# Joint epidemic and spatio-temporal approach for modelling disease outbreaks

Lisa Koeppel, MSc.

CHICAS, Lancaster Medical School
Lancaster University
l.koeppel@lancaster.ac.uk

Submitted for the degree of Doctor of Philosophy
at Lancaster University,

October, 2022

# Abstract

When forecasting epidemics, the main interests lie in understanding the determinants of transmission and predicting who is likely to become infected next. However, for vector-borne diseases, data availability and alteration can constitute an obstacle to doing so: climate change and globalized trade contribute to the expansion of vector habitats to different territories and hence the distribution of many diseases. As a consequence, in the face of a rapidly changing environmental and ecological climatic conditions, previously well-fitted models might become obsolete soon. The demand for precise forecast and prediction of the spread of a disease requires a model that is flexible with respect to the availability of vector data, unobserved random effects and only partially observed data for diseases incidence. Thus, we introduce a combination of a mechanistic SIR model with principled data-based methods from geostatistics. We allow flexibility by replacing a parameter of a continuous-time mechanistic model with a random effect, that is assumed to stem from a spatial Gaussian process. By employing Bayesian inference techniques, we identify points in space where transmission (as opposed to simply incidence) is unusually high or low compared to a national average. We explore how well the spatial random effect can be recovered within a mechanistic model and only partially observed outbreak data available. To this end, we extended the Python probabilistic programming library PyMC3 with our own sampler to effectively impute missing infection and removal time data.

# Contents

# List of Tables

# List of Figures

# List of Abbreviations

| | |
|---|---|
| $n$ | Population size |
| $\mathcal{S}(t)$ | Set of susceptible individuals |
| $\mathcal{I}(t)$ | Set of infected individuals |
| $\mathcal{R}(t)$ | Set of removed/recovered/detected individuals |
| $\mathbf{I}$ | Infection times |
| $\mathrm{I}_i$ | Infection time of individual $i$ |
| $\mathrm{I}_i^-$ | Time shortly before individual $i$ became infected |
| $\min\{\mathbf{I}\}$ | Time point of first infection |
| $i_0$ | First infected individual |
| $n_{\mathrm{I}}$ | Number of total infections in the outbreak |
| $\mathbf{R}$ | Removal times (correspond to detection times in this thesis) |
| $\mathbf{R_M}$ | Detection times that are assumed to be missing |
| $\mathbf{R_T}$ | Detection times that are assumed to reflect reality |
| $Q$ | Infectious period |
| $t_{\mathrm{end}}$ | End of observation period |
| $\lambda_j(t)$ | Infectious pressure on susceptible $j$ at time $t$ |
| $\alpha$ | Basic infectious pressure parameter |
| $\kappa$ | Spatial kernel parameter |

| | |
|---|---|
| $\gamma$ | Removal rate parameter |
| $\psi$ | Transmission rate parameter |
| $\boldsymbol{\theta}$ | Vector of parameters |
| $\boldsymbol{\theta}_{\setminus\{\psi\}}$ | Set of parameters $\boldsymbol{\theta}$ without $\psi$ |
| $d_{ij}$ | Euclidean distance between locations of individual $i$ and $j$ |
| $\|\|x - x'\|\|$ | Euclidean distance between $x$ and $x'$ |
| $\ell_j$ | Spatial location of individual $j$ |
| $v(\ell_j)$ | Susceptibility of individual $j$ |
| $\Sigma$ | Covariance matrix |
| $\Sigma^{-1}$ | Inverse of matrix |
| $\Sigma_{ij}$ | (i,j)th entry of $\Sigma$ |
| $\Sigma_{AB}$ | Sliced covariance matrix into four block matrices with respect to $A$ and $B$ |
| $\mathbf{c}(\cdot)$ | Covariance function |
| $\mathbf{m}(\cdot)$ | Mean function |
| $\mathcal{N}(0,1)$ | Standard normal distribution |
| $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ | Multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\Sigma$ |
| $\mathbb{E}$ | Expectation |
| $\mathbb{V}$ | Variance |
| Cov | Covariance |
| $p_X$ | Probability density function of random variable $X$ |
| $y^T$ | Transpose of $y$ |
| $\mathbb{R}^d$ | Real numbers |
| $\mathbb{R}^d$ | $d$-dimensional real space |

# Acknowledgements

First of all I would like to thank my supervisors Chris Jewell and Peter Neal for the financial support, the many discussions and their incredible knowledge and programming skills.

I would like to thank Andy Mitchell at APHA for the provision of the 2007 BTV-8 outbreak records on which we build our methodology. I furthermore acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and I thank the climate modeling group *The Institute Pierre Simon Laplace (IPSL)* for producing and making available their model output. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. World Climate Research Program (WCRP) data was downloaded from the Copernicus Climate Change Service (C3S) Climate Data Store (CDS) (https://cds.climate.copernicus.eu/ cdsapp#!/dataset/projections-cmip5-monthly-single-levels?tab=overview).

Furthermore, I am very grateful for the financial and ideological support by the Foundation of the German Economy (sdw) which had been supporting me since my second semester.

Last but not least, I am eternally grateful for my friends, for your constant support, motivation and crazy laughter; I would not have conquered this journey without you! In specific Poppy Miller and Cyrus Childs, Farhad Mohami, Daniela Schlüter, Irene Kyomuhangi, Rachel Tribbick, Annabelle Edwards, the Lancaster

Tennis club, Şevki Aygün, Laura-Inés Böhler, Markus Sauter, Melina Blessing, Mary Gaeddert, Matthias Schlösser and in particular thank you to Katha Schmitzen and Tobias Siems - I am eternally grateful for your patience and support throughout.

This thesis is dedicated to my parents and my grandma for their constant belief and unconditional love.

# Declaration

This thesis has not been submitted in support of an application for another degree at this or any other university. It is the result of my own work and includes nothing that is the outcome of work done in collaboration except where specifically indicated. Many of the ideas in this thesis were the product of discussions with my supervisors, Dr. Chris Jewell and Professor Peter Neal.

# Chapter 1

# Introduction and Background

## 1.1 Epidemics and overview

Each disease is unique with respect to its pathogen, its symptoms and especially its transmission pathway. According to (for Disease Control and Prevention, 2012), a *disease outbreak* is defined as an oft sudden increase of incident cases above the normally expected levels. While the term outbreak is more used for a limited geographical area, an *epidemic* affects are larger region culminating in a *pandemic* if the epidemic has spread to several countries or continents.

In epidemiology the interests lie in studying the patterns and the determinants of diseases and their transmissions. Of particular importance is the class of *infectious diseases* as they can lead to fast spreading, major outbreaks. Recent examples include the Foot-and-Mouth (FMD) epidemic in Great Britain 2001, the worldwide prevalent Severe Acute Respiratory Syndrome (SARS) in 2003, the Ebola outbreak in West Africa 2014, the emerging Zika virus disease in South America in 2015, or the Coronavirus disease 2019 (COVID-19) pandemic, leading to high morbidity and mortality rates as well as economic and social effects.

Infectious diseases are caused by pathogenic microorganisms, namely bacteria, viruses, parasites or fungi. We distinguish between *directly* transmitted diseases, e.g. influenza, and *indirectly* transmitted diseases, e.g. Malaria or Bluetongue, as displayed in Figure 1.1. In the case of directly transmitted diseases, pathogens are

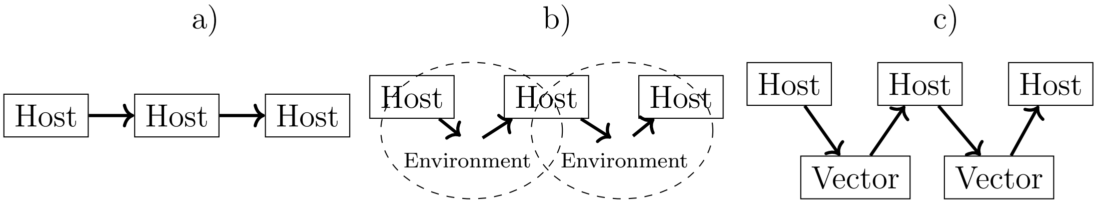


Figure 1.1: Different infectious disease transmission pathways. They are clustered in a) direct transmission and indirect transmission which is split into further classification b) through the environment and c) through a vector.

not able to survive long outside of an host's organism resulting in the diseases being passed between individuals through close contact in the presence of an infectious case. In contrast, the pathogens of *indirectly* transmitted diseases might even have part of their life cycle outside of the host's body and thus is not passed on from one infected individual to the next. Indirect transmission can either occur through shared environment, which is contaminated by an infected individual and in turn able to infect susceptible individuals, or by means of a vector. In the latter case, a vector takes in the pathogen through an interaction with an infected host (e.g. a blood meal). Inside the vector, the pathogen multiplies over time and can be injected into a previously disease free host during a subsequent blood meal. Therefore, susceptible hosts can only acquire the disease in the presence of the transmitting vector. It is to mention that examples of diseases with a mixture of direct and indirect transmission exist: the Zika virus is transmitted by the *Aedes aegypti* mosquitoes, but also directly through sexual contact and blood transfusions (Brauer et al., 2016). Further, bovine Tuberculosis (bTB) is directly transmitted amongst cattle and indirectly via environment contamination. Moreover, badgers play an important vector role in bTB transmission, as they can acquire and pass on the disease through the environment, but also through close contact with cattle (Woodroffe et al., 2016).

In this thesis we focus on vector-borne diseases. According to the WHO, vector-borne diseases account for more than 17% of all infectious diseases leaving over one billion people infected and one million people dead every year (World Health Organization, 2014). By now, more than half of the world's population is at risk, endangered by a great range of illnesses that can lead to severe sickness, hospitalization and permanent damage such as mutilations and distortions (World Health

Organization, 2014). Examples of vectors and their transmitting diseases include mosquitoes (Dengue Fever, Malaria, Bluetongue), flies (Leishmaniasis, Sleeping sickness), bugs (Chagas disease), ticks (Lyme disease, Bovine Theileriosis), fleas (Plague) and freshwater snails (Schistosomiasis).

A large proportion of the global burden from vector-borne diseases concentrates in tropical and subtropical areas, where mosquitoes experience perfect breeding conditions and where access to sanitary facilities and clean drinking water is not always secured. However, climate change and globalized trade contribute to the extension of vector habitats to different territories and hence the distribution of many diseases (see e.g. Fischer et al. (2011); Medlock and Leach (2015); Campbell-Lendrum et al. (2015)). For instance, Dengue Fever has experienced a 30-fold increase over the past 50 years with a presence in more than 65% of all countries worldwide (Murray et al., 2013; World Health Organization, 2014). Further, socio-economic, environmental and ecological factors are linked to the emergence of new diseases (Jones et al., 2008), for example Schmallenberg (Beer et al., 2013) or the Bluetongue disease (Purse et al., 2005) causing a large epidemic in Europe in ruminants.

In this thesis, we shall be concerned with the mathematical modelling of vector-borne diseases. In situations with little knowledge about the vector population and partially observed disease data, we aim to understand determinants of transmission and to predict underlying vector risk surfaces. In specific, we develop an approach that combines a mechanistic spatio-temporal epidemiological model with a spatial empirical model. The flexible Bayesian framework and Markov Chain Monte Carlo algorithms allow data augmentation and posterior inference on model parameters. Our research displayed in this thesis can be applied to human as well as veterinary diseases. Although we focus in particular on vector-borne diseases, our approach leaves space to be adopted to other diseases as well.

The remainder of this introductory chapter is organized as follows: In Section 1.2 we give an outline of the Bayesian paradigm before we give an overview of MCMC methods used in our context in Section 1.3. Thereafter we introduce a stochastic processes for spatial variation, namely the Gaussian process (Section 1.4). In Section 1.5 we state an overview of epidemic modelling and embed our research into the current literature. Finally, in Section 1.6 we introduce the outbreak we will test our methodology on and motivate the remainder of this thesis.

## 1.2 Bayesian inference

Assume we have observed data which we perceive as the outcome of a random process. We reconstruct this random process by means of a family of statistical models which is parametrized through $\boldsymbol{\theta} \in \Theta$, whereby $\Theta \subseteq \mathbb{R}^d$. On these grounds, we can draw inference based on the given data. For example, we can estimate the parameters or make predictions over future observations or investigate arbitrary latent processes.

Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be our dataset and $f(\boldsymbol{y}; \boldsymbol{\theta})$ the likelihood of the data with respect to the parameters $\boldsymbol{\theta}$. In the Bayesian setting these parameters are modelled as a random variable $\boldsymbol{\theta}$. Therewith we can express the joint distribution of the parameters and the data as

$$f(\boldsymbol{y}, \boldsymbol{\theta}) = f(\boldsymbol{y} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta}) = f(\boldsymbol{\theta} \mid \boldsymbol{y}) f(\boldsymbol{y}) \tag{1.1}$$

In a Bayesian setting we also call the conditional distribution $f(\boldsymbol{y} \mid \boldsymbol{\theta})$ the likelihood function. Here we use the notation $f(\boldsymbol{y} \mid \boldsymbol{\theta})$ to emphasize that we condition on the *random* parameters $\boldsymbol{\theta}$, as opposed to $f(\boldsymbol{y}; \boldsymbol{\theta})$ where we take the parameters $\boldsymbol{\theta}$ as constant. The marginal distribution $f(\boldsymbol{y})$ is then called *marginal likelihood.* Further, let $f(\boldsymbol{\theta})$ denote the *prior distribution*, which incorporates the a priori belief about the parameters in the absence of the data. Similarly, the conditional distribution $f(\boldsymbol{\theta} \mid \boldsymbol{y})$ of the parameters given the data is called *posterior distribution.*

In Bayesian statistics we are primarily interested in the posterior distribution because we get a reassessment of the parameters given the data. Rearranging Equation (1.1) leads to Bayes' theorem (Bayes, 1763)

$$f(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{f(\boldsymbol{y} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta})}{f(\boldsymbol{y})} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}} \tag{1.2}$$

From Equation (1.2) it becomes clear that inference about the parameters is greatly influenced by two quantities: the likelihood and the prior distribution. As samples size becomes large, the data, and therefore the likelihood as a function of the data, will have an increasing effect on the posterior, similarly, if the prior is very concentrated around one point, the a priori knowledge will contribute more towards the posterior. The selection of the prior distribution can be informed by knowledge about the subject-matter from expert knowledge. In this case, we call the prior *informative.* On the other hand, if we lack information or possess only little knowledge about the parameters, we usually choose a *noninformative* prior, for example one with a flat shape. When the prior $f(\boldsymbol{\theta})$ and the posterior $f(\boldsymbol{\theta} \mid \boldsymbol{y})$ belong to the same family of probability distributions, we call the prior a *conjugate prior.* Such a choice of prior is largely used to facilitate mathematical computations as we will see in later sections.

In order to obtain the posterior distribution analytically, the marginal likelihood $f(\boldsymbol{y})$, or marginal distribution of the data, needs to be calculated. However, in most practical matters this computation is intractable, because it involves calculating an integral , i.e.

$$f(\boldsymbol{y}) = \int f(\boldsymbol{y} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

For many years this constituted an obstacle for the use of Bayesian inference techniques for practical applications. However, by means of Markov Chain Monte Carlo methods we can draw samples which are approximately distributed according to the desired distribution. Especially the method by Metropolis-Hastings (Hastings, 1970) circumvents the impediment of having to calculate the normalizing factor. It relies on the fact that one only needs to be able to calculate the distribution of interest up to a constant of proportionality. Thus, with the marginal likelihood as normalizing constant,

$$f(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto f(\boldsymbol{y} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta})$$

enables us to draw samples from the posterior distribution of the parameters given the data. This was a driving force for the use of Bayesian inference techniques in applied and fundamental research and it has become increasingly popular in particular with the advancement in technology and the increase in computational efficiency. We defer to introduce how to generate posterior samples by means of MCMC to Section 1.3.

## 1.3 Markov chain Monte Carlo

In Section 1.2 we introduce the Bayesian paradigm whose foundation for inference constitutes the posterior distribution. However, to be able to draw samples directly from the posterior, the normalization constant needs to be calculated which is often not analytically tractable. By means of Markov chain Monte Carlo (MCMC) we circumvent this intractability.

The idea of MCMC sampling is that we create a Markov chain whose invariant distribution is the desired analytically intractable distribution. The most well known MCMC method is the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970), which makes constructing such a Markov chain fairly easy. It suffices to only be able to calculate the posterior distribution up to a proportionality constant. Recently the Hamiltonian Monte Carlo method (Neal et al., 2011) has also received wider attention in practical applications. In the following chapters we describe these algorithms and introduce PyMC3 (Salvatier et al., 2016), a statistical computing library for the programming language Python, by means of which the research of this report has been conducted.

### 1.3.1 Metropolis-Hastings

Metropolis et al. (1953) introduced the concept of MCMC sampling first in the context of physics before it was later generalized by Hastings (1970). The idea of the Metropolis-Hastings algorithm is to explore the state space by means of a user-defined proposal distribution. This proposal distribution is defined to be sufficiently simple and straightforward to sample from directly. In each iteration of the algorithm, we decide whether the proposed value is being accepted or rejected with respect to a certain acceptance probability. We keep record of these samples which, if the chain runs long enough, are assumed to be samples from our target distribution.

Algorithm 1 represents the Metropolis-Hastings algorithm. To be more specific, let $f(\theta \mid \boldsymbol{y})$ be the posterior distribution we want to obtain samples from. Suppose $\theta^{(i)}$ is the value of $\theta$ in the $i$-th iteration. We sample a new value $\theta^{*}$ from a proposal

**Algorithm 1** Metropolis-Hastings

```
Input: $n$ number of samples, $q(\cdot)$ proposal, $f(\theta, \boldsymbol{y})$ joint distribution
Initialize $\theta^{(0)}$
for $i = 1, \ldots, n$ do
    Propose a sample $\theta^*$ from $q(\cdot \mid \theta^{(i)})$
    $\alpha(\theta^*, \theta^{(i)}) \leftarrow \min\left\{1, \frac{f(\boldsymbol{y}, \theta^*) q(\theta^{(i)} | \theta^*)}{f(\boldsymbol{y}, \theta^{(i)}) q(\theta^* | \theta^{(i)})}\right\}$    ▷ calculate acceptance probability
    Obtain a sample $u$ from $U[0, 1]$
    if $u < \alpha(\theta^*, \theta^{(i)})$ then
        Update $\theta^{(i+1)} \leftarrow \theta^*$    ▷ acceptance
    else
        Update $\theta^{(i+1)} \leftarrow \theta^{(i)}$    ▷ rejection
    end if
end for
```

distribution $q(\cdot \mid \theta^{(i)})$. Whether we accept $\theta^*$ or not is governed by the acceptance probability

$$\min\left\{1, \frac{f(\theta^* \mid \boldsymbol{y}) q(\theta^{(i)} \mid \theta^*)}{f(\theta^{(i)} \mid \boldsymbol{y}) q(\theta^* \mid \theta^{(i)})}\right\} \tag{1.3}$$

According to Equation (1.2), we may also write this in terms of the joint density $f(\boldsymbol{y}, \theta)$. The fact that we are not obliged to computing the normalization constant, greatly facilitates posterior inference via the Metropolis-Hastings algorithm. For our new sample $\theta^{(i+1)}$ we store the current sampled value $\theta^{(i)}$ in case of rejection and the new sampled value $\theta^*$ in case of acceptance and repeat the procedure.

The proposal distribution is chosen so as to be easy to sample from. In the case of a symmetric proposal distribution the acceptance probability of Equation (1.3) can be simplified, because the ratio $\frac{q(\theta^{(i)} | \theta^*)}{q(\theta^* | \theta^{(i)})}$ equals to unity. This form is called Metropolis algorithm and dates back to Metropolis et al. (1953) (before it was extended to include non-symmetric proposal distributions in the form of the Metropolis-Hastings algorithm (Hastings, 1970)). A popular selection for a symmetric proposal distribution constitutes a Gaussian distribution which is centred at the current sample value (*Random Walk Metropolis algorithm*). The choice of the normal distribution is advantageous as it is easy to implement and can be optimized straightforwardly. In this thesis we use a truncated Gaussian proposal distribution and adapt its variance as we iterate through the algorithm (see Section 2.4).

If we have a higher dimensional, say $d$-dimensional, posterior distribution, Algorithm 1 performs a so-called single block update. That is, the new proposed sample $\theta^*$ is a vector of dimension $d$ and in the acceptance-rejection step the proposed values for all $d$ dimensions get accepted or rejected at once. A different way to approach this can be by means of the so-called *Metropolis within Gibbs* update. One way to do this is to sequentially update only a certain number of dimensions at a time conditioned on the current state of the remaining dimensions. To be more specific, assume $\boldsymbol{\theta}_K^{(i+1)} = (\theta_1^{(i+1)}, \theta_2^{(i+1)}, \ldots, \theta_k^{(i+1)})$ is a $k < d$ dimensional vector of the new updated values and $\boldsymbol{\theta}_{\sim K}^{(i)} = (\theta_{k+1}^{(i)}, \theta_{k+2}^{(i)}, \ldots, \theta_d^{(i)})$ is the vector of the remaining – yet to be updated – dimensions of size $d - k$. Then, for a new update step, we sample $\theta^*$ from the one dimensional proposal distribution $q(\cdot \mid \boldsymbol{\theta}_K^{(i+1)}, \boldsymbol{\theta}_{\sim K}^{(i)})$. Similarly we calculate the acceptance probability of Equation (1.3) conditioned on $\boldsymbol{\theta}_K^{(i+1)}, \boldsymbol{\theta}_{\sim K}^{(i)}$.

The outcome of the Metropolis-Hastings algorithm are realizations of a Markov chain, which exhibits $f(\theta \mid \boldsymbol{y})$ as its invariant distribution. This and further mild conditions, namely aperiodicity and irreducibility, guarantee the convergence of the Markov chain towards $f(\theta \mid \boldsymbol{y})$ (Tierney, 1994). This holds regardless of the initial state of the Markov chain. Thus, by running the chain long enough, we eventually obtain an approximate sample from $f(\theta \mid \boldsymbol{y})$.

As the initial state of the Markov chain may not be represented very well by the posterior, we need a certain *burn-in phase* in order to reach a good first sample to start with. This means that we discard a certain user-defined amount of first links of the Markov chain.

We often need to perform lots of iterations to obtain a representative sample from the posterior, because there is dependence between samples. A mechanism called *subsampling* or *thinning* is a way to get a sample of a manageable size with reduced dependence between each sample. In this procedure, we obtain samples by taking every $t$-th link of the Markov chain, whereby $t$ is chosen by the user.

How to choose the updating step and the variance parameter in a random walk Metropolis algorithm may be adjusted according to the acceptance ratio, that is the proportion of samples accepted. Roberts et al. (1997) showed that for a class of multi-parameter densities it is optimal to tune the acceptance rate to 23.4% and subsequent work has shown this result to robust and a very good general guide.

### 1.3.2 Hamiltonian Monte Carlo and NUTS

As introduced in the last section, the Metropolis-Hastings algorithm is a simple way of creating a Markov chain where the invariant distribution corresponds to the target distribution. However, a major problem of the random walk is two-fold. First, if the variance is chosen too small, we will progress only very slowly and might be unable to explore the state space exhaustively. Second, if it is chosen too large, we might reject too many proposed states, which also slows down the exploration of the state space. Proper tuning of the algorithm becomes even more difficult in high dimensions.

A work around here is a different MCMC method based on a physical analogy, the Hamiltonian Monte Carlo algorithm. It's idea is to imagine a Metropolis-Hastings sampler with a proposal that first draws an auxiliary variable $\rho$ (with the same dimension as $\theta$) according to a density $g$ and then feeds a deterministic function $T$ with $\theta$ and $\rho$ in order to obtain the newly proposed state $\theta^*$. This can be described by a density transformation which may be performed by a diffeomorphism $T : (\theta, \rho) \mapsto (\theta^*, \rho^*)$. Let $J$ be the absolute value of the Jacobian determinant of $T$, the acceptance probability can then be set up as

$$\min\left\{1, \frac{f(\theta^*, \boldsymbol{y})g(\rho^*)}{f(\theta, \boldsymbol{y})g(\rho)} J(\theta, \rho)\right\} \tag{1.4}$$

where $(\theta^*, \rho^*) = T(\theta, \rho)$ (Green, 1995).

Hamiltonian Monte Carlo (HMC) is an MCMC scheme that enjoys an ever increasing popularity (Nishio and Arakawa, 2019) and is one of the main samplers used in this thesis. It facilitates the proposition of distant states from the target distribution while still maintaining high acceptance rates thus can circumvent the slow exploration of the state space, which may result from random-walk proposals.

The following discourse on HMC is based on Neal et al. (2011) and Hoffman and Gelman (2014).

### The physical analogy

We want to sample values of $\theta$ in exact proportion to the height of its probability distribution $f(\theta \mid \boldsymbol{y})$ given some data $\boldsymbol{y}$. For the Hamiltonian Monte Carlo algorithm, we mirror the distribution and take the log to obtain $-\log(f(\theta \mid \boldsymbol{y}))$. We imagine a particle that slides frictionless over the surface with changing heights. When we flick the particle in a random direction, it will flow and eventually turn around. In this system, the particle experiences two types of energies: the potential energy $P(\theta)$ which relates to its position and is our variable of interest and the kinetic energy $K(\rho)$, which depends on the momentum $\rho$ of the particle.

The potential energy $P(\theta)$ and the kinetic energy $K(\rho)$ define a system of energies, the *Hamiltonian.* In our context, it corresponds to the negative joint loglikelihood, i.e.

$$H(\theta, \rho) = \quad \underbrace{-\log(f(\boldsymbol{y} \mid \theta))}_{=P(\theta)} \quad \underbrace{-\log(g(\rho))}_{=K(\rho)}$$

where we agree that $\log(0) = -\infty$.

We can learn about the shape of the surface if we keep track of the particle's position along its path, before we give it another push in a random direction. The *Hamiltonian equations* describe how $\theta$ and $\rho$ change over time when the particle moves and slides over the surface

$$\begin{aligned} \frac{\partial \theta_i}{\partial t} &= \frac{\partial H}{\partial \rho_i} \\ \frac{\partial \rho_i}{\partial t} &= -\frac{\partial H}{\partial \theta_i} \end{aligned} \tag{1.5}$$

whereby $\theta = (\theta_1, ..., \theta_d)$ and $\rho = (\rho_1, ..., \rho_d)$.

### Obtaining new samples

Generating transitions for $\theta$ follows a two step process before we consider an acceptance step: First, we sample a new momentum variable $\rho$ according to $g$. Given $(\theta, \rho)$, as a second step we follow the dynamics according to the Hamiltonian equations (Equation (1.5)) for a fixed time $s$ and obtain a new state $(\theta^*, \rho^*)$.

When 'transitioning backwards' that is starting at $(\theta^*, \rho^*)$, it is very unlikely to end up at $(\theta, \rho)$ again. A way to get around this is to negate the momentum variable after having solved the Hamiltonian equations. This defines a map $T : (\theta, \rho) \mapsto (\theta^*, -\rho^*)$, which is a diffeomorphism. Besides, because of the volume preservation property of the Hamiltonian dynamics, the Jacobian determinant of $T$ equals 1.

If we choose $g$ to be symmetric, i.e. $K(\rho) = K(-\rho)$,then

$$H(\theta^*, \rho^*) = -\ln(f(\theta, \boldsymbol{y})) - \ln(g(\rho^*)) = -\ln(f(\theta, \boldsymbol{y})) - \ln(g(-\rho^*)) = H(\theta^*, -\rho^*)$$

and together with Equation (1.4) the acceptance probability reads

$$\min\left\{1, \exp\left(H(\theta, \rho) - H(\theta^*, \rho^*)\right)\right\}$$

where we agree that $\exp(-\infty) = 0$.

It is an intrinsic property of the Hamiltonian dynamics that it leaves the value of $H$ invariant at all times. Thus, in an ideal world, where we are able to follow these dynamics accurately, we will never reject a proposed state, as $H(\theta^*, \rho^*) = H(\theta, \rho)$ and thus the acceptance probability equals 1.

Unfortunately, it is generally infeasible to simulate the Hamiltonian dynamics exactly. Thus, we need to rely on an approximation scheme. In order to maintain a viable acceptance probability, this scheme should also define a diffeomorphism and hopefully maintain the volume preservation property. A common solver that meets these requirements is the so-called leapfrog integrator and is depicted in Algorithm 2. However, due to numerical approximation errors, we have to give up the invariance with respect to $H$.

To be more specific, we start at the current values of $(\theta, \rho)$. We first draw a new sample for the momentum variable. For ease of computation in practice it is common to use $\rho^* \sim N(0, \Sigma)$. The leapfrog integrator discretizes time by using a small step size $\epsilon$ and computes the trajectory for $L$ steps at times $\epsilon, 2\epsilon, \ldots, L\epsilon$. In every iteration it alternates between half steps for the momentum variable $\rho$ and a full steps for the position variable $\theta$ as displayed in Algorithm 2.

Since the position variable is of our interest, in case of acceptance, we store the

**Algorithm 2** Leapfrog integrator

```
Input: $\epsilon$ step size, $L$ number of steps, $(\theta, \rho)$ current samples whereby $\rho \sim N(0, \Sigma)$
for $i = 1, \dots, L$ do
    $\rho^* \leftarrow \rho - \frac{\epsilon}{2}\frac{\partial P}{\partial \theta}$
    $\theta^* \leftarrow \theta + \epsilon \Sigma^{-1} \rho^*$
    $\rho^* \leftarrow \rho^* - \frac{\epsilon}{2}\frac{\partial P}{\partial \theta^*}$
    $\rho \leftarrow \rho^*$
    $\theta \leftarrow \theta^*$
end for
Return $(\theta^*, \rho^*)$
```

new sample $\theta^*$ and in case of rejection we store the current sample $\theta$, before we continue to draw a new sample for the momentum variable.

In summary, we can construct a Markov chain that alternates between updates for the auxiliary momentum variables, which are typically independent Gaussians, and a Metropolis update of the position variable, which is the variable of our target distribution. For the latter, a new value is proposed by calculating a trajectory due to the Hamiltonian. In most cases, the implementation must be discretized using the leapfrog algorithm, which depends on a step size and a desired number of steps. This way, proposals can be distant, yet with a high probability of acceptance.

**Limitations of the Hamiltonian Monte Carlo algorithm**

The greatest limitation of the algorithm is that fact that it is only suitable for continuous state spaces with the additional condition that the gradient of the log probability with respect to the state variables needs to be able to be computed quickly with sufficient accuracy. Else, the computational time needed may make the sampler inefficient.

For some distributions the sampler has difficulties exploring the whole state space. For example if $f(\theta, \boldsymbol{y})$ is not strictly positive, we might run into problems since the Hamiltonian dynamics are unable to pass regions where it is zero.

The overall performance of the Hamiltonian Monte Carlo algorithm is very sensitive with respect to the step size and the desired number of steps in the leapfrog method. If the number of steps is chosen to be too small, the exploration of the state space can be impaired as in the case of a random walk. On the other hand, if the number of steps is chosen too large, it can lead to unnecessary computation due to u-turns. Also, if the step size is too small, too many small steps will be taken. If it is too large, the leapfrog integration will be less accurate leading to more rejected proposals.

A possible work around would be to find a rule to adjust the step size dynamically in order to avoid u-turns. However, it is hard to find one that respects the dynamics equally in both directions, forward and backward. Ultimately, this would violate the necessary reversibility of the dynamics.

#### No-U-Turn Sampler (NUTS)

The No-U-Turn Sampler (NUTS) is an extension of the Hamiltonian Monte Carlo sampler Hoffman and Gelman (2014). It addresses the aforementioned issues regarding the tuning of the number of steps and the step size. Its idea is to automatically tune the Hamiltonian algorithm and to avoid for the Hamiltonian trajectories to go backwards and thus making u-turns. Therewith it reduces the dependence between samples and the computational cost.

It starts with proposing a new momentum variable. In order to find out when a u-turn takes place, NUTS simulates in both directions of time, forwards and backwards at each iteration. This can be represented by a balanced tree that grows in depth by one and doubles the number of nodes per iteration. As it also doubles computation time, the algorithm terminates when either the maximal tree depth allowed is reached, or the trajectory has started to turn back on itself. Finally, the transitions are sampled from the binary tree with respect to multinomial sampling with a bias towards more distant nodes.

The NUTS performs at least as well as the Hamiltonian Monte Carlo sampler and automatically tunes the parameters and thus fully circumvents the hand-tuning of the sampler.

For a vivid comparison of the random walk Metropolis algorithm, the Hamiltonian Monte Carlo and the NUTS sampler, see for example McElreath (2017).

### 1.3.3 PyMC3

PyMC3 (Salvatier et al., 2016) is a Python package for Bayesian statistical modelling and probabilistic machine learning. It is open source and allows to fit Bayesian models using MCMC methods like Metropolis but also state of the art gradient-based Hamiltonian Monte Carlo (HMC) and the No U-turn Sampler (NUTS). By relying on the python library Theano as the computational back end it renders computation optimization, dynamic C compilation as well as simple GPU integration.

The interface is programmed such that high-dimensional and complex models can be fitted with only a little specialized knowledge of the complex internals and the underlying fitting algorithms. Thus PyMC3 is bridging the gap between statistical theory and the user application. However, so far, its implementation does not include epidemic models because they do not fall into the category of hierarchical models: they cannot be represented as a product of conditional distributions with respect to single infection times. Consequently, model specification is not straightforward within the PyMC3 environment. Our research aims to address this problem.

Note, the programming code for the extension of the PyMC3 framework and the models can be found at `https://fhm-chicas-code.lancs.ac.uk/koeppell/py-mc-3-epi-extension`.

## 1.4 Gaussian process

In this section we turn to another stochastic process, namely the Gaussian process. They enjoy increasing popularity in spatial statistics (Diggle and Ribeiro, 2007; Diggle and Giorgi, 2019). A Gaussian process is a semiparametric approach to modelling and is commonly used in machine learning for 1-d regression problems. We start by first recalling the definition of a normal distribution and some of its properties before we proceed to introduce Gaussian processes and how they can be used for prediction.

### 1.4.1 Multivariate normal distribution

**Definition 1:** A univariate random variable $X$ follows a *1-dimensional normal distribution* (or *Gaussian distribution*) with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ if its probability density function $p_X$ reads

$$p_X(x \in \mathbb{R} \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Extending this concept to a multivariate setting with $d$ random variables $X = (X_1, ..., X_d)$ it is not sufficient to only consider their marginal means and variances. The random variables $X_1, ..., X_d$ may be correlated, which is why the marginal variance for each of the random variables in general does not comprise enough information about the joint probability distribution, Therefore, instead of dealing with a single vector for the variance parameter, we consider the *covariance matrix* $\Sigma$, whose $i, j$-th entry is the covariance between the $i$-th and $j$-th random variable, that is $\Sigma_{ij} = \text{Cov}[X_i, X_j] = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])]$. The covariance matrix has two properties: $\Sigma$ is symmetric and $\Sigma$ is positive-semidefinite. The latter states that for any vector $y \in \mathbb{R}^d$ we have $y^T \Sigma y \geq 0$ (Bishop, 2006). Furthermore, we note that since $\text{Cov}[X_i, X_i] = \mathbb{V}[X_i]$, its diagonal elements consist of the marginal variances.

Assuming that the determinant of $\Sigma$ is not equal to 0 and thus the inverse $\Sigma^{-1}$ exists, we can proceed with the following definition:

**Definition 2:** A $d$-dimensional random variable $\boldsymbol{X}$ follows a *multivariate normal distribution* if its probability density function $p_X$ reads

$$p_X(\boldsymbol{x} \mid \boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{|2\pi\Sigma|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{1.6}$$

for $\boldsymbol{x} \in \mathbb{R}^d$, the mean vector $\boldsymbol{\mu} \in \mathbb{R}^d$, the covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ and $|\cdot|$ denotes the determinant of a matrix.

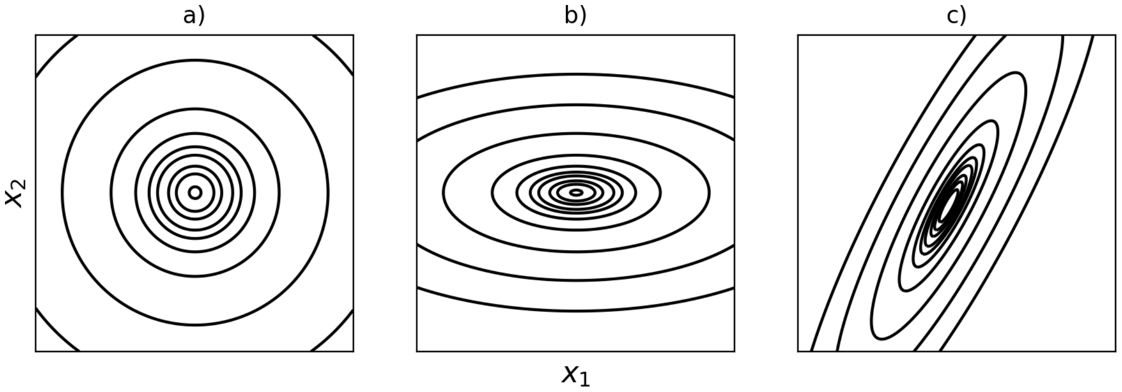
The multiplicative factor in front of the exponent is a normalization constant for the density function to integrate to one. The interesting and distinct characteristic of the distribution in Equation (1.6) is the exponent, because it can be viewed as a certain measure of distance. In fact, $d_M(\boldsymbol{x}, \boldsymbol{\mu}) = \sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$ is called the *Mahalanobis distance* between the points $\boldsymbol{x}$ and $\boldsymbol{\mu}$.

The Mahalanobis distance takes into account the different standard deviations of the $X_i$s, and also their correlations. It is mainly governed by the inverse of the covariance matrix, which is also referred to *precision matrix.*

To illustrate the relationship of $\Sigma$ and the Mahalanobis distance, we consider the bivariate case as depicted in Figure 1.2 a) - c). Given a center point, all points with the same Mahalanobis distance from this centre lie on an ellipse. These ellipses constitute the contour lines and produce a picture that is familiar from the bivariate normal distribution. In particular, in a) the random variables are uncorrelated and have equal variances, hence we receive a circular round shape. In b) the random variables are still uncorrelated but the variance of $X_1$ is higher then the variance of $X_2$ leading to an oval shape stretched horizontally. Finally, c) extends case b) and further illustrates the case with an entry in the off-diagonal denoting a correlation between the two variables. This leads to a rotation of the space.

### 1.4.2 Slicing of the multivariate normal distribution

Let $\boldsymbol{X}$ be an $(n+m)$ - dimensional random variable that follows a multivariate Gaussian distribution $\boldsymbol{X} \sim \mathcal{N}(0, \Sigma)$ with mean vector $\mathbf{0}$ and covariance matrix $\Sigma$.



$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_1^2 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \quad \sigma_2^2 < \sigma_1^2 \qquad \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_1^2\sigma_2^2 c \\ \sigma_1^2\sigma_2^2 c & \sigma_2^2 \end{bmatrix}$$

$$-1 < c < 1$$

Figure 1.2: Examples of the Mahalanobis distance for 3 different settings for 2-dimensional random variables. a) uncorrelated with equal variances, b) uncorrelated with different variances c) (in this example positively) correlated with different variances.

With the following partitions $\boldsymbol{A} = (X_1, X_2, \dots, X_n)^T$ and $\boldsymbol{B} = (X_{n+1}, X_{n+2}, \dots, X_{n+m})^T$ of dimensions $n$ and $m$ respectively, we have

$$\begin{bmatrix} \boldsymbol{A} \\ \boldsymbol{B} \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix} \right)$$

whereby the covariance matrix is sliced into four block matrices with respect to $\boldsymbol{A}$ and $\boldsymbol{B}$. The elements of these block matrices are the covariances evaluated for every two random variables in the corresponding partitions. For example, $\Sigma_{AB}$ is a matrix of size $(n \times m)$ and

$$\Sigma_{AB} = \begin{bmatrix} \mathrm{Cov}[X_1, X_{n+1}\ ] & \mathrm{Cov}[X_1, X_{n+2}\ ] & \dots & \mathrm{Cov}[X_1, X_{n+m}\ ] \\ \mathrm{Cov}[X_2, X_{n+1}\ ] & \mathrm{Cov}[X_2, X_{n+2}\ ] & \dots & \mathrm{Cov}[X_2, X_{n+m}\ ] \\ & & \vdots & \\ \mathrm{Cov}[X_n, X_{n+1}\ ] & \mathrm{Cov}[X_n, X_{n+2}\ ] & \dots & \mathrm{Cov}[X_n, X_{n+m}\ ] \end{bmatrix}$$

A nice property of the multivariate normal distribution is that the marginal distribution of the partition $\boldsymbol{A}$ (and $\boldsymbol{B}$ respectively) is given by the slices

$$\boldsymbol{A} \sim \mathcal{N}(\mathbf{0}, \Sigma_{AA})$$

Further, if we condition $\boldsymbol{B}$ on the partition $\boldsymbol{A}$, the resulting conditional distribution is again multivariate normal (Bishop, 2006). To be precise, we get

$$(\boldsymbol{B} \mid \boldsymbol{A} = \boldsymbol{a}) \sim \mathcal{N}(\ \Sigma_{BA}\Sigma_{AA}^{-1}\boldsymbol{a}\ ,\ \Sigma_{BB} - \Sigma_{BA}\Sigma_{AA}^{-1}\Sigma_{AB}) \tag{1.7}$$

This concept is the central pillar of prediction with Gaussian processes, as we will see later in Section 1.4.6.

