Cullen and Frey graph for Norway . . . . .	135
6.32	A negative binomial fit to the number of cases in German municipalities	136
6.33	A normal fit to the number of cases in German municipalities . . . . .	136

6.34	Histogram for the number of cases in German municipalities with a normal and a negative binomial distribution overlayed. . . . .	137
6.35	The predicted number of infections in Germany according to the temporal model. The vertical line indicates where the test data begins. . . . .	139
6.36	The predicted number of infections in Germany according to the temporal model. . . . .	140
6.37	The 7-day incidence of the actual number of infections and the predicted number of infections. The vertical line indicates where the test data begins. . . . .	141
6.38	The posterior temporal trend for the number of infections. . . . .	141
6.39	The predicted number of infections in Norway according to the temporal model. The vertical line indicates where the test data begins. . . . .	144
6.40	The predicted number of infections in Norway according to the temporal model. . . . .	144
6.41	The 7-day incidence of the actual number of infections and the predicted number of infections. The vertical line indicates where the test data begins. . . . .	145
6.42	The posterior temporal trend for the number of infections. . . . .	145
7.1	A hexagon map of all bakeries in Germany. . . . .	148
7.2	A heat map of all bakeries in Germany. . . . .	149
7.3	A choropleth map of all bakeries in Germany. . . . .	149
7.4	The seven-day incidence in Munich compared to Germany. . . . .	150
7.5	Predicted numbers in Sweden using an ar1 model with a test size of 28. 151	
8.1	Relative risk of contracting Covid-19 in Norway. . . . .	156
8.2	Posterior mean of the municipality-specific relative risks $\zeta = \exp(\xi)$ compared with the whole of Norway (left) and posterior probability $\mathbb{P}(\zeta_i > 1   \mathbf{y})$ . . . . .	156
8.3	Relative risk of contracting Covid-19 in Germany. . . . .	159
8.4	Posterior mean of the municipality-specific relative risks $\zeta = \exp(\xi)$ compared with the whole of Germany (left) and posterior probability $\mathbb{P}(\zeta_i > 1   \mathbf{y})$ . . . . .	160
10.1	A normal fit to the number of cases in German municipalities . . . . .	171
10.2	A Poisson fit to the number of cases in German municipalities . . . . .	171
10.3	A Poisson fit to the number of cases in German municipalities . . . . .	172
10.4	A normal fit to the number of cases in Norwegian municipalities . . . . .	172
10.5	A Poisson fit to the number of cases in Norwegian municipalities . . . . .	173
10.6	A negative binomial fit to the number of cases in Norwegian municipalities	174

10.7	A normal fit to the number of cases in Norwegian municipalities . . . . .	174
10.8	A Poisson fit to the number of cases in Norwegian municipalities . . . . .	175
10.9	Histogram for the number of cases in Norwegian municipalities with a normal and a negative binomial distribution overlayed. . . . .	175
10.10	Values of the DIC and the WAIC when changing the value for $\sigma_0$ . The black line highlights the values for $\sigma_0 = 1$ . . . . .	176
10.11	Values of the MAE when changing the value for $\sigma_0$ . The black line highlights the values for $\sigma_0 = 1$ . . . . .	176
10.12	Spatial field for a Besag model and a Leroux model. . . . .	177
10.13	Spatial fields for a BYM2 model. . . . .	177

# List of Tables

5.1	An excerpt from the Covid-19 data for Norway. Does not contain all variables. The number of infections are the cumulative number of infections. . . . .	85
5.2	An excerpt from the Covid-19 data for Germany. Does not contain all variables. . . . .	85
5.3	An excerpt from the long version of the Norwegian Covid-19 data. Does not contain all variables. . . . .	93
5.4	The variables contained in the final dataset. . . . .	95
5.5	The variables contained in the final dataset. . . . .	97
5.6	The variables contained in the final dataset. . . . .	99
6.1	The AIC for different distributions for Germany and Norway . . . . .	107
6.2	The German municipalities with the most infections as of May 2nd 2021. . . . .	110
6.3	The Norwegian municipalities with the most infections as of May 2nd 2021. . . . .	110
6.4	The performance measures for the model without a spatial component. . . . .	111
6.5	The fixed effects for the model. Values are rounded. A * denotes a significant effect. . . . .	111
6.6	The performance measures for the model without a spatial component. . . . .	112
6.7	The fixed effects for the model. Values are rounded. A * denotes a significant effect. . . . .	112
6.8	Results of the Moran test for Germany and Norway. . . . .	113
6.9	The performance measures for the best performing model of each type. . . . .	114
6.10	The fixed effects for the model. Values are rounded. A * denotes a significant effect. . . . .	115
6.11	The performance measures for the best performing model of each type. . . . .	116
6.12	The fixed effects for the model. Values are rounded. A * denotes a significant effect. . . . .	117
6.13	The MAE for the BYM2 model and the predictive models. . . . .	125
6.14	The MAE for the BYM2 model and the predictive models. . . . .	130
6.15	The AIC for different distributions for Germany and Norway . . . . .	137

6.16	The performance measures for different types of temporal models for Germany . . . . .	139
6.17	The fixed effects for the model. Values are rounded. A * denotes a significant effect. . . . .	142
6.18	The performance measures for different types of temporal models for Norway . . . . .	143
6.19	The fixed effects for the model. Values are rounded. A * denotes a significant effect. . . . .	146
8.1	The performance measures for different types of temporal models for Germany . . . . .	162
10.1	A list of all the key-value pairs used to query OpenStreetMap, except the ones used for residential buildings . . . . .	178
10.2	A list of all the key-value pairs used to query OpenStreetMap for residential buildings . . . . .	179
10.3	A list of all the key-value pairs that were combined to create variables	180

# Declaration

I herewith formally declare that I have written the submitted thesis independently. I did not use any outside support except for the quoted literature and other sources mentioned in the thesis. I clearly marked and separately listed all the literature and all the other sources which I employed when producing this academic work, either literally or in content.

*Tromsø, June 21, 2021*

---

Nico Hahn