### 1.4.3 Sampling from a multivariate normal distribution

In this thesis, we will be confronted with sampling from a Gaussian process, which is why this chapter focuses on a sampling algorithm from such a multivariate normal distribution.

Let us first look at the linear transformation of a multivariate normal distribution. For this purpose, let $\boldsymbol{X} \sim N(\mathbf{0}, I)$ be a standard multivariate normal distributed random variable with mean vector $\mathbf{0}$ and the identity matrix $I$ as covariance matrix. We apply the transformation $\boldsymbol{Y} = \boldsymbol{\mu} + L\boldsymbol{X}$ with an invertible matrix $L$. By means of the linearity of the expectation and the bilinearity of the covariance we get

$$\begin{aligned}
\mathbb{E}[\boldsymbol{Y}] &= \boldsymbol{\mu} + L\mathbb{E}[\boldsymbol{X}] = \boldsymbol{\mu} \\
\mathbb{V}[\boldsymbol{Y}] &= \mathbb{E}[(\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}])(\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}])^T] \\
&= \mathbb{E}[(\boldsymbol{Y} - \boldsymbol{\mu})(\boldsymbol{Y} - \boldsymbol{\mu})^T] \\
&= \mathbb{E}[(L\boldsymbol{X})(L\boldsymbol{X})^T] \\
&= \mathbb{E}[L\boldsymbol{X}\boldsymbol{X}^T L^T] \\
&= L\mathbb{E}[\boldsymbol{X}\boldsymbol{X}^T]L^T \\
&= LL^T
\end{aligned} \tag{1.8}$$

It is easy to show, that a linear transformation of a multivariate normal random variable is again a multivariate normal random variable. Therefore, with the fact that $\boldsymbol{Y}$ is normal and (1.8), it follows that $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, LL^T)$. We can now make use of the Cholesky decomposition (Press et al., 1992), which states that for any symmetric, positive definite matrix $\Sigma$, there exists a composition

$$\Sigma = LL^T$$

with L being a lower triangular matrix with strictly positive entries on the diagonal.

Assuming we are able to sample from a 1-dimensional normal distribution (see for example Box and Muller (1958)) and we have a fixed covariance matrix, then Algorithm 3 describes how to sample from a multivariate normal distribution. The algorithmic complexity is of $O(n^3)$ for the Cholesky decomposition, $O(n)$ to draw $n$ one-dimensional samples and $O(n^2)$ for the matrix-vector multiplication. This results in an algorithm with cubic complexity overall.

**Algorithm 3** Sampling from a multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$

```
Input: $\Sigma$: $n \times n$ covariance matrix, $\boldsymbol{\mu}$: $1 \times n$ vector
$L \leftarrow$ CholeskyDecomposition($\Sigma$)
for $j = 1, 2, \ldots, n$ do
    Draw a sample $x^{(j)}$ from $\mathcal{N}(0, 1)$
end for
$\boldsymbol{x} \leftarrow (x^{(1)}, x^{(2)}, \ldots, x^{(n)})^T$
Store $\boldsymbol{y}^{\leftarrow}\boldsymbol{\mu} + L\boldsymbol{x}$
Return $\boldsymbol{y}$
```

### 1.4.4 Defining a Gaussian process

An intuitive idea to describe a Gaussian process is to consider it as a distribution over functions. More formally, we cite the definition of Rasmussen and Williams (2006):

**Definition 3:** A *Gaussian process* is a collection of random variables, any finite number of which have a joint Gaussian distribution.

Let this collection of random variables be a real valued process $F(x)$ that projects elements from a space $\mathcal{X}$ onto the real line. We write

$$F \sim GP[\mathbf{m}(x), \mathbf{c}(x, x^{'})]$$

whereby its *mean function* $\mathbf{m} : \mathcal{X} \rightarrow \mathbb{R}$ is given by

$$\mathbf{m}(x) = \mathbb{E}[F(x)]$$

and its *covariance function* $\mathbf{c} : \mathcal{X}^2 \rightarrow \mathbb{R}$ is given by

$$\begin{aligned}\mathbf{c}(x, x') &= \mathbb{E}\big[\big(F(x) - \mathbb{E}[F(x)]\big)\big(F(x') - \mathbb{E}[F(x')]\big)\big] \\ &= \mathbb{E}\big[\big(F(x) - \mathbf{m}(x)\big)\big(F(x') - \mathbf{m}(x')\big)\big].\end{aligned}$$

The mean function may encode a priori information for the expectation of the function. However, in the following we assume without loss of generality $\mathbf{m}(x) = 0$ for all $x \in \mathcal{X}$.

The covariance function determines the entries of the covariance matrix. It gives information about, given two points $x$ and $x'$, what the relationship of their function values $F(x)$ and $F(x')$ are. In other words, it tells us about the correlation of $F(x)$ and $F(x')$ with respect to $x, x' \in \mathcal{X}$. Hence, the covariance function is responsible for the smoothness of the process. This measure of similarity and smoothness can be implemented in many ways and we are free to choose any function $\mathbf{c} : \mathbb{R}^2 \rightarrow \mathbb{R}$ that is symmetric in its arguments and positive-semidefinite. In other words, for $d$ real numbers $x_1, \ldots, x_d$ it must hold, $\mathbf{c}(x_i, x_j) = \mathbf{c}(x_j, x_i)$, and the matrix $\Sigma$ with entries $\Sigma_{ij} = \mathbf{c}(x_i, x_j)$ must be positive semi-definite. In the next section we present a selection of different types of covariance functions, though a more detailed overview of covariance functions and their construction can be found in Rasmussen and Williams (2006).

### 1.4.5 Covariance functions

In this section we list some common examples for covariance functions. In each of the examples, $\rho$ is a scaling parameter, also referred to as *lengthscale* parameter, that governs how fast the correlation decreases; the parameter $\sigma^2$ governs the variance. To illustrate the various covariance functions, Figure 1.3 depicts the decay with respect to distance of these functions and some realizations of a Gaussian process with respect to different covariance functions.

- *Squared exponential (SE)* or *Gaussian* covariance function

  $$\mathbf{c}(x, x') = \sigma^2 \exp\Big(-\frac{||x - x'||^2}{2\rho^2}\Big)$$

  For points that lie very close, the squared exponential covariance function takes values close to $\sigma^2$. It then decreases exponentially as the distance of the points increases. Since realizations of a process with this function are infinitely differentiable (Rasmussen and Williams, 2006), this covariance function entails a very smooth process.

- *Exponential* covariance function

  $$\mathbf{c}(x, x') = \sigma^2 \exp\Big(-\frac{||x - x'||}{\rho}\Big)$$

  The exponential covariance function is also $\sigma^2$ when points are equal, but decreases at a faster rate for small distances compared to the *SE* covariance function. It leads to realizations that are continuous but not differentiable (Rasmussen and Williams, 2006), and is a very rough process.

- *Matérn class*

  $$\mathbf{c}(x, x') = \sigma^2 \frac{2^{1-v}}{\Gamma(v)} \Big(\frac{\sqrt{2v}||x - x'||}{\rho}\Big)^v K_v\Big(\frac{\sqrt{2v}||x - x'||}{\rho}\Big),$$

  where $K_v$ is the modified Bessel function of the second kind (Abramowitz and Stegun, 1965). It is a generalization of the above covariance functions, and thus more flexible but also more complex. For half integer values of $v$, that is $v = m + 0.5, m \in \mathbb{N}$, the function takes a simpler form: it decomposes into a product of an exponential and a polynomial of order m. The resulting realizations are $v - 1$ times differentiable (Rasmussen and Williams, 2006), which accounts for their smoothness. Examples are

- $v = 1/2$: This simplifies to the exponential covariance function
- $v = 3/2$: The functional form of a Matern(3/2) covariance function reads

$$\mathbf{c}(x,x') = \sigma^2\left(1 + \frac{\sqrt{3}||x-x'||}{\rho}\right)\exp\left(-\frac{\sqrt{3}||x-x'||}{\rho}\right)$$

- $v \to \infty$: In the limit, it simplifies to the *SE* covariance function

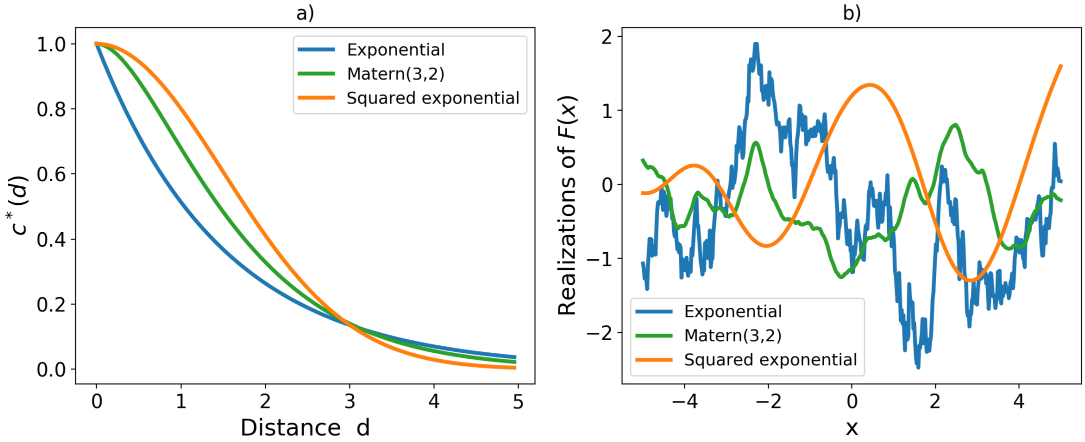


Figure 1.3: Comparison of different covariance functions. a) Decay of the covariance functions with increasing distance of the points. b) Realizations from processes with the according covariance function with a function approximation of 500 points. In both plots, the lengthscale is $\rho^2 = 1.5$ and the prior variance is $\sigma^2 = 1$.

Another way of classifying covariance functions is by certain properties:

If a function $\mathbf{c}(x, x')$ can be reduced to be a function of the difference between the two points, that is

$$\mathbf{c}(x,x') = c^*(x-x') \quad \forall x, x' \tag{1.9}$$

then the covariance function is called *stationary*. In other words, the output of $c$ is only depending on the vector between these two points. Hence, it is independent of where in space the points lie.

An even stronger characteristic is if the covariance function can be expressed as a function of only the Euclidean distance between the points, that is

$$\mathbf{c}(x, x') = c^*(||x - x'||) \quad \forall x, x' \tag{1.10}$$

then the covariance function $c$ is called *isotropic*. In this case the output of $c$ is only depending on the distance between the two points, that is the length of the vector between the two points. Hence, $c$ is translation invariant but also invariant of the direction of the vector. Since Equation (1.10) is implying Equation (1.9), any isotropic covariance function is also stationary.

All covariance functions of Example 1.4.5 are isotropic and thus stationary because they only depend on the Euclidean distance of the points. Note that there exist a variety of other non-stationary covariance functions. In fact, one can combine or modify existing covariance functions to make new ones (see for example Rasmussen and Williams (2006)). This variability can be made use of and should be considered when constructing a model for a specific application because the type of covariance function should be chosen with respect to a priori information about the process.

For our context, the Matern(3/2) covariance function is of particular interest because of its simple form and its degree of smoothness: it is not unrealistically smooth as the squared exponential function, but also it is not unrealistically rough as the exponential function.

In geostatistics, instead of working with a covariance function, one can also use the so-called *(semi-)variogram* (see for example Diggle and Ribeiro (2007)), as there exists a one-to-one mapping between the two. This exceeds the scope of this thesis, why we will not go into further detail here.

### 1.4.6 Predictions using Gaussian processes

In this chapter we are particularly interested in the posterior distribution of a Gaussian process after having observed the values at a finite number of points. In other words, we are interested in the distribution over functions that are restricted to pass through (or lie close) to the observations. In Figure 1.4 this concept is illustrated: in a) we see random draws from a Gaussian process evaluated at 500

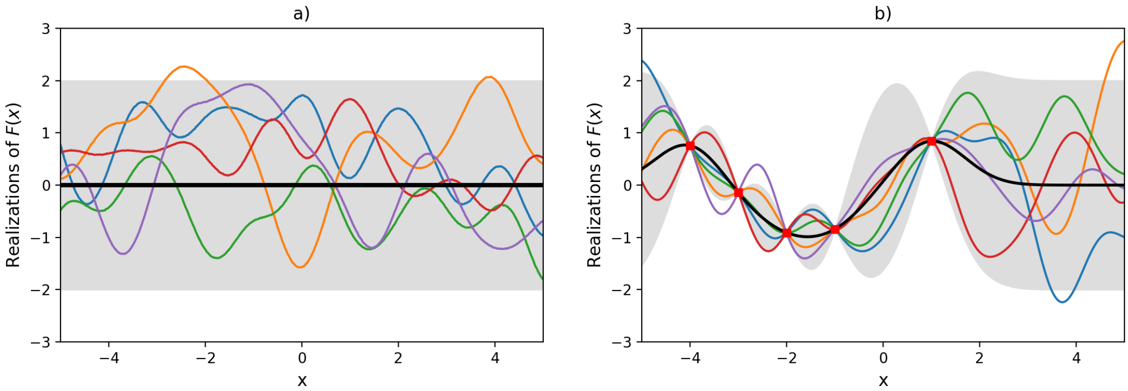


Figure 1.4: a) Sample paths from a Gaussian process (Prior) b) Sample paths from a Gaussian process after the noiseless observation. In both pictures, the black line denotes the mean function and the grey area denotes the 95% credible region evaluated pointwise.

points with mean function $m(x) = 0$ for all $x \in \mathbb{R}$ and a for illustration purposes, a squared exponential covariance function with parameters $\sigma^2 = 1$ and $\rho^2 = 0.5$. The black line denotes the mean function, the grey area denotes the area between the mean and two times the standard deviation. Since every marginal distribution is normal, this corresponds to the 95% credible region evaluated at each point. Figure 1.4 b) displays sample paths of this Gaussian process after having observed the red points. It becomes clear that the more points are observed the more certain we are about the curvature of the functions. As the variability decreases close to points and increases as you move away from points, at areas with less observations (e.g. at the borders) the uncertainty about the functions is higher and the credible region is broader.

To express this in mathematical notation, let $\boldsymbol{A}$ denote a finite set of observations, usually referred to as training points. Conditioning on the training points $\boldsymbol{A}$ we are interested in predicting the values for the new unobserved points $\boldsymbol{B}$. According to the Definition 3 of Gaussian process, the joint distribution of $\boldsymbol{A}$ and $\boldsymbol{B}$ is a multivariate normal. By means of the slicing properties of a multivariate normal distribution from Section 1.4.2, we can state the conditional distribution, or in other words the posterior given the observations $\boldsymbol{A}$, as

$$(\boldsymbol{B} \mid \boldsymbol{A} = \boldsymbol{a}) \sim \mathcal{N}(\ \Sigma_{BA}\Sigma_{AA}^{-1}\boldsymbol{a}\ ,\ \Sigma_{BB} - \Sigma_{BA}\Sigma_{AA}^{-1}\Sigma_{AB})$$

$\boldsymbol{B}$ is always assumed to be a finite set by which the posterior function is approximated. For every point we achieved a normal distribution as posterior distribution given the observations and thus we can calculate estimators and credible regions. In a geospatial setting, this type of Gaussian regression and prediction is also referred to as *Kriging* (Krige, 1951).

In reality, observations are assumed to be measured with noise. To take account of that, we assume that the noise at each point is normal distributed with mean 0 and variance $\sigma^2$, that is $\boldsymbol{\mathcal{E}} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$. It follows that the noisy version of the observation $\boldsymbol{A}$ is distributed $(\boldsymbol{A} + \boldsymbol{\mathcal{E}}) \sim \mathcal{N}(\mathbf{0}, \Sigma_{AA} + \sigma^2 I)$.

Therefore, the joint distribution of the training and predictive points changes to

$$\begin{bmatrix} \boldsymbol{A} + \boldsymbol{\mathcal{E}} \\ \boldsymbol{B} \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \Sigma_{AA} + \sigma^2 I & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix} \right)$$

It is to note that the noise term $\sigma^2 I$ is only added to the covariance matrix of the training points as we do not want to predict with an error. Using Equation (1.7) the posterior distribution given the noisy observation $\boldsymbol{A} + \boldsymbol{\mathcal{E}}$ becomes

$$(\boldsymbol{B} \mid \boldsymbol{A}+\boldsymbol{\mathcal{E}} = \boldsymbol{a}+\boldsymbol{\varepsilon}) \sim \mathcal{N}(\ \Sigma_{BA}(\Sigma_{AA}+\sigma^2 I)^{-1}(\boldsymbol{a}+\boldsymbol{\varepsilon})\ ,\ \Sigma_{BB}-\Sigma_{BA}(\Sigma_{AA}+\sigma^2 I)^{-1}\Sigma_{AB})$$

Figure 1.4 displays the prediction using the assumption of noiseless data. Using the same prior assumptions in Figure 1.5 a), that is $m(x) = 0$ for all $x \in \mathbb{R}$ and a squared exponential covariance function with parameters $\sigma^2 = 1$ and $\rho^2 = 0.5$, Figure 1.5 illustrates the prediction this time with noisy training data. Whereas a) is identical to Figure 1.4a), in b) we notice that the credible region around the observed points is broader compared to Figure 1.4b). That means, that the choice of functions is less limited to pass exactly through the observed points.

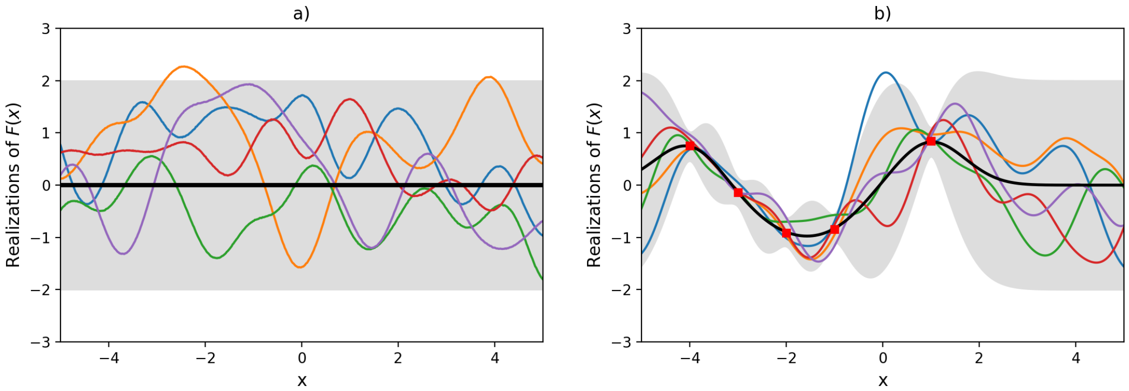


Figure 1.5: a) Sample paths from a Gaussian process (Prior) b) Sample paths from a Gaussian process after the noisy observation. In both pictures, the black line denotes the mean function and the grey area denotes the 95% credible interval evaluated pointwise.

## 1.5 Epidemic modelling

### 1.5.1 History of epidemic modelling

The use of mathematics for modelling infectious diseases and practical disease control has a long history. A nice overview can be found for example in the paper of Brauer (2017) or Serfling (1952).

Analysing data of infectious diseases began in 1662 with John Graunt with his work on "The Bills of Mortality". These records comprised weekly data starting from 1592 on the counts and the cause of deaths of citizens in London parishes. In his analysis, Graunt estimated the risk of dying due to a certain disease setting the grounds for the theory of competing risks. Later in 1766, the first approach of a mathematical model for disease data was taken by the Swiss mathematician Daniel Bernoulli where he analysed the prevention of (at that time endemic) smallpox disease by inoculation. In his work he calculated life expectancies and death rates and for the first time defined a force of infection parameter, the annual rate of becoming infected, and a case fatality parameter, the proportion of infections resulting in death (Dietz and Heesterbeek, 2000).

About 100 years later, before people had knowledge about the transmission process of infectious diseases, John Snow studied the temporal and spatial pattern of the cholera epidemic in London. In his work in 1855, he identified the water pump at Broad Street as major force of cholera infection. A similar understanding of the spatial spread of typhoid was achieved by William Budd in 1873. References to these works can be found in Brauer (2017)

Finally, in the early twentieth century, the foundations to compartmental models were developed by public health physicians, namely R.A. Ross, W.H. Hamer, A.G. McKendrick, and W.O. Kermack.

In the year of 1906, Hamer considered the spread of infections to be dependent on the number of susceptible and infected individuals (Hamer, 1906). Ross studied the transmission dynamics of malaria between mosquitoes and humans and his work was honoured with the second Nobel Prize in Medicine. Contrary to the common belief that malaria can only be eradicated with the elimination of all mosquitoes, Ross showed with a compartmental model that it would suffice to reduce the number of mosquitoes below a critical threshold (Ross, 1911). This set the basis for what is nowadays known as *basic reproduction number*. It is defined as the average number of secondary cases that arise from one infected individual in an otherwise susceptible population. Although McKendrick had developed a stochastic version of the general epidemic model (McKendrick, 1926), it did not receive much attention at that time. Instead a model which attract more attention was the deterministic model of Kermack and McKendrick (1927). It is considered to be the first complete mathematical model to describe the spread of an infectious disease consisting of ordinary differential equations for the general epidemic model.

A stochastic model in discrete time which has received great attention in the literature, is the chain-binomial model by Reed and Frost in 1928 (Fine, 1977). In 1949 Bartlett proposed a stochastic version of the aforementioned primal model by Kermack and McKendrick (1927), which became known as the *general stochastic epidemic* model (Bartlett, 1949).

This has entailed a great variety of research on epidemic modelling since. Hence, elaborating further on the developments to be found in the literature would exceed the scope of this report. Instead, we highlight key references which give an overview of the mathematical modelling of infectious diseases. For the work on epidemic modelling before 1975, the book of Bailey (1975) is widely known. Further, the monograph by Anderson and May (1991) is one of the most cited references on epidemic models and covers almost exclusively deterministic models. Recent books include Diekmann and Heesterbeek (2000), which focuses on deterministic models, while the monographs by Andersson and Britton (2012) and Greenwood and Gordillo (2009) focus on stochastic modelling. The books by Daley and Gani (2001), and Keeling and Rohani (2007) introduce aspects of both approaches.

### 1.5.2 Epidemiological models

By means of mathematical modelling we can obtain a better understanding of disease transmission dynamics and the infection process. Using underlying assumptions of the disease, we can express the epidemic process as a simplified reality, that captures the essential features of this system. Fitting such a model leads us to estimate disease transmission parameters of interest. Further, models can provide predictions of the disease spread in space and time, and thus, may be of great help to guide policy makers in their decision making for targeted countermeasures, such as for instance vaccination, trading restrictions zones, fogging of areas or even culling of farms.

There exists a wide variety of different approaches to infectious disease modelling. In the literature these approaches are usually divided into mechanistic and empirical (or statistical) models (see for example Diggle and Giorgi (2019)), which is subject to oversimplification. Both, statistical and mechanistic models in epidemiology, are utilized to analyse data for gaining better understanding of diseases. The aim of a particular research study is to seek for knowledge which is provided by the information we generate from the model. Since our research deals with such a combination of these approaches, we elaborate shortly on their differences.

The aim of mechanistic models is to mimic a real process to reflect this underlying data generating process. Hence, the structural composition of such a model is informed by the contextual behaviour and seeks to conform with this. Examples of mechanistic models include modelling the temporal and spatial progression of a disease outbreak or modelling structural biological cell processes. The model is built on the assumption that the complex system can be understood by the functionalities of its individual parts, hence the structure of the model can be responsive to the resolution of the data available; from a simple compartmental SIR model (as given in Section 1.5.4) through to a full individual based modelling framework. It is to mention that contrary to common belief, mechanistic models are not deterministic by nature. Instead of using for example differential equations we can also utilize stochastic approaches as we will see later in Section 1.5.4.

Instead of assuming a fixed model outline, for empirical models, we decide from the data what the model should look like. That is, we test whether our model assumption leads to a good explanation of the data or whether a different model with maybe different covariates or unexplained random effects would explain the observation better. Here, statistical testing and residual analysis are usually applied to obtain the model of best fit. The aim is then to estimate covariate-effects on an outcome variable, for example the effect of temperature on malaria incidences, or map unexplained (spatial) variation due to unknown factors that have not yet been accounted for in the model, for example an unexplained spatial effect in disease occurrence. Since choosing the right model is data-driven, the flexibility of the model approach poses a risk for overfitting complex models to sparse data.

In both cases, after we assume that our model is of great fit for the purpose, we can find parameter estimates (either using a Bayesian or likelihood-based approach) and predict; whether it is the probability of infection due to certain features (covariates) of an individual by means of an empirical model, or the number of infected individuals after some time by means of a mechanistic model.

Of course it cannot be denied that this partition is oversimplified. An empirical model might be informed a priori about possible relationships between certain covariates and the response variable. On the other hand, to fit a mechanistic model we use statistical inference techniques. Nevertheless, we want to stick to this separation as we see importance in considering their differences when choosing the appropriate modelling approach, especially with respect to answering a research question and analysing problems.

### 1.5.3 Stochastic versus deterministic modelling

While in the beginning of epidemic modelling there was more interest in deterministic models, stochasticity was subsequently added. The reason for having used deterministic rather than stochastic models lies in their advantage of their simpler analysis. They can have an increased level of complexity and yet - as long as numerical solutions are available - can be analysed. By placing our trust on the law of large numbers, they are suitable for large populations with a high level of disease incidence.

However, there are a number of arguments why a stochastic model should be preferred to a deterministic one. Firstly, in their nature disease outbreaks are always stochastic: Starting with an initial setting and reversing time several times, one cannot expect the exact same observation, that is the same number of individuals being infected at the exact same time. Hence, the notion of *probability of infection* plays an important role as we cannot state the definite outcome.

Further, only stochastic models allow us to model stochastic phenomena such as the probability of a major epidemic outbreak. Additionally, by means of a stochastic model we cannot only make inference on model parameters, but also gain information about the uncertainty of these estimates. Finally, in the era of a tremendous increase in the speed and memory of computers, highly intense computations are now possible in contrast to past decades. This gives rise to enormous computational simulations of stochastic models.

Considering these above arguments, in this thesis we will only deal with stochastic modelling approaches of infectious diseases.

### 1.5.4 Mechanistic models

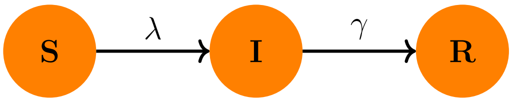


Figure 1.6: Graphical representation of a stochastic SIR compartment model

We start with a mechanistic modelling context in which we only consider the so-called compartmental models. Such models describe disease dynamics in a population that is divided into a mutually exclusive discrete set of compartments or states. The transition rules between different compartments is subject to prior information about the disease process. Since individuals transition between different states, such models are also referred to as state-transmission models.

As displayed in Figure 1.6, a common example is the SIR model, where the population consists of susceptible (S), infected (I) and removed/recovered (R) individuals. Susceptible individuals can get infected with rate $\lambda$ and only thereafter can get removed from the system with rate $\gamma$, for instance by culling of farms or death, or recover from the disease with a lifelong immunity. For the model outline, let $(S, I)$ denote the number of susceptible and infected individuals in the respective compartment. Considering an outbreak as a sequence of subsequent events (infection or removal), the time until the next event follows an exponential distribution with rate $\lambda SI + \gamma I$. Either the next event is an infection, that is $(S, I) \to (S-1, I+1)$ with probability $\frac{\lambda SI}{\lambda SI+\gamma I}$, or the next event is a removal, that is $(S, I) \to (S, I-1)$ with probability $\frac{\gamma I}{\lambda SI+\gamma I}$. This leads to a Markov process as the underlying distribution is exponential; the next event it is only dependent of the current number of individuals in the compartments and independent of the past.

Compartment models can be modified in many ways relevant to a particular disease process: Individuals can transition to a previous compartment, for instance when infected individuals become susceptible again (SIS model), or a new compartment can be added to the model, for example an exposed compartment prior to infection, where the individual is infected but not yet infectious (SEIR model).

The simplest general setting is to assume that we have a homogeneously mixing population where every individual is facing the same force of infection. In this setting, only the total number of individuals in each compartment at each time is important; since all individuals are identical, there is no need to track specific individuals. In this sense, the model can be thought of as representing a mass-action process. This approach is especially suitable for large populations with a high number of infections. In such a setting individual characteristics are aggregated out and stochastic effects do not play an important role. Relying on these assumption, the epidemic process can be described by the deterministic model of Kermack and McKendrick (1927), using a set of differential equations to describe the transition process. However, as discussed earlier, disease outbreaks are stochastic in their nature. Thus, we seek for ways to introduce stochasticity to such a model.

One approach is to start with a deterministic model and add observational noise. For example in the deterministic model for plague transmission via human ectoparasites from Dean et al. (2018), parameter samples of their respective posterior distributions are used to solve a set of differential equations. Its deterministic solution is then used as the mean of a Poisson distribution modelling the mortality count. In a different research from Samat and Percy (2012) to study dengue disease in Malaysia, a (SIR-SI) model is proposed consisting of a discrete spatio-temporal susceptible - infective - recovered model for human populations and a susceptible - infective model for mosquito populations. They use a set of differential equations to describe the disease transmission process. In order to reflect the randomness inherent in the data, one of the parameters in one of the differential equations, namely the number of new infectives, is assumed to be Poisson distributed with a mean depending on elements of the transmission.

Although such an approach can bring computational advantages as one can rely on techniques used for deterministic models, it does not incorporate the individual nature. Such individual stochastic effects, also referred to as demographic stochasticity, can be implemented by assuming underlying distributions for the transition process for each individual. Supposing that these distributions are all exponentials with respective transition rates, we get a Markov process. That is, due to the memorylessness property of the exponential distribution, it is only important in which compartment an individual is at a given time in order to determine where and when the individuals is moving next. Examples in the literature for homogeneous mixing in a Markovian epidemic modelling manner can be found in O'Neill and Roberts (1999); Xiang and Neal (2014).

### 1.5.5 Homogeneous vs heterogeneous mixing in mechanistic models

It can be argued whether the assumption of homogeneous mixing of the population is realistic. So far we have assumed that the probability of interaction does not depend on the location or social group behaviour, and every individual is facing the same amount of infectious pressure. However, for example clustering of individuals, social interaction or intrinsic conditions may increase the risk of infection for some susceptibles and influence the dynamics of disease spread.

In fact, to take account for heterogeneities in population-based epidemiological studies has shown to be important (Coutinho et al., 1999). For example, White et al. (2010) reasons that due to environmental conditions, vector and host-related factors individual heterogeneity arises in malaria transmission and failing to account for them can lead to biased estimates of malaria vaccine efficacy.

Thus, in the following we want to introduce some approaches on how heterogeneity in their given context is modelled.

*Network models* provide a powerful tool when individuals are in contact with only a small proportion of the population. They can aid to investigate disease dynamics, if the focus lies especially on the interaction between individuals or different groups. For example, in the case of livestock diseases, an outbreak can be investigated with respect to the underlying trade network (Koeppel et al., 2018; Lentz et al., 2016). In such an analytic approach, a farm is represented as a node, and an edge between two nodes exists if there is a possibility of disease transmission between the two farms; that is, if the respective farms traded with each other. On a similar basis, Britton and O'Neill (2002) consider a Markovian stochastic epidemic model in which the underlying social structure of a population is described by a Bernoulli random graph. A different theoretical approach is random geometric graphs theory (Penrose et al., 2003). Here, a spatial Poisson process is used to create a point pattern and the points that lie within a certain distance are then connected. As an example, Preciado and Jadbabaie (2009) study the dynamics of a viral spreading process and describe a spreading model in random geometric graphs. Further types of networks commonly used within an epidemiological context can be found in Keeling and Rohani (2007).

A different from of heterogeneity in the spread of diseases can arise due to spatially varying factors, which cause differences between individuals or populations at different geographical locations. For example, climatic and environmental conditions can have an impact on the habitat of vectors and their distribution, and thus on the infection risk they pose. Further, the social structure of human populations might vary spatially due to a higher contact rate in cities than in smaller villages. To capture such features, there are many different forms of *spatial models* that can be adapted to the scale of data we possess, the amount of knowledge of the population's behaviour. An overview is given in the book by Keeling and Rohani (2007) out of which we briefly mention the most important ones.

Within spatial models, *metapopulation models* provide a powerful framework to account for features that occur naturally at some locations but not at others. In such a model the population is divided disjointly into smaller populations with own independent dynamics. The model allows additionally for different rates of infection between the populations. Further inside to applications of this approach especially within ecology can be found in Hanski et al. (2004).

If the spatial location of the individuals is seen to be important, though a natural partitioning in discrete subpopulations is not possible, we can take advantage of *lattice-based models*. They are essentially a special form of metapopulation models because a lattice or grid partitions the population and all individuals within one grid are grouped together to form a subpopulation. The difference is that the interaction between the subpopulations (grid cells) is tightly constrained such that interactions occur only localized with the nearest or nearest and next-nearest cells. For example Kao (2003) used a model of foot-and-mouth disease transmission on a hexagonal lattice of farming premises to investigate whether alternative policies would have resulted in significantly better control of the epidemic.

Although lattice models provide a framework to take account of the spatial location, the discretization of space can be a disadvantage as the exact position of an individual is limited by the size of the grid it is located in. A workaround would be to consider space and population on a continuous scale by decreasing the grid size infinitely small (Keeling and Rohani, 2007). This leads to continuous-space models that uses partial differential equations or integro-differential equations. Although this approach allows for flexible modelling, it is mathematically complex and can be computationally expensive.

The purest form of heterogeneity are *individual-based models* which can include complex and detailed individual behaviour and features and is the focus of this report. They are linked with the above concepts: when we consider a fully connected graph where every node depicts one individual or a metapopulation model where every subpopulation has population size 1. In a fully individual model, we can incorporate spatial influences on a susceptible individual due to the certain location and the environmental factors at that particular location. For example, one individual might face a higher mosquito density due to advantageous mosquito habitat conditions at its location, compared to little mosquito presence at a location of an individual living in a climatic much cooler environment. To incorporate the varying susceptibility of an individual that influenced *T. orientalis* (Ikeda) spreading, (Jewell and Brown, 2015) included parameters for the occurrence of the transmitting vector, but also environmental factors.

Moreover, such an individual-based model allows for individual transmission rates $\lambda_{ij}$ for a susceptible $j$ who receives infectious pressure from an infected individual $i$. This is especially valuable when relying on the assumption that geographically close infected individuals exert more infectious pressure than infected individuals further away. This feature of spatial variation in infectious pressure can be modelled using a so-called *transmission kernel* or *spatial kernel.* It is a function of distance, which decreases as distance gets large. For example in Jewell et al. (2009) and Xiang and Neal (2014), the spatial kernel represents the environmental transmission rate between farms i and j for foot-and-mouth disease using a Cauchy-type kernel, whereas Gibson (1997) uses a power-law decay to model the spatial spread of Citrus Tristeza virus.

So far, the above approaches were based on a Markov process with a Poisson point process for infections (exponential waiting times between events) and exponential infectious periods. The choice of an exponential distribution is rather motivated by its simplicity for the model because the memorylessness property enables us to reduce the process history to only the previous event to compute the current one. But it can be argued whether it is realistic. To circumvent this problem, some literature introduce a new exposed state prior to being infected to avoid exponential infectious periods (see for example Jewell et al. (2009); Lekone and Finkenstädt (2006)). However, as this is not the focus of this report, we will not go into further detail here.

### 1.5.6 Empirical models

Geostatistical models are typically empirical in character. In geostatistics, we decide from the data, what the model should be and chose the appropriate model that explains the observed data best.

Here, the class of generalized linear models (GLM) introduced by Nelder and Wedderburn (1972) is of particular focus. They use the framework of a linear regression, where the response variable follows a Normal distribution, and extend it to response variables with distributions from the exponential family. The linear predictor consists of a linear combination of explanatory variables, or covariates, and is linked to the expectation of the response variable by means of a link function. In case of linear regression, this link function is simply the identity function. We assume that the response variables are independent and that the covariates are measured exactly, that is without measurement errors.

Sometimes, however, the data shows greater variability than what we would be able to explain from fitting a GLM with fixed effects. To take account of an unobserved factor which has not been considered by any of the included covariates, we can add random effects to the linear predictor. This yields a model which includes both, fixed and random effects, and thus is called generalized linear mixed model (GLMM). In geostatistics, modelling a random effect for unobserved spatial correlation by means of a Gaussian process received much attention over past years (see Diggle and Ribeiro (2007)).

We shortly want to introduce the reader to the concept of a geostatistical generalized linear mixed model. Suppose at $n$ locations $\ell_1, \ell_2, \ldots, \ell_n$, we have corresponding observations $Y_1, Y_2, \ldots, Y_n$. For each location $\ell_i, i = 1, \ldots, n$, we further associate $k$ explanatory variables $\boldsymbol{X_i} = (X_{i1}, X_{i2}, \ldots, X_{ik})^T$. For a geostatistical generalized linear mixed model we define the linear predictor $\eta_i$ as

$$\eta_i = \boldsymbol{X_i}^T \boldsymbol{\beta} + s(\ell_i)$$

where $\boldsymbol{\beta} = \beta_1, \beta_2, \ldots, \beta_k$ are the unknown parameters for the fixed effects and the unknown spatial effect $s(\ell_i)$ is modelled through a Gaussian process $\big(s(\ell)\big)_{\ell \in \mathbb{R}^2}$ with constant 0 mean function, and some stationary and isotropic covariance function. We link the linear predictor consisting of the fixed and random effects to the observed outcome random variable $Y_i$ using an appropriate link function $g : \mathbb{R} \to \mathbb{R}$ with

$$\mathbb{E}(Y_i) = g(\eta_i)$$

In epidemiology especially Poisson and Binomial models are of importance. By means of these distributions we can model disease or vector counts and relate this number to observed features at the location of a population or individuals, and unobserved effects. For example, Diggle et al. (2007) used a Binomial model with an logit link function to create prevalence maps of the parasitic *Loa loa* infection in West Africa. In a different study, McCann et al. (2017) modelled the abundance of adult malaria vectors inside people's houses using a Poisson model with a log link function. They included environmental covariates specific to the location of the house and features of the household affecting Malaria transmission (e.g. the use of insecticides or different wall types). Next to a spatial random effect, they also included an unobserved independent random effect specific to each household in the linear predictor.

In geostatistical modelling we may observe an additional summand in the linear predictor, the so-called *nugget effect.* This is an independent random effect specific to the location $\ell_i$. It consists of two components, a microscale variation and a measurement error. In this thesis we neglect this effect as it would add complexity and further non-identifiabilities when combining mechanistic and empirical models.

Including terms for unobserved (spatial) effects gives rise to great flexibility in constructing a suitable model. Sometimes the available dataset might be incomplete and come with a low resolution spatially and/or temporally which is true especially for low-to-middle-income-countries where disease registries are non-existent or have a partially complete geographical distribution (Diggle and Giorgi, 2019). In these situations, such a geostatistical model finds great use due to its flexible structure.

### 1.5.7 Combination of mechanistic and empirical models

While mechanistic models are being constructed using information out of the contextual knowledge, empirical models acquire information from data. As Diggle and Giorgi (2019) reason, mechanistic models can oversimplify the process and thus, including a data-based empirical model can improve their fit. In this thesis we want to introduce a combination of a mechanistic SIR model with principled data-based methods from geostatistics. The idea behind this approach is to allow a parameter of a continuous time mechanistic model to be replaced by a term consisting of covariates and a latent spatial Gaussian process which operates on a continuous space.

We list some literature on combining a mechanistic with an empirical model which we will put into context relating our research in Section 1.5.8.

Mugenyi et al. (2017) investigated a deterministic malaria SIS model where they estimated the force of infection using the point prevalence. This prevalence was modelled by means of a generalized linear mixed model to account for individual-

and household-specific clustering. That way, they were able to model the intrinsic observed and unobserved heterogeneity of the individual risk of malaria infection within a compartment model framework.

Samat and Percy (2012) studied dengue disease in Malaysia with a discrete spatio-temporal combination of a SIR model for human populations and SI model for mosquito populations (SIR-SI) model. The model uses deterministic differential equations to describe the disease transmission process, and adds a stochastic component by assuming the number of new infectives to be Poisson distributed. The mean of this distribution is defined by a linear predictor term consisting of an intercept and a spatial random effect using a conditional autoregressive CAR prior.

A time continuous susceptible-infected-detected (SID) model is proposed by Jewell and Brown (2015) to model host incidence cases of *Theileria Orientalis* in New Zealand. They include a term for indirect observations of vector activity in the intensity functions of the infection process modelled as Poisson processes. By means of Bayesian inference methods, he is able to map a seasonal discrete-space latent risk surface and is able to give prediction estimates of the future incident counts of the outbreak.

For our research, this model by Jewell and Brown (2015) constitutes the basis and our aim is to further extend it to a continuous spatial scale.

### 1.5.8 Modelling vector-borne diseases

Vector-borne diseases are characterized through the indirect transmission of pathogens by a vector specific to the disease (Figure 1.1). An essential approach in protecting humans and veterinary health from the burden caused by vector-borne disease agents is controlling the vector population. Ross (1911) modelled the interaction between mosquitoes and humans for Malaria infection, where he showed it would suffice to reduce the number of mosquitoes below a critical threshold to eradicate Malaria. Later, George Macdonald developed this methodology further and linked metrics for measuring the important components of transmission to the model.

However, vector habitats, vector behaviour and thus the distribution of the pathogen are influenced by environmental factors and meteorology: climate change and globalized trade contribute to the extension of vector habitats to different territories and hence the distribution of many diseases (see e.g. Fischer et al. (2011); Medlock and Leach (2015); Campbell-Lendrum et al. (2015)). Only concentrating on meteorological data, that influence the vector habitat has its limitation, as Ostfeld et al. (2005) points out clearly; while there might be a high density in vector presence due to advantageous conditions for the vector's habitat, the actual number of infected vectors can still be small. On the other hand, when only concentrating on disease incidence cases, the underlying risk might not be displayed correctly; for instance due to a small susceptible population size in a high risk region.

Hence, we seek for a joint approach that combines the meteorological data and the host incidence cases. As farmers are obliged to report a suspicion of notifiable diseases, veterinary host incidence data is usually available from government databases. In contrast, collecting data about all possible vector populations and their habitat with the vector's rapid adaptation to changing ecological and environmental conditions is unfeasible (Braks et al., 2011; Purse et al., 2005; Parham et al., 2015). Further, in the face of a rapidly changing environmental and ecological climatic conditions previously well fitted models might become obsolete soon. Hence, the need for a model which is flexible with respect to the availability of the data increases.

Here, a Poisson or Binomial model (see Section 1.5.6) can be of great use. However, such models are concerned with predicting the intensity or the probability of disease infection and not necessarily the spatio-temporal progression of the outbreak. Thus, stochastic mechanistic vector-borne disease transmission models seem to be more suitable in this context. Szmaragd et al. (2009) proposed a model for bluetongue virus transmission within and between farms in Great Britain. However, their approach only accounts for the farm's local temperature and no other meteorological data that might influence the vector's distribution. Similarly, Sumner et al. (2017) studied bluetongue virus and Schmallenberg virus by assuming that the vector-mortality rate is temperature-dependent and added a function for seasonal vector activity based on previous knowledge.

So far, the flexibility of including meteorological data, that influence vector behaviour, into compartment models seems to remain a challenge. Thus, combining a rather rigid mechanistic approach with flexible methods from geostatistics can provide a suitable solution.

One such an approach is proposed by Samat and Percy (2012), who developed a spatio-temporal combination of a stochastic SIR model for human populations and a deterministic SI model for mosquito populations. The number of new infections is modelled using a Poisson distribution, where the mean consists of a linear predictor term that includes an intercept and a spatial random effect. They propose to include other covariates, which for example can be of meteorological character. Although this approach provides a great combination of an epidemic model with an empirical model by including spatial components within a compartment model, it has two major disadvantages: It does not include individual stochastic effects, and it operates on a discrete time scale and on discrete space.

On a continuous time scale, Jewell and Brown (2015) link data on host incidence cases with indirect observations of vector activity and predict the spatio-temporal variation in disease transmission in response to a sudden incursion of the novel strain of *Theileria Orientalis* in New Zealand. In the absence of detailed information on vector ecology, their Bayesian dynamical model learns sequentially with the progression of the epidemic, tracks the disease process and is able to forecast ongoing disease spread. This approach is based on estimating a discrete-space latent risk surface for the vector presence, which provides a good approximation of key components of vector presence. However, it is limited by its level of discretization: there might be considerable variability in risk over a district with districts possibly being too coarse a scale for reasonable estimation of the risk.

To the knowledge of the author, so far there exists no disease modelling approach that operates on a continuous time scale as well as on a continuous spatial scale. In this thesis, we for the first time propose a model that accounts for continuity in both, time and space, by including a Gaussian processes framework into a compartmental SIR model. This way, we are able to account for variability in risk for reasonable estimation within a larger district. Using these augmented data, the prediction of the likelihood of the disease arriving in different parts of the country together with the vector population's propensity to spread the disease once it has arrived will be more accurate.

As described in Jewell et al. (2009), a Bayesian approach to data assimilation and forecasting allows for flexibility with the choice of the model, and enables us to include missing data, such as infection times or vector population related information, or latent variables. In our approach we can take great advantage of well-developed, complex frameworks like Gaussian processes. Furthermore, we are able to marginalize over unobserved or latent quantities in order to estimate model parameters, which is otherwise difficult in the frequentist framework. Moreover, such marginalizations take into account the uncertainty arising from an estimation of these unknowns.

## 1.6 Bluetongue virus outbreak

In this section we introduce the real world outbreak by which we test our methodology in this thesis. The dataset of the bluetongue outbreak in 2007 in the United Kingdom is particularly suitable for developing and testing our model, being a well-characterised outbreak of a notifiable vector-borne disease occurring when little in the way of prophylactic control had been implemented. First, we will elaborate on the background of the disease and the outbreak, before we describe the datasets of the outbreak and the meteorological data we will be working with. This way we point out characteristics of the datasets that will motivate the forthcoming chapters in this thesis.

### 1.6.1 Bluetongue virus

Bluetongue is a serious illness affecting ruminants and is caused by the bluetongue virus (BTV) with so far 24 different known serotypes (Friedrich-Loeffler-Institut, 2014). It is often subclinical or unapparent, though for sheep and cattle the disease is acutely progressing. Symptoms generally include high fever, excessive salivation, nasal discharges, swelling of the head and neck, and potentially reddening of the tongue, hence the disease' distinct name. This leads to high mortality in susceptible animals and production loss. Vaccination of ruminants against bluetongue exists and is an effective protection against the spread of the disease (Science Media Centre, 2007).

The virus is transmitted by biting midges of the genus Culicoides. Their biting activity is usually highest between dusk and dawn and about seven days after a bloodmeal from an infected animal, the midges are able to transmit the virus further to susceptible animals in a subsequent bloodmeal. Meterological factors like hot temperature and humidity are favourable for midge reproduction and activity. Therefore, bluetongue infections are highly seasonal and most prevalent during warm, wet months in late summer and autumn. Furthermore, wind can carry the midges over great distances due to their small size (1-3 mm), leading to the introduction of the disease into previously disease-free regions. Historically, Bluetongue was first recognized in South Africa and only prevalent in tropical and subtropical areas. However, climate change and globalized trade contributed drastically to the increased spread of the disease to temperate regions (Maclachlan, 2011).

One of the most severe outbreaks in Northern Europe 2006 - 2008 was caused by the Bluetongue virus serotype 8 (BTV-8). The route of BTV-8 introduction to Central Europe in 2006 is still unknown, but independently, BTV-8 found a new vector, midges of the group Culicoides obsoletus. They are indigenous in Central and Northern Europe while being highly adapted to ruminants. In 2006 cases were reported in the tri-border region between Germany, the Netherlands and Belgium to all neighbouring countries. In the following year 2007, first cases were confirmed in Germany which confirmed that the virus was able to overwinter. Subsequently the disease appeared again in all countries where BTV-8 was prevalent in 2006 and finally with additional cases in UK in fall 2007. With major economic losses, enhanced by trade restrictions, the disease was only eradicated in 2008 after a major vaccination campaign (Ganter, 2014).

### 1.6.2 BTV-8 outbreak progression in UK

Although Bluetongue virus 8 was already present in mainland Europe in 2006, it only reached UK in 2007. The introduction to the UK is hypothesised to be due to wind carrying the midges from mainland Europe to the East coast of England. Seemingly unlikely at first, Gloster et al. (2008) states that such a long-distance transport of BTV-infected vectors by wind has already been recorded in history. For this to happen they agree that the following conditions must be met:

- First, hot temperatures need to be present for the enhancement of virus replication within the midges and midge reproduction resulting in high abundances of infected midges.
- Second, suitable winds directed from infected areas on the continent to the South-East coast in England must be present being able to transport midges over long distances.
- Third, a susceptible host population needs to be present at their arrival destination for the midges to bite and infect before they die.

In a retrospective analysis, Burgin et al. (2009) investigated how these conditions were met for the years 2006 – 2008. Their findings were in line with the appearance of cases in the UK.

While in 2006 the hot temperatures were ideal for a spread of the disease on the continent, only a limited number of suitable days were present for windborne carriage of the infected midge population to the UK. However, in 2007 after having overwintered on the continent, hot temperatures enhanced the rapid replication of the infected midge population. Moreover, Gloster et al. (2008) identified at least 40 opportunities where meteorological conditions were favourable for windborne introduction of BTV from the Ostend area of the continent to the South East coast of England. In their risk assessment they predicted that the greatest risk of BTV-8 introduction through midges transported by wind was between May and October 2007 (Gloster et al., 2007). In fact, the first case in the UK was detected on September 15, 2007 in a cow on a farm in Suffolk. Within only a few days, BTV-8 was confirmed on a second animal on the same farm and other farms in Lowestoft and Essex. Following this, more cases appeared in Eastern England with simultaneous appearances of cases that were about 100 km distant and for which no direct link to infected farms could be found. As animal movements from infected areas were restricted and rigourous testing of livestock from disease-free areas was performed, it seems very unlikely that the introduction of BTV-8 into the country was due to trading with infected livestock. This suggests that the possibility of potential windborne spread of BTV to the UK.

Although hot temperatures in 2008 provided ideal breeding conditions for midges, there existed a significant difference to the previous years: Starting from April of 2008, an extensive vaccination campaign drastically decreased the number of susceptible livestock. Areas of infected farms from the previous year and on the South East coast of England were given priority in case infected midges were able to overwinter or in case of new windborne Bluetongue introduction. In the following months it was extended to farms located further North and West. These counter measures were so successful that no new BTV-8 cases were reported in UK in 2008 or subsequent years.

### 1.6.3 The datasets

#### 1.6.3.1 Host data

The following data on livestock farms was used for the analysis in this thesis:

- Farm demographic data for Great Britain from the Animal & Plant Health Agency (APHA). This includes the OS map reference of cattle, sheep, deer, and goat farm population data for the whole of Great Britain (England, Scotland, Wales).

- Bluetongue case record data for Great Britain from APHA. This includes the OS map reference for the farms, as well as the farm detection date (i.e. time of first suspicion and restriction imposed on the farm).

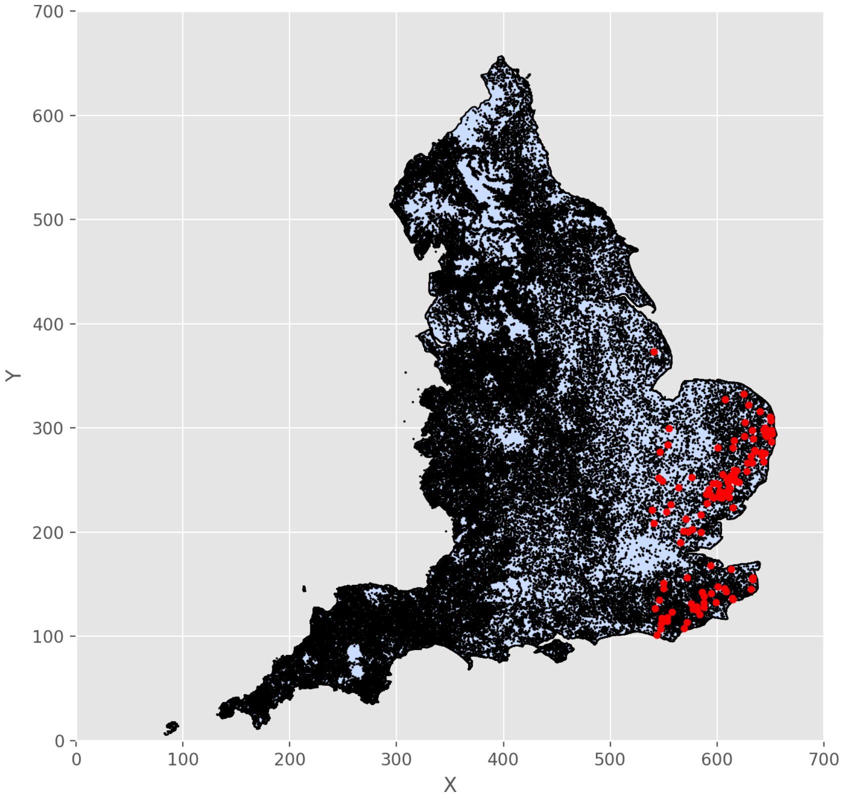


Figure 1.7: The spatial distribution of farms in Great Britain and farms with BTV-8 confirmation in red.

By means of the OS map reference, we linked cases to demographic data. The dataset comprises information on the location of 94310 farms in Great Britain which is depicted in Figure 1.7. Since the outbreak never spread further than South East England, we restricted the dataset to farms in this area resulting in 14266 farms (Figure 1.8).

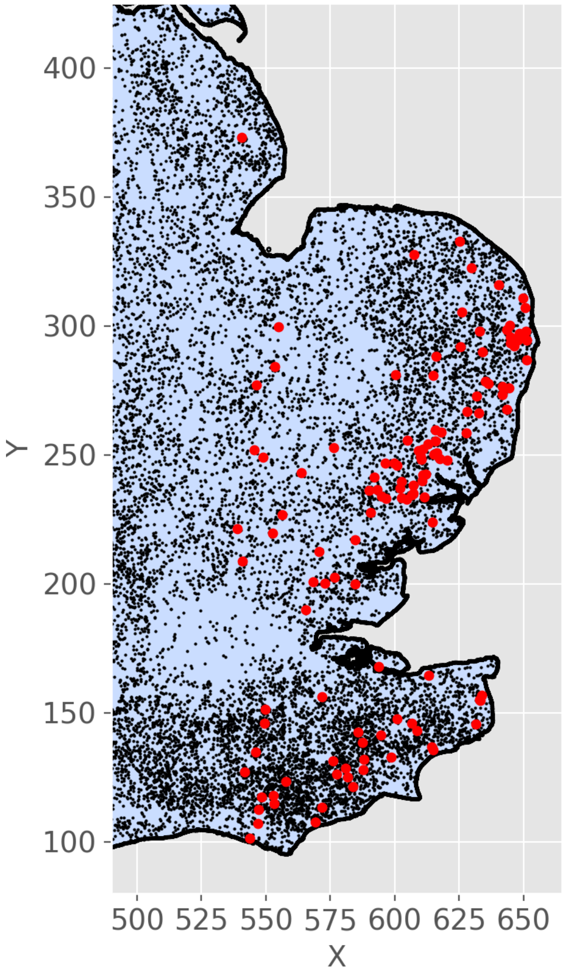


Figure 1.8: The spatial distribution of farms in Great Britain restricted to the smaller area of South East England. Red points denote infected farms.

In the dataset 129 BTV-8 cases were reported in GB. In two cases we identified two cases on the same location. Since we perform our analysis on farm level, it diminishes the number to 127 infected farms, which aligns with reports from Burgin et al. (2009). For two further cases their detection date was recorded in May

2008 which cannot be matched to information found in the literature. These cases appear with a time gap of more than 240 days after the last case, thus we discarded them from our dataset to avoid any bias for our epidemiological model. Figure 1.9 displays the chronology of the BTV-8 outbreak in Great Britain. The outbreak started on 20 September and lasted until 3 December 2007. It is to note that three days after the introduction of the first case, 64 cases were reported on the same day. This is presumably due a sudden increase in testing and thus an accumulation in reporting and/or entering into the database. We therefore assume that the disease had been present on these farms before. The remaining ongoing of the outbreak is characterized by local spreading to farms close by and simultaneously random distant appearances that would then form new small outbreak foci.

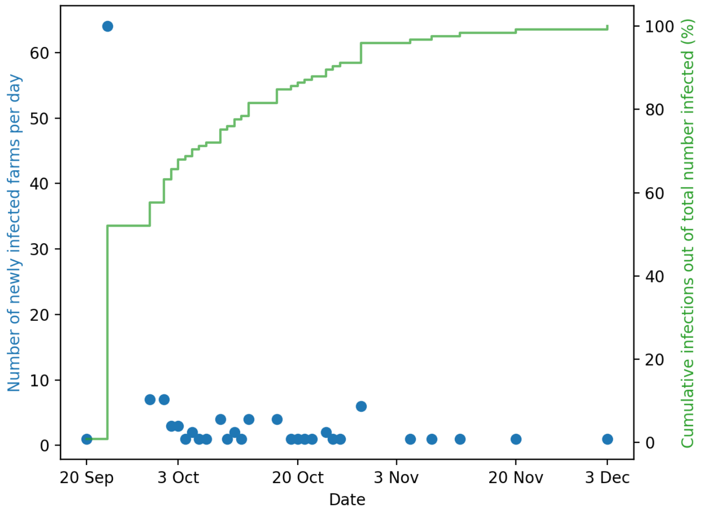


Figure 1.9: Chronology of the number of detected BTV-8 infections in Great Britain.

### 1.6.3.2 Environmental information

As mentioned before, midges are very susceptible to environmental conditions. Thus, for each farm we mapped meteorological data obtained from World Climate Research Programme (WCRP) (2016): CMIP5 monthly data on single levels, IPSL-CM5A-MR (IPSL, France), Copernicus Climate Change Service (C3S)

Climate Data Store (CDS). to include in our model. We only consider those covariates we believe midges are most affected by: temperature, precipitation, humidity, wind speed, and u/v wind direction. The data contains gridded 0.25 x 0.25-degree estimates. This corresponds to a spatial resolution of about 27-28 km for the length of the grid cell. Since our model will not be time dependent, we averaged the value of the monthly estimates for September, October, November, December in 2007 and mapped the farm locations to the nearest point in the grid.

We performed principal component analysis on the six environmental covariates and identified that the top two principal components (PCs) explain 97% of the variation in the data, out of which the first PC accounts for 67.64% and the second PC accounts for 28.97%. As the contribution of the third PC is very small (2.6%), we will be using only the top two PCs to inform our model on meteorological data. Their spatial distribution is displayed in Figure 1.10. It becomes apparent, that the principal components give a coarse representation of the environmental information of the South East of England.

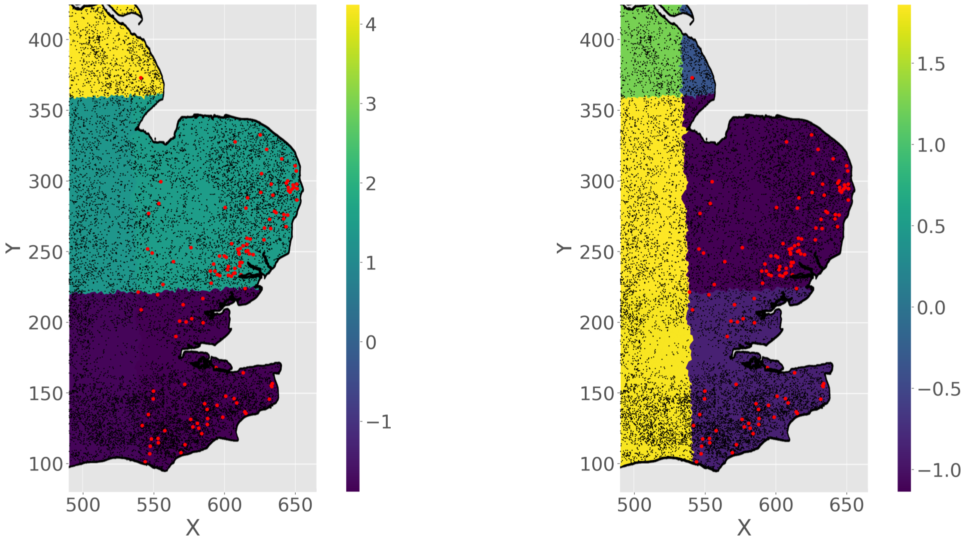


Figure 1.10: Spatial distribution of the first principal component (left) and the second principal component (right) obtained from the PCA of the environmental covariates.

The ethical approval for working with these datasets in context of this thesis was obtained by the Lancaster Medical School on 04 July 2019, Reference: FHMREC18100.

## 1.7 Remainder of this thesis

First in Chapter 2 we explain how we link data on host incidence cases with indirect observations of vector activity by constructing a joint epidemic and geostatistical model. We investigate parameter estimates of the joint model as well as the spatial variation of the underlying unknown vector risk surface.

In Chapter 3 we describe how we can deal with the computational demand when dealing with big data. Although we have already diminished the number of farms in the United Kingdom to only the ones located in South East England (Figure 1.8), we still face a challenge with an individual based model including roughly 15,000 farms. We investigate how computational efficiency can be reached in the framework of this epidemiological problem.

In the Chapter 4 we turn to the problem of the aggregated detection dates in our dataset: Half of the infected farms are recorded to be detected on the same day (Figure 1.9), which is presumable a consequence of the data collection and does not reflect reality. We thus treat them as unobserved and investigate the model parameter estimates when marginalizing over these latent quantities using a Bayesian approach.

All of the above chapters will be accompanied by extensive simulation studies in order to investigate the model's behaviour and limits.

Finally, in Chapter 5 we test our fully developed methodology by means of the real world disease outbreak BTV-8 in UK in 2007, before we end this thesis with a discussion and concluding remarks in Chapter 6.

# Chapter 2

# The joint model framework

In this chapter we specify our joint model. First, we display how we model the infection process and derive its likelihood. Afterwards, we elaborate on the joint model framework and simplify it for our use. We then construct an MCMC algorithm to efficiently draw samples from the posterior distribution. Thereafter, we perform a simulation study to investigate the model's behaviour and conclude with a discussion.

## 2.1 Infection process

As introduced in Section 1.5.4, we describe the outbreak dynamics by means of a mechanistic Susceptible - Infected - Removed/Recovered (SIR) model (Figure 1.6) where the population is divided into mutually exclusive compartments. The time until an infection is represented by a Markov property: The probability of infection only depends on the current number of individuals and is independent of the past. This can be also rephrased as the present observation is dependent on the observations of the previous time point and is independent of the observations before. The only continuous distribution that satisfies such a memorylessness property is the exponential distribution. Thus, in the following, we first introduce the exponential distribution to model the time until infection and gradually extend the concept for our use. Being part of our modelling framework, we derive the likelihood of the infection process for a whole population, which enables us to do inference on its parameters.

A reference to an article where exponential distributions were used to model epidemics would be for example Andersson and Britton (2012), and a selection of articles in an applied setting is Jewell et al. (2009), Jewell and Brown (2015), Xiang and Neal (2014).

### 2.1.1 Exponential distribution

Our aim is to model the temporal progression of epidemics. This means that we want to model the time until an infection occurs. These times are also referred to as *waiting times*.

Since the infection process is Markovian, the natural choice to be looking for modelling such waiting times is the exponential distribution as it has the memorylessness property: the waiting time until the next event in the future is independent of how much time we have already waited. In other words, for a waiting time $T \in [0, \infty)$ and any given time $t'$ with $0 \leq t' \leq t$ we have

$$\mathbb{P}\left(T \leq t \mid T \geq t'\right) = \mathbb{P}\left(T \leq t - t'\right) \tag{2.1}$$

The probability that we at most wait $t$ given that we have already waited $t'$ is the same as the probability that we at most wait $t - t'$. The only continuous probability distribution that fulfills this property is the exponential distribution (Forbes et al., 2010). Hence, for any random variable $T \in [0, \infty)$ that fulfills (2.1), the probability density is given by

$$\mathbb{P}\left(T \leq t\right) = \int_0^t \lambda e^{-\lambda x} dx$$

for all $t \in [0, \infty)$ and a fixed $\lambda \in (0, \infty)$. The parameter $\lambda$ can be interpreted as a rate or intensity parameter and governs the length of the waiting time:

$$\begin{aligned}\mathbb{E}(T) &= \int_0^\infty x \cdot \lambda e^{-\lambda x} dx \\ &= \big[-xe^{-\lambda x}\big]_0^\infty + \int_0^\infty e^{-\lambda x}\end{aligned}$$

$$
\begin{aligned}
&= (0+0) + \big[ -\frac{1}{\lambda} e^{-\lambda x} \big]_0^\infty \\
&= \frac{1}{\lambda}
\end{aligned}
$$

The mean of the waiting time states how long we have to wait on average for an event to occur. Hence, a large $\lambda$ leads to a short waiting time, a small $\lambda$ entails a longer waiting time on average. In our case for the waiting time until infection, a higher number of infected people corresponds to a higher infection pressure and thus a higher $\lambda$ and the waiting time to one's own infection is shorter. Analogously, a lower number of infected people reduces the infection pressure and thus, $\lambda$ becomes smaller and the waiting time longer.

### 2.1.2 Simultaneous exponential distributions

In the previous section we introduced the exponential distribution to represent the waiting time until an infection occurs. In an outbreak scenario we are interested in modelling several waiting times simultaneously. Consider a population of individuals where we want to model the waiting times until an infection of any individual occurs. We can then raise the questions of who will become first infected and when? The following theorem clarifies this.

**Theorem 1:** Let $X_1, X_2, \dots, X_m$ be independent exponentially distributed random variables with intensities $\lambda_1, \lambda_2, \dots, \lambda_m$ respectively. Further let
$Z := \min\{X_1, X_2, \dots, X_m\}$.
a) $Z$ is a exponentially distributed with rate $\lambda := \lambda_1 + \lambda_2 + \cdots + \lambda_m$.
b) The probability that the event of $Z$ belongs to $X_i$ is

$$\mathbb{P}\left(\{X_i < X_j, \text{for all } j \neq i\}\right) = \frac{\lambda_i}{\lambda}$$

*Proof.* a)

$$
\begin{aligned}
\mathbb{P}\left(\{Z \leq y\}\right) &= 1 - \mathbb{P}\left(\{Z > y\}\right) \\
&= 1 - \mathbb{P}\left(\{\min\{X_1, X_2, \dots, X_m\} > y\right) \\
&= 1 - \mathbb{P}\left(X_1 > y, X_2 > y, \dots, X_m > y\right)
\end{aligned}
$$

$$
\begin{aligned}
&= 1 - \prod_{i=1}^{m} \mathbb{P}\left(X_i > y\right) && \text{(Independence)} \\
&= 1 - \prod_{i=1}^{m} \exp\left(-\lambda_i y\right) \\
&= 1 - \exp\left(-\sum_{i=1}^{m} \lambda_i y\right)
\end{aligned}
$$

b) We are interested in $\mathbb{P}\left(\{X_i < X_j, \text{for all } j \neq i\}\right)$. We show that it holds for the case of two independent exponential distributions, that is $\mathbb{P}\left(\{X_1 < X_2\}\right)$. We write

$$
\begin{aligned}
\mathbb{P}\left(\{X_1 < X_2\}\right) &= \int_0^\infty \int_x^\infty \lambda_1 e^{-\lambda_1 x} \cdot \lambda_2 e^{-\lambda_2 y} \, dy \, dx \\
&= \int_0^\infty \left[\lambda_1 e^{-\lambda_1 x} \cdot (-e^{-\lambda_2 y})\right]_x^\infty \, dx \\
&= \int_0^\infty \lambda_1 e^{-(\lambda_1+\lambda_2)x} \, dx \\
&= \frac{\lambda_1}{\lambda_1+\lambda_2} \int_0^\infty (\lambda_1+\lambda_2) e^{-(\lambda_1+\lambda_2)x} \, dx \\
&= \frac{\lambda_1}{\lambda_1+\lambda_2} \left[-e^{(\lambda_1+\lambda_2)x}\right]_0^\infty \\
&= \frac{\lambda_1}{\lambda_1+\lambda_2}
\end{aligned}
$$

□

Theorem 1 tells us that considering several exponential distributions simultaneously is the same as considering one exponential distribution with the sum of its rates. Further, the infection event can be attributed to an individual with its probability being the proportion of its rate out of the sum of all rates.

### 2.1.3 Piece-wise constant rates

In the previous section we considered a collection of independent exponential distributions by which we model the waiting time until the first event. So far we assumed that all distributions are independent of each other. However, this is often too simple for practical matters, as in an outbreak scenario, the waiting time can be influenced by the infection events of the other exponential distributions.

Consider an outbreak scenario where the number of infected individuals increases, that is we observe an event for an exponential distribution as described in the previous chapter. As now there are more infected individuals, the infectious pressure on a single individual increases and we expect a shorter waiting time until its infection. Thus, we want to condition the waiting time of one individual on the events of all other individuals. This leads us to a conditional distribution for the time until infection where the rate changes each time we observe an event in any of the other distributions. Given a series of infection event times, what is the distribution of the waiting time conditioned on these external events?

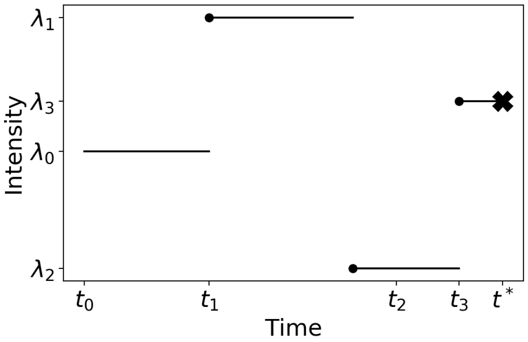


Figure 2.1: Waiting time with changing rates over time. The cross marks the event.

For this let us consider a small example depicted in Figure 2.1. The waiting process until infection starts as an exponential distribution with rate $\lambda_0$ at time point $t_0$. At time point $t_1$, we observe an external event that leads to an increase in the rate to $\lambda_1$. We recall that the exponential distribution has the memorylessness property (Equation (2.1)). It does not matter how much time has already passed for the probability of the next event to happen. We can thus "reset" the waiting process at time point $t_1$ and start a new exponentially distributed waiting time with the altered rate $\lambda_1$. We can analogously take advantage of the memorylessness property at time points $t_2$ and $t_3$, where we each start a new exponential distribution with the altered rate $\lambda_2$ and $\lambda_3$, respectively. Finally at time point $t^*$ the individual becomes infected and the waiting process terminates. We notice that the rate is piece-wise constant in between consecutive events and jumps as soon as a new external event occurs (at time points $t_1, t_2, t_3$) and thus a new exponential distribution was started.

The probability that the event happens at $t^*$ entails that it happened after $t_3$ as $t^* > t_3$. Further, the probability that the event did not happen entails that the event also did not happen until $t_2$ and until $t_1$ because $t_1 < t_2 < t_3$. In other words

$$\begin{aligned}\mathbb{P}\left(X = t^*\right) &= \mathbb{P}\left(X = t^*, X \geq t_3\right)\\ \mathbb{P}\left(X \geq t_3\right) &= \mathbb{P}\left(X \geq t_1, X \geq t_2, X \geq t_3\right)\end{aligned}$$

The probability that the event did not occur until a certain time is called *survival* and is expressed as one minus the cumulative distribution function of the exponential distribution for the time since starting the process. On the contrary, for the occurrence of the event, we state the density of the exponential distribution evaluated at the amount of time that has passed since starting the process.

We are now in the position to state the distribution of $X$.

$$\begin{aligned}&\mathbb{P}\left(X = t^* \mid t_1, t_2, t_3\right)\\ =&\mathbb{P}\left(X = t^*, X \geq t_1, X \geq t_2, X \geq t_3\right)\\ =&\mathbb{P}\left(X \geq t_1\right) \cdot \mathbb{P}\left(X \geq t_2 \mid X \geq t_1\right) \cdot \mathbb{P}\left(X \geq t_3 \mid X \geq t_2\right) \cdot \mathbb{P}\left(X = dt^* \mid X \geq t_3\right)\\ =&e^{-\lambda_0 t_1} \cdot e^{-\lambda_1(t_2-t_1)} \cdot e^{-\lambda_2(t_3-t_2)} \lambda_3 e^{-\lambda_3(t^*-t_3)}\end{aligned}$$

Generalizing this concept to $n$ external events at time points $t_1 < \cdots < t_n$, we can state the probability for a susceptible person to get infected at $t^*$ as

$$\exp\Big(-\sum_{i=0}^{n-1} \lambda_i(t_{i+1} - t_i)\Big) \cdot \lambda_n \exp(-\lambda_n(t^* - t_n)) \tag{2.2}$$

with $t_0 := 0$.

### 2.1.4 Likelihood of the infection process

In the previous section we introduced the theory behind the waiting time process which enables us to model a time until infection with varying rates over time. The amount of variation of the rates may be determined by parameters. We are interested in drawing inference on these parameters given only the observed event times in reality. Hence, we rely on the likelihood function, which is the interface between the observed event times and the unknown parameters: it is a function of the parameters given the data. Therefore, we now derive the likelihood given observed event times; that is, the value of the density function evaluated at the observation.

In Section 2.1.3 we introduced the waiting time of an individual conditioned on the waiting times of other individuals. As we have seen from the example in Figure 2.1, the rate of the conditional waiting time changes over time. A different way to perceive this is to model the rate as a function of time $t \mapsto \lambda(t)$ with $\lambda(t) \in [0, \infty)$ for all $t \in [0, \infty)$, that is piece-wise constant and continuous from above. Assume $t_1 < t_2 < \cdots < t_n < t^*$, where $t^*$ denotes the time of the event. Then we write

$$\lambda(t) = \begin{cases} \lambda_0 & t < t_1 \\ \lambda_1 & t_1 \leq t < t_2 \\ & \vdots \\ \lambda_{n-1} & t_{n-1} \leq t < t_n \\ \lambda_n & t_n \leq t < t^* \\ 0 & t^* \leq t \end{cases} \tag{2.3}$$

After the event at $t^*$ occurred, the rate or intensity function becomes 0 as no further event is possible anymore.

In the following, let $\lambda(\mathrm{I}^-)$ denote the rate that was present right before the infection occurred. For example, in Equation (2.3), this would equal to $\lambda_n$. Using (2.3) we can rewrite Equation (2.2) as follows

$$\exp\left(-\int_{t_0}^{t^*} \lambda(t)dt\right) \lambda(\mathrm{I}^-)$$

with $t_0 := 0$.

It may happen that the end of our observation period $t_{end}$ has been reached before an event was observed. In this case, the density term for the infection drops and we get

$$\exp\left(-\int_{t_0}^{t_{end}} \lambda(t)dt\right)$$

We are now in the position to derive the likelihood of the infection process for all individuals in a whole population. Let $\mathbf{I}$ be the vector of observed infection times and $\mathrm{I}_j$ be the entry for individual $j$. In case of no infection until the end of the observation period $t_{end}$, we set the infection time to infinity. Please note that the first infected individual, denoted by $i_0$, does not receive any infectious pressure.

$$\begin{aligned} f(\mathbf{I} \mid \boldsymbol{\theta}) &= \prod_{\substack{j:\mathrm{I}_j<\infty \\ j\neq i_0}} \left[ \exp\left( - \int\limits_{\min\{\mathbf{I}\}}^{t_{\text{end}}} \lambda_j(t)dt \right) \lambda_j(\mathrm{I}_j^-) \right] \times \prod_{j:\mathrm{I}_j=\infty} \exp\left( - \int\limits_{\min\{\mathbf{I}\}}^{t_{\text{end}}} \lambda_j(t)dt \right) \\ &= \exp\left( -\sum_j \int\limits_{\min\{\mathbf{I}\}}^{t_{\text{end}}} \lambda_j(t)dt \right) \prod_{j:\mathrm{I}_j<\infty} \lambda_j(\mathrm{I}_j^-) \end{aligned} \tag{2.4}$$

Please note, that this theory on the infection process may also be perceived as the sum of conditional Poisson processes. A Poisson process is a counting process with exponentially distributed waiting times between events. In this case, we model the time until infection for a single individual through a Poisson process conditioned on all other Poisson processes. The intensity function is set to 0 as soon as the infection occurred and therefore, the Poisson process of a single individual may jump at most once. A more detailed coverage of this concept is provided in the Appendix (A.1).

## 2.2 The joint model specification

As discussed in Sections 1.5.7 and 1.5.8, our research combines a mechanistic model with an empirical modelling approach: the idea is to allow a parameter of a continuous time mechanistic model to be replaced by a term consisting of covariates and a random effect, which operates on a continuous spatial scale. In our context of vector-borne disease modelling, these added covariates and random effects represent factors that influence the vector's habitat or the vector's activity. Note that this approach is also applicable, for example, to directly transmitted diseases. In such a scenario, the latent spatial random effect may reflect areas of unknown spatial transmission risk.

In order to investigate the need of integrating a Gaussian process into the model, we can plot the infected individuals at equally distant, consecutive time points and look at the spatial distribution of the infections. If we notice a pattern of delayed new infections close to previously infected individuals, we can assume a spatial influence over time between the infected individuals. In some areas the infection might spread faster (or slower), with equal distance between the individuals. This can further provide proof that an underlying risk surface has impact on the disease transmission dynamics.

We use a SIR state transition model to describe the epidemic process as introduced in Section 1.5.4. The population of $n$ individuals, labelled $1, 2, \ldots, n$, is divided into three mutually exclusive sets: susceptible $\mathcal{S}$, infected $\mathcal{I}$ and removed or recovered $\mathcal{R}$ individuals. We assume that the population is closed, that is, there are no births, deaths, immigration or emigration during the course of the epidemic. This entails that the number of susceptibles, infected and removed individuals at any time $t$ sums to $n$. Such an assumption is reasonable if the disease outbreak lasts only for a relatively short period in which major demographic changes will most likely not occur.

Transitioning between different states is only allowed when susceptibles get infected, or when infecteds die or acquire life-long immunity to the disease. In the latter case, removed individuals still count as part of the system (because the population is closed) but do not play any further role in the spread of the disease. In fact, any contact between an infected individual and another infected or removed individual does not have any effect as individuals can only get infected once.

We model the infection and removal process per individual (individual-based model). The removal process is governed by the infectious period which is the time between infection and removal events and is distributed according to a random variable $Q \sim \text{Gamma}(\delta, \delta\gamma)$ with probability density $f_Q(x) = (\delta\gamma)^{\delta} x^{\delta-1} \exp(-\delta\gamma x))/\Gamma(\delta)$. The reciprocal of $\gamma$ reflects the average length of the infectious period with $\mathbb{E}(Q) = \delta/(\delta\gamma) = 1/\gamma$.

Note that a $\text{Gamma}(\delta, \delta\gamma)$ distribution with an integer shape parameter $\delta$ is the sum of $\delta$ exponential distributions each with mean $\delta\gamma$. This is quite convenient for epidemic modelling as it allows us to combine a chain of events (e.g. different stages of the infection process). Thus, for $\delta = 1$, we get an exponential distribution with mean $1/\gamma$ and variance $1/\gamma^2$. If we increase $\delta$, the expectation stays constant, however, the variance decreases: $\mathbb{V}(Q) = \delta/(\delta\gamma)^2$. In the literature (Xiang and Neal, 2014), $\gamma = 4$ was found to be a good value for the shape parameter for the Gamma distribution of the infectious period, which is why we adapt it for our method.

Whilst each infectious period is independent of the rest of the process, the same does not hold for the time until infection. The infectious pressure on every individual depends on several factors and on the infection state of other individuals. It changes over time and is different for every individual as introduced in the previous section.

To be more specific, we suppose that a susceptible $j$ is more likely to be infected if other infected individuals are close. This can be modeled by a spatial kernel $K$ which describes the infectious pressure depending on the distance to other infecteds. For two individuals, $d_{ij}$ denotes their Euclidean distance and this is equivalent to $||\ell_i - \ell_j||$ where $\ell_i$ is the spatial location of individual $i$. We make use of an exponential kernel $K(d_{ij}; \kappa) = \exp(-d_{ij} \cdot \kappa)$ with scale parameter $\kappa$ where the infectious pressure is high for close contact and decreases as the distance becomes large. Here $\kappa$ governs the speed of the decay, that is for larger $\kappa$ the infectious pressure decreases more rapidly with distance (see Figure 2.2). The total infectious pressure exerted on a susceptible individual $j$ from all surrounding infected individuals at time $t$ is then $\psi \sum_{i \in \mathcal{I}(t)} K(d_{ij}; \kappa)$, where $\mathcal{I}(t)$ denotes the

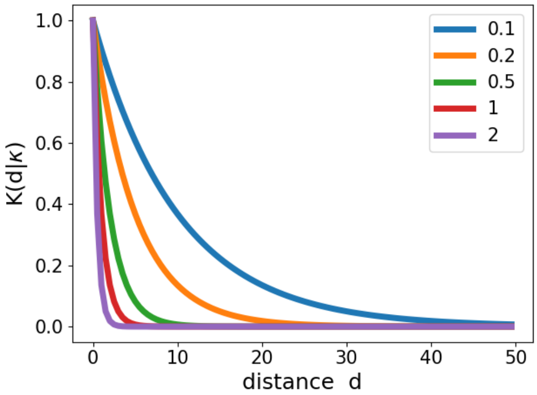


Figure 2.2: The decay of the spatial kernel function $\exp(-d_{ij} \cdot \kappa)$ for different values of $\kappa$.

set of indices from all infected individuals at time $t$ and $\psi$ is an infection rate parameter.

Furthermore, the spatial location of a susceptible can play a role in the risk of infection. Advantageous ecological and environmental conditions may increase vector activity or vector reproduction, thus imposing a greater risk by for example an increased vector presence at a specific location. In the following, let $\ell_j$ be the location of an individual $j$ expressed in spatial coordinates (X and Y coordinates). We want to incorporate indirect but measurable vector related information at location $\ell_j$ into our model. We define

$$v(\ell_j) = \boldsymbol{X}_{\boldsymbol{\ell_j}}^T \boldsymbol{\beta} + \mathbf{s}(\ell_j)$$

$v(\ell_j)$ incorporates explanatory variables $\boldsymbol{X}_{\boldsymbol{\ell_j}} = (X_{\ell_j 1}, \ldots, X_{\ell_j k})^T$ with respect to location $\ell_j$, and corresponding unknown parameters $\boldsymbol{\beta} = \beta_1, \ldots, \beta_k$. Examples for such covariates may include remotely sensed environmental data, such as temperature, humidity and precipitation, but also vector density if such information is available.

$\mathbf{s}(\ell_j)$ describes the remaining spatial effect that was not yet accounted for by any of the covariates. To be more specific, the collection $\mathbf{s} = \big(\mathbf{s}(\ell)\big)_{\ell \in \mathbb{R}^2}$ of continuous random variables are assumed to constitute a Gaussian process. Consequently, for the random variables at a finite set of observation points with locations

$\boldsymbol{L} = (\ell_1, \ldots, \ell_n)$, we have $\mathbf{s}_L = (\mathbf{s}(\ell_1), \ldots \mathbf{s}(\ell_n))^T \sim \mathcal{N}(\mathbf{0}, \Sigma)$ with mean vector $\mathbf{0}$ and covariance matrix $\Sigma$. $\Sigma_{ij}$ is calculated using a Matern(3/2) covariance function such that

$$\Sigma_{ij} = \sigma^2 \left(1 + \frac{\sqrt{3}d_{ij}}{\rho}\right) \exp\left(-\frac{\sqrt{3}d_{ij}}{\rho}\right)$$

As discussed in Section 1.4.5, the choice of a Matern(3/2) covariance function is suitable due to its simple form and its degree of smoothness: it is not unrealistically smooth as the squared exponential function, but also not unrealistically rough as the exponential function.

We further include a basic infectious pressure parameter $\alpha$, which captures infections that cannot otherwise be explained by the spatial kernel. Another important quantity in infectious disease modelling is the transmission rate $\psi$, which can be higher for some diseases like Measles and lower for others, like Leprosy.

We are now in the position to state the infection process as introduced in Section 2.1. Assume in the beginning of an outbreak, one initial infective introduces the disease into an otherwise totally susceptible population. For a susceptible individual $j$, we model the infection process through a conditional Poisson process, where the intensity function is given by the force of infection on $j$

$$\lambda_j(t) = \begin{cases} \psi \cdot \left(\alpha + \exp(v(\ell_j)) \sum\limits_{i \in \mathcal{I}(t)} \exp(-d_{ij} \cdot \kappa)\right) & j \in \mathcal{S}(t) \\ 0 & \text{else} \end{cases} \tag{2.5}$$

Conditioned on all others, each individual undergoes an infection process whereby its intensity function reduces to 0 after infection. Until infection the intensity function changes its values according to the other individuals with piece-wise constant intensity in between successive events.

We allow $v(\ell_j)$ to be real-valued and by taking the exponential function to map $v(\ell_j)$ to a positive number. Note that $v(\ell_j)$ can be thought of as the susceptibility of individual $j$ and that Equation (2.5) is only positive for susceptible individuals. Consequently, for every individual we can observe at most one event denoting the time of its infection, leading to an intensity function which is constantly 0 afterwards.

When individuals move in space, for example through livestock trade or random movement of humans, the infectious pressure can alter, because the spatial kernel $K$ is defined on their Euclidean distance. However, in our research we do not include any term representing the movement of individuals which is a reasonable assumption when considering notifiable livestock diseases: as soon as such a disease is detected, a farmer is obliged to report it to national authorities leading in a comprehensive animal movement ban. If such a movement ban is only partial, one could extend the model using data on for example the animal trading network. In the case of human diseases, there exist modelling approaches for movement of individuals that are based on the concept of flux models where the infectious contacts of individuals are determined by a mobility kernel, that is, a function representing the probability that two individuals at different locations make contact (Riley et al., 2015).

## 2.3 The likelihood

Given data from an outbreak, we are interested in making inference about the unknown parameters in the model set up defined in the previous section. We only consider completed epidemics: by the end of the data observation all infected individuals have recovered/removed and there are no more infected individuals among the population. In practice, we assume that (for our veterinary example) individuals are removed as soon as they are detected. As soon as an outbreak is observed by noticing respective symptoms, the corresponding animals or farms are being taken out of the system. In contrast, the true times of infection are rarely observed and thus we treat them as missing data for our purposes.

In order to construct the likelihood let $\mathbf{R} = (\mathrm{R}_1, \mathrm{R}_2, \ldots, \mathrm{R}_n) \in \{[0, t_{\text{end}}] \cup \{\infty\}\}^n$ denote the observed removal times and $\mathbf{I} = (\mathrm{I}_1, \mathrm{I}_2, \ldots, \mathrm{I}_n) \in \{(-\infty, t_{\text{end}}] \cup \{\infty\}\}^n$ the infection times of the $n$ individuals. Here, $t_{end}$ states the end of the observation period. If an individual is not observed as a removal by $t_{end}$ it does not get infected and we set its infection and removal time to $\infty$. Note that $i_0 \in \{1, \ldots, n\}$ is the first infective who introduces the disease into the population. Its infection time $\mathrm{I}_{i_0}$ is part of $\mathbf{I}$ and thus, is unknown and can become negative due to the updating step during the MCMC.

For the infection process let us recall the likelihood (Equation (2.4)) and the specification of the intensity function $\lambda_j(t)$ for a susceptible individual $j$ at time $t$ (Equation (2.5)).

The removal process of an infected individual is determined though the length of its infectious period $Q$, which we modelled $Q \sim \text{Gamma}(\delta, \delta\gamma)$.

The likelihood of the infection process and the removal process reads

$$\begin{aligned} f(\mathbf{R}, \mathbf{I} \mid \boldsymbol{\theta}) = \exp\Big( -\sum_j \int_{\min\{\mathbf{I}\}}^{t_{\text{end}}} \lambda_j(t) dt \Big) \prod_{\substack{i:\mathrm{I}_i < \infty \\ i \neq i_0}} \lambda_i(\mathrm{I}_i^-) \\ \times \prod_{i:\mathrm{I}_i < \infty} (\delta\gamma)^{\delta} (\mathrm{R}_i - \mathrm{I}_i)^{\delta - 1} \exp(-\delta\gamma(\mathrm{R}_i - \mathrm{I}_i))/\Gamma(\delta) \end{aligned} \tag{2.6}$$

$\boldsymbol{\theta} = (\alpha, \beta, \psi, \kappa, \gamma, \delta, \mathbf{s}_L)$ is the vector of parameters. The expression $\mathrm{I}_j^-$ denotes the time just right before individual $j$ got infected. $\min\{\mathbf{I}\} = I_{i_0}$ refers to the time point of the first infection and is the start of the outbreak.

Equation (2.5) states the intensity function of the infection process for a susceptible individual. This intensity function is defined in such a way that it only changes after an infection or removal of any of the other individuals occurred. Thus, it is piecewise constant in between events. Further it is to note that the infectious pressure of an infected individual $i$ onto a susceptible individual $j$ is constant over time. Consequently, if we know for every susceptible individual $j$, how much time and the amount of infectious pressure that was exerted from every infected individual on $j$, we are in a position to simplify the integral term in Equation (2.6).

In order to derive such an expression, let $a \wedge b := \min\{a, b\}$. With

$$T_{ij} = (\mathrm{R}_i \wedge \mathrm{I}_j) - (\mathrm{I}_i \wedge \mathrm{I}_j) \tag{2.7}$$

we obtain total duration in which an infected individual $i$ exerted infectious pressure on a susceptible individual $j$. Equation (2.7) becomes clear when considering the three different cases:

- $\mathrm{I}_i < \mathrm{I}_j < \mathrm{R}_i$
  $j$ experiences infectious pressure from $i$ since $I_i$ got infected until its own infection $\mathrm{I}_j$;
  duration of exerted infectious pressure on $j$: $\mathrm{I}_j - \mathrm{I}_i$

- $\mathrm{I}_i < \mathrm{R}_i < \mathrm{I}_j$
  $j$ experiences infectious pressure throughout the whole infectious period of $i$;
  duration of exerted infectious pressure on $j$: $\mathrm{R}_i - \mathrm{I}_i$

- $\mathrm{I}_j < \mathrm{I}_i$
  $j$ became infected before $i$;
  duration of exerted infectious pressure on $j$: $\mathrm{I}_j - \mathrm{I}_j = 0$

For the amount of infectious pressure from all infected individuals $\{i : \mathrm{I}_i < \infty\}$

onto a susceptible individual $j$, we define

$$\Lambda_j = \int \lambda_j(t)dt = \alpha(\mathrm{I}_j - \min\{\mathbf{I}\}) + \sum_{i:\mathrm{I}_i<\infty} T_{ij} \cdot \exp\left(v(\ell_j)\right) K(d_{ij};\kappa) \tag{2.8}$$

With Equations (2.7) and (2.8), we simplify the likelihood from Equation (2.6) for piecewise constant intensity functions to

$$\begin{aligned} f(\mathbf{R},\mathbf{I} \mid \boldsymbol{\theta}) = \exp\left[-\psi \cdot \sum_j \Lambda_j\right] &\prod_{\substack{i:\mathrm{I}_i<\infty \\ i\neq i_0}} \lambda_i(\mathrm{I}_i^-) \\ \times \prod_{i:\mathrm{I}_i<\infty} &(\delta\gamma)^\delta (\mathrm{R}_i - \mathrm{I}_i)^{\delta-1} \exp(-\delta\gamma(\mathrm{R}_i - \mathrm{I}_i))/\Gamma(\delta) \end{aligned}$$

## 2.4 Constructing an MCMC algorithm

We are interested in obtaining samples from the posterior distribution $f(\boldsymbol{\theta} \mid \mathbf{R})$. In order to do so, we follow a Bayesian approach to draw samples from the joint posterior distribution

$$f(\mathbf{I}, \boldsymbol{\theta} \mid \mathbf{R}) \propto f(\mathbf{R} \mid \boldsymbol{\theta}, \mathbf{I}) f(\mathbf{I} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta}) \tag{2.9}$$

Since the true times of infection are unobserved, we treat them as missing data. We would like to marginalize them out in order to draw inference over the parameter in the model. However, marginalization involves intractable computation, because we have to sum over all possible infection times of every individual within our population. This is an infeasible task. Therefore, we take samples from the joint posterior (Equation (2.9)) and discard the infection times to mimic a marginalization.

Setting up an MCMC algorithm to obtain samples from $f(\mathbf{I}, \boldsymbol{\theta} \mid \mathbf{R})$ can be done using a Metropolis within Gibbs algorithm. That is, we update the infection times $\mathbf{I}$ given $\boldsymbol{\theta}$ and $\mathbf{R}$, and the parameters $\boldsymbol{\theta}$ given $\mathbf{I}$ and $\mathbf{R}$ in turns.

By means of the Python library PyMC3, the parameters $\boldsymbol{\theta}$ can be updated using the integrated NUTS (No U-turn) sampler which relies on the Hamiltonian Monte Carlo algorithm. With Hamiltonian Monte Carlo methods we can update a set a parameters simultaneously and can speed up the computation time (see Section 1.3.2). Compared to a random walk Metropolis update, Hamiltonian Monte Carlo circumvents the slow exploration of the state space and is especially suited for high-dimensional problems.

For updating the infection times, one has to think about how to do this in an efficient manner. Jewell et al. (2009) updated them one at a time sequentially, whereas Neal and Roberts (2005) updated them one at a time randomly chosen per MCMC iteration. The conditional distribution of the infection times given the parameters is too complicated to be able to sample directly from it. Taking into account that the infectious period is Gamma$(\alpha, \beta)$ distributed, Neal and Roberts (2005) used an independence sampler by sampling $Q^*$ from Gamma$(\alpha, \beta)$ and proposing a new infection time through $\mathrm{I}_j^* = \mathrm{R}_i - Q^*$.

For our research, it was found to be more effective to use a Random Walk Metropolis (RWM) update step rather than an independence sampler proposal. For every infected individual $i$ with $\mathrm{R}_i < \infty$ in our dataset and with current infection time $\mathrm{I}_i$, the new value $\mathrm{I}_i^*$ is proposed by sampling from a truncated Normal distribution $\mathcal{N}(\mathrm{I}_i, \sigma^2)|_{(-\infty, \mathrm{R}_i)}$. The truncation is necessary because the infection time of an individual cannot lie past its removal time. As mentioned earlier in Section 1.3.1, the variance $\sigma^2$ should be chosen in such a way that about a quarter of the proposed samples are accepted. We follow the approach suggested by Xiang and Neal (2014) and adapt $\sigma^2$ according to whether a proposed move is accepted or not. To be precise, let $J$ be the iteration number. Then after a proposed infection time $\mathrm{I}_i^*$ we update $\sigma^2$ according to

$$\sigma^2 = \begin{cases} \sigma^2 + 3\frac{\sigma^2}{100\sqrt{J}} & \mathrm{I}_i^* \text{ was accepted} \\ \sigma^2 - \frac{\sigma^2}{100\sqrt{J}} & \mathrm{I}_i^* \text{ was rejected} \end{cases}$$

Shifting the infection times in such a manner is not built in into the PyMC3 framework intrinsically. Therefore, we extended the features of PyMC3 with our own implementation of an adaptive Random Walk Metropolis (-Hastings, due to the truncation of the normal distribution) sampler for shifting the infection times. The programming code is available at `https://fhm-chicas-code.lancs.ac.uk/koeppell/py-mc-3-epi-extension`.

The pseudocode in Algorithm 4 below summarises the MCMC algorithm to obtain samples from $f(\mathbf{I}, \boldsymbol{\theta} \mid \mathbf{R})$. One way to initialize the infection times is to include prior belief about the length of the infectious period. Here, we assumed that it lasts $d$ days on average.

**Algorithm 4** Sampling from $f(\mathbf{I}, \boldsymbol{\theta} \mid \mathbf{R})$ using Hybrid NUTS-Metropolis within Gibbs

```
Input: $s$ number of samples, $n_{\mathrm{I}}$ number of total infections in the outbreak, $d$
  prior belief about length of infectious period, $\mathbf{1}_{n\times 1}$ an $n \times 1$ vector of all ones
Initialize: Set $\mathbf{I}^{(0)} \leftarrow (\mathbf{R} - d\mathbf{1}_{n\times 1})$
for $k = 1, \ldots, s$ do
    Obtain a sample $\boldsymbol{\theta}^{(k+1)}$ from $(\boldsymbol{\theta} \mid \mathbf{R}, \mathbf{I}^{(k)})$ using the NUTS sampler
    for i do $= 1, \ldots, n_{\mathrm{I}}$:
        Obtain a sample $\mathrm{I}_i^{(k+1)}$ from $(\mathrm{I}_i \mid \mathbf{R}, \mathbf{I}_{-i}^{(k)}, \boldsymbol{\theta}^{(k+1)})$
        using adaptive Metropolis-Hastings
    end for
    $k \leftarrow k + 1$
end for
```

## 2.5 Parameter reduction to increase MCMC efficiency

Several experiments revealed that poor mixing of the infection times leads to poor mixing of the other parameters. Thus, it is important to investigate MCMC methods with greater focus on updating the infection times than updating the parameters $\theta$. One approach is to analytically integrate out parameters if possible. This can improve the mixing of the MCMC algorithm because we diminish the dimension of the sampling target space. This approach was applied in Neal and Roberts (2005), as well as in Kypraios (2007) accompanied with extensive simulation studies, and subsequently used in Xiang and Neal (2014).

To integrate out parameters, we make use of conjugate priors. We know from Bayes' theorem that for a parameter $\theta$ and data $\boldsymbol{y}$,

$$f(\theta)f(\boldsymbol{y} \mid \theta) = f(\boldsymbol{y})f(\theta \mid \boldsymbol{y}).$$

A prior $f(\theta)$ is called *conjugate* for the likelihood $f(\boldsymbol{y} \mid \theta)$ if it belongs to the same family of distributions as the posterior distribution $f(\theta \mid \boldsymbol{y})$. Our aim is to integrate out $\theta$, that is $\int f(\boldsymbol{y} \mid \theta)f(\theta)d\theta = f(\boldsymbol{y})$ and note that this yields the marginal likelihood. Thus, if we know the functional form of our posterior from the conjugacy, integration can be easily performed. First, we rearrange the joint distribution into the functional parts of the parameter of interest and the rest. Since the normalization constant of the posterior is uniquely defined, we can expand the expression by multiplying and dividing by it, which yields the posterior distribution and the marginal likelihood, our desired integral.

Recall from Equation (2.5), the intensity function for a susceptible individual $j$ at time $t$ is $\lambda_j(t) = \psi\big(\alpha + \exp(v(\ell_j)) \sum_{i \in \mathcal{I}(t)} \exp(-d_{ij}\kappa)\big)$. It becomes apparent that the two parameters $\kappa$ and $\psi$ have an inherent colinearity: as soon as $\kappa$ increases, the spatial kernel decreases and in order to account for the lower infectious pressure $\psi$ has to be increased as well. In this setting, the parameter $\kappa$ already incorporates the information of $\psi$ due to their relationship and having information about $\kappa$ is informative about $\psi$. Thus, $\psi$ constitutes a perfect candidate for being integrated out. We assume that we have independent Gamma priors for each of the parameters with Gamma$(n_v, \delta_v)$ denoting the prior for parameter $v$ and $v = \kappa, \psi, \gamma$. The joint distribution ordered for all terms of $\psi$ equals

$$f(\mathbf{R}, \mathbf{I}, \boldsymbol{\theta}) \propto \psi^{n_{\mathrm{I}}-1} \times \exp\left(-\psi \cdot \sum_j \sum_{i:\mathrm{I}_i<\infty} T_{ij}\left(\alpha + e^{v_j - \kappa d_{ij}}\right)\right) \times \psi^{n_\psi - 1}\exp(-\delta_\psi \psi) \\ \times C$$

with

$$C = \prod_{\substack{i:\mathrm{I}_i<\infty \\ i \neq i_0}} \left(\alpha + \sum_{i \in \mathcal{I}(t_i^-)} e^{v_i - \kappa d_{ij}}\right) \times \prod_{i:\mathrm{I}_i<\infty} (\delta\gamma)^\delta (\mathrm{R}_i - \mathrm{I}_i)^{\delta-1} \exp(-\delta\gamma(\mathrm{R}_i - \mathrm{I}_i))/\Gamma(\delta) \\ \times \exp(-\mathbf{s}_L^T \Sigma^{-1} \mathbf{s}_L/2) \times \kappa^{n_\kappa - 1}\exp(-\delta_\kappa \kappa) \times \gamma^{n_\gamma - 1}\exp(-\delta_\gamma \gamma)$$

Rearranging yields

$$f(\mathbf{R}, \mathbf{I}, \boldsymbol{\theta}) \propto \psi^{(n_{\mathrm{I}}+n_\psi-1)-1} \times \exp\left[-\psi \cdot \left(\delta_\psi + \sum_j \sum_{i:\mathrm{I}_i<\infty} T_{ij}\left(\alpha + e^{v_j - \kappa d_{ij}}\right)\right)\right] \times C \tag{2.10}$$

Looking at the functional form of Equation (2.10), we notice that it equals the kernel of a Gamma distribution with shape $\delta_\psi + \sum_j \sum_{i:\mathrm{I}_i<\infty} T_{ij}\left(\alpha + e^{v_j - \kappa d_{ij}}\right)$ and rate $n_{\mathrm{I}}+n_\psi-1$. Therewith we conclude that the desired marginal likelihood reads

$$f(\mathbf{R},\mathbf{I},\boldsymbol{\theta}_{\setminus\{\boldsymbol{\psi}\}}) \propto \frac{C\cdot\Gamma(n_{\mathrm{I}}+n_\psi-1)}{\left[\delta_\psi+\sum_j\sum_{i:\mathrm{I}_i<\infty}T_{ij}\left(\alpha+e^{v_j-\kappa d_{ij}}\right)\right]^{n_{\mathrm{I}}+n_\psi-1}}$$

whereby $\boldsymbol{\theta}_{\setminus\{\boldsymbol{\psi}\}}$ denotes the set of parameters $\boldsymbol{\theta}$ without $\psi$.

#### 2.5.0.1 Summary

In summary, our model outline comprises Gamma distributions for the infectious periods and a single conditional Poisson process for the infection of every individual in our population. A piecewise constant intensity function allows us to compute the likelihood in simple arithmetic terms. Based on a Bayesian approach, we construct an MCMC algorithm to draw samples from the joint conditional distribution of the infection times and the parameters, given the data of the removal times. For updating the parameters we choose the NUTS sampler. Updating the infection times is done in a sequential manner with our own implementation of an adaptive Metropolis-Hastings algorithm using a truncated Gaussian proposal. In order to facilitate MCMC mixing we integrate out the infection rate parameter $\psi$ from the joint distribution of the parameters and the data utilizing the prior conjugacy property. This yields our desired target distribution up to a proportionality constant

$$\begin{aligned} f(\mathbf{I},\boldsymbol{\theta}_{\setminus\{\boldsymbol{\psi}\}}\mid\mathbf{R}) \propto & \left[\frac{\Gamma(n_{\mathrm{I}}+n_\psi-1)\times\prod\limits_{\substack{i:\mathrm{I}_i<\infty\\ i\neq i_0}}\left(\alpha+\sum\limits_{i\in\mathcal{I}(t_i^-)}e^{v_i-\kappa d_{ij}}\right)}{\left[\delta_\psi+\sum\limits_j\sum\limits_{i:\mathrm{I}_i<\infty}T_{ij}\left(\alpha+e^{v_j-\kappa d_{ij}}\right)\right]^{n_{\mathrm{I}}+n_\psi-1}}\right] \\ & \times\prod_{i:\mathrm{I}_i<\infty}(\delta\gamma)^\delta(\mathrm{R}_i-\mathrm{I}_i)^{\delta-1}\exp(-\delta\gamma(\mathrm{R}_i-\mathrm{I}_i))/\Gamma(\delta) \\ & \times\exp(-s_L^T\Sigma^{-1}s_L/2)\times\kappa^{n_\kappa-1}\exp(-\delta_\kappa\kappa)\times\gamma^{n_\gamma-1}\exp(-\delta_\gamma\gamma) \end{aligned} \qquad (2.11)$$

## 2.6 Simulation studies outline

This thesis is accompanied with simulation studies to evaluate how well the parameters for transmission, the environmental covariates and the spatial random effect, that is the Gaussian risk surface, can be estimated from simulated outbreak data. By means of our inference algorithm we recover the spatial risk surface at a finite number of locations of the individuals in our population. This enables us to interpolate the spatial random effect for the remaining area. Such a prediction, also referred to as *Kriging* (e.g. Diggle and Giorgi (2019)), is common in geospatial statistics within the generalized linear model framework. Its use in this epidemiological context slightly differs from the generalized linear model framework: instead of linking the linear predictor directly to the expectation of our model distribution, we incorporate the effects as part of the intensity function within the infection process.

In this section we describe how we simulate epidemics and explain the basic set up of the simulations including the interpolation of the spatial random effect for the whole area.

### 2.6.1 Epidemic outbreak simulation

Recall that an outbreak involves a sudden increase of incident cases of the normally expected levels in that area (Chapter 1.1). We will refer to this in the remainder of this thesis, in particular in Chapter 4 and 5, where we observe a jump in the number of infected individuals instead of following a fast rising but smoother curve.

Throughout the simulation study, we used a Doob-Gillespie algorithm (Algorithm 5) to simulate an epidemic outbreak in continuous time. An outbreak comprises the development of random times which either indicate an infection or act as a countdown towards removal. The algorithm alternates between two steps: first, it draws an exponentially distributed time based on the cumulative rates of all competing random times. Second, it normalizes the involved rates to obtain probabilities which are then utilized to decide which event (infection or (partial) removal) has occurred.

Let us recall that we model the infection process by means of a Poisson process

**Algorithm 5** Doob-Gillespie algorithm

```
Input: $n$ population size; $t$ current time; $\lambda_j$ infectious pressure on individual $j$; $4\gamma$ removal rate; $Z$ array of infection/removal state with length $n$; Update_infection_rates() function to update $\lambda_j$ for all susceptible individuals $j$
Output: $S, I, R$ number of susceptibles, infected and removed individuals; T time points
Initialize: Set $S = n - 1, I = 1, R = 0, t = 0$, $Z = \text{zeros}(n)$
Obtain a random integer $k$ from $U_{\{1,\dots,n\}}$                ▷ first infective
$Z[k] = 1$
while I > 0 do
    Update_infection_rates()
    $C = 4\gamma I + \sum_{j:Z[j]=0} \lambda_j$
    Obtain a sample $t^*$ from $Exp(C)$                   ▷ obtain new event time
    Update $t \leftarrow t + t^*$
    Obtain a sample $u$ from $U_{[0,1]}$              ▷ decide which event happened
    if $u < \frac{4\gamma I}{C}$ then                                   ▷ removal
        $Sum = 0$
        for $i$ in $\{i$ with $Z[i] > 0\}$ do
            $Sum = Sum + 4\gamma$
            if $u < \frac{Sum}{C}$ then
                if $Z[i] < 4$ then                           ▷ partial removal
                    $Z[i] = Z[i] + 1$
                else                                            ▷ full removal
                    $Z[i] = -1$
                    Update $I \leftarrow I - 1$
                    Update $R \leftarrow R + 1$
                    T.append($t$)
                end if
                break for loop
            end if
        end for
    else                                                           ▷ infection
        $Sum = 4\gamma I$
        for $i$ in $\{i$ with $Z[i] == 0\}$ do
            $Sum = Sum + \lambda_i$
            if $u < \frac{\text{Sum}}{C}$ then
                $Z[i] = 1$
                Update $S \leftarrow S - 1$
                Update $I \leftarrow I + 1$
                T.append($t$)
                break for loop
            end if
        end for
    end if
    Store.append($(S, I, R)$)
end while
```

with piecewise constant intensity functions between events (Equation (2.5)). The infectious period follows a Gamma distribution with $Q \sim \text{Gamma}(4, 4\gamma)$ which equals the sum of four independent exponential distributions each with rate $4\gamma$. We can make use of this fact by introducing artificial stages to the infectious period and keeping record of which state an individual is in before it gets removed. In other words, when infected, an individual enters stage 1 and for $k = 1, \ldots, 4$, it spends an exponential time in stage $k$ (with rate $4\gamma$) before moving on to the next stage, or in case of $k = 4$, before recovering from the disease (getting removed). The exit time from stage 4 is the removal time and there is no difference in infectivity between these different stages.

As displayed in Algorithm 5, we draw the time until the next event (including artificial partial removals) from an exponential distribution with cumulative rate $C = 4\gamma I + \sum_{j:Z[j]=0} \lambda_j$. Here, $\lambda_j$ is the sum of the infectious pressure on individual $j$ from all infected individuals in the population and the entry $Z[j]$ denotes the current infection/removal state of an individual $j$, that is $Z[j] = 0$: $j$ is susceptible, $Z[j] = k > 0$: $j$ is in its artificial partial removal state $k$ and infectious, and $Z[j] = -1$: $j$ is removed. It then converts rates into probabilities by taking the proportion of the individual rates with respect to the cumulative rate $C$. By means of a sample from a standard uniform distribution $u$, we decide which event happened. In case of a partial removal we update the artificial status of the chosen individual, in case of an infection we update the time and numbers in each class. In every iteration the infectious pressure on susceptibles is updated by means of the function Update_infection_rates(). The infectious pressure follows the intensity function of Equation (2.5). It is constant in between events and only alters with the occurrence of a removal (less infectious pressure) or an infection (additional infectious pressure). Thus, it suffices to update the infectious pressure on each susceptible individual $j$ at the beginning of each iteration. Finally, this algorithm runs until no infected individual is left.

### 2.6.2 Basic setup

We performed simulation studies with different simulated disease outbreaks. For the initialization of each outbreak the first infected individual was chosen uniformly and its infection time marked the start of the observation period, that is $t = 0$. The rest of the population was assumed to be entirely susceptible. The parameters for the simulations in this thesis were chosen specific to each individual outbreak and are provided accordingly. The epidemic was simulated using a Doob-Gillespie algorithm (Algorithm 5), where the infectious process, that is transitioning from the susceptible to infected state, was governed by the intensity function described in Equation (2.5).

The data we obtained from a simulation are the locations $\boldsymbol{L} = (\ell_1, \ldots, \ell_n)$ and a vector $\mathbf{R} = (\mathrm{R}_1, \ldots, \mathrm{R}_n) \in \{[0, t_{\text{end}}]\}^n$ of the removal times of the $n$ individuals in the population, where the observation period is $[0, t_{\text{end}}]$. A removal time is set to be $t_{\text{end}}$ if an individual escaped infection (and was therefore never removed). In practice the true times of infection $\mathbf{I} \in \{(-\infty, t_{\text{end}}]\}^n$ are not observed and hence treated as missing data for our purposes. Since this quantity is unknown, the inference algorithm allows an infection to have happened before the beginning of the observation period.

We follow a Bayesian approach to draw samples from the joint posterior distribution $f(\mathbf{I}, \boldsymbol{\theta}_{\setminus\{\psi\}} \mid \mathbf{R})$ (Equation (2.11)) with a Hybrid NUTS/Metropolis within Gibbs Sampler as specified in Section 2.4. We state the different prior choice with each example and performed sensitivity analyses to investigate the robustness of the prior. The length of the burn-in phase was set according to the inspection of the MCMC outputs.

Due to reasons of display, we show results from simulation studies that are based on a single dataset (produced by single set of parameters). We want to emphasize that we performed further simulations with differing outbreaks and differing parameters to be able to make generalized statements on the inference setting.

#### 2.6.2.1 Risk surface prediction

With samples from the posterior distribution of $\mathbf{s}_L$ for the observed locations $\boldsymbol{L}$, conditioned on the observed removal times $\mathbf{R}$ we are able to predict the value of the Gaussian process for any new unobserved location in the grid of our simulation study. As a result, we can create infographic maps that inform about areas of high susceptibility.

Let $\mathbf{s}_{\boldsymbol{L}} = (\mathbf{s}(\ell_1), \ldots, \mathbf{s}(\ell_k))$ be a realization of the Gaussian process at $k$ training locations $\boldsymbol{L} = (\ell_1, \ldots \ell_k)$. For a sequence of unobserved locations $\boldsymbol{L_*} = (\ell_{k+1}, \ell_{k+2}, \ldots \ell_{k+p})$, we are now interested in predicting the distribution of $\mathbf{s}_{\boldsymbol{L_*}} = (\mathbf{s}(\ell_{k+1}), \ldots, \mathbf{s}(\ell_{k+p}))$ conditioned on $\mathbf{s}_L$. From Section 1.4 we know that the conditional distribution is of the form

$$(\mathbf{s}_{\boldsymbol{L_*}} \mid \mathbf{s}_{\boldsymbol{L}}) \sim \mathcal{N}(\ \Sigma_{L_*L}\Sigma_{LL}^{-1}\mathbf{s}_L\ ,\ \Sigma_{L_*L_*} - \Sigma_{L_*L}\Sigma_{LL}^{-1}\Sigma_{LL_*}) \tag{2.12}$$

We perform the simulation study with $n$ different locations. Assume further we have $m$ samples

$$\boldsymbol{s_L} = \begin{bmatrix} s_1^{(1)}, \ldots, s_n^{(1)} \\ \vdots \\ s_1^{(m)}, \ldots, s_n^{(m)} \end{bmatrix}$$

from $(\mathbf{s}_{\boldsymbol{L}} \mid \mathbf{R})$ for each of the sites. Utilizing the i-th sample $\boldsymbol{s^{(i)}} = (s_1^{(i)}, \ldots, s_n^{(i)})$ as training data, we can calculate the predictive distribution of $(\mathbf{s}_{\boldsymbol{L_*}} \mid \mathbf{s}_{\boldsymbol{L}} = \boldsymbol{s^{(i)}})$ with Equation (2.12) and hence obtain the predicted mean of the Gaussian process at the new locations $\boldsymbol{L_*}$

$$\boldsymbol{\hat{\mu}_{L_*}^{(i)}} = \Sigma_{L_*L}\Sigma_{LL}^{-1}\boldsymbol{s^{(i)}}$$

This collection of mean values can be used to create infographic maps highlighting areas of high spatial transmission risk. This can aid policy makers in their decision process for appropriate counter measures fighting a disease outbreak. One way of getting suitable point estimates at each location, is to average the predicted means $\boldsymbol{\hat{\mu}_L} = \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\hat{\mu}_{L_*}^{(i)}}$. However, because the marginal distributions at each location are Gaussian distributions, the samples can take values from the whole of the real axis, in particular negative values. For informing policy makers on previously unknown high risk areas we would like to talk in the notion of probability. Therefore, a different way to plot such a risk map is to consider p-values. We are particularly interested in the p-values of $\hat{\mu}$ with respect to the prior mean in our model. Hence, to get information on how likely it is to have values of 0 and greater we take into account how spread out the distribution is for a single location. We define

$$\hat{\mu}^*_{\ell_*} := \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\{\hat{\mu}^{(i)}_{\ell_*} > 0\}}$$

as the exceedance probability for every location $\ell_*$, where $\mathbb{1}$ serves as the indicator function. In other words, for each location, it gives us the relative frequency that the predicted mean is greater than the average of the spatial random effect, which is 0.

Risk maps with respect to $\hat{\mu}$ and $\hat{\mu}^*$ look very similar apart from the scale of the colorbar, thus we will only depict the prediction plots with respect to $\hat{\mu}^*$ for better interpretability in this thesis.

### 2.6.3 Model diagnostics and comparators

For comparisons of the sampling efficiency and performance we need diagnostic tools and make use of the autocorrelation function and the effective sample size. A more detailed coverage can be found for example in Rubin (2013).

For samples from independent identically distributed random variables the central limit theorem provides information about the uncertainty in the model. Central limit theorems exist for Markov processes but with larger variance (due to the positive correlation). The effective sample size (ESS) allows us to compare the variance for a dependent sample with an independent sample. If the MCMC chain has variance $V$ and independent samples have variance $S$, we calculate ESS $= n \cdot S/V$, where n is the sample size from the MCMC.

In the remainder of this thesis will use the effective sample size and the autocorrelation function for comparison of inference outputs.

## 2.7 Simulation study: Investigating non-identifiabilities

In this section we explore the performance of our joint model. We analyse our model's behaviour and investigate non-identifiabilities of the parameters.

To do so, we first analyse risk surfaces for the random effect matching our simulation specification, that is Matern(3/2) risk surfaces, and investigate their recovery. We then explore how well our algorithm is performing with respect to misspecification of the covariance function by choosing surfaces that do not match the inital model description. Thereafter we include solely environmental information and finally consider the full model specification by incorporating a spatial random effect and fixed effects.

The basic setup was identical in all simulations in this section and comprised a population of size $n = 100$ with locations $\boldsymbol{L} = (\ell_1, \dots, \ell_{100})$. The locations were sampled independently on a 100 x 100 square area in a uniform manner with the exception of example 2.1 d).

In all simulations we applied the infectious pressure with respect to Equation (2.5) and fixed the basic infectious pressure $\alpha$ to $10^{-6}$. The length of the infectious period was modelled as $\gamma = 1/7$ throughout. In other words, on average individuals were infectious for 7 days. The parameters $\psi$ and $\kappa$ were chosen specific to each individual outbreak and marked accordingly in Tables 2.1, 2.4 and 2.6.

If not stated otherwise, we chose Gamma(0.1, 0.1) priors for the parameters $\kappa, \psi$ and $\gamma$. We remind the reader that we analytically integrated out the parameter $\psi$ (see Equation (2.11)) For the estimation of the spatial effect, we fixed the values for the hyperparameters of the Matern(3/2) covariance function either to the true value or state the priors used. For each example we drew 50,000 samples together with an additional 10% for the burn-in (5000 samples). Inspection of the MCMC outputs suggested this was enough.

For the purpose of model investigation, we selected outbreak examples where about 60% of the population was infected. This provided enough information about the outbreak dynamics, that is a reasonable number of infected individuals, to be able to infer parameter estimates. On the other hand, disease-free areas remained to investigate the impact of the risk surface as the driver of infection.

As there is a lot of randomness in outbreak simulations, we also investigated the behaviour of outbreaks scenarios with the same infection parameters, varying spatial location of the first infected individual, and different random risk surfaces. Similar results were observed and the results proved to be robust with respect to the above mentioned randomness. Thus, we restrict ourselves to the example selection in this thesis to illustrate features of the model and sampler behaviour.

### 2.7.1 Simulating from the model

We now turn to the investigation of how well we can recover the risk surface when we simulate the data from the model directly. As specified in Section 2.2, we model the spatial random effect using a Gaussian process with a Matern(3/2) covariance function. Recall that the Matern(3/2) function depends on the distance between points, $\mathbf{c}(x_i, x_j) = \sigma^2\big(1 + \frac{\sqrt{3}d_{ij}}{\rho}\big)\exp\big(-\frac{\sqrt{3}d_{ij}}{\rho}\big)$. It is determined by the marginal variance parameter $\sigma^2$ and the lengthscale $\rho$.

We simulated three different Matern(3/2) risk surfaces with constant $\sigma^2 = 3$ and varying lengthscale $\rho = 20, 40, 100$ on a $100 \times 100$ grid (Table 2.1). We did not alter $\sigma^2$ to make the simulations comparable as a higher $\sigma^2$ would lead to higher and lower values in the risk surface which can resemble a risk surface with a lower lengthscale.

In this section we only included a random spatial effect for simplicity, that is $v(\ell_j) = \mathbf{s}(\ell_j)$ for all $j = 1, \ldots, 100$. The more complex case, when additionally adding fixed effects is covered in Section 2.7.3. Recall that the population size is 100. Therefore, we have 100 points of the risk surface at the locations $\boldsymbol{L} = (\ell_1, \ldots, \ell_{100})$ of our individuals and $100^2$ points from the grid. Thus, for the inference we sample from the $100 + 100^2$ dimensional normal distribution with mean vector equal to 0 and covariance matrix $\Sigma$, whereby the entries of $\Sigma$ follow the above mentioned Matern(3/2) covariance function using the Euclidean metric.

Example d) in Table 2.1 illustrates an outbreak scenario with a spatially non-uniform distribution of points. The locations were sampled around the margins to mimic high and low point density regions determined by environmental factors, for example valleys and mountains.

In the following, we first investigate the mixing behaviour of the model and then consider the accuracy of the posterior parameter estimates. Finally, we analyze the recovery of the risk surface and conclude with a remark.

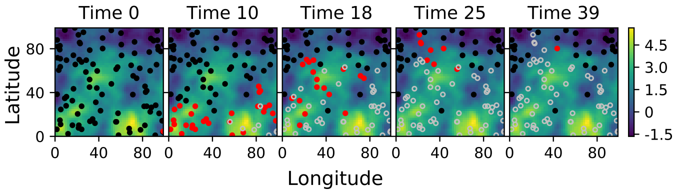
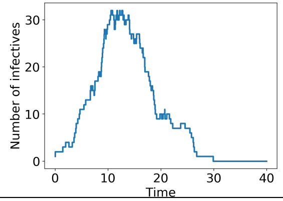
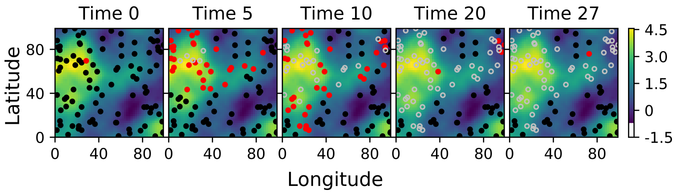
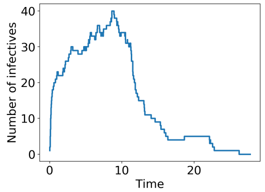
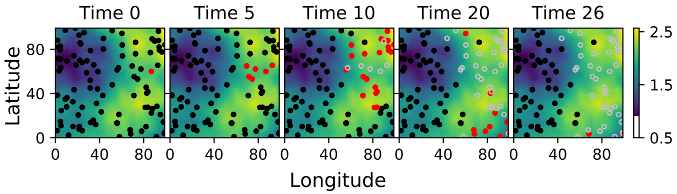
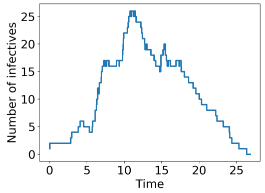
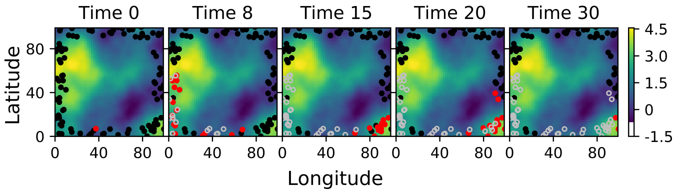
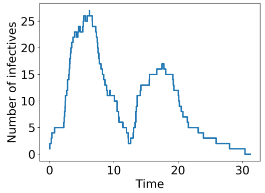
| | Characteristics | Spatial outbreak progression | Number of infectives | Outbreak size | Parameters |
|---|---|---|---|---|---|
| a) | Matern(3/2)<br>$\sigma^2 = 3$<br>$\rho = 20$ |  |  | 64 | $\kappa = 0.3$<br>$\psi = 0.3$ |
| b) | Matern(3/2)<br>$\sigma^2 = 3$<br>$\rho = 40$ |  |  | 62 | $\kappa = 0.3$<br>$\psi = 0.4$ |
| c) | Matern(3/2)<br>$\sigma^2 = 3$<br>$\rho = 100$ |  |  | 54 | $\kappa = 0.35$<br>$\psi = 0.7$ |
| d) | Example b) with varied population locations |  |  | 48 | $\kappa = 0.3$<br>$\psi = 0.4$ |

Table 2.1: Overview of different simulated epidemic outbreaks a) - d). Black and red points represent susceptible and infected individuals respectively, grey circles denote removed individuals.

#### 2.7.1.1 Mixing behaviour

In the following we discuss the features of the mixing behaviour of the joint modelling approach. We exemplify the characteristics of the inference by means of the outbreak simulations from Table 2.1 as reference. Different starting points for the MCMC chain did not alter the results and the mixing behaviour obtained from other outbreak simulations looked similar. Thus we refrain from displaying further figures with similar characteristics from different simulations for the discussion of the model behaviour. Table 2.2 gives an overview over the different inference scenarios.

For simplicity reasons we first fixed the parameters of the covariance functions to their true values when making inference. We were particularly interested in the posterior mean infectious time as a key summary statistic for the infection times I. We chose Gamma(0.1, 0.1) priors for the spatial kernel parameter and for the removal rate parameter. Although they have a lot of mass on 0, they are relatively flat for values greater than zero. The impact with respect to a more informative priors is discussed further on.

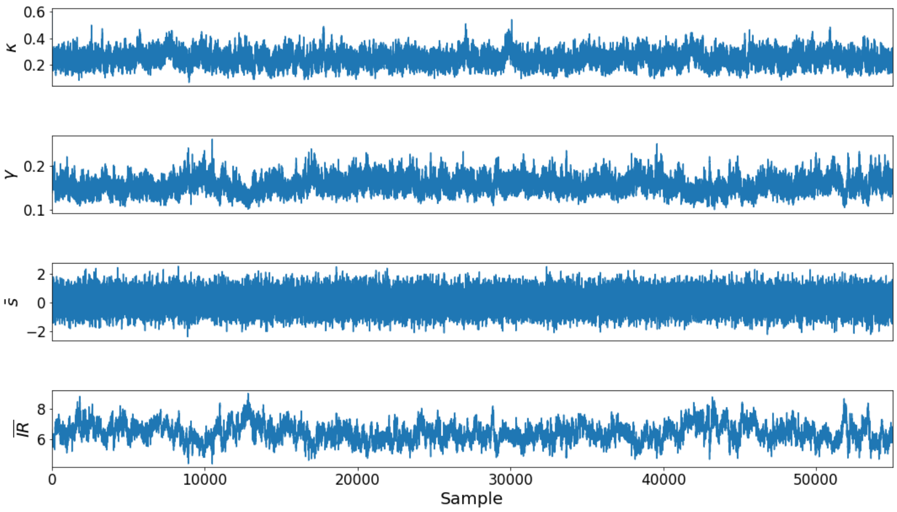


Figure 2.3: MCMC traceplot of posterior samples with respect to the outbreak simulations of Table 2.1. We fixed the parameters of the Matern(3/2) covariance function to their true values. The burn-in phase of the first 5000 samples is included in the plot. Parameters shown are the infection and removal process parameters $\kappa$ and $\gamma$, the mean of the spatial random effect $\bar{\mathrm{s}}$ and the mean infectious period $\overline{\mathrm{IR}}$.

| | $\kappa$ | $\gamma$ | $\overline{\text{IR}}$ | $\sigma^2$ | $\rho$ | $\bar{\text{s}}$ |
|---|---|---|---|---|---|---|
| **Outbreak a)** | | | | | | |
| Fixed I, $\rho, \sigma^2$ | 16,891 | 25,375 | - | - | - | 46,464 |
| Fixed $\sigma^2, \rho$ | 401 | 168 | 110 | - | - | 16,045 |
| Fixed $\mathbf{s}$ | 719 | 99 | 60 | - | - | - |
| Fixed $\rho$<br>Set $\sigma^2 = 1$ | 1,162 | 354 | 237 | - | - | 57,614 |
| Fixed none<br>$\sigma^2 \sim$ Uniform$(0, 10)$<br>$\rho \sim$ Gamma$(200, 10)$ | 181 | 109 | 63 | 232 | 27,088 | 8,305 |
| Set $\mathbf{s} = \mathbf{0}$ | 2359 | 793 | 688 | - | - | - |
| **Outbreak b)** | | | | | | |
| Fixed $\rho, \sigma^2$ | 310 | 137 | 127 | - | - | 20,878 |
| Fixed $\sigma^2, \rho$<br>$\gamma \sim$ Gamma$(14, 100)$ | 404 | 241 | 199 | - | - | 26,566 |
| **Outbreak c)** | | | | | | |
| Fixed $\rho, \sigma^2$ | 133 | 125 | 66 | - | - | 841 |
| Fixed none<br>$\sigma^2 \sim$ Uniform(0,10)<br>$\rho \sim$ Gamma(1000, 10) | 169 | 156 | 74 | 1,493 | 99,745 | 594 |
| Fixed none<br>$\sigma^2 \sim$ Uniform(0,10)<br>$\rho \sim$ Gamma(200, 10) | 167 | 179 | 96 | 252 | 15,722 | 761 |
| Fixed none<br>Set $\sigma^2 = 1$<br>$\rho \sim$ Gamma(200, 10) | 228 | 179 | 131 | - | 49,742 | 3,642 |
| **Outbreak d)** | | | | | | |
| Fixed $\rho, \sigma^2$ | 602 | 632 | 483 | - | - | 72,046 |

Table 2.2: Effective sample size (ESS) of different MCMC chains from outbreak simulations from Table 2.1. If not otherwise specified, we chose Gamma(0.1, 0.1) priors. If fixed, the parameter was set to its true value and denoted by - in the table.

Figure 2.3 displays the MCMC traceplot of the infection and removal process parameters, the mean of the spatial random effects and the mean infectious period for outbreak a) in Table 2.1. We noticed that the mixing of the MCMC algorithm was reasonable with the mean spatial random effect performing best and the mixing of the mean infectious period performing worst. This behaviour is also reflected in the autocorrelation of the MCMC chains with the slowest decay for $\overline{\mathrm{IR}}$ and the fasted decay for $\bar{\mathrm{s}}$. Moreover, the effective sample size (ESS) of the chains confirmed the above observation with $\kappa : 401, \gamma : 186, \bar{\mathrm{s}} : 16,045$ and $\overline{\mathrm{IR}} = 110$.

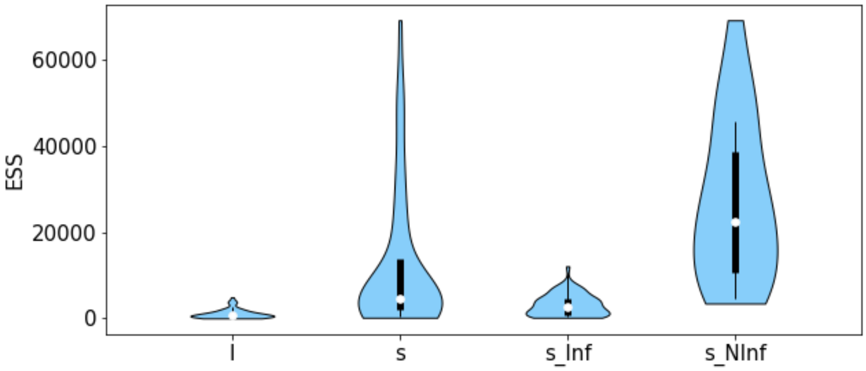


Figure 2.4: Violin plots of the effective sample sizes (ESS) of the set of infection times I and the spatial random effects **s** with respect to the data of the outbreak a) (Table 2.1). s_Inf and s_NInf denote the subsets of random effects at the locations of the infected and non-infected individuals, respectively. In the inference the hyperparameters of the Gaussian process were fixed to their true value.

The mean infectious period and the mean spatial random effect are summary statistics which is why we investigated the individual infection time and random effect mixing behaviour in more detail through violin plots of their effective sample sizes in Figure 2.4. These plots confirmed that in general the mixing of the infection times was very slow and that there was a great variability in particular in the mixing of the spatial effect. It became clear that for the non-infected individuals the ESS of the spatial random effect was on average much higher than for the infected individuals. This suggested that the shifting of the infection times plays a major role in the behaviour of the MCMC mixing. In fact, when fixing the infection times to their true value, we obtained a smoother mixing and comparable high values for the ESS of the other parameters $\kappa : 16,891, \gamma : 25,375, \bar{\mathrm{s}} : 46,464$

in contrast to not fixing the infection times ($\kappa : 401, \gamma : 186, \bar{\text{s}} : 16,045$). Here, the dimension of the sample space has been reduced massively, that is almost halved, which let the sampler traverse more easily through the state space. When fixing the random effects and thus, decreasing the dimension of the sample space by the population size, and not only the number of the infected individuals, the mixing was still a lot slower ($\kappa : 719, \gamma : 99, \overline{\text{IR}} : 60$) than when fixing the infection times. This again indicated that the shifting of the infection times poses a hardship for the sampler.

Such a slow mixing behaviour when shifting the infection times has also been observed before in the literature, see for example Neal and Roberts (2005). An overview on efficiently updating the infection times in SIR-type of models and quantitatively assessment of the fit of proposed models is also provided in Aristotelous (2020). In the course of the simulation study we investigated different updating schemes for shifting the infection times. The best performing approach, whose results are displayed in the simulation studies, was a random walk update with varying variance (see Section 2) but for the purpose of completeness we also want to mention other attempts:

- We used a Metropolis sampler for all parameters instead of a the hybrid version (NUTS and Metropolis). There is a tendency that using the same sampler worsens mixing behaviour. This can largely explained by the fact that the NUTS sampler is optimized for high dimensional problems.

- A non-centered parametrization approach: In a Bayesian hierachical model, we term the parametrization for the data Y and parameters Θ *centered*, if the missing data lie (are centered) between the observations Y and the parameters Θ. If we can reparametrize the model in such a way that the centered node is a deterministic function of the parameters and some unknown data, we term this non-centered approach. In the literature it was shown that a different parametrization might boost MCMC efficiency (Papaspiliopoulos and O Roberts, 2003), however for our model we did not observe any discernible difference.

- Sequential updates, that is updating all infection times sequentially, or block updates, that is updating a certain percentage of the infection times before they are subject to the accept-reject step, did not improve sampling performance

- Estimating the variance of the random walk update during the burn-in phase and fixing it, improved computation time slightly but not the mixing behaviour.

- Choosing an exponentially distributed infectious period (vs. the sum of several identically exponentially distributed infection times) increased the variance of the distribution and added additional uncertainty into the estimator.

- Running the chain for longer and greater thinning can reduce the autocorrelation between the samples. However, this becomes a matter of time feasibility and availability of the computational resources, which are often limited.

For sensitivity analysis, we investigated the mixing behaviour when a priori information on the parameters estimates were translated into the model by informative priors. We inferred that in general informative priors improve mixing of the MCMC chain, but they have only smaller effects. For example, the ESS for the parameter estimates in outbreak b) improved from $\kappa = 310, \gamma = 137, \bar{\mathrm{s}} = 20,878, \overline{\mathrm{IR}} = 127$ with a Gamma(0.1,0.1) prior on $\gamma$ to $\kappa = 404, \gamma = 241, \bar{\mathrm{s}} = 26,566, \overline{\mathrm{IR}} = 199$ using a Gamma(14,100) prior. Similar effects were observed when placing an informative prior on $\kappa$ in outbreak a).

So far for simplicity reasons we had fixed the hyperparameters $\sigma^2$ and $\rho$ of the Gaussian process. When estimating them additionally, the mixing for $\sigma^2$ and $\rho$ is reasonable. The sampler seemed to have more difficulties estimating $\sigma^2$. The additional parameter estimates slowed down the mixing of the other parameters in outbreak a), whereas it improved mixing in outbreak c).Fixing $\sigma^2$ to 1 lead to an overall improved mixing of the remaining parameters. This held true, even in the case of a prior for $\rho$ which is centered away from the true parameter value in outbreak c). In fact, the mixing in the inference for outbreak a) and c) was better when setting $\sigma^2$ to 1 than when fixing it to the true parameter.

Putting a prior on $\rho$ which is centered away from the true parameter value slowed down the mixing of the parameter $\rho$ with yet a reasonable performance.

The mixing of the other parameters seemed to be robust towards that prior misspecification.

In conclusion, mixing of the MCMC chain is robust towards the estimation of the hyperparameters even in the case of wrongly specified priors and performs well for fixing $\sigma^2$ to 1.

### Parameter non-identifiability

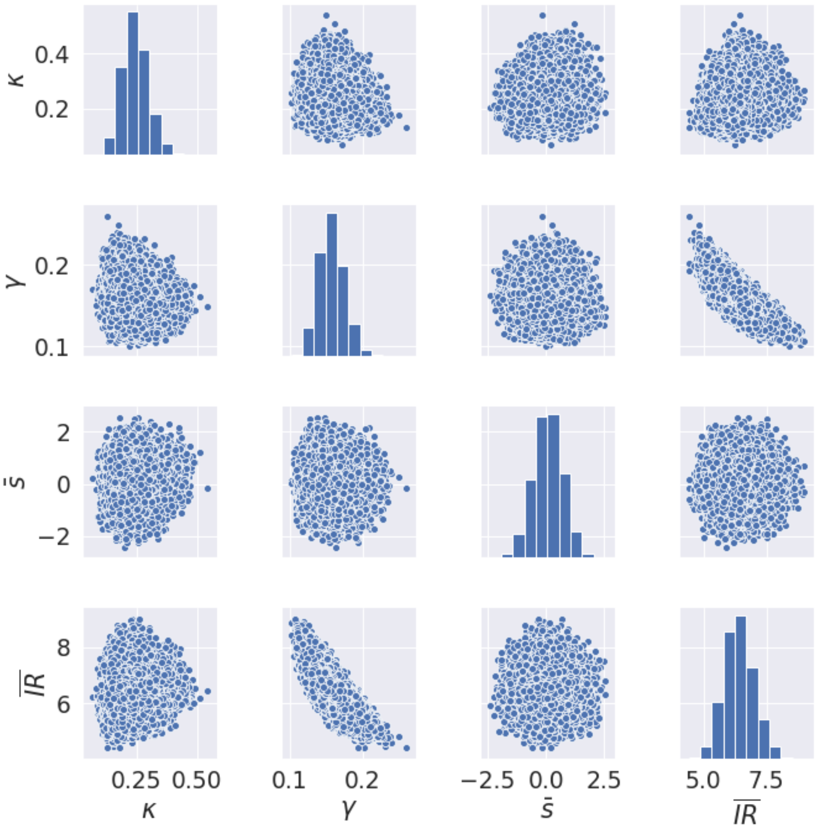


Figure 2.5: Pairwise correlation of posterior samples from an MCMC chain of outbreak simulation a) (Table 2.1). For the inference we fixed the hyperparameters of the Gaussian process $\bar{s}$ to their true value and chose Gamma(0.1, 0.1) priors for the spatial kernel parameter $\kappa$ and the infectious period $\gamma$. The mean infectious period is denoted by $\overline{\text{IR}}$.

The plot in Figure 2.5 gives us insight into the pairwise correlation of the parameters. This plot is with reference to the outbreak simulation a) and the MCMC chain with fixed Gaussian process hyperparameters and Gamma(0.1, 0.1) priors for the spatial kernel parameter $\kappa$ and the infectious period $\gamma$. There is a clear negative correlation between $\gamma$ and the mean infectious period, which was expected as the reciprocal of $\gamma$ is the mean length of the infectious period. Apart from that, the displayed parameters do not seem to be correlated.

In some simulations we noted a slight non-identifiability between the spatial kernel lengthscale parameter $\kappa$ and the Gaussian process $\mathbf{s}$. This can be explained by the fact that both parameters account for spatial variation in the data. We sought to find the relationship between these two and investigated the lengthscale versus the Pearson correlation coefficient of $\bar{\text{s}}$ and $\kappa$, the lengthscale of the Gaussian process versus the effective sample size of $\kappa$, as well as the Pearson correlation coefficient of the lengthscale and $\kappa$. The results indicated no obvious trend.

**Effect of distribution of locations**

For outbreaks it might happen that the spatial location of individuals is more dense or more sparse in some areas than others due to environmental factors, such as mountains, lakes, etc.. When investigating the effect of the spatial location of individuals on the model inference we observed that the distribution of the less centralized locations in outbreak d) improved mixing of the MCMC chain (ESS for $\kappa : 602, \gamma : 632, \bar{\text{s}} : 72,046, \overline{\text{IR}} : 483$) compared to the outbreak b) on the same risk surface with a different distribution of points over the whole observation area (ESS for $\kappa : 310, \gamma : 137, \bar{\text{s}} : 20,878, \overline{\text{IR}} : 127$). A possible explanation could be that the spatial distribution of the points can heavily influence the infection behaviour within the population: the spatial kernel determines that the next infections must be close around a current infected individual. The fixed locations along the margins of the observation area limit the possibilities of the next infection and thus the ongoing of the outbreak. This is also reflected in the bimodal infection curve, with the infection process first proceeding to the left and then to the right of the first infected individual. Therefore, the spatial location can add information a priori into the model making it easier for the sampler.

| | $\kappa$ | $\gamma$ | $1/\overline{\text{IR}}$ | $\overline{\text{IR}}$ | $\sigma^2$ | $\rho$ | $\bar{\text{s}}$ |
|---|---|---|---|---|---|---|---|
| **Outbreak a)** | | | | | | | |
| *True params* | 0.3 | 0.14 | 0.154 | 6.475 | 3 | 20 | 1.692 |
| Fixed I, $\rho, \sigma^2$ | 0.264 | 0.154 | - | - | - | - | 0.127 |
| Fixed $\sigma^2, \rho$ | 0.246 | 0.157 | 0.156 | 6.418 | - | - | 0.091 |
| Fixed $\sigma^2, \rho$<br>$\kappa \sim$ Gamma(30,100) | 0.277 | 0.162 | 0.160 | 6.246 | - | - | 0.127 |
| Fixed **s** | 0.262 | 0.156 | 0.155 | 6.470 | - | - | - |
| Fixed $\rho$<br>Set $\sigma^2 = 1$ | 0.215 | 0.169 | 0.167 | 5.977 | - | - | 0.021 |
| Fixed none<br>$\sigma^2 \sim$ Uniform$(0, 10)$<br>$\rho \sim$ Gamma$(200, 10)$ | 0.252 | 0.160 | 0.158 | 6.314 | 3.956 | 20.096 | 0.139 |
| Set $\mathbf{s} = \mathbf{0}$ | 0.203 | 0.191 | 0.189 | 5.285 | - | - | - |
| **Outbreak b)** | | | | | | | |
| *True params* | 0.3 | 0.14 | 0.141 | 7.078 | 3 | 40 | 2.001 |
| Fixed $\rho, \sigma^2$ | 0.205 | 0.155 | 0.150 | 6.683 | - | - | 0.453 |
| Fixed $\sigma^2, \rho$<br>$\gamma \sim$ Gamma$(14, 100)$ | 0.209 | 0.148 | 0.146 | 6.836 | - | - | 0.460 |
| **Outbreak c)** | | | | | | | |
| *True params* | 0.35 | 0.14 | 0.146 | 6.857 | 3 | 100 | 1.770 |
| Fixed $\rho, \sigma^2$ | 0.697 | 0.153 | 0.151 | 6.630 | - | - | 0.880 |
| Fixed none<br>$\sigma^2 \sim$ Uniform(0,10)<br>$\rho \sim$ Gamma(1000, 10) | 0.361 | 0.151 | 0.149 | 6.709 | 2.662 | 100.088 | 0.606 |
| Fixed none<br>$\sigma^2 \sim$ Uniform(0,10)<br>$\rho \sim$ Gamma(200, 10) | 0.360 | 0.147 | 0.145 | 6.877 | 1.846 | 19.938 | 0.058 |
| Set $\sigma^2 = 1$<br>$\rho \sim$ Gamma(200, 10) | 0.343 | 0.151 | 0.149 | 6.698 | - | 19.900 | 0.011 |
| **Outbreak d)** | | | | | | | |
| *True params* | 0.3 | 0.14 | 0.153 | 6.545 | 3 | 40 | 2.001 |
| Fixed $\rho, \sigma^2$ | 0.349 | 0.203 | 0.199 | 5.021 | - | - | 0.273 |

Table 2.3: Posterior mean estimates of outbreak simulations from Table 2.1. If not specified else, we chose Gamma(0.1, 0.1) priors. If fixed, the parameter was set to its true value and denoted by - in the table.

#### 2.7.1.2 Parameter estimates

Table 2.3 gives an overview over the parameter estimates for different inference scenarios. The simulation study revealed that some parameters were difficult to estimate and the parameter estimates depended on the outbreak simulation as well as on the a priori information we put into the inference through fixing parameters or deploying informative priors. In more detail:

- **A priori information:** Fixing parameters or placing more informative priors on parameters in the inference, lead to tighter posteriors and thus less uncertainty in the estimates. This can be seen for example in the 95% credible intervals when fixing $\rho$ and $\sigma^2$ compared to when additionally fixing the infection times. The credible intervals of $\gamma, \kappa$ and $\bar{\mathrm{s}}$ shrunk from [0.124, 0.193], [0.146, 0.354],[-1.157, 1.299] to [0.136, 0.174], [0.204, 0.322] and $[-1.094, 1.354]$, respectively.

- **Removal rate $\gamma$:** Throughout all inference settings, the removal rate parameter $\gamma$ was estimated higher than the true value. For outbreak a) this was not surprising, as the reciprocal of the mean infectious period was also higher for this particular outbreak simulation. However, for the other simulations the mean infectious period was also overestimated. If it is possible to obtain a reasonable estimate of the length of the infectious period from expert opinion, we could place a more informative prior on gamma, such as Gamma(14, 100) with mean set to the estimate and variance kept small to 0.0014. For example, for outbreak b) the posterior mean of $\gamma$ shifted closer to the true value as well as the 95% credible interval of $\gamma$ decreased from [0.104, 0.218] to [0.111, 0.193]

- **Spatial kernel $\kappa$:** In all the examples, the spatial kernel parameter $\kappa$ was underestimated. This might be a prior effect due to the large mass on 0 of a Gamma(0.1, 0.1) prior. Choosing a more informative Gamma(30,100) prior for $\kappa$ in outbreak a) improved the posterior mean and credible intervals for $\kappa$ (0.277, CI: [0.202, 0.351] in contrast to 0.246, CI: [0.146, 0.354]). Additionally, it shifted the posterior mean of $\bar{\mathrm{s}}$ closer to its true value from 0.091 to 0.127, though the estimate of $\gamma$ worsened. Thus, we can conclude that with a more informative prior on $\kappa$, we add more spatial information improving the estimates of $\kappa$ and $\bar{\mathrm{s}}$.

- **Nonidentifiability between $\kappa$ and $\bar{\mathrm{s}}$**: The deviation of the $\kappa$ estimate and the great underestimation of $\bar{\mathrm{s}}$ might be partially explained by the non-identifiability between the spatial kernel parameter $\kappa$ and the mean of the spatial random effect $\bar{\mathrm{s}}$. Both parameters account for spatial variation. Fixing $\mathbf{s}$ to its true values in outbreak a) led to a better posterior mean of $\kappa$ than when fixing the hyperparameters of the Gaussian process alone. On the other hand, putting a more informative prior on $\kappa$ let to better estimates of $\bar{\mathrm{s}}$ than when only fixing the hyperparameters. However, we could not determine any correlation structure between the two parameters (see Section 2.7.1.1).

- **Estimating the lengthscale of the Gaussian process**: The value of $\bar{\mathrm{s}}$ intrinsically depends on $\sigma^2$ and $\rho$. A priori we do not have any information on these hyperparameters because we want to model an *unknown* random effect. However, very low estimates of the lengthscale would result in different independent point risks and very large estimates would lead to a flat risk surface. Thus, we need to adapt the prior of the lengthscale to the scale of the observation area we are working on. Outbreak c) illustrates that the posterior lengthscale estimate is very prior sensitive. However, the prior choice for $\rho$ placing mass away from the true parameter, e.g. Gamma(200,10) with mean 20 whereas the true value is at 100, mainly affects its own estimate and $\bar{\mathrm{s}}$. All other parameter estimates are still reasonable.

- **Fixing the marginal variance $\sigma^2$ of the Gaussian process**: From the different inference scenarios of outbreak c) we can infer that $\sigma^2$ is really difficult to estimate. However, we are not interested in the value of $\sigma^2$ directly, but in the areas of higher and lower spatial risk in general. We investigated the impact of fixing the marginal variance to 1. This influenced the estimate of $\bar{\mathrm{s}}$ and thus the estimates of the other parameters as well. For outbreak a), for instance, this lead to the overall worst parameter estimates out of all scenarios. However, on the contrary, for outbreak c) it lead to quite reasonable estimates except $\bar{s}$ as expected, even though, the prior for the lengthscale was chosen far away from the true value.

We noticed that recovering the true parameters of the Gaussian process within an epidemiological setting can pose a real challenge. Thus, a fundamental question that arises in this context is: Are the overall dynamics of the outbreak still captured despite the deviating parameters estimates from the true values?

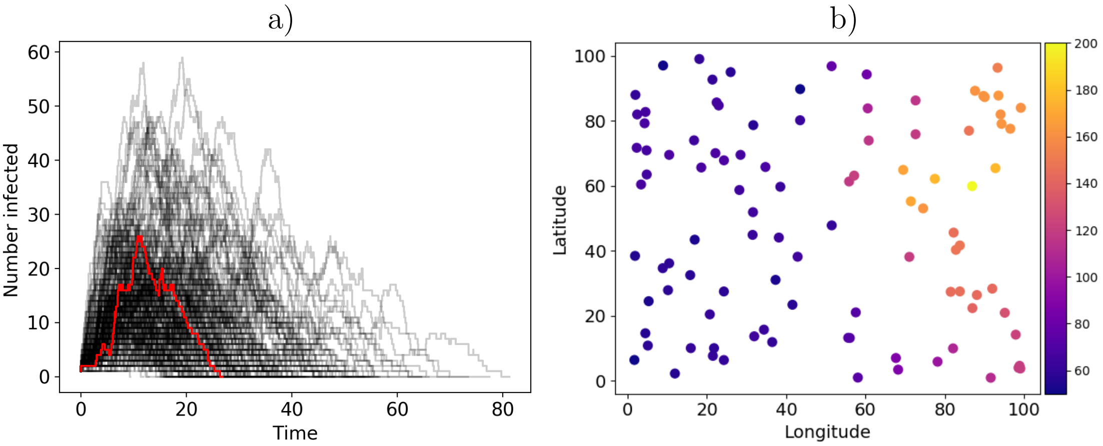


Figure 2.6: Investigation of the outbreak dynamics by the model. We simulated 200 disease outbreaks from the posterior distribution given outbreak c). The model captured the temporal dynamics (a)) as well as the spatial distribution (b)) of the outbreak. The red curve in a) displays the true outbreak data and the colormap in b) represent the frequency of how often an individual got infected within the 200 simulations.

Using Outbreak c) we simulated 200 outbreaks from posterior samples of the MCMC chain for which we had fixed $\sigma^2$ to 1 (true value 3) and had chosen a shifted prior for the lengthscale, that is Gamma(200,10) with mean 20, where the true value was 100 (Table 2.1). This resulted in an posterior mean estimate for the lengthscale that is far away from the true value (19.9). Despite the difficulty of recovering the true parameters of the Gaussian process, it became clearly visible in Figure 2.6 that the model was able to capture the dynamics of the outbreak. The true infection curve (red curve in Figure 2.6 a)) lies in the middle of the shaded outbreak curves representing the temporal distribution of the outbreak. Furthermore, we notice that points infected by the posterior outbreaks are in line with the spatial distribution of the original outbreaks (Table 2.1).

#### 2.7.1.3 Prediction

In this section we compare the true risk surfaces with the predicted risk surfaces given the data of the outbreaks. Figures 2.7 and 2.8 display the predicted risk surfaces for Outbreaks a) - d) of Table 2.1.

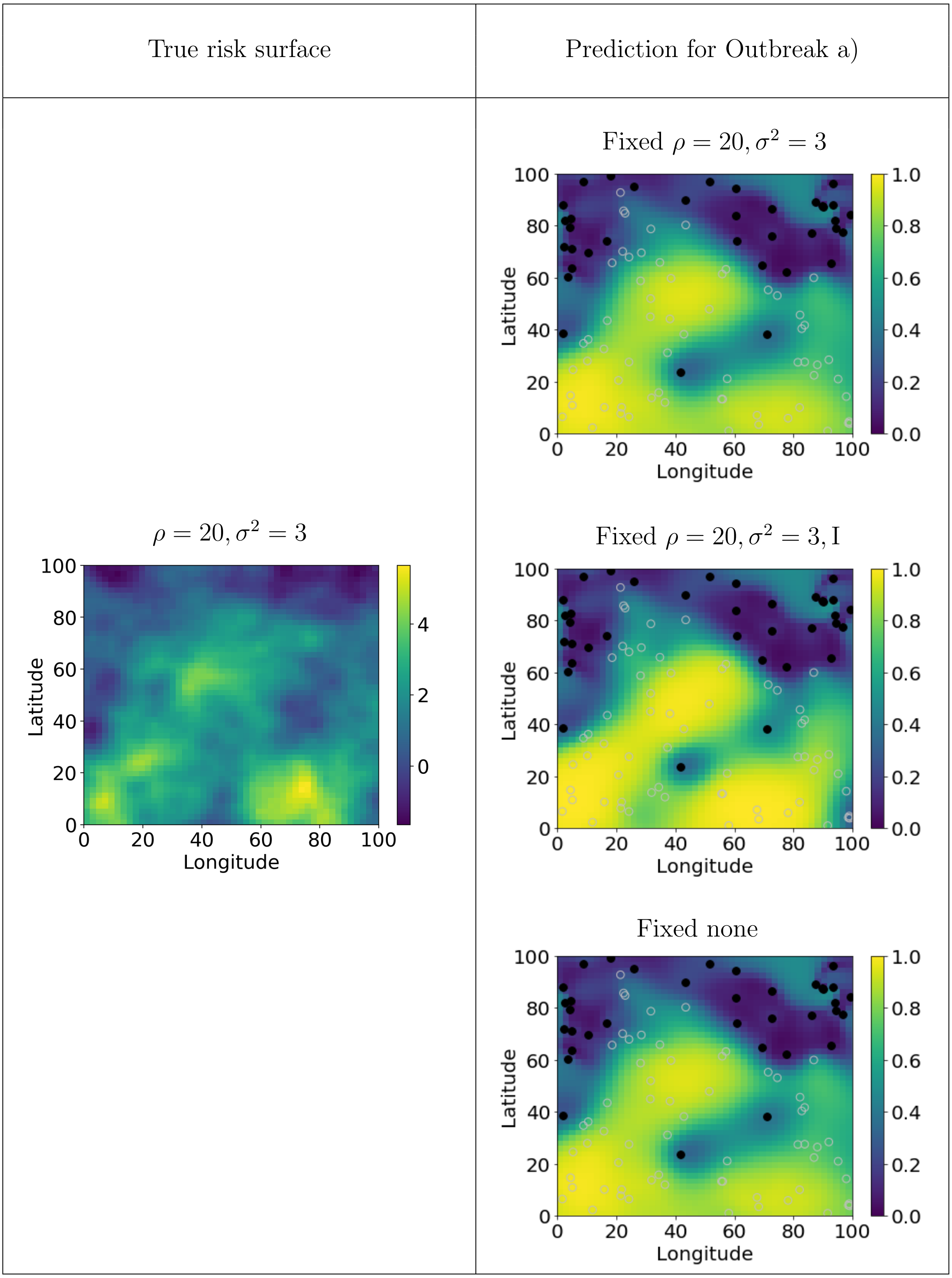
As a first step we fixed the hyperparameters $\rho$ and $\sigma^2$ of the Gaussian process to their true values and investigated how well the risk surface of the random spatial effect can be recovered by our method. When excluding the special case of Outbreak d), we observed that the high risk area from the simulation was mostly identified in all of the predicted risk surfaces. The model obtains most information on the risk surface in areas of infections, thus some high risk areas were not able to be identified, e.g. in Outbreak c) at the center bottom. On the other hand, the model was able to capture also lower risk areas exemplified clearly in prediction by Outbreak a) and in Outbreak b) on the left bottom corner. For better inference on such areas, it is necessary to include other spatial covariates as additional data.

The prediction of Outbreak c) in 2.8 has a very narrow colormap scale representing minimal variation in the severity of the spatial risk. In comparison to the other true risk surfaces, the scale of the colormap is also narrower because of a larger lengthscale ($\rho = 100$), thus less variation in the prediction was expected. However, with more spatial information from other covariates we might be able infer a better risk surface at areas of no infection, e.g. on the top left.

The prediction of the risk surface in Outbreak d) (Figure 2.8) serves as an exception to the remaining outbreaks. Here the non-centered spatial distribution of the individuals' locations affected the prediction of the random risk surface immensely. The model does not have information in areas where no individuals reside and thus cannot capture any features of the surface anywhere else. In other words, if the accumulation of points is too sparse with respect to the varying features of the surface, the model is not able to capture these variations.

Given that we had fixed $\rho$ and $\sigma^2$ so far, we can also conclude that the method performs similarly well when including their estimation in the inference. This is represented through the prediction plots of Outbreak a) which look very similar.

In order to analyze the impact of the slow mixing, we also investigated the difference in the risk surface when fixing and estimating the infection times. For



True risk surface

Prediction for Outbreak a)

$\rho = 20, \sigma^2 = 3$

Fixed $\rho = 20, \sigma^2 = 3$

Fixed $\rho = 20, \sigma^2 = 3$, I

Fixed none

Figure 2.7: Comparison of the simulated (left) and predicted (right) Matern(3/2) risk surfaces for Outbreak a). Different parameters were fixed according to the inference scenarios. The grey circles and black points denote individuals that were removed and survived the outbreak, respectively.

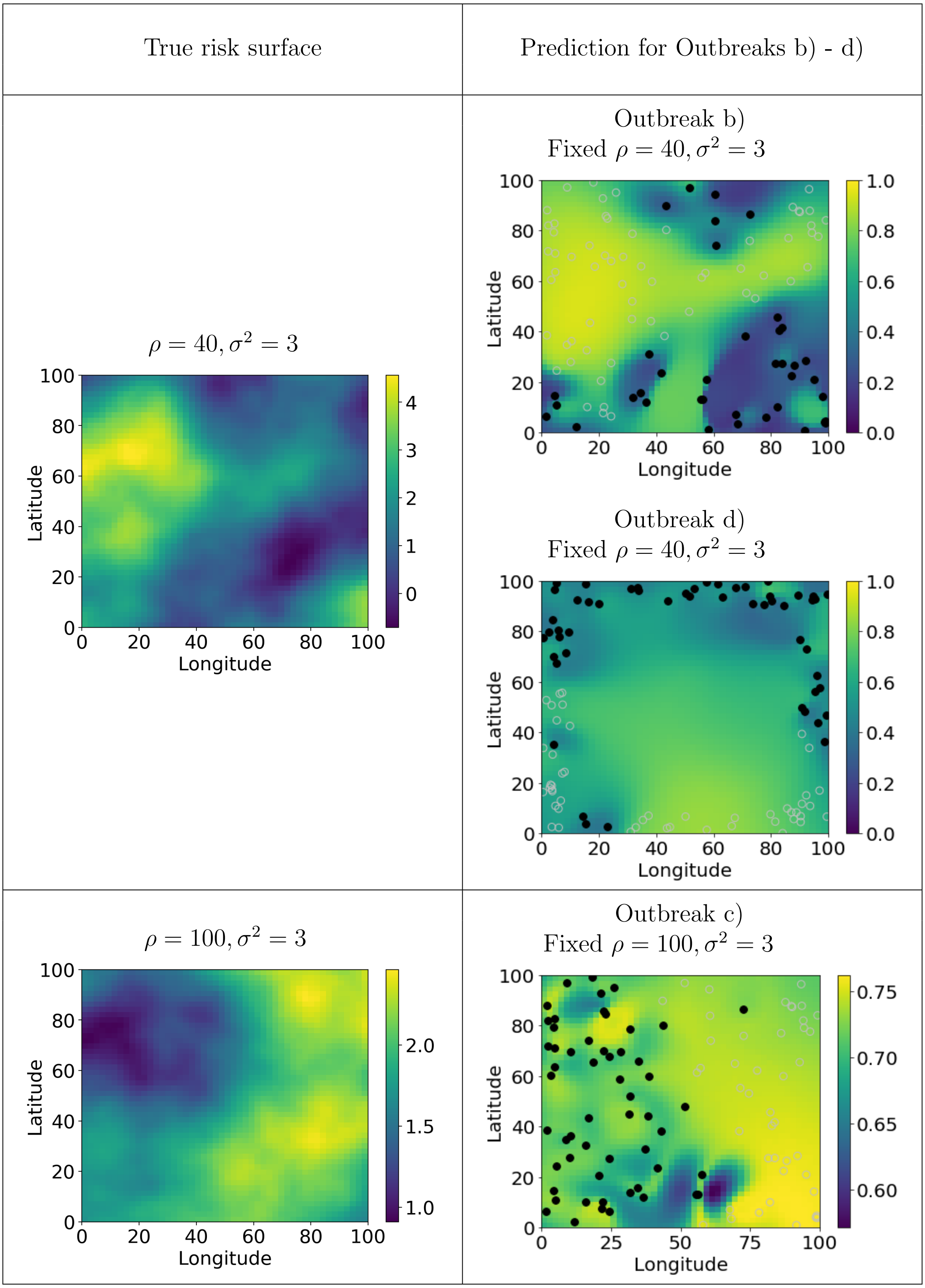


Figure 2.8: Comparison of the simulated (left) and predicted (right) Matern(3/2) risk surfaces for Outbreaks b) - d). The hyperparameters of the Gaussian process were fixed to the true values in the inference. The grey circles and black points denote individuals that were removed and survived the outbreak, respectively.

Outbreak a) in Figure 2.7 the different prediction plots are displayed. Due to their high similarity, we can place a lower weight on the slower mixing of the infection times after all.

#### 2.7.1.4 Why is a vector model needed?

In this section we want to elaborate why such a joint approach is needed and why an epidemic model alone does not suffice. To answer this, we used the outbreak example a) from Table 2.1 and made inference without estimating a spatial random effect, that is setting the spatial random effect to 0.

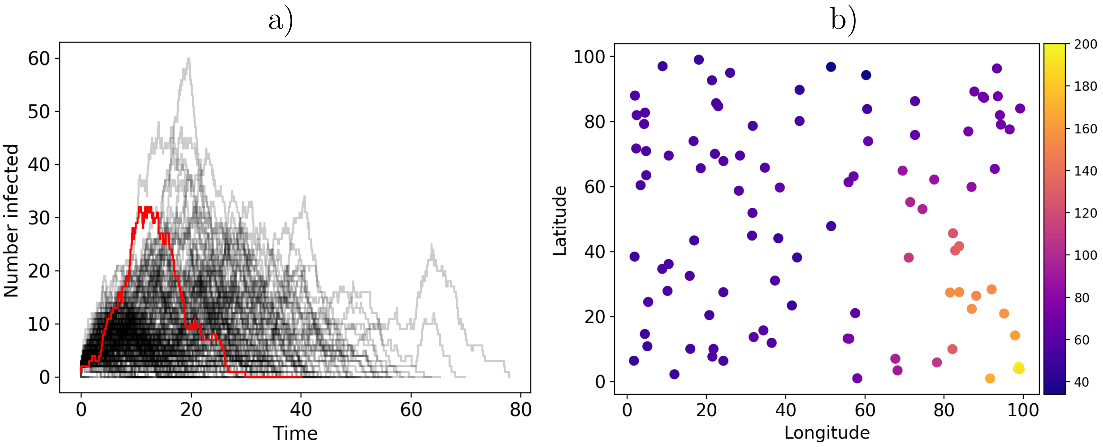


Figure 2.9: Investigation of the outbreak dynamics by the model. We simulated 200 simulated disease outbreaks from the posterior distribution given the data of outbreak c). a) displays the temporal dynamics with the red curve being the true data, and b) the spatial distribution of the outbreaks with the colormap representing the frequency of how often an individual got infected within the 200 simulations.

Compared to the model when fixing the hyperparameters of the Gaussian process to their true value and estimating the random effect, we obtained noticeably poorer estimates for the other parameters (see Table 2.3), in particular the mean infectious period was estimated more than a whole day shorter (6.5 vs 5.3). Moreover, the spatial kernel was estimated much lower to seemingly account for the missing infectious pressure.

The same holds for the dynamics of the outbreak. Figure 2.9 pictures 200 simulated outbreaks from the posterior samples of the MCMC chain. We can clearly see that the true infection curve (red line) does not represent the overall infection curves (Figure 2.9 a) ). Furthermore, when comparing the spatial distribution of the outbreaks, that is the frequency of individuals getting infected in Figure 2.9 b), we can clearly see that the outbreaks have a tendency of going north. The true outbreak spread (Table 2.1 a) ) however, was also driven by the underlying risk surface, which is located at the bottom center and thus would imply a tendency towards the west as well.

This example displays that if we neglect the impact of the spatial random effect on the outbreak, it leads to poor parameter estimates and different outbreak dynamics.

#### 2.7.1.5 Remark 1

From this simulation study of the joint approach we can conclude:

- Including stronger prior information improves the identifiability of the parameters.

- As expected, prior information on the parameters increases the identifiability of the parameters. However, such data is often difficult to obtain, in particular for a random effect. In this case it is advisable to adapt the spatial scale parameter of the Gaussian process to the appropriate scale of the observation area by means of an informative prior. Besides, we found out that the variance parameter of the Gaussian process, $\sigma^2$, was very difficult to estimate. Moreover, fixing $\sigma^2$ did not influence the mixing behaviour. Thus, it is advisable to fix this parameter to decrease the sampling space and facilitate inference.

- The mixing of the MCMC method was reasonable with the infection times performing worst and the spatial random effect performing best. The latter is also represented by the similar prediction plots when fixing or estimating the parameters in the inference.

- Although mixing might not be optimal, this can still be acceptable because we are not interested in the exact estimation of the parameters and rather in the forecast or prediction of an ongoing disease spread. Obtaining similar outbreak characteristics from posterior samples showed that we captured the outbreak dynamics.

- Shifting the infection times posed a hardship for the sampler and slowed down the mixing also for the other parameters. This became clear in the effective sample size, in particular, when fixing the infection times to their true values. Several approaches for improving the shifting of the infection times did not lead to better results.
- The removal rate, also given by its reciprocal the infectious period, may be estimated roughly through expert opinion or from literature and can be translated into a suitable prior for $\gamma$.
- In the course of this simulation study we inferred that the spatial location of the individuals impact model inference, if the number of individuals cannot represent the variation of the risk surface adequately.

In summary, simulating from the model and making inference and prediction, leads to reasonable results as anticipated. Although the mixing of the infection times is rather slow, it does not pose a problem for capturing the dynamics of the model. Stronger prior information about the parameters leads to better identifiability of the parameters. In our outbreak setting this could be realized by gathering information about the length of the infectious period and adapting the prior for lengthscale to the observation area. Furthermore, other data on spatial covariates may contribute to the better estimation of the risk surface in areas of no disease occurrence. In the remainder of this thesis, we will not analyze the mixing in detail and refer to Section 2.7.1.1 for a thorough discussion.

### 2.7.2 Exploring further risk surfaces

So far, in Section 2.6.2, we worked with Matern(3/2) risk surfaces, as this reflects our simulation of the covariance function in the model setup of the Gaussian process. However, usually we do not have much information about the spatial effect a priori. Thus, we want to explore how the algorithm is performing in case of misspecification of the model. Hence, in the following sections we explore three other risk surfaces: a flat risk surface, a surface with a high risk area in the upper left corner and a sinusoidal risk surface - all estimated with a Matern(3/2) covariance function. In all examples we follow the procedure of simulation, inference and prediction described in Section 2.6.2. We neglect the discussion about the mixing of the algorithms as it was similar to the model performance described in Section 2.7.1.1.

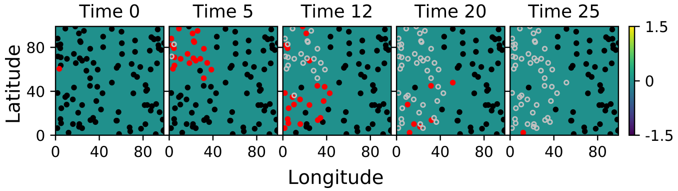
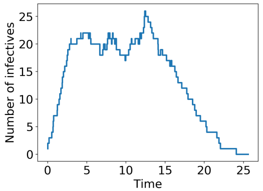
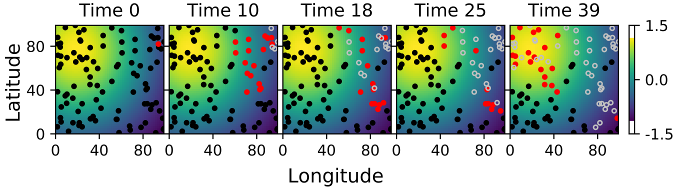
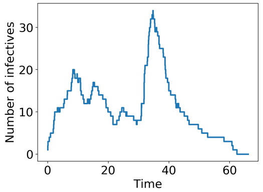
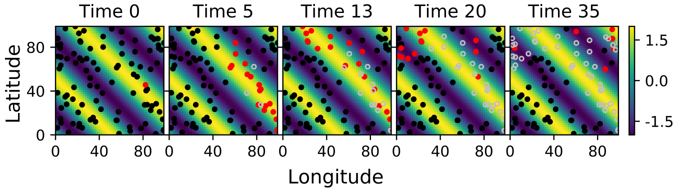
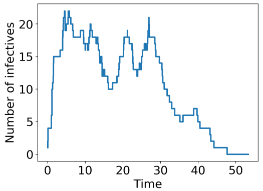
| | Risk surface | Spatial outbreak progression | Number of infectives | Outbreak size | Parameters |
|---|---|---|---|---|---|
| e) | Flat risk surface |  |  | 49 | $\kappa = 0.42$<br>$\psi = 10$ |
| f) | Shifted and scaled bi-variate Normal |  |  | 64 | $\kappa = 0.5$<br>$\psi = 0.4$ |
| g) | Sinusoidal |  |  | 56 | $\kappa = 0.4$<br>$\psi = 0.4$ |

Table 2.4: Overview of simulated epidemic outbreaks e) - g). Black and red points represent susceptible and infected individuals respectively, grey circles denote removed individuals.

#### 2.7.2.1 Flat risk surface

We simulated an outbreak without a spatial random effect and misspecified our model by trying to recover a flat risk surface with a Matern(3/2) covariance function. The outbreak simulation is displayed in Table 2.4 e). In contrast to all other outbreaks so far, the driving force of infection was solely the distance between the individuals governed by $\kappa$ and the infection rate parameter $\psi$.

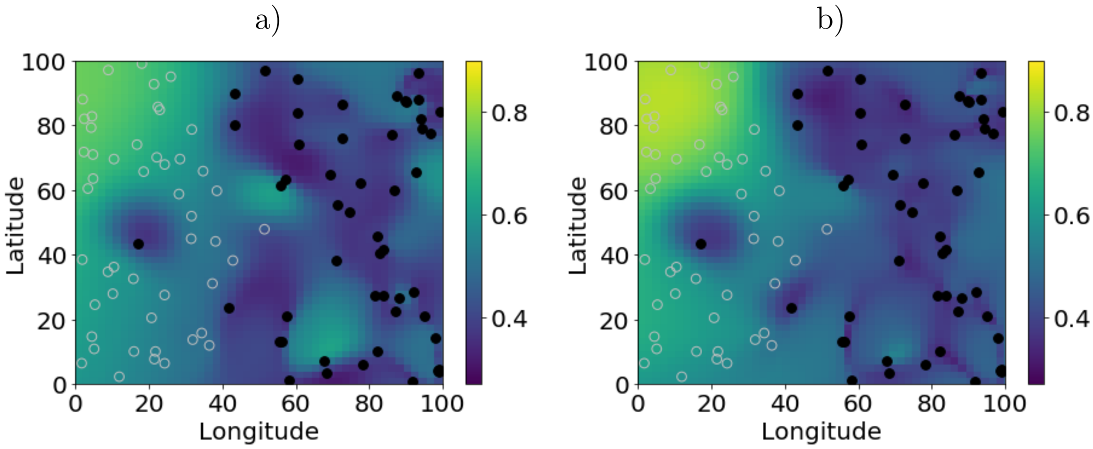


Figure 2.10: Comparison of the predicted risk surfaces for Outbreak e) with a) fixed $\sigma^2 = 1, \rho = 40$, and b) estimated hyperparameters ($\sigma^2 \sim U(0, 10)$ and $\rho \sim \text{Gamma}(40, 1)$).

To stabilize inference, we first fixed the hyperparameters $\sigma^2 = 1$ and $\rho = 40$ (estimated based on the size of the surface). Predicting the spatial random effect from posterior samples showed a higher probability at all infected individual's locations (Figure 2.10 a)). In the upper left corner there was slight increase in the posterior probability for the spatial risk surface. This might be attributable to the fact that the outbreak started there (see Table 2.4 e)). However, the weight given to the spatial random effect, that is the posterior probability for values greater than 0, is not as strong as compared for example to the prediction of Outbreak a) (Figure 2.7). Furthermore, the posterior estimate for the mean $\bar{\text{s}} = 0.013$ is very close to the true value of 0 denoting a small effect. The posterior mean estimate of $\kappa = 0.589$ was overestimated, which led to a notable decrease of the infectious pressure attributed to the spatial kernel. We additionally sampled from the posterior distribution of the infection rate parameter $\psi$ (Section 2.5), and

| | *Mean posterior estimate* | | | | | | |
|---|---|---|---|---|---|---|---|
| | $\kappa$ | $\gamma$ | $1/\overline{\text{IR}}$ | $\overline{\text{IR}}$ | $\sigma^2$ | $\rho$ | $\bar{\text{s}}$ |
| **Outbreak e)**<br>True parameters | 0.42 | 0.14 | 0.138 | 7.27 | - | - | 0 |
| Fixed $\rho = 40, \sigma^2 = 1$ | 0.589 | 0.154 | 0.152 | 6.578 | - | - | 0.013 |
| Fixed none<br>$\sigma^2 \sim \text{Uniform}(0, 10)$<br>$\rho \sim \text{Gamma}(40, 1)$ | 0.305 | 0.151 | 0.150 | 6.660 | 2.219 | 40.829 | 0.095 |
| **Outbreak f)**<br>True parameters | 0.5 | 0.14 | 0.085 | 11.650 | - | - | 0.162 |
| Fixed $\rho = 50, \sigma^2 = 3$ | 0.535 | 0.074 | 0.073 | 13.630 | - | - | 0.698 |
| Fixed $\rho = 100, \sigma^2 = 1$ | 0.431 | 0.078 | 0.078 | 12.823 | - | - | 0.243 |
| Fixed none<br>$\sigma^2 \sim \text{Uniform}(0, 10)$<br>$\rho \sim \text{Gamma}(40, 1)$ | 0.480 | 0.080 | 0.080 | 12.530 | 4.39 | 40.266 | 1.109 |
| **Outbreak g)**<br>True parameters | 0.4 | 0.14 | 0.098 | 10.227 | - | - | -0.050 |
| Fixed $\rho = 30, \sigma^2 = 3$ | 0.327 | 0.089 | 0.088 | 11.304 | - | - | 0.070 |
| Fixed none<br>$\sigma^2 \sim \text{Uniform}(0, 10)$<br>$\rho \sim \text{Gamma}(40, 1)$ | 0.339 | 0.090 | 0.089 | 11.196 | 6.274 | 36.209 | 0.378 |

| | *Effective sample size* | | | | | | |
|---|---|---|---|---|---|---|---|
| | $\kappa$ | $\gamma$ | $\overline{\text{IR}}$ | $\sigma^2$ | $\rho$ | $\bar{\text{s}}$ | |
| **Outbreak e)** | | | | | | | |
| Fixed $\rho = 40, \sigma^2 = 1$ | 408 | 223 | 188 | - | - | 7706 | |
| Fixed none<br>$\sigma^2 \sim \text{Uniform}(0, 10)$<br>$\rho \sim \text{Gamma}(40, 1)$ | 177 | 170 | 99 | 221 | 5968 | 1319 | |
| **Outbreak f)** | | | | | | | |
| Fixed $\rho = 50, \sigma^2 = 3$ | 51 | 61 | 29 | - | - | 1131 | |
| Fixed $\rho = 100, \sigma^2 = 1$ | 88 | 82 | 44 | - | - | 6133 | |
| Fixed none<br>$\sigma^2 \sim \text{Uniform}(0, 10)$<br>$\rho \sim \text{Gamma}(40, 1)$ | 36 | 94 | 44 | 157 | 3030 | 100 | |
| **Outbreak g)** | | | | | | | |
| Fixed $\rho = 30, \sigma^2 = 3$ | 250 | 175 | 96 | - | - | 6523 | |
| Fixed none<br>$\sigma^2 \sim \text{Uniform}(0, 10)$<br>$\rho \sim \text{Gamma}(40, 1)$ | 137 | 92 | 54 | 370 | 1353 | 1029 | |

Table 2.5: Mean posterior estimates and effective sample sizes for outbreak simulations e) - g) from Table 2.4. If not otherwise specified, we chose Gamma(0.1, 0.1) priors.

noticed a large increase in infectious pressure from originally 10 to 31.24. This in particular explains the model dynamics displayed in Figure 2.11: the disease was a lot more infectious to account for the missing infectious pressure, but the spatial kernel inhibited the spread in most cases leading to smaller outbreak size (a) and b)) and only local spread in the upper left corner (c)).

Setting priors with a large variance on the hyperparameters of the Gaussian process results in an even more pronounced hotspot in the upper left corner. The posterior estimate for $\rho$ was very close to its mean despite the large variance (40). The marginal variance $\sigma^2$ was estimated higher with a uniform prior than the previously fixed value (1), attributing a greater variability to the spatial random effect. This is a result of the misspecification of the model. Yet, the mean spatial effect was still close to 0.

In contrast to the fixed hyperparameter case, $\kappa$ was underestimated attributing more infectious pressure to individuals lying even further away (Table 2.5). The posterior mean of the infectious rate parameter ($\psi = 5.14$) revealed that the distance to other individuals and the underlying spatial random effect played the major role in transmission, instead of the infectiousness of the disease.

In Figure 2.11 it is visible that the dynamics of the outbreak was captured in the mean when estimating the hyperparameters, despite the deviating parameter estimates.

In both inference scenarios the removal rate was slightly lower than the true value, shortening the infectious period and giving more weight of the infectious pressure to the other parameters.

This example of a simulated flat risk surface exemplified the overfit of the model by integrating spatial information that were never used in the data generating process. When controlling the marginal variance of the Gaussian process (fixing $\sigma^2$ to 1), we obtain a relatively flat risk surface such that the random spatial effect was not as pronounced, but the outbreak dynamics are only partially captured. When setting priors on the hyperparameters, we obtain a more pronounced risk surface, not reflecting the truth.

Thus we can conclude, for this case, if we focus on capturing the dynamics of the outbreak, using non-informative priors for the hyperparameters of the Gaussian

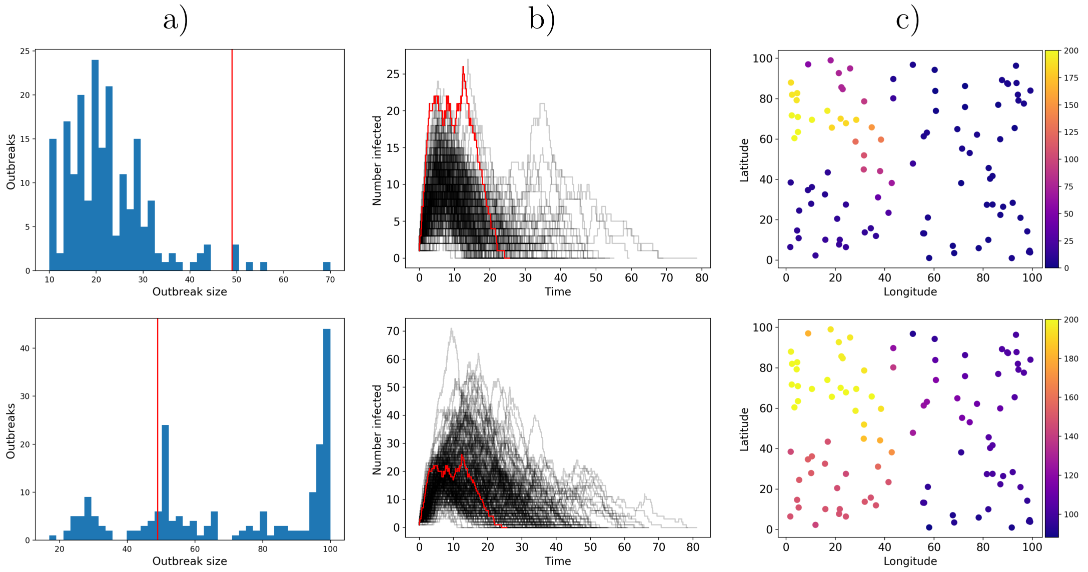


Figure 2.11: Comparison of outbreak dynamics of the MCMC chain with fixed hyperparameters (top row) and estimated hyperparameters (bottom row) for Outbreak e). 200 outbreaks were simulated from posterior samples. Column a) displays a histogram of the outbreak size, column b) the temporal progression, and column c) the spatial progression of the outbreak. The red lines refer to the original outbreak simulation.

process is favourable, whereas estimating the spatial random effect works better if we fix these parameters.

#### 2.7.2.2 Shifted and scaled normal risk surface

This risk surface was created by means of a bivariate normal distribution with mean $\mu = (20, 80)$ and covariance matrix $\Sigma = [[45, 0], [0, 45]]$. The density $f(\ell \mid \mu, \Sigma)$ was evaluated at all locations $\ell$ in the grid and transformed by $f(\ell \mid \mu, \Sigma) \cdot 55000 - 1.5$. It yielded a surface with a high risk area in the upper left corner of the grid.

The simulated outbreak (Table 2.4 f)) can be characterized by two outbreak waves. The first wave started with the first infected individual in the right upper corner where the risk surface displays lower values. The spread evolved locally on the right hand side due to the restriction of the spatial kernel. As soon as an individual within the high risk area was infected, the outbreak kicked off leading to a second outbreak wave and an infection of all individuals in the high risk zone within a short time interval (about 5 days).

For inference, we first tried to fit the Matern covariance function with an informed guess about the parameters to the risk surface. We first fixed $\rho = 50$ and $\sigma^2 = 3$ due to the slow decay of risk across space.The posterior mean estimates were somewhat close to the true values, except that the mean of the spatial kernel was estimated higher. This decreases when setting the marginal variance $\sigma^2 = 1$, as expected. Apart from that, note that the parameter estimates did not vary much in the different inference scenarios despite the change of hyperparameters for the Gaussian process. Even estimating the posterior distribution of the hyperparameters led to similar results.

Thus, we can infer that the parameter estimates seem to be robust towards varying the hyperparameters. Figure 2.12 confirms this observation, as the prediction plots capture the same features with only varying intensities. By highlighting the upper part of the grid, the algorithm captures the simulated risk coming from the spatial random effect. Yet, it also attributes a lot of infectious pressure to the whole of the top half, which is an overestimation of the infectious pressure and does not reflect the true risk surface.

Our model is set up in such a way that host incidence cases inform on vector presence. Hence, if the disease has not arrived in particular areas and no individual was infected there (in this example in the bottom half of the picture), we cannot draw conclusions on the vector population and the values will average them out to the mean of the spatial effect.

Although the model proved to be robust towards varying the hyperparameters of the covariance function, it also had difficulties to capture details of the risk surface with the misspecification of the function family.

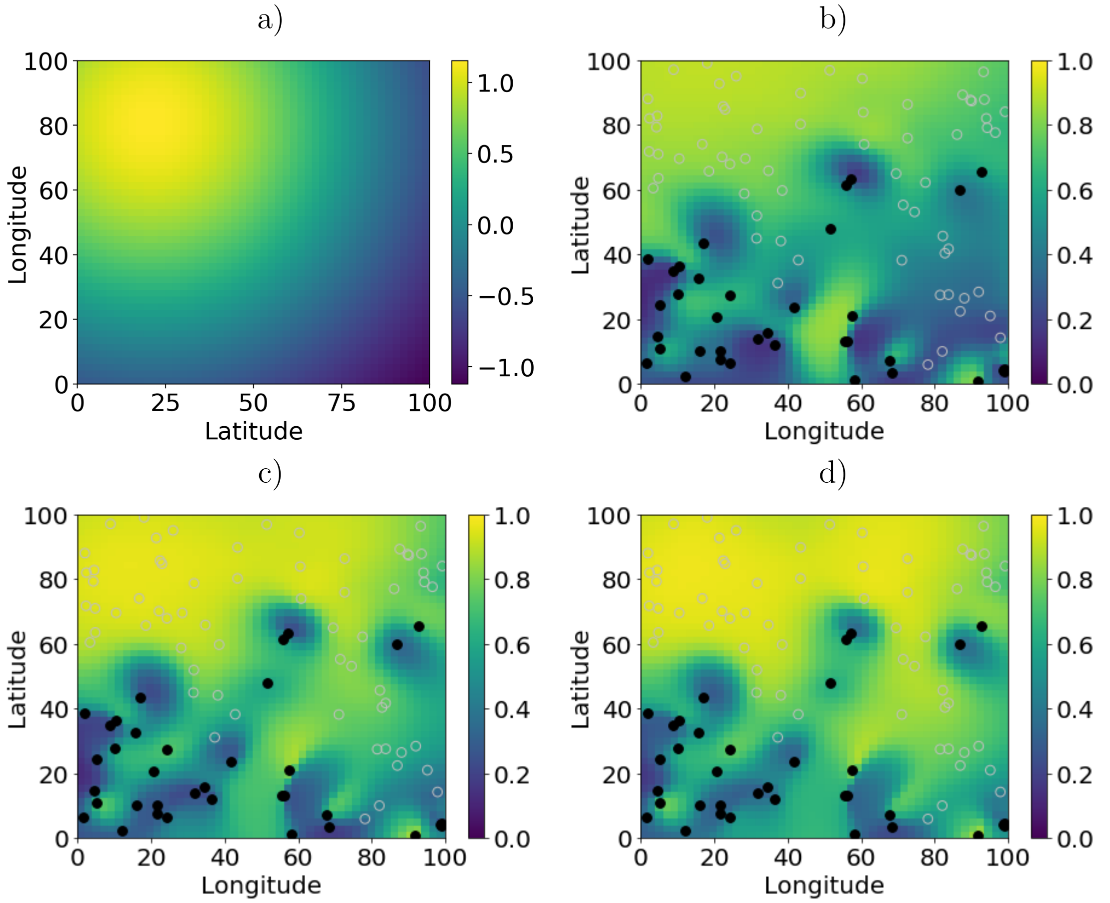


Figure 2.12: Comparison of the a) simulated and predicted risk surfaces with b) Fixed $\sigma^2 = 3, \rho = 50$, c) Fixed $\sigma^2 = 1, \rho = 100$, and d) Estimated hyperparameters with priors $\sigma^2 \sim U(0, 10)$ and $\rho \sim \text{Gamma}(40, 1)$ with respect to outbreak simulation f)

#### 2.7.2.3 Sinusoidal risk surface

The sinusoidal risk surface at coordinate $(x, y)$ is defined as

$$2\ \sin\left(1.1\pi\sqrt{(x-3)+(y-3)+300}\right)$$

The risk surface is characterized by parallel, alternating stripes of high and low risk. They are diagonal from the left bottom corner to the right upper corner of the grid.

The outbreak depicted in Table 2.1 g), is confined to a certain area of high spatial risk. By the end of the outbreak individuals lying on the left stripe of high risk were infected, in contrast to the individuals lying on the bottom risk stripe. This can be explained by the partially sparse distribution of susceptibles and the local transmission restriction. The low risk valley between the high risk areas may have further reduced the spread of the disease.

For the inference and the prediction, we first fixed the lengthscale of the Gaussian process to $\rho = 30$ because it takes about 30 units to get from the highest point to the lowest point of the colorbar considering the width of the stripes. The marginal variance was fixed to a higher value of $\sigma^2 = 3$ in order to give the model a chance to capture this variation in the surface. In general the posterior mean estimates reflected the true values reasonably well. Although the true mean spatial infectious pressure has a negative value (-0.05), it is still very close to 0, just like the estimate (0.07).

The results obtained when setting flat priors on the hyperparameters are similar with the main difference that the marginal variance was estimated about twice as high (6.3). This is reflected in a slightly higher mean spatial risk.

Comparing the simulated sinusoidal risk surface and predicted risk surface given only the removal data of the outbreak (Figure 2.13), we can conclude that the risk surface was well recovered where possible. Due to the information obtained from the infectives, the upper stripe of high risk was able to be recovered well. However, we did not have any information in the lower left corner, which is why the Gaussian process was not able to capture any variation in the surface. The choice of covariance function is certainly not optimal in this scenario. By informing the inference and prediction algorithm with a periodic covariance function, chances are

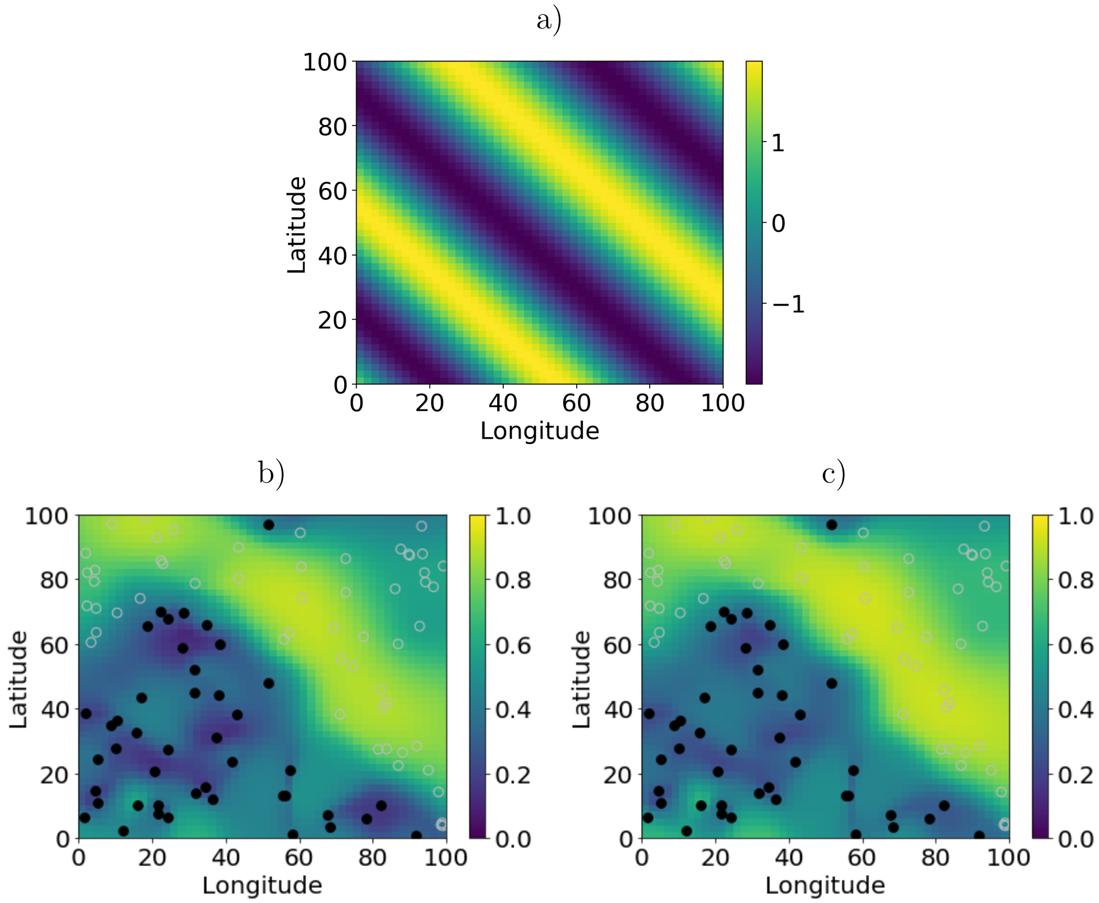


Figure 2.13: Comparison of the a) simulated sinusoidal risk surface from Outbreak g) with the predicted risk surfaces. In b) fixed the hyperparameters $\sigma^2 = 3, \rho = 30$, and in c) estimated the hyperparameters using priors $\sigma^2 \sim U(0, 10)$ and $\rho \sim \text{Gamma}(40, 1)$).

higher to reconstruct the period structure reflected in the simulated risk surface.

#### 2.7.2.4 Remark 2

In this chapter we investigated the behaviour of this joint epidemiological and empirical modelling approach with different spatial risk surfaces not simulated from the model. It became clear our model showed robustness towards varying the hyperparameters of the covariance function. It was able to identify the spatial high risk areas where possible, that is where enough information of the outbreak was given, and showed a poor performance of recovering the underlying true risk surface in the remaining area. In this case further spatial covariates are needed to inform the model.

In the special case where we tried to recover a non-existing underlying spatial random effect, we overfitted the model and obtained deviating posterior estimates as well as slightly different outbreak dynamics. Here, fixing the parameters leads to a better estimation of the spatial random effect, whereas non-informative priors are expedient for capturing the dynamics of the outbreak.

### 2.7.3 Including observed spatial information

In the previous section it became evident, that additional spatial covariates can inform the model to better identify the spatial random effect, in particular in areas of no infection. Thus, in this section we investigate the model behaviour when including observed spatial information into the epidemiological model. These data can be for example environmental or meteorological covariates influencing the vector's habitat and disease transmission. A different example for COVID-19 could be spatially varying covariates, such as socioeconomic or age distributions, which could provide information on the population risk of disease transmission and severity.

In this section, we first analyse the model's ability to identify coefficients for fixed effects only before we additionally recover the spatial random effect. All results are displayed in Table 2.7.

#### 2.7.3.1 The outbreak

The outbreak creation was performed by means of the Matern(3/2) risk surface with a lengthscale $\rho = 20$ from Outbreak a). Furthermore, the two risk surfaces of Sections 2.7.2.2 and 2.7.2.3 (see Outbreaks f) and g)) served as the fixed effects with corresponding coefficients $\beta_0$ and $\beta_1$ respectively for our model. The outbreak progression and its features are displayed in Table 2.6. The disease appeared first in the middle on the top and spread first towards the left upper corner before almost all individuals on the left half of the plot got infected. The spread can be attributed to the influence of the high risk areas of the fixed effects at first with the random effect contributing rather at the bottom towards the outbreak. As it is displayed in the joint spatial risk surface, the individuals on the top right corner were not infected, because the low spatial risk prevented the outbreak from spreading to that corner.

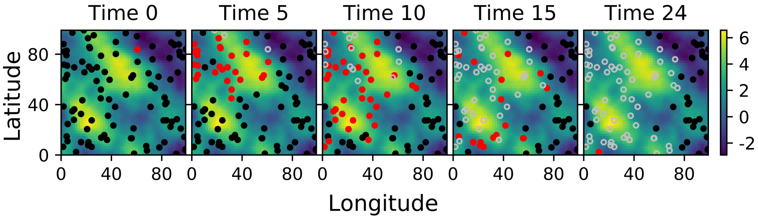
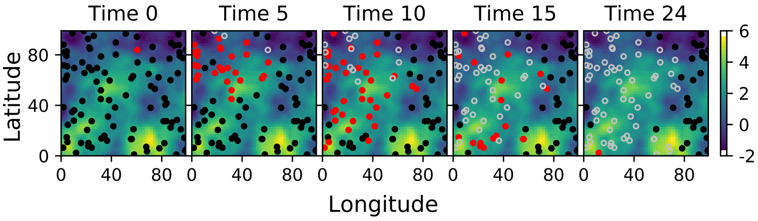
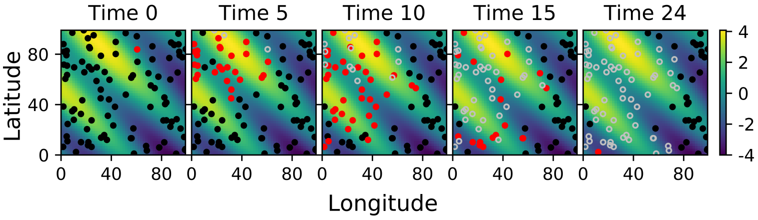
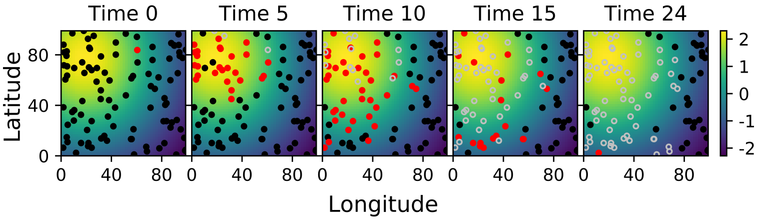
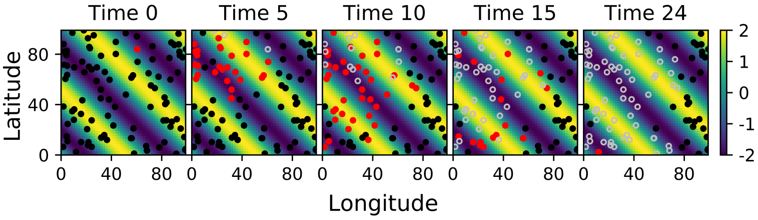
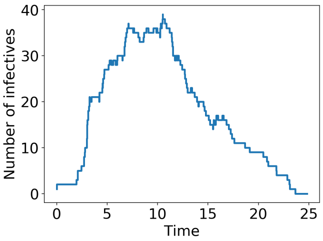
| | Outbreak h) |
|---|---|
| Joint spatial risk |  |
| Only random effect risk |  |
| Only fixed effects risk |  |
| Only fixed effect $X_0$ |  |
| Only fixed effect $X_1$ |  |
| Number of infectives |  |
| Outbreak size | 64 |
| Parameters | $\kappa = 0.3$ $\gamma = 0.143$ $\psi = 0.3$ $\beta_0 = 1$ $\beta_1 = 2$ |

Table 2.6: Overview of simulated epidemic outbreak h). Black and red points represent susceptible and infected individuals respectively, grey circles denote removed individuals.

| | $\kappa$ | $\gamma$ | $\overline{\text{IR}}$ | $\beta_0$ | $\beta_1$ | $\sigma^2$ | $\rho$ | $\bar{\text{s}}$ |
|---|---|---|---|---|---|---|---|---|
| True parameters | 0.3 | 0.143 | 6.94 | 1 | 2 | 2 | 20 | 1.914 |
| $\beta_0, \beta_1 \sim N(0, 2)$<br>Fixed $\mathbf{s}$ | 0.223<br>[0.122, 0.334] | 0.148<br>[0.115, 0.182] | 6.820<br>[5.512, 8.101] | 0.845<br>[0.419, 1.292] | 1.686<br>[1.134, 2.240] | - | - | - |
| $\beta_0, \beta_1 \sim N(0, 100)$<br>Fixed $\mathbf{s}$ | 0.219<br>[0.121, 0.325] | 0.150<br>[0.114, 0.187] | 6.736<br>[5.381, 8.137] | 0.843<br>[0.405, 1.284] | 0.167<br>[1.135, 2.187] | - | - | - |
| $\beta_0, \beta_1 \sim N(0, 2)$<br>Fixed<br>$\rho = 20, \sigma^2 = 3$ | 0.250<br>[0.138, 0.370] | 0.139<br>[0.107, 0.174] | 7.266<br>[5.764, 8.796] | 0.984<br>[0.256, 1.696] | 1.016<br>[-0.245, 2.345] | - | - | 0.061<br>[-1.213, 1.316] |
| $\beta_0, \beta_1 \sim N(0, 2)$<br>$\sigma^2 \sim$ Uniform(0,10)<br>$\rho \sim$ Gamma$(200, 20)$ | 0.250<br>[0.129, 0.382] | 0.139<br>[0.103, 0.178] | 7.319<br>[5.510, 8.977] | 0.998<br>[0.206, 1.857] | 1.054<br>[-0.400, 2.652] | 4.222<br>[0.390, 8.897] | 20.077<br>[17.256, 22.757] | 0.071<br>[-1.435, 1.564] |
| $\beta_0, \beta_1 \sim N(0, 2)$<br>Fixed $\sigma^2 = 1$<br>$\rho \sim$ Gamma$(40, 1)$ | 0.184<br>[0.098, 0.274] | 0.159<br>[0.119, 0.201] | 6.386<br>[4.988, 7.846] | 0.602<br>[0.160, 1.053] | 0.754<br>[-0.149, 1.637] | - | 39.412<br>[27.668, 51.733] | -0.0005<br>[-1.115, 1.129] |

Table 2.7: Mean posterior estimates and 95% credible intervals for simulated Outbreak f) from Table 2.6. If not otherwise specified, we chose Gamma(0.1, 0.1) priors.

#### 2.7.3.2 Fixing spatial random effect

In the inference scenarios we first fixed the random spatial effect to its true value in order to investigate how well the model can identify the coefficients of the spatial fixed effects $(\beta_0, \beta_1)$. We chose normally distributed priors with mean 0 and varying variance (2 and 100) for the coefficients (see Table 2.7).

It became clear that a less informative prior on the fixed effect parameters did not influence parameter estimates and did not affect the mixing behaviour much. All estimates were reasonable and the true parameter was always contained in the 95% credible interval. In other simulations the wider variance in the prior led to wider confidence intervals. Note that the ratio $\beta_0/\beta_1 = 0.5$ was inferred by the ratio of the according posterior means $0.84/1.69 = 0.50$.

Thus, we can infer that the model can identify the parameters well and that a less informative prior on the fixed effect parameters has almost no influence the model inference.

#### 2.7.3.3 Fixing Gaussian process hyperparameters

As a next step, we fixed the hyperparameters of the Gaussian process to its true values ($\sigma^2 = 3$ and $\rho = 20$) and estimated the spatial random risk surface $\mathbf{s}$. Comparing the scenario to fixing $\mathbf{s}$ (Table 2.7), the parameters were estimated equally reasonable, except that the sinusoidal risk surface was given less impact in the infectious pressure; its posterior mean coefficient $\beta_1$ deviates further from the true value (1.01 compared to 1.69, true value: 2), the credible interval widens but still contains the true value ([-0.24, 2.35]). A noticeable change is that the ratio between the posterior means of $\beta_0$ and $\beta_1$ increases from 0.5 to almost 1. Furthermore, the mean spatial random effect was underestimated despite fixing the hyperparameters of the Gaussian process to their true value. With more parameters accounting for spatial influence on the infectious pressure $(\kappa, \beta_0, \beta_1, \mathbf{s})$ the sampler has a harder time to identify the spatial parameters. The outbreaks dynamics in terms of outbreak size, temporal progression and spatial spread are still captured by the model as displayed in Figure 2.15.

The predicted surface for the random spatial effect is displayed in Figure 2.14 b). The three hot spots are clearly identified by the model.

#### 2.7.3.4 Estimation of Gaussian process hyperparameters

In this section we additionally estimate the Gaussian process parameters (Table 2.7). An informative prior on the lengthscale $\rho$ and a uniform prior on $\sigma^2$ led to very similar results as fixing the hyperparameters. However, in a real scenario draft we do not have information about the random spatial effect. Thus, we checked the model inference when assuming the mean lengthscale of $\rho$ to lie at 40 due to the size of the observation area and use a Gamma(40,1) prior for a wider variance. We further fixed $\sigma^2$ to 1 to stabilize the inference. In this example, the posterior mean parameters deviated further away from the true values, however the changes are not drastic. The lengthscale was estimated higher which can be attributed to a prior effect. Since the marginal variance parameter $\sigma^2$ was fixed to 1, the mean spatial random effect was estimated lower than in the other scenarios, but this observation is in line with previous simulation studies.

Analyzing the outbreak dynamics, in Figure 2.15 it becomes clear, that the outbreak size, the temporal progression and the spatial spread of the outbreak is captured by the model.

We were also interested in the prediction of the random spatial effect when having estimated all parameters (Figure 2.14 c)). The prediction plot of the scenario with an informative prior on the lengthscale looks very similar to fixing the hyperparameters to their true value (Figure 2.14 b)). This is not surprising as the posterior estimates are very similar.

When choosing a prior for the lengthscale which lies further away from the mean, we still capture the high risk area with a little less detail due to the broader lengthscale (Figure 2.14 d)).

The fixed effects coefficients $\beta_0, \beta_1$ add more information to the model. If we fix them to their true values, we obtain a prediction plot of the spatial random risk which clearly focuses on the infection pressure on the bottom (Figure 2.14 e)). In comparison to all other predictions of Figure 2.14, on the top left quarter, e) has the lowest infection pressure. Thus, it can differentiate better between the risk coming from the spatial random effect and the other parameters of the epidemiological model. In the area of no infection, the predictions look very similar regardless of the additional information on the fixed effects.

#### 2.7.3.5 Remark 3

We can conclude that if we only want to estimate the fixed effects and leave out the estimation of the spatial random effect, the model performs reasonable well. A less informative prior on the fixed effect parameters has almost no influence towards the model inference. The more spatial parameters are estimated in the model, the harder it is to identify them. We noted that the model responds to a change of prior, but not drastically. When we consider the predictive distribution of the risk surface and the outbreak dynamics, the exact parameter estimates are less important; instead we can infer that our results reflect the model scenery we are working in. To sum up, the model responses to the change of prior but is still able to identify the outbreak dynamics and the random risk surface where possible.

## 2.8 Discussion

In this chapter we developed and elaborated on the joint epidemiological and empirical modelling approach for disease outbreaks. We inferred the likelihood and constructed an efficient MCMC algorithm for inference.

In the course of the simulation study, we inferred that the mixing performance of the sampler was reasonable, yet suboptimal. In particular the data augmentation of the infection times slowed down the sampler. This effect has been observed in the literature before and a possible solution (if time and computational resources allow) would be to let the chain run longer and to perform greater thinning for less autocorrelation of the samples. Stronger prior information about the parameters stabilized the inference which could be obtained through experts' opinion and information about the epidemiology of a particular disease.

In summary, the model showed robustness towards varying the hyperparameters of the covariance function of the Gaussian process. It was also able to predict high risk areas of risk surfaces not simulated from the model. However, sufficient information on the outbreak needs to be presented in order to improve prediction results, that is the outbreak size and a sufficient density and spatial spread of the individuals to capture the variation of the spatial risk surface adequately.

The more information about spatial fixed effects we inserted into the model, the better the random effect could be predicted. However, this came with the cost of non-identifiability of the parameters as all of them compete for the spatial infectious pressure. Yet as displayed in the simulation study, the dynamics of the outbreak in terms of outbreak size, temporal progression and spatial distribution were able to be captured. We were hoping to identify the unknown spatial transmission better in areas of no infection by including fixed effects into the model. In our simulation study this was not the case. With limited amount of data on the outbreak available, including more data on fixed effects might ease this problem. Stated from a practical point of view, if we had additional information about the risk surface (e.g. from vector trapping studies), this would help us to pin down the latent risk surface more accurately.

We performed the simulation study on a population size of 100 individuals for the ease of computational efficiency. Here, we chose outbreaks with about 60%

infections to obtain enough information for inference. In the following chapters we also investigate outbreaks with less percentages of infections.

We can conclude, it is a valuable approach for informing about the ongoing transmission of the disease, covariates that are drivers of morbidity and mortality and unknown spatial hot spots of disease transmission.

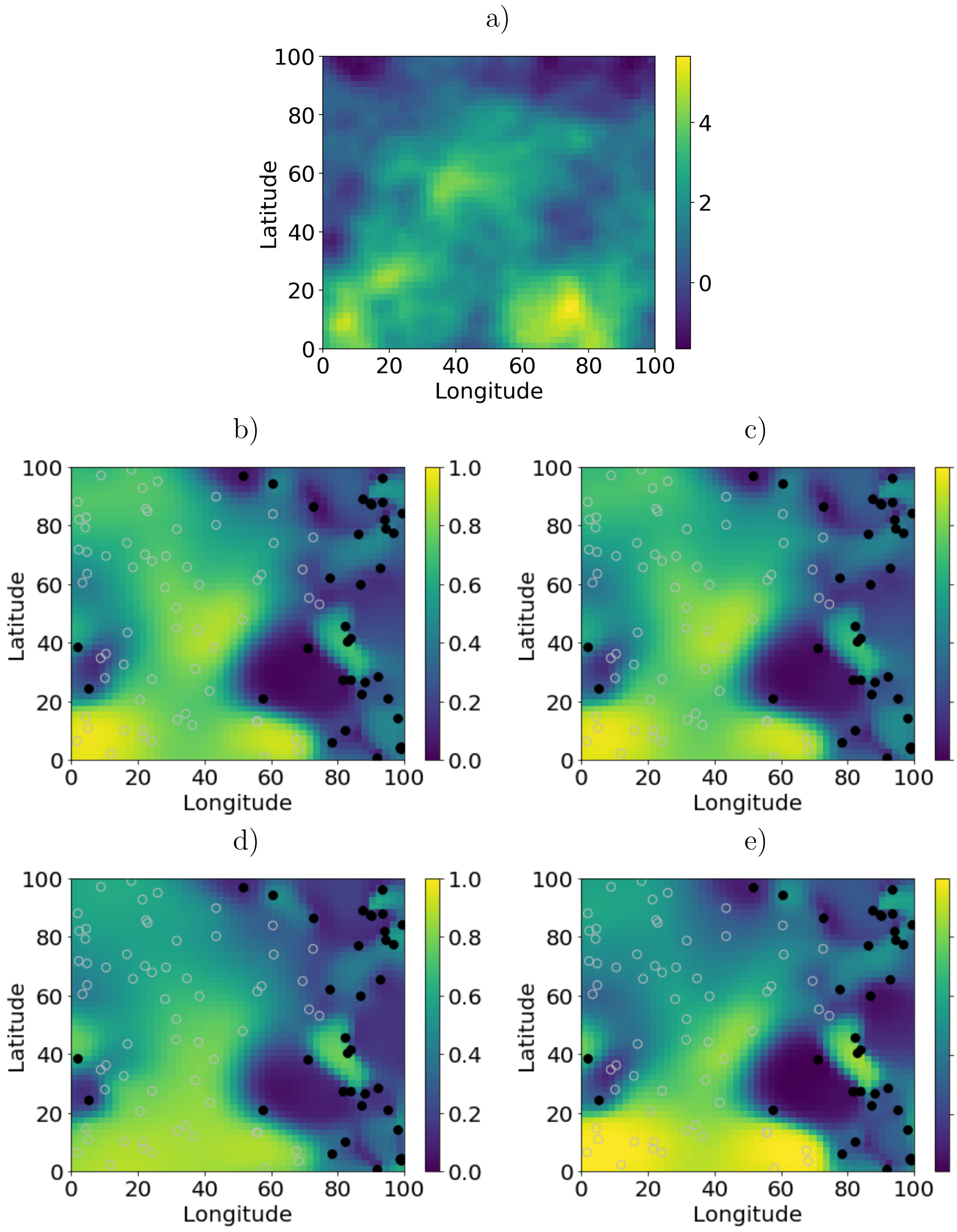


Figure 2.14: Comparison of the a) simulated random spatial risk surface from Outbreak h) with the predicted risk surfaces. Inference scenarios are b) hyper-parameters fixed to their true values $\rho = 20, \sigma^2 = 3$ c) $\rho$ ∼Gamma(200, 20), $\sigma^2 \sim U(0, 10)$, and d) fixed $\sigma^2 = 1$ and $\rho$ ∼Gamma(40,1). In a)-d) we used N(0,2) priors for c and Gamma(0.1, 0.1) else. In e) has the same scenario as c) with additionally fixing $\beta_0$ and $\beta_1$ to their true values.

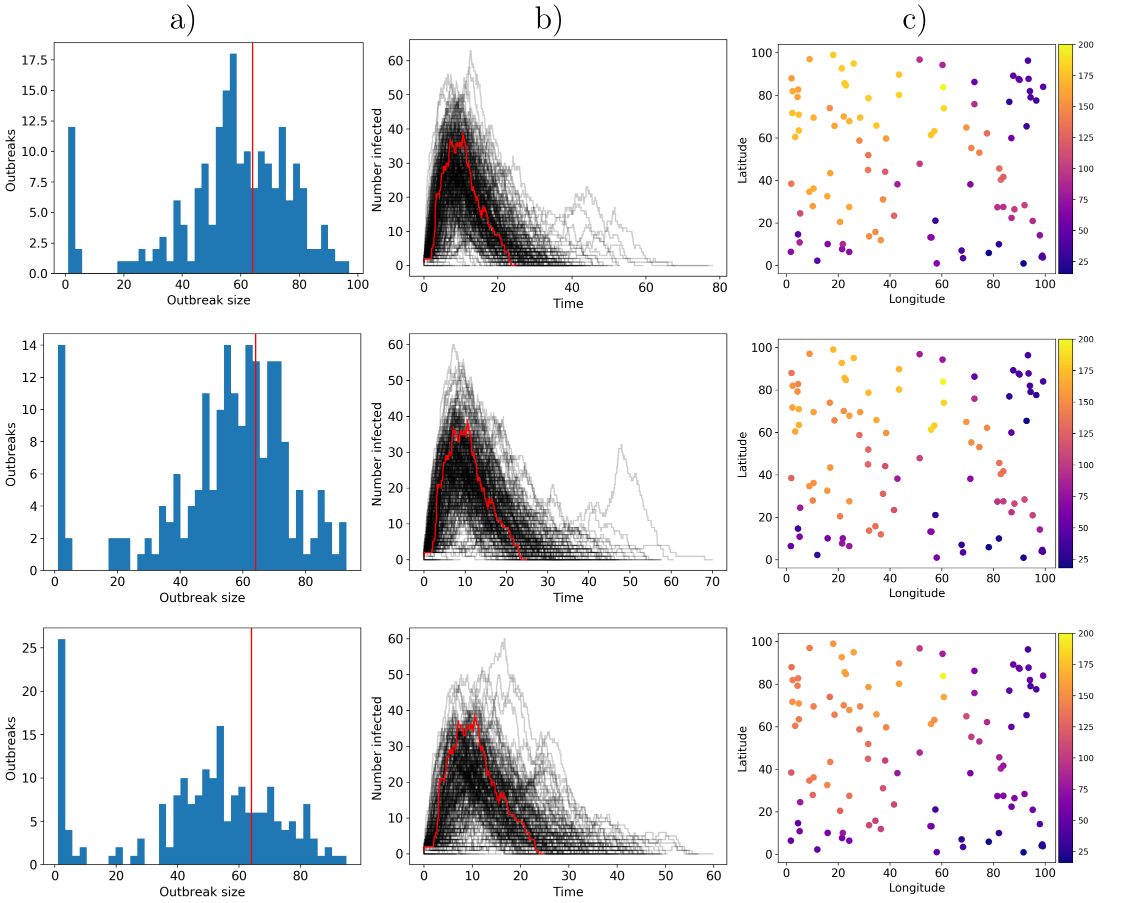


Figure 2.15: Analysis of outbreak dynamics of the MCMC chain for Outbreak h). Fixed hyperparameters (top row), estimated hyperparameters $\rho \sim$Gamma(200,10) and $\sigma^2 \sim U(0, 10)$ (middle row), and estimated hyperparameters $\rho \sim$Gamma(40,1) and $\sigma^2 = 1$ (bottom row). 200 outbreaks were simulated from posterior samples. Column a) displays a histogram of the outbreak size, column b) the temporal progression, and column c) the spatial progression of the outbreak. The red lines refer to the original outbreak simulation.

# Chapter 3

# Scaling to a larger population size

So far, the size of the population with 100 individuals was relatively small. When working with real disease data the population size can be up to 100 to 1000 times greater. Such big data pose a challenge for computer algorithms with respect to run time and memory usage. Thus, efficient algorithms that suit the available computational resources are needed.

One such example in big data settings are high dimensional matrices. The efficiency of the matrix operations mainly depends on the *sparsity* of a matrix: the proportion of zero entries among all entries in the matrix. A matrix is called *sparse* if it contains mostly zero values, and *dense* if it contains mostly non-zero values.

In sparse matrices only non-zero elements need to be stored with a desired storage complexity of at most $O(n)$ in contrast to $O(n^2)$ for dense matrices. Depending on the desired matrix operation thereafter different data structures for sparse storage can be used. Examples include:

- *Coordinate list (COO)*: a list storing three element tuples of [row, column, value] for each non-zero entry. This allows for fast access and value modification.

- *Compressed sparse row (CSR)*: three one-dimensional arrays storing the non-zero values, the column indices and the extents of rows by means of row pointers (in contrast to the row indices). This allows for fast row access and slicing, and matrix vector product, but slow column operations.

- *Compressed sparse column (CSC)*: similar to CSR except that the row indices and the column extents are stored. This allows for fast column access and slicing, and matrix vector products, but slow row operations.

Scaling up our methodology to a big data setting, in this chapter we explain how we make use of the structure of sparse matrices in two different scenarios: 1.) calculating the infectious pressure, in particular the value of the spatial kernel, and 2.) representing Gaussian processes by Gaussian Markov random fields to estimate the spatial random effect. Note, that the goal of our work is not to approximate Gaussian processes via Gaussian Markov Random fields, as covered for example in Lindgren et al. (2011). Instead, we explicitly use the term *represent* to emphasize that we use the spatial structure of a Gaussian process to derive spatial association through a Gaussian random Markov field. This allows us to circumvent calculations with the dense covariance matrix by dealing with the sparse precision matrix instead.

## 3.1 Efficient computation of the infectious pressure

First, we recall that the infectious pressure on an individual $j$ is modelled through a Poisson process with intensity function $\lambda_j(t) = \psi \cdot \Big(\alpha + \exp(v(\ell_j)) \sum_{i \in \mathcal{I}(t)} K(d_{ij}; \kappa)\Big)$. It depends on the number of infected individuals at time $t$ and thus alters after each event (infection or removal). An efficient way to calculate the sum over all infected individuals is via matrix representation of the spatial kernel. We define the symmetric matrix $\boldsymbol{D} = [d_{ij}]_{i,j \in \{1,\dots,n\}}$ where the element $d_{ij}$ is the Euclidean distance between individual $i$ and $j$. We then apply the function $K(\cdot\,; \kappa)$ elementwise and write

$$K(\boldsymbol{D}; \kappa) = \begin{bmatrix} K(d_{11}; \kappa) & K(d_{12}; \kappa) & \dots & K(d_{1n}; \kappa) \\ K(d_{21}; \kappa) & K(d_{22}; \kappa) & \dots & K(d_{2n}; \kappa) \\ & & \vdots & \\ K(d_{n1}; \kappa) & K(d_{n2}; \kappa) & \dots & K(d_{nn}; \kappa) \end{bmatrix}$$

Summing over the corresponding row of a susceptible and only taking into account the columns of infected individuals at time $t$ yields the total infectious pressure from other infected individuals. In Chapter 2 we introduced $K$ to be an exponential distance kernel. Therefore, all entries of $K(\boldsymbol{D}, \kappa)$ are non-zero and the matrix is dense. When sampling with MCMC methods, $(n^2 + n)/2$ entries have to be calculated for every sample of $\kappa$, because $\boldsymbol{D}$ is symmetric . With a large population size of $10^4$ or more individuals this can constitute a high computational burden.

In order to reduce the complexity $O(n^2)$ to at most $O(n)$ we make use of the fact, that spatial kernels are decaying functions. The spatial kernel function

$$K(d_{ij}; \kappa) = \begin{cases} \exp(-d_{ij} \cdot \kappa) & d_{ij} < c \\ 0 & \text{else} \end{cases} \tag{3.1}$$

where $c$ serves as a threshold for "cutting off" the spatial kernel as illustrated in Figure 3.1, yields a sparse matrix: entries are 0 for all pairs of individuals with distance greater than $c$. The threshold $c$ is determined through prior information about the infectiousness of the disease and the spatial decay of infectious pressure.

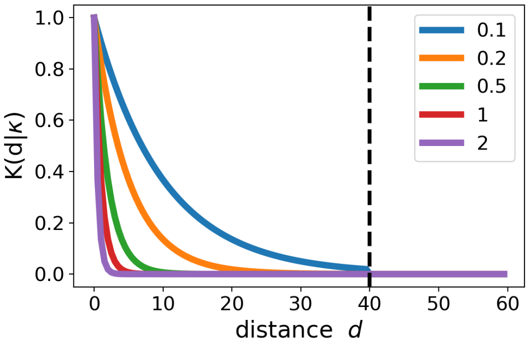


Figure 3.1: The decay of the spatial kernel function $K(d|\kappa)$ for different values of $\kappa$. The dotted black line denotes the distance after which the value of the spatial kernel is 0.

Equation (3.1) is advantageous for efficient computing because it is clear a priori which elements are zero and which need computing.

## 3.2 Gaussian Markov Random Fields

In spatial statistics we often want to infer a spatially continuous random effect over an area of interest. Here, Gaussian processes constitute are a popular tool where the correlation structure between any two points is determined by the covariance function. However, sampling from a Gaussian process has complexity $O(n^3)$ due to the Cholesky decomposition of the covariance matrix (Section 1.4.3). This becomes infeasible for large $n$ and is thus informally denoted as the *big $n$ problem*.

An alternative are the computationally more attractive Gaussian Markov Random Fields (GMRF). They are jointly Gaussian and have a Markov property in space: given an underlying network, the distribution of data at a certain location in space conditioned on the values of its neighbours is independent of the value of all other locations. This network structure is displayed in the precision matrix where an entry is nonzero if and only if the nodes are neighbours in the network. With only a small number of neighbours for each location, the precision matrix becomes sparse. For MCMC based inference this constitutes computational advantages because the Gaussian densities can be computed in linear time with respect to the number of locations $n$.

In this section we first explain the concept of how to define such an underlying network structure and the concept of conditional independence in undirected networks before we move on to the definition of GMRF and how to sample from it. A thorough coverage of this topic including proofs and applications can be found in Rue and Held (2005).

### 3.2.1 Voronoi tessellation and Delaunay triangulation

In this section we are interested in imposing a neighbourhood structure of a set of points in order to create an underlying graph.

**Definition 4:**
Given a set of locations $L = \{\ell_1, \ell_2, \dots, \ell_n\}$ in an area $A \subset \mathbb{R}^2$. A *graph* over $L$ is a tuple $\mathcal{G} = (L, \mathcal{E})$, whereby $L$ is the set of nodes or points, and $\mathcal{E} \subseteq \{(\ell_i, \ell_j) \in L^2, \ell_i \neq \ell_j\}$ is the set of edges in the graph.

The graph is called *undirected*, if $(\ell_i, \ell_j) \in \mathcal{E} \Leftrightarrow (\ell_j, \ell_i) \in \mathcal{E}$ for all $i, j \in \{1, \dots, n\}$. We call $\ell_i$ *a neighbour of* $\ell_j$ (and write $\ell_i \sim \ell_j$) with respect to $G$ if there exists an edge $(\ell_i, \ell_j) \in \mathcal{E}$.

In the following we will exclusively deal with undirected graphs. If we are free to select $L$ in order to cover a certain area, we might choose the points to lie on the corners of a regular grid. Then it is straightforward to define the respective four neighbouring points by means of the edges of the grid cells. However, for disease outbreaks, the points, for example the locations of livestock farms, are fixed and typically not arranged on a grid but rather irregularly scattered within the area of interest. Thus, we deploy an alternative method by first partitioning the plane into regions based on the Euclidean distance of the farms and thereafter defining a neighbourhood structure by means of neighbouring regions.

This section belongs to the field of computational geometry. More details including various properties of the below concepts can be found in Okabe et al. (2009).

**Definition 5:**
Let $L = \{\ell_1, \ell_2, \dots, \ell_n\}$ be points in a region $A \subset \mathbb{R}^2$ and let $|| \cdot ||$ denote the Euclidean distance. The *Voronoi region* $V(\ell)$ of a point $\ell$ is defined as

$$V(\ell) = \{x \in A \mid ||x - \ell|| \leq ||x - \ell'|| \quad \forall \ell' \in L\}$$

and the *Voronoi diagram* or *Voronoi tessellation* of $L$ is

$$V(L) = \bigcup_{\ell \neq \ell'} V(\ell) \cap V(\ell')$$

The Voronoi region of a point $\ell$ contains all points in $A$ that are at least as close to $\ell$ than to any other point in $L$. It also follows that $\bigcup_{\ell \in L} V(\ell) = A$. Voronoi regions can be unbounded if they lie on the border of $A$. Figure 3.2 a) exemplifies a Voronoi tessellation.

**Definition 6:**
Given the set of Voronoi regions $\{V(\ell)\}_{\ell \in L}$, the graph $\mathcal{G} = (L, \mathcal{E})$ with $(\ell, \ell') \in \mathcal{E}$ iff $V(\ell)$ and $V(\ell')$ share an edge, is called *Delaunay triangulation* of the convex hull of $L$.

The Delaunay triangulation is the straight-line dual of the Voronoi diagram and as displayed in Figure 3.2 b), the edges $\mathcal{E}$ of its graph $\mathcal{G}$ cover the convex hull of $L$ with triangles.

Given a set of irregularly scattered points $L$ on our area of interest $A$, our aim was to define a neighbourhood structure for $L$. The set of edges $\mathcal{E}$ from the Delaunay triangulation of $L$ enables us to do so, as $\ell \sim \ell' \leftrightarrow (\ell, \ell') \in \mathcal{E}$.

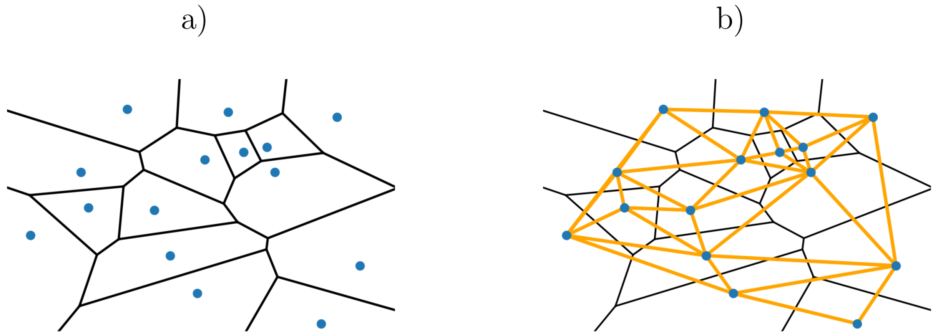


Figure 3.2: a) Voronoi tessellation and b) Delaunay triangulation of 15 spatially uniformly distributed points.

### 3.2.2 Conditional Independence

Two random variables are called *independent* if their joint distribution factorizes into marginal distributions. The concept of conditional independence involves a third random variable, as we see in the following definition.

**Definition 7:**
Let $A, B$ and $C$ be three random variables. We call $A$ and $B$ *conditionally inde-*

*pendent* given $C$ and write $A \perp\!\!\!\perp B \mid C$ iff

$$f(A, B|C) = f(A|C)f(B|C)$$

The joint conditional distribution of $A$ and $B$ given $C$ factorizes into conditional marginal distributions for any value of $C$. It should be noted that two random variables can be conditionally independent but unconditionally dependent.

For our purpose in the context of Gaussian Markov Random Fields, we are interested in verifying conditional independence in undirected graphs (Definition 4). The following theorem, which we state without proof (see e.g. Bishop (2006)), provides such a tool.

**Theorem 2:**
Let $\mathcal{G} = (L, \mathcal{E})$ be an undirected graph and let $A, B$ and $C$ be three disjoint subsets of $L$. A *path in* $\mathcal{G}$ is a sequence of nodes such that every two successive nodes in the sequence are neighbours with respect to $\mathcal{G}$. To see whether $A \perp\!\!\!\perp B \mid C$, it suffices to check whether all paths from any node in $A$ to any node in $B$ pass through a node in $C$.

Restating Theorem 2, when we remove all nodes of $C$ together with any edges connected to $C$ from the graph $\mathcal{G}$, the conditional independence property holds if no path from $A$ to $B$ exists in the resulting graph.

For a single node in an undirected graph, it is straightforward to see that conditioning on all its neighbours, it is independent from all other points as illustrated in Figure 3.3. This is one of the pillars for the theory of Gaussian Markov Random Fields as we will see in the next section.

### 3.2.3 Definition of Gaussian Markov Random Fields

A *Gaussian Markov Random Field (GMRF)* is defined on a graph $\mathcal{G} = (L, \mathcal{E})$. Each node (or location) $\ell_i \in L$ with $i = 1, \ldots, n$ is assigned a random variable $Y(\ell_i)$. For simplicity in the following we write $Y_i := Y(\ell_i)$ and $\boldsymbol{Y_{-i}} := \{Y(\ell_1), \ldots, Y(\ell_{i-1}), Y(\ell_{i+1}), \ldots, Y(\ell_n)\}$. Let $\mathcal{N}_i$ denote the set of neighbouring

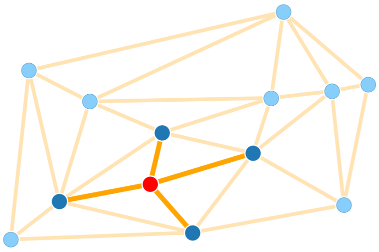


Figure 3.3: Conditional independence property for undirected graphs. When conditioning on its neighbours, the red point is conditionally independent on the remaining points (light blue): There exists no path from the red point to a light blue point without passing through any of the red point's neighbours.

nodes of $i$ with respect to $\mathcal{G}$ and $j \sim i$ represents that $\ell_j$ and $\ell_i$ are neighbours. For a GMRF the following conditional independence property and the Gaussian distribution hold

$$\big(Y_i | \boldsymbol{Y_{-i}} = \boldsymbol{y_{-i}}\big) \overset{d}{=} \big(Y_i | Y_j = y_j, j \in \mathcal{N}_i\big) \sim N\left(\mu_i + \alpha \sum_{j \in \mathcal{N}_i} b_{ij}(y_j - \mu_j), \tau_i\right) \tag{3.2}$$

whereby $\mu_i$ are respective marginal means parameters, $\alpha$ is a spatial correlation parameter, the $b_{ij}$'s determine the weights of the spatial influence with $b_{ii} = 0$, and $\tau_i$ are the marginal conditional variances. Here, $\overset{d}{=}$ denotes equality in distribution.

According to (Besag, 1974) the joint distribution is then uniquely specified as

$$\boldsymbol{Y} = \big(Y_1, Y_2, \ldots, Y_n\big) \sim N\Big(\mu, (I - \alpha B)^{-1} M\Big) \tag{3.3}$$

$\mu = (\mu_1, \ldots, \mu_n)$ is the vector of the marginal means, $I$ is a $n \times n$ identity matrix and $M$ is an $n \times n$ diagonal matrix with marginal conditional variances $\tau_i$ on the diagonal. $B$ is an $n \times n$ weight matrix with entries $b_{ij}$ controlling the impact of the neighbouring values (see Equation (3.2)). In order to define a proper probability distribution, the elements of $B$ must satisfy certain conditions:

i) the rows must sum up to 1, i.e. $\sum_j b_{ij} = 1$

ii) $b_{ij}\tau_j = b_{ji}\tau_i$ to ensure symmetry of the variance-covariance matrix

iii) $b_{ii} = 0$

iv) $b_{ij} = 0 \Leftrightarrow i \not\sim j$

Rewriting the covariance matrix $(I - \alpha B)^{-1} M = \left(M^{-1}(I - \alpha B)\right)^{-1}$, we notice that the covariance matrix may be specified through its inverse, the precision matrix
$Q := M^{-1}(I - \alpha B)$. Because $M^{-1}$ is a diagonal matrix and thus, $M^{-1}I$ is diagonal, the off-diagonal entries only depend on $\alpha M^{-1} B$. Furthermore, with the constraint iv) on the entries of $B$, it follows

$$Q_{ij} = 0 \quad \Leftrightarrow \quad \ell_i \not\sim \ell_j, i \neq j$$

In other words, whether $\ell_i$ and $\ell_j$ are neighbours can be directly inferred from the precision matrix $Q^{-1}$ and vice versa. If the locations $\ell_i$ and $\ell_j$ are not neighbours, and thus $Y_i$ and $Y_j$ are independent given $\boldsymbol{Y_{-ij}}$, the respective entry in the precision matrix is 0. This is also true vice versa, if $Q_{ij} = 0$ then the according locations are not neighbours and $Y_i$ and $Y_j$ are conditionally independent.

This close connection between the underlying graph of a Gaussian Markov Random Field and its precision matrix accounts for its computational advantages: with only small numbers of neighbours for example defined through a Delaunay triangulation (Section 3.2.1), the precision matrix becomes sparse.

### 3.2.4 CAR Models

As introduced earlier the weight matrix $B$ must fulfil constraints i) – iv) to obtain a proper definition for the probability distribution. These weights can be determined in several ways and define specific types of a GMRF; for example, it can be a function of distance, where the function value decreases with increasing distance of two points normalized by the row sum of the weight matrix.

A popular way in spatial statistics is to set the mean of the conditional distribution as the average of the value of the neighbouring nodes. This leads to a model known as *conditional autoregressive (CAR)* model and is specified by

$$b_{ij} = \begin{cases} 0, & i \not\sim j, i \neq j \quad \text{or} \quad i = j \\ \frac{1}{|\mathcal{N}_i|}, & i \sim j \end{cases}$$

with the choice of conditional marginal covariance

$$\tau_i := \frac{1}{\sigma^2 |\mathcal{N}_i|}$$

$\sigma^2$ is a shared parameter among all nodes and $|\mathcal{N}_i|$ represents the number of elements in $\mathcal{N}_i$. The formulation of $\tau_i$ is functional, because we assume that having more neighbours leads to more information on the estimation of $Y_i$ and thus, a higher precision. It can be easily checked that constraints i) – iv) are satisfied using this choice of weights and conditional marginal covariance. Without loss of generality, in the following we set $\boldsymbol{\mu} = 0$. The full conditional distribution specification for the CAR model can be written as follows

$$\left(Y_i | \boldsymbol{Y_{-i}} = \boldsymbol{y_{-i}}\right) \sim N\left(\frac{\alpha}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} y_j, \frac{1}{\sigma^2 |\mathcal{N}_i|}\right) \tag{3.4}$$

In order to express the joint probability distribution as in Equation (3.3), we first simplify the precision matrix $Q = M^{-1}(I - \alpha B)$. Let us rewrite the diagonal matrix $M^{-1} = \sigma^2 D$, where $D$ is a diagonal matrix with $|\mathcal{N}_i|$ in the $i$-th row. Similarly, the weight matrix $B$ can be expressed through the product $D^{-1}A$, whereby $A$ denotes the adjacency matrix with entries $a_{ii} = 0, a_{ij} = 1 \Leftrightarrow i \sim j$ and 0 else. This is true because the inverse of $D$ scales the adjacency matrix such that the row sum is equal to 1.

It follows for the precision matrix:

$$\begin{aligned} Q &= M^{-1}(I - \alpha B) \\ &= \sigma^2 D(I - \alpha D^{-1} A) \\ &= \sigma^2 (D - \alpha D D^{-1} A) \\ &= \sigma^2 (D - \alpha A) \end{aligned} \tag{3.5}$$

Consequently, the joint distribution of the CAR model becomes

$$\boldsymbol{Y} \sim MVN\Big(0, \big(\sigma^2(D - \alpha A)\big)^{-1}\Big)$$

So far with the Markov property defined on the network we conditioned only on adjacent neighbours of a particular location. Such a model is called a first-order conditional autoregressive model. This can of course be extended to a model of higher order neighbours (which includes conditioning on the neighbours of the neighbours and their neighbours and so forth) resulting in a $k$-order CAR model.

### 3.2.5 Sampling from a GMRF

Gaussian Markov Random Fields are multivariate normal distributions. In a similar fashion to Algorithm 3, we can sample from the Gaussian distribution which now is defined by its precision matrix.

**Algorithm 6** Sampling from a multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, Q^{-1})$

```
1: $Q$: $n \times n$ precision matrix, $\boldsymbol{\mu}$: $1 \times n$ vector
2: $L \leftarrow \text{CholeskyDecomposition}(Q)$
3: for $j = 1, 2, \ldots, n$ do
4:     Draw a sample $x^{(j)}$ from $\mathcal{N}(0, 1)$
5: end for
6: $\boldsymbol{x} \leftarrow (x^{(1)}, x^{(2)}, \ldots, x^{(n)})^T$
7: Solve $L^T \boldsymbol{z} = \boldsymbol{x}$ for $\boldsymbol{z}$
8: Store $\boldsymbol{y}^{(i)} \leftarrow \boldsymbol{\mu} + \boldsymbol{z}$
```

Because $L^T$ is an upper triangular matrix, solving the system of linear equations in line 7 of Algorithm 6 can be done through backwards substitution. This algorithm yields samples from the desired distribution:

$$\text{Cov}[\boldsymbol{y}] = \text{Cov}[L^{-T}\boldsymbol{x}] = L^{-T}\text{Cov}[\boldsymbol{x}]L^{-1} = L^{-T}IL^{-1} = (LL^T)^{-1} = Q^{-1}$$

Analogously to Algorithm 3, the bottleneck of the overall computational complexity is the matrix decomposition in line 2. In general the Cholesky decomposition has complexity $O(n^3)$, however a sparse precision matrix can reduce this to $O(n^{3/2})$ for spatial GMRFs (see Section 2.4 of Rue and Held (2005) for algorithmic details). For repetitive sampling when drawing inference on the GMRF, line 2 has to be executed for each sample separately because the parameters of the precision matrix are random variables themselves and are updated as well.

The question arises, is there a way to draw repetitive samples from a GMRF that circumvents the Cholesky decomposition and that is computationally more efficient? The answer to this is yes.
Instead of sampling directly from the Gaussian distribution, we make use of MCMC inference techniques. In the following we will show how we can efficiently evaluate the log-density of a Gaussian distribution while fully exploiting the sparse structure of the precision matrix $Q$. This is particularly suitable within an inference framework such as the Python library PyMC3 as it is easy to implement and makes use of the integrated NUTS sampler (Joseph, 2016).

We are interested in calculating

$$\log(f(\boldsymbol{Y} \mid \sigma, \alpha)) = \frac{1}{2}\left[\log\left(\det(Q)\right) - \boldsymbol{Y}^T Q \boldsymbol{Y}\right] + \text{const}$$

In order to facilitate the computation of the determinant of $Q$ we reformulate with respect to the notation of Equation (3.5):

$$\begin{aligned}
\log\left(\det(Q)\right) &= \log\left(\det(\sigma^2(D - \alpha A))\right) \\
&= \log\left((\sigma^2)^n \det(D - \alpha A)\right) \\
&= n\log(\sigma^2) + \log\left(\det(D - \alpha A)\right)
\end{aligned}$$

Further, Jin et al. (2005) showed that

$$\det(D - \alpha A) = c\prod_{i=1}^{n}(1 - \alpha\lambda_i)$$

with $\lambda_1, \ldots, \lambda_n$ being the eigenvalues of $D^{-\frac{1}{2}}AD^{\frac{1}{2}}$, and $c$ a multiplicative constant. We can then simplify

$$\begin{aligned}\log\left(\det(D-\alpha A)\right) &= \log\left(c\prod_{i=1}^{n}(1-\alpha\lambda_i)\right)\\ &= \log(c)+\log\left(\prod_{i=1}^{n}(1-\alpha\lambda_i)\right)\\ &= \log(c)+\sum_{i=1}^{n}\log(1-\alpha\lambda_i)\end{aligned}$$

After dropping additive constants, we are left with calculating

$$\frac{1}{2}\left[n\log(\sigma^2)+\sum_{i=1}^{n}\log(1-\alpha\lambda_i)-\boldsymbol{Y}^TQ\boldsymbol{Y}\right]$$

For repetitive sampling we only need to calculate the eigenvalues of $D^{-\frac{1}{2}}AD^{\frac{1}{2}}$ once. This is because the adjacency matrix $A$ and the diagonal matrix $D$ are independent of the parameters $\alpha$ and $\sigma^2$. This can be done ahead of time with cubic complexity for dense matrices by means of the Jacobian eigenvalue algorithm and thus at most cubic for sparse matrices.

The complexity of evaluating the log-density is of order $O(n)$: for the multiplication of $\boldsymbol{Y}^TQ\boldsymbol{Y}$ we exploit the sparse structure of $Q$; multiplying a vector with a sparse matrix with $n$ non-zero entries requires $n$ multiplications and thereafter $n-1$ summations. Thus, the computation of $\boldsymbol{Y}^TQ\boldsymbol{Y}$ can be done in linear time. Additionally, calculating $n\log(\sigma^2)$ and $\sum_{i=1}^{n}\log(1-\alpha\lambda_i)$ can each be done in at most linear time.

Embedding the log-density evaluation into the HMC framework leads to an overall complexity of $O(n^3+dn^{5/4})$ for $d$ samples (Hoffman and Gelman, 2014). Compared to $O(dn^3)$ for sampling from a Gaussian process, this signifies a major speed-up in terms of the computational complexity.

### 3.2.6 An intuition on why GMRF's represent GP's well

Consider a 2 dimensional map that represents the heat of a surface and assume that heat only flows through this surface and not beneath or above it. This can be modelled through a Gaussian process. Here, the covariance function may describe the conductivity of the surface in a homogeneous, but also a nonhomogeneous sense. Figure 3.4 a) gives an example of such a map.

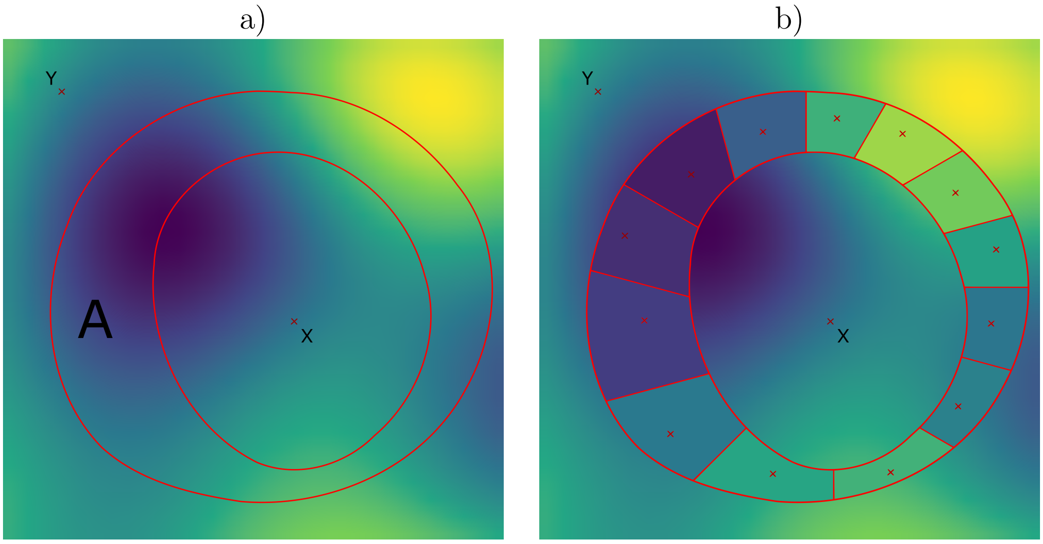


Figure 3.4: Intuition on the GMRF's representation of GPs. The points $X$ and $Y$ are conditionally independent given the area $A$ for a) a continuous heat surface and b) a tessellation thereof.

Since the locations $X$ and $Y$ are spatially separated through area $A$, it is intuitively clear that $X$ and $Y$ are conditionally independent given $A$. In our context, this supposes that there is no heat radiation able to cross area $A$.

Next, we partition $A$ into smaller areas and choose one representative location per segment as in Figure 3.4 b). Depending on the fineness of the partitioning, $X$ and $Y$ exhibit more or less approximate independence given only the representatives of $A$'s tessellation. For example, if the correlation of two locations decays only slowly with distance, the segments can be chosen relatively large.

Similarly, two random variables in a Markov Random Field are conditionally independent given a set of other random variables that separate them from each other. Thus, the above discretization of GP's through representatives with respect to a partitioning, yields a GMRF. Here, the Voronoi tessellation is the appropriate tool to define the segments that make up the partition based on a given set of locations.

## 3.3 Simulation study: GMRFs in the epidemiological model framework

In this section we investigate the model's behaviour when representing Gaussian processes by GMRFs and when then scaling up the population size. We compare the prediction with Gaussian processes to the prediction with GMRFs and afterwards analyze the inclusion of the GMRFs into the epidemiological modelling framework first on a small population and finally on a larger population. We conclude this simulation study with a discussion.

### 3.3.1 Estimation of a Gaussian process using a GMRF

In this section we compare how well the prediction obtained by a GMRF can represent a Gaussian process risk surface. For this reason, we take Outbreak h) as an example (Section 2.7.3) whose random spatial effect was simulated using a Gaussian process with a Matern(3/2) covariance function. We predicted the risk surface with a GMRF. In Figure 3.5 the comparison of a) the original simulated plot, b) its prediction via a Gaussian process, and in c) and d) its prediction via a GMRF is displayed. For comparability of the risk surface predictions, we chose the same prior distributions for the remaining epidemic parameters in all of these settings.

The Voronoi tessellation of the space by the GMRF c) gives a more coarse discretization than the prediction by the Gaussian process b). Compared with the true risk surface, in c) we identify the high risk areas at the bottom left quadrant and the lower risk in the top left corner. From the single high risk areas only the bottom left corner was identified. This is particularly pronounced when plotting only the extremes, that is only the areas with a probability of more than 80% of lying above the mean value 0, i.e. $\{\mathbf{s} \mid \mathbb{P}(\mathbf{s}) > 0.8\}$. The computational benefits are particularly pronounced: the inference using the Gaussian process took 3 hours 4 minutes, the GMRF only needed 1 hour 27 minutes on an Intel i7 4x2Ghz CPU with 8GB RAM. While the computation time is halved for this example, it will be scaled down further with the sparse structure of the GMRF considering an increased population size. The exact scaling depends on the number of individuals

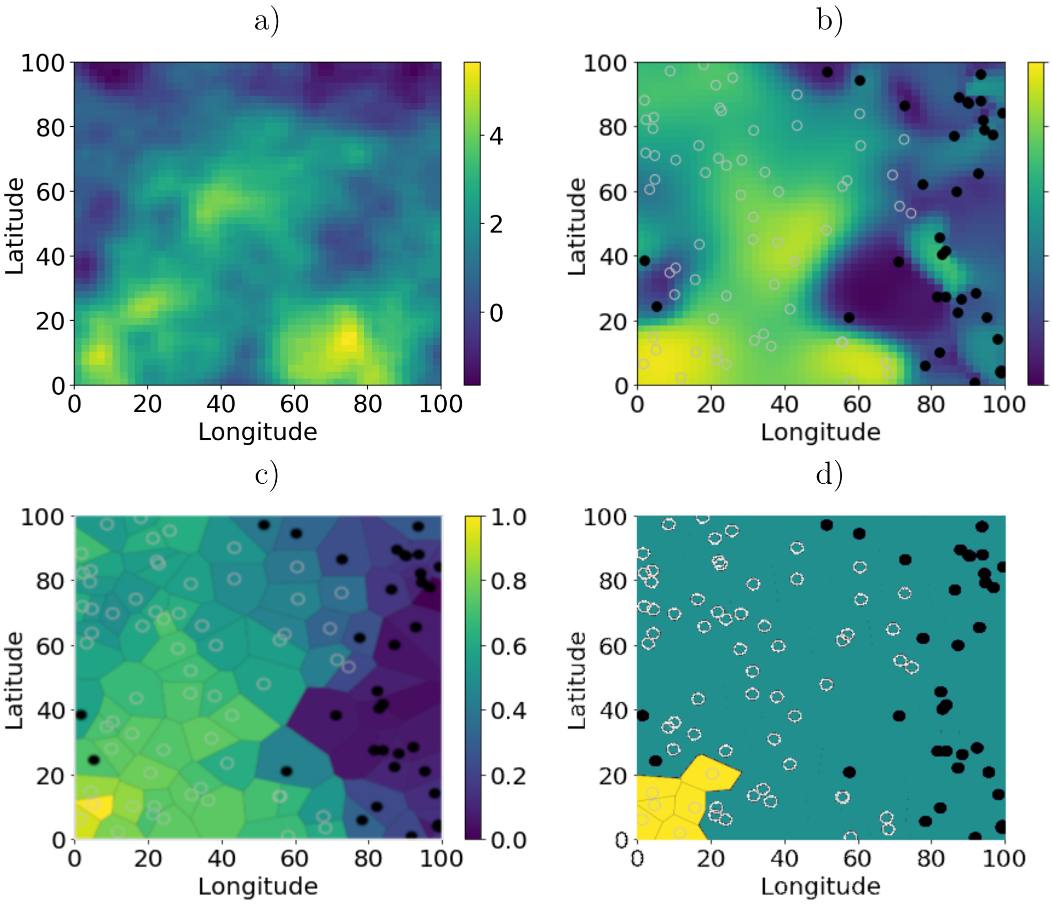


Figure 3.5: Comparison of the a) simulated random spatial risk surface based on a Gaussian process from Outbreak h) with the predicted risk surfaces from different inference scenarios. b) Gaussian process parameterized by $\rho \sim$Gamma(200, 20), $\sigma^2 \sim U(0, 10)$, c) GMRF with $\sigma^2 \sim$ Gamma(2,2) and fixed $\alpha = 0.9$. d) same inference as c) with prediction of $\{\mathbf{s} \mid \mathbb{P}(\mathbf{s}) > 0.8\}$. In b)-d) we used N(0,2) priors for the fixed effect coefficients $\beta_0, \beta_1$ and Gamma(0.1, 0.1) else.

and their spatial location.

In this example we used a first-order conditional autoregressive model for inference. Increasing the dependence on further neighbours in the model may improve the prediction of the GMRF. However, we stick to the approach of depending on the first order neighbours (Section 3.2.4) in order to make full use of the sparsity and the computational benefits of using GMRFs, in particular with the larger population sizes in the following sections. Thus, we conclude that the GMRF is still able to identify higher risk regions, though with less detail than the Gaussian process but with the gain of computational resource usage.

### 3.3.2 Inference behaviour of GMRFs

We now explore the model's sensitivity when integrating GMRFs into the epidemiological model framework. Therefore, we try to recover the spatial risk (prediction) when having simulated from the model. We do not consider misclassification of the model nor prior choices of the epidemic parameters since this was already done in the previous Section 2.7.2.

In Outbreak i) (Table 3.1) we simulated all three risk surfaces (with a random effect and two fixed effects) from a GMRF using $\sigma^2 = 0.1$ and $\alpha = 0.99$ (see Equation (3.4)). We selected high spatial correlation (Figure 3.1) in order to mimic the risk surfaces generated by the Gaussian process in Chapter 2. All three risk surfaces have about the same range. The joint spatial risk surface is therefore determined by the coefficient of the fixed effect covariate ($\beta_0 = 3$). The outbreak has a steep increase in number of infected individuals in the beginning, which coincides with the location of the high risk regions of the joint spatial risk surface.

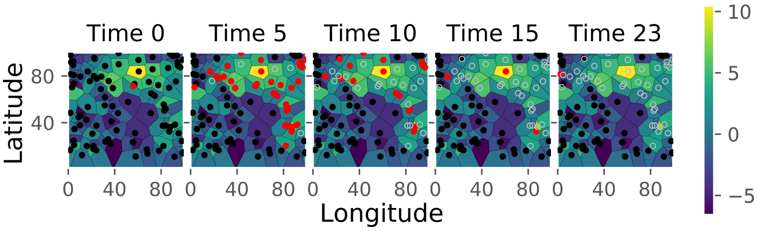
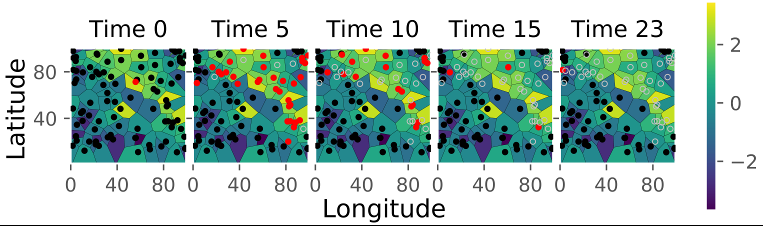
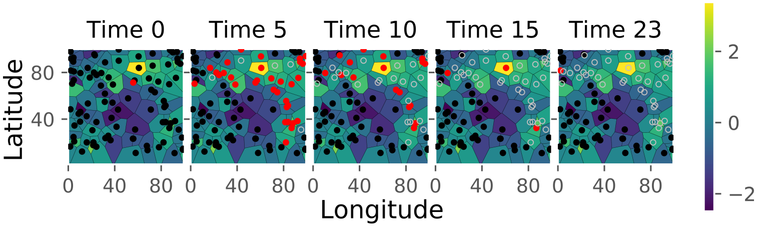
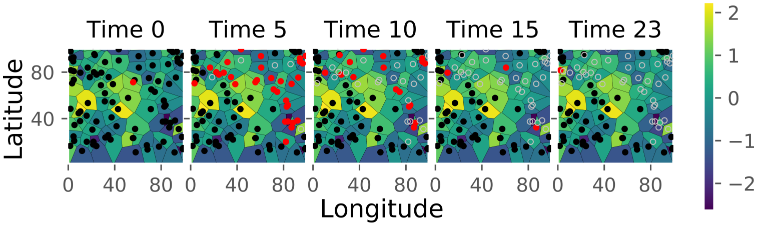
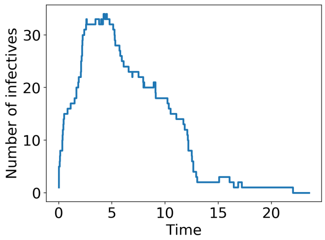
| | Outbreak i) |
|---|---|
| Joint<br>spatial risk |  Time 0<br>Time 5<br>Time 10<br>Time 15<br>Time 23<br>Latitude<br>80<br>40<br>0 40 80<br>Longitude<br>10<br>5<br>0<br>−5 |
| Only<br>random effect<br>risk |  Time 0<br>Time 5<br>Time 10<br>Time 15<br>Time 23<br>Latitude<br>80<br>40<br>0 40 80<br>Longitude<br>2<br>0<br>−2 |
| Only<br>fixed effect<br>$X_0$ |  Time 0<br>Time 5<br>Time 10<br>Time 15<br>Time 23<br>Latitude<br>80<br>40<br>0 40 80<br>Longitude<br>2<br>0<br>−2 |
| Only<br>fixed effect<br>$X_1$ |  Time 0<br>Time 5<br>Time 10<br>Time 15<br>Time 23<br>Latitude<br>80<br>40<br>0 40 80<br>Longitude<br>2<br>1<br>0<br>−1<br>−2 |
| Number of<br>infectives |  <br>Outbreak size:<br>43 |

Table 3.1: Overview of simulated epidemic outbreak i). Black and red points represent susceptible and infected individuals respectively, grey circles denote removed individuals.

| | $\kappa$ | $\gamma$ | $\overline{\text{IR}}$ | $\beta_0$ | $\beta_1$ | $\sigma^2$ | $\alpha$ | $\bar{\text{s}}$ |
|---|---|---|---|---|---|---|---|---|
| **Outbreak i)**<br>True parameters | 0.3 | 0.143 | 7.006 | 3 | 1 | 0.1 | 0.99 | 0.038 |
| $\sigma^2 \sim$ Gamma(2,2)<br>Fixed $\alpha = 0.99$,<br>Fixed $\beta_0 = 3, \beta_1 = 1$ | 0.367 | 0.144 | 6.993 | - | - | 0.398 | - | 0.160 |
| $\sigma^2 \sim$ Gamma(2,2)<br>$\alpha \sim U(0,1)$<br>Fixed $\beta_0 = 3, \beta_1 = 1$ | 0.367 | 0.126 | 7.970 | - | - | 0.341 | 0.988 | 0.615 |
| $\sigma^2 \sim$ Gamma(2,2)<br>Fixed $\alpha = 0.99$ | 0.305 | 0.143 | 7.047 | 2.528 | 0.250 | 0.959 | - | 0.099 |
| $\sigma^2 \sim$ Gamma(2,2)<br>$\alpha \sim U(0,1)$ | 0.342 | 0.136 | 7.389 | 2.981 | 0.445 | 0.451 | 0.986 | 0.474 |

Table 3.2: Mean posterior estimates for simulated Outbreak i) from Table 3.1.If not otherwise specified, we chose $\beta_0, \beta_1 \sim N(0, 2)$ and Gamma(0.1, 0.1) priors for the remaining parameters.

#### 3.3.2.1 Sampling performance

The sampling behaviour is analogous to the description in Chapter 2 where we displayed the characteristics of the joint model. Adding the shifting of the infection times and more parameters to the inference, slowed down the sampler as observed before, yet the performance is reasonable.

In the first inference scenarios, we set the coefficients of the fixed effects to their true values and concluded that all parameters were estimated reasonably well, except $\sigma^2$ was overestimated (see Table 3.2).

Several model parameters account for the spatial variation as observed before in Section 2.7.1.1. Thus, non-identifiabilities between the parameters $\sigma^2, \alpha, \beta_0, \beta_1$ and $\kappa$ were expected. The scale up and the resulting cut off of the spatial kernel might have stabilized the estimation of its parameter $\kappa$.

The additional estimation of the spatial correlation parameter $\alpha$ influenced the overall parameter estimation slightly but not drastically. $\alpha$ itself was estimated very close to the true value with small credible regions despite the uninformative uniform prior (e.g. 0.988 [0.955, 0.999]) with true value 0.99 when fixing $\beta_0$ and $\beta_1$). Although the sampling was stabilized more by fixing $\alpha$, the estimation of $\beta_0$ and $\beta_1$ was better without fixing $\alpha$. This again can be attributed to the non-identifiability between the parameters.

We further tested the robustness of the model towards changing $\sigma^2$. A larger $\sigma^2$ in the simulation leads to a decreased range of the spatial random effect, e.g. 1.3 for $\sigma^2 = 5$ compared to 7 with $\sigma^2 = 0.01$ in Outbreak i). While with a Gamma(2,2) prior on $\sigma^2$ the larger $\sigma^2$ parameter was underestimated (0.8), the other parameters were estimated reasonably well. Placing a more informative prior on $\sigma^2$ with mean 5 led to a better estimate 4.7 for $\sigma^2$ as expected. Note, that the posterior estimates of the other parameters did not vary significantly and furthermore, the prediction plots and the outbreak dynamics for both inference scenarios look very similar.

In other simulations with different parameter values, we noticed similar non-identifiability and autocorrelation behaviour. The parameter estimates responded to a change of prior, but not in a drastically.

#### 3.3.2.2 Prediction performance and outbreak dynamics

By means of our inference scenarios, we recover high risk areas of the spatial random risk surface where possible (Figure 3.6), i.e. where we had enough information from the outbreak data. While the high risk area at the top and on the right were captured, the single hot spot in the middle could not be identified due to the lack of infections in that area.

Fixing the covariate coefficients leads to more information about the model which is why the prediction of the high risk areas is more pronounced in b) than when estimating all parameters in d). The small shift of uncertainty in the prediction is pronounced when comparing the regions with a risk of higher than 0.8 and lower than 0.2: in comparison to the plot of the more informed model in c), the high risk area shrinked by a small proportion in e). Nevertheless, in general the prediction plots are very similar confirming that they are robust towards estimating more parameters. The same holds true for a different prior choice of $\sigma^2$, such as a distribution with mean away from the true value, or a more informative prior towards $\alpha$ as the highest and lowest risk areas were still captured.

For the inference note that the spatial random effect has similar hot spots at the top center as the fixed effect $X_0$. Despite this non-identifiability issue and the more pronounced effect by $\beta_0 = 3$, the model was able to capture the risk area on the top for the unknown spatial random effect.

Similarly to the prediction plots that looked similar, the outbreak dynamics were also captured by the model and looked similar for the different inference scenarios.

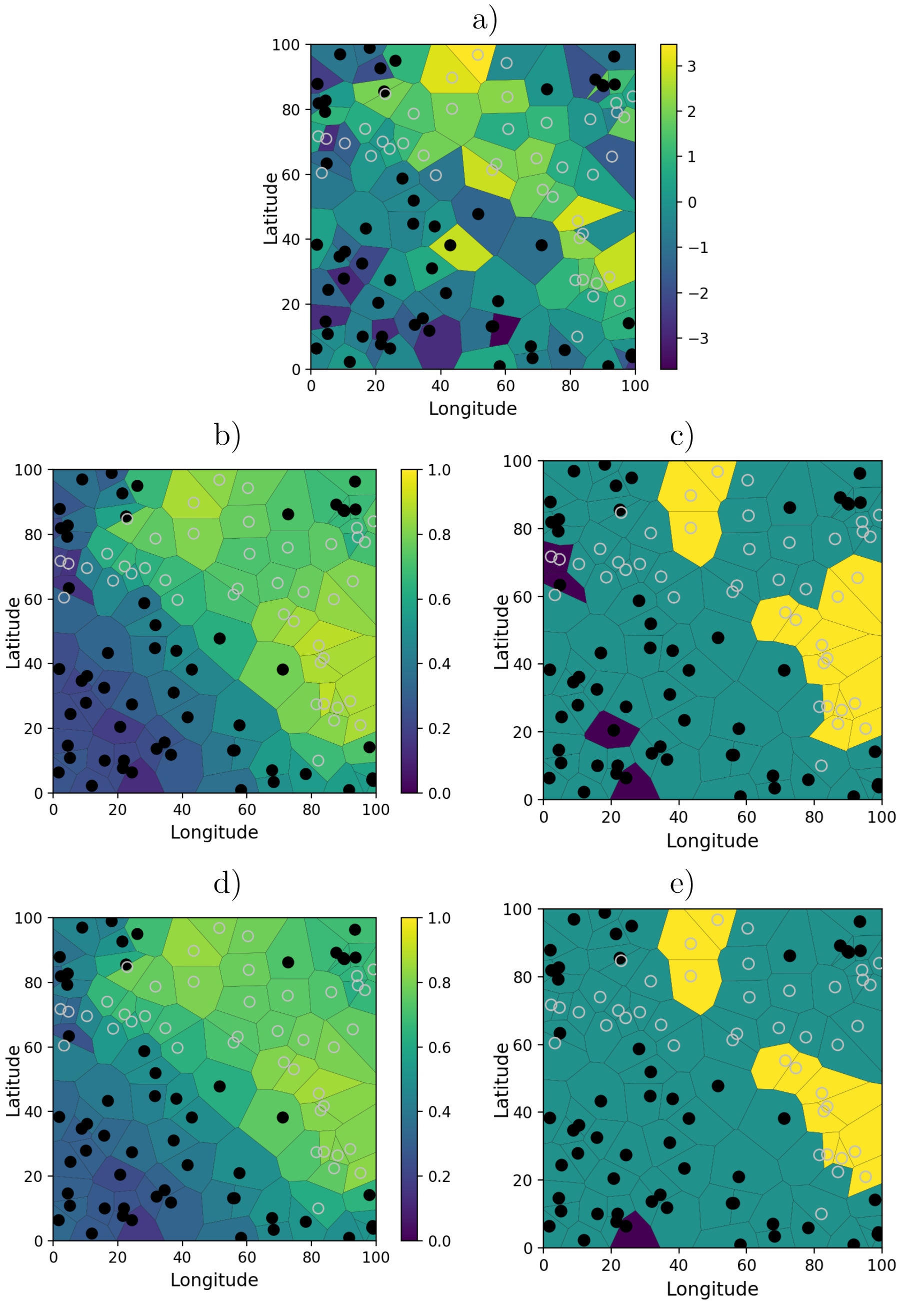


Figure 3.6: Comparison of a) the simulated random spatial risk surface of Outbreak i) with risk surface predictions from inference scenarios b) GMRF with $\alpha$ and $\beta_0, \beta_1$ fixed d) priors $\alpha \sim U(0,1)$ and $\beta_0, \beta_1 \sim N(0,2)$. c) and e) highlight areas of higher (yellow $= \{\mathbf{s} \mid \mathbb{P}(\mathbf{s}) > 0.8\}$) and lower (dark blue $= \{\mathbf{s} \mid \mathbb{P}(\mathbf{s}) < 0.2\}$) posterior probability for b) and d), respectively.

### 3.3.3 Scaling up the population size

So far we have dealt with outbreaks where the outbreak covered most of the observation area. Scaling up the population size with still a relatively small number of infected individuals can influence in particular the spatial spread. In this setting, it is important to consider the background infection pressure $\epsilon$ since it plays a major role in the spatial (distant) spreading of the outbreak. A lower $\epsilon$ leads to an outbreak cluster where the distance between individuals governs the spread, while a larger $\epsilon$ can cause different outbreak clusters at distant random locations.

Thus, in the following, we investigate the influence of the background infection pressure on the inference. For simplicity we only display simulations that were generated with solely a random spatial risk surface and no fixed effects. All the results previously discussed on fixed effects also hold true for an increased population size.

#### 3.3.3.1 Small background infection pressure

We scaled up the population size to spatially uniformly distributed 10,000 individuals and simulated an outbreak with $\epsilon = 0$ denoting no background infection pressure (Outbreak j), Figure 3.8). Note that $\epsilon$ is independent of the introduction of the disease, as we start to model and investigate the outbreaks after the first infection had taken place. The underlying spatial random effect was simulated using a GMRF with parameters $\sigma^2 = 0.05$ and $\alpha = 0.98$ (Figure 3.7).

The final size of the outbreak was 134 and the temporal spread of the outbreak is characterized by two waves as displayed in Figure 3.8. This resulted from the spread first proceeding south and then towards east due to the higher spatial risk. Figure 3.9 shows that the outbreak is characterized by only one outbreak cluster with a high density region of infections leaving the most of the observation area untouched by the disease.

In its nature the Outbreak j) resembles the outbreaks we covered in the previous chapters. However, here we see many non-infected individuals surrounding the outbreak herd. Thus, we performed inference on a smaller area cropped to the outbreak and first used the same inference setting for the full observation area for comparison. The smaller area was chosen such that $\{(X, Y) \mid$

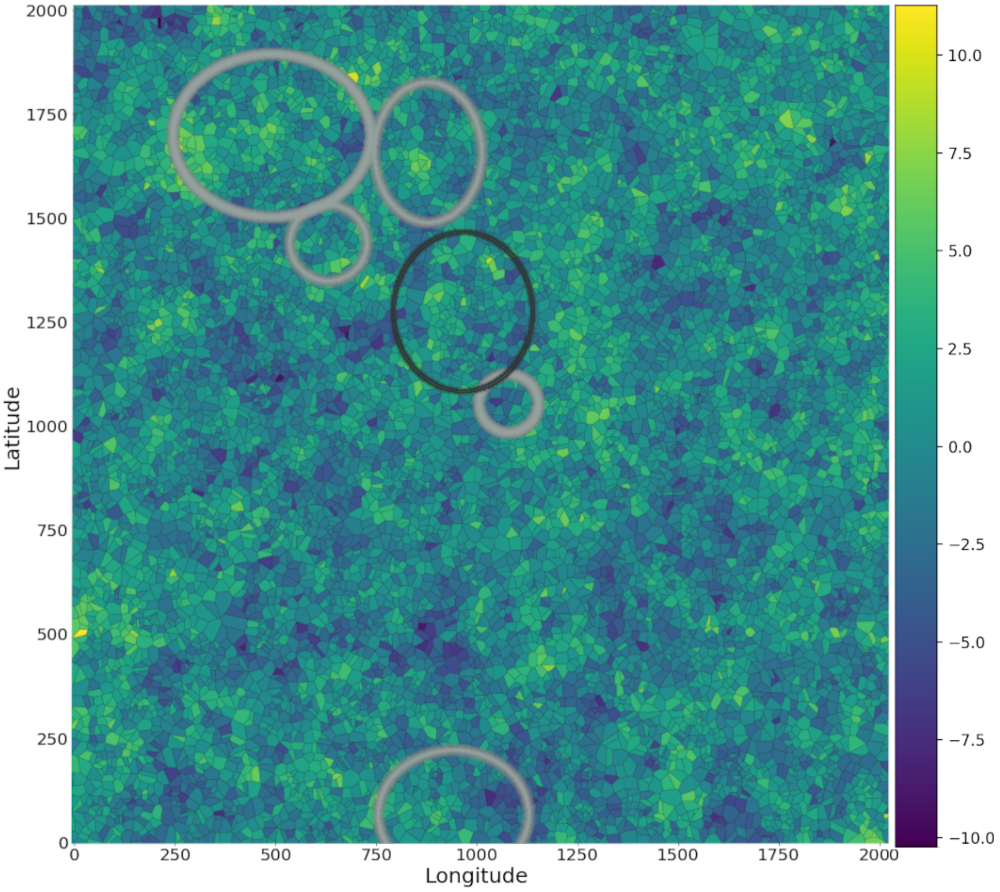


Figure 3.7: GMRF risk surface simulated with $\alpha = 0.98$ and $\sigma = 0.05$ with population size 10,000. The black and the grey circles point towards the location of Outbreak j) and k), respectively.

$500 < X < 1400,\ 750 < Y < 1750\}$ for the longitude and latitude yielding 2251 individuals, that is about a quarter of the whole area. This smaller area is indicated by the rectangle in Figure 3.9.

For both scenarios we chose $\kappa, \gamma \sim$ Gamma(0.1, 0.1), $\sigma^2 \sim$ Gamma(2,2) and $\alpha \sim U(0,1)$ priors.

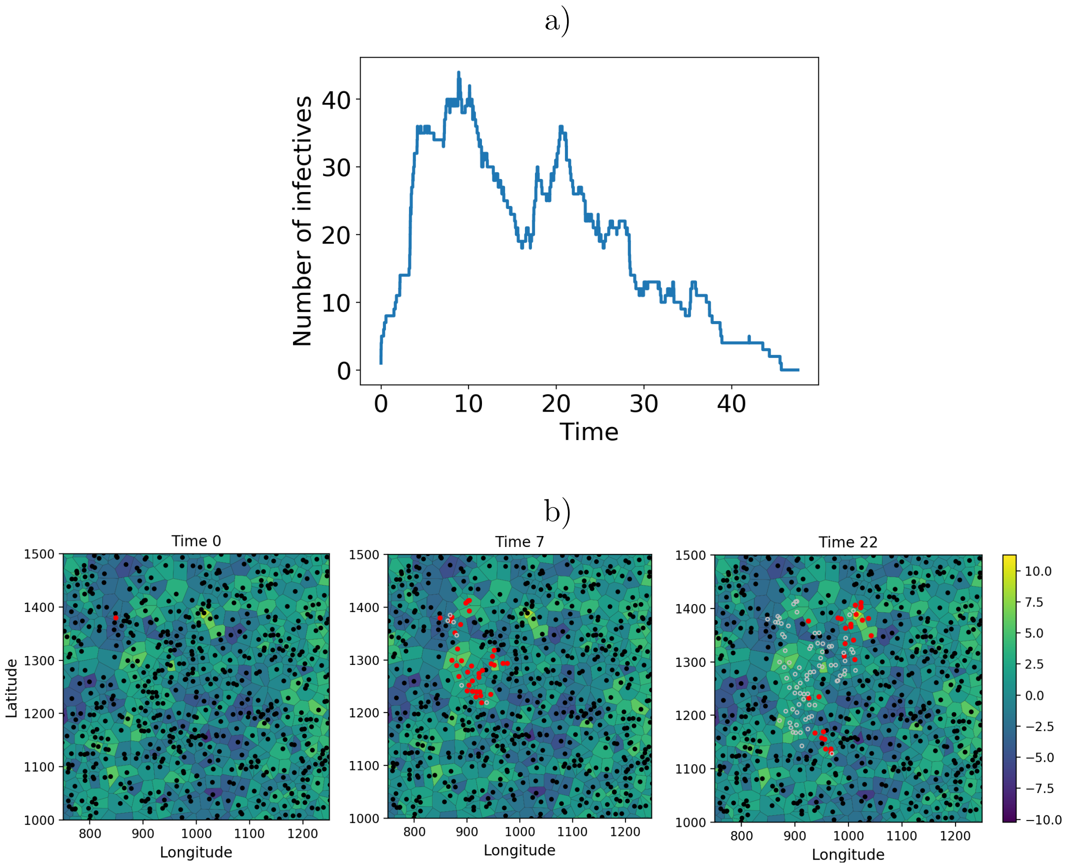


Figure 3.8: Spread of the Outbreak simulation j). a) Temporal spread b) Spatial spread at different time points. The outbreak is depicted on top of the random spatial risk surface and the plots are zoomed in to the area of infection, whereby red points denote infected, black denote susceptible and grey circles detected individuals, respectively.

Inference on the smaller area led to reasonable parameter estimates (Table 3.3) as well as recovered outbreak dynamics (Figure 3.11), as expected based on previous simulation studies. Performing inference with the same settings on the larger population size, the effective sample size worsened but led to similar parameter estimates and comparable high risk regions (Figure 3.10). In order to distinguish between the two scenarios in more detail, in the plot we chose the probability of the posterior mean to be greater than 98%. A threshold of 80% as in figures before would lead to very similar high risk areas for both scenarios. There is not much data on the remaining non-infected area, as the infections feed the model with

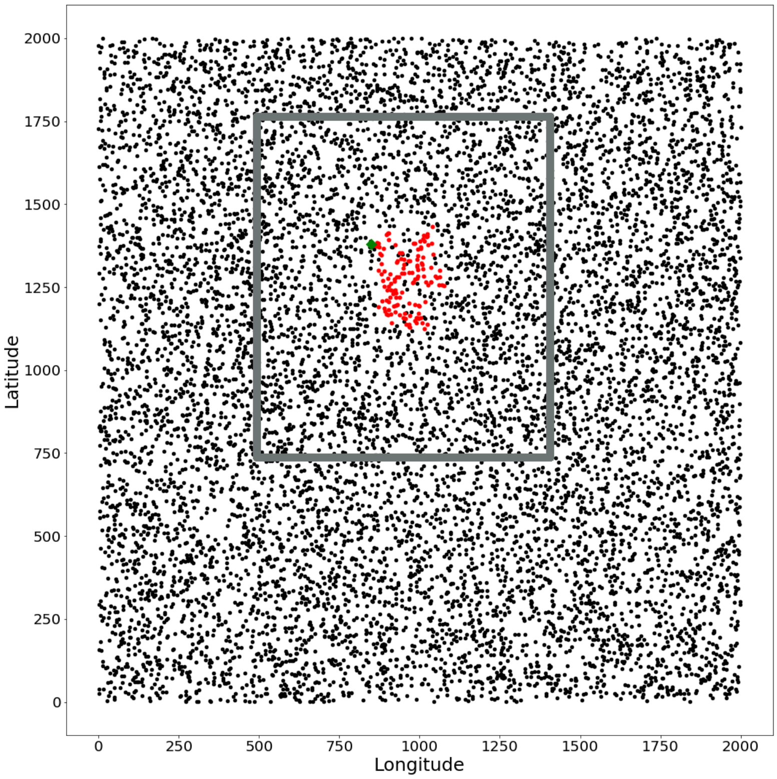


Figure 3.9: Simulated outbreak with 10,000 individuals and small background infection pressure. The red points denote all infected individuals of the outbreak and the green diamond on the top left corner represents the first infected individual. The square denotes the cropped area for inference on a smaller population.

| Parameter | True value | Small area Posterior mean | Large area Posterior mean | Small area ESS | Large area ESS |
|---|---|---|---|---|---|
| $\kappa$ | 0.27 | 0.166 | 0.183 | 262.3 | 89.1 |
| $\gamma$ | 0.142 | 0.143 | 0.140 | 110.0 | 72.1 |
| $\overline{\text{IR}}$ | 6.728 | 6.998 | 7.159 | 68.9 | 37.0 |
| $\sigma^2$ | 0.05 | 0.138 | 0.156 | 186.1 | 37.0 |
| $\alpha$ | 0.98 | 0.999 | 0.999 | 2188.2 | 343.2 |

Table 3.3: Mean posterior estimates and effective sample size for simulated Outbreak j)

most information, which is why the model could not estimate heterogeneity there.

The outbreak dynamics of both population sizes are similar (Figure 3.11) and

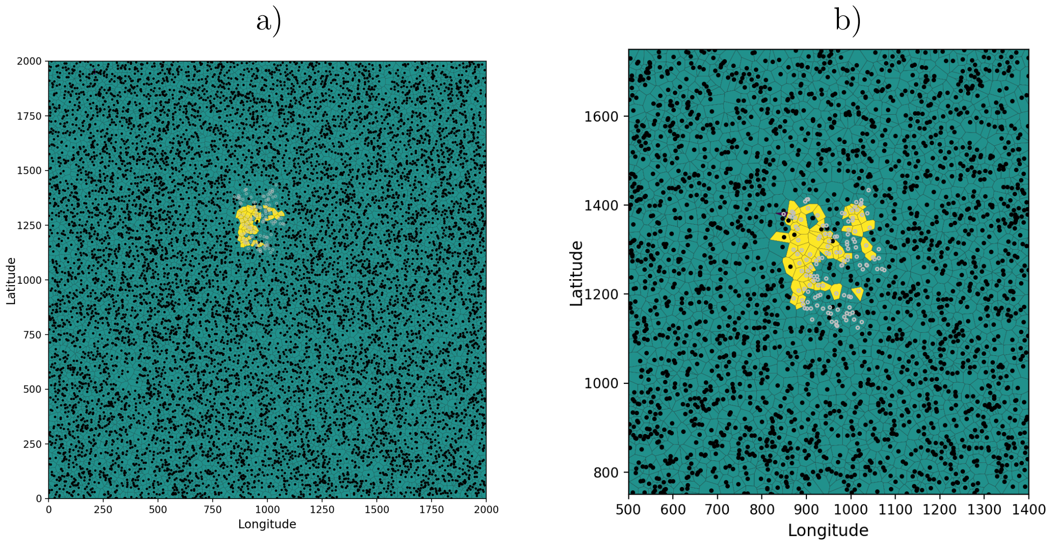


Figure 3.10: Estimated high risk regions $\{\mathbf{s} \mid \mathbb{P}(\mathbf{s}) > 0.98\}$ for Outbreak simulation j). a) Prediction on larger population b) Prediction on smaller population.

matched the simulated outbreak. Due to the fact that the larger dataset enabled a further outbreak spread, there were slightly more outbreaks with a larger final size.

We can conclude, that cropping the area to a region that is smaller to the area of infection stabilizes the inference and saves time and computational resources with comparable results.

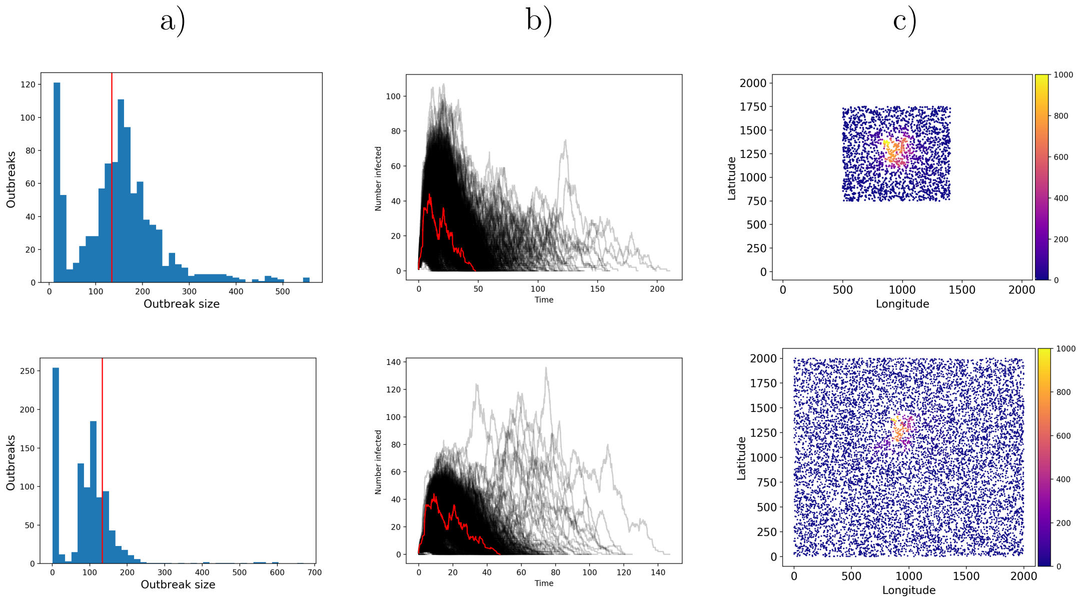


Figure 3.11: Analysis of outbreak dynamics of the MCMC chain for Outbreak j). 1000 outbreaks were simulated from posterior samples. Top row is regarding the smaller population, bottom row regarding the larger population. Column a) displays a histogram of the outbreak size, column b) the temporal progression, and column c) the spatial progression of the outbreak. The red lines refer to the original outbreak simulation.

#### 3.3.3.2 Greater background infection pressure

In this section we investigate inference when using a larger background infection pressure. For this reason we created Outbreak k) on the same observation area with 10,000 individuals and spatial risk surface as in Section 3.3.3.1 (Figure 3.7). The exact simulation parameters are stated in Table 3.4. The simulation had an outbreak of size 187 and led to several spatial outbreak clusters (Figure 3.12). In fact, we can identify five different clusters marked with circles. With $\epsilon = 10^{-6}$, and $\psi = 1$, on average $10^{-6} \times 10000 \times 47 = 4.7$ infections arise due to the underlying infection pressure given the outbreak duration of 47 days. Thus, every cluster after the introductory case presumably started with one infected individual due to the background infectious pressure $\epsilon$. It then evolved locally or died out due to the fact that no other individual nearby was infected, such as for example the single infected individual in the center of the picture for the latter case. In the

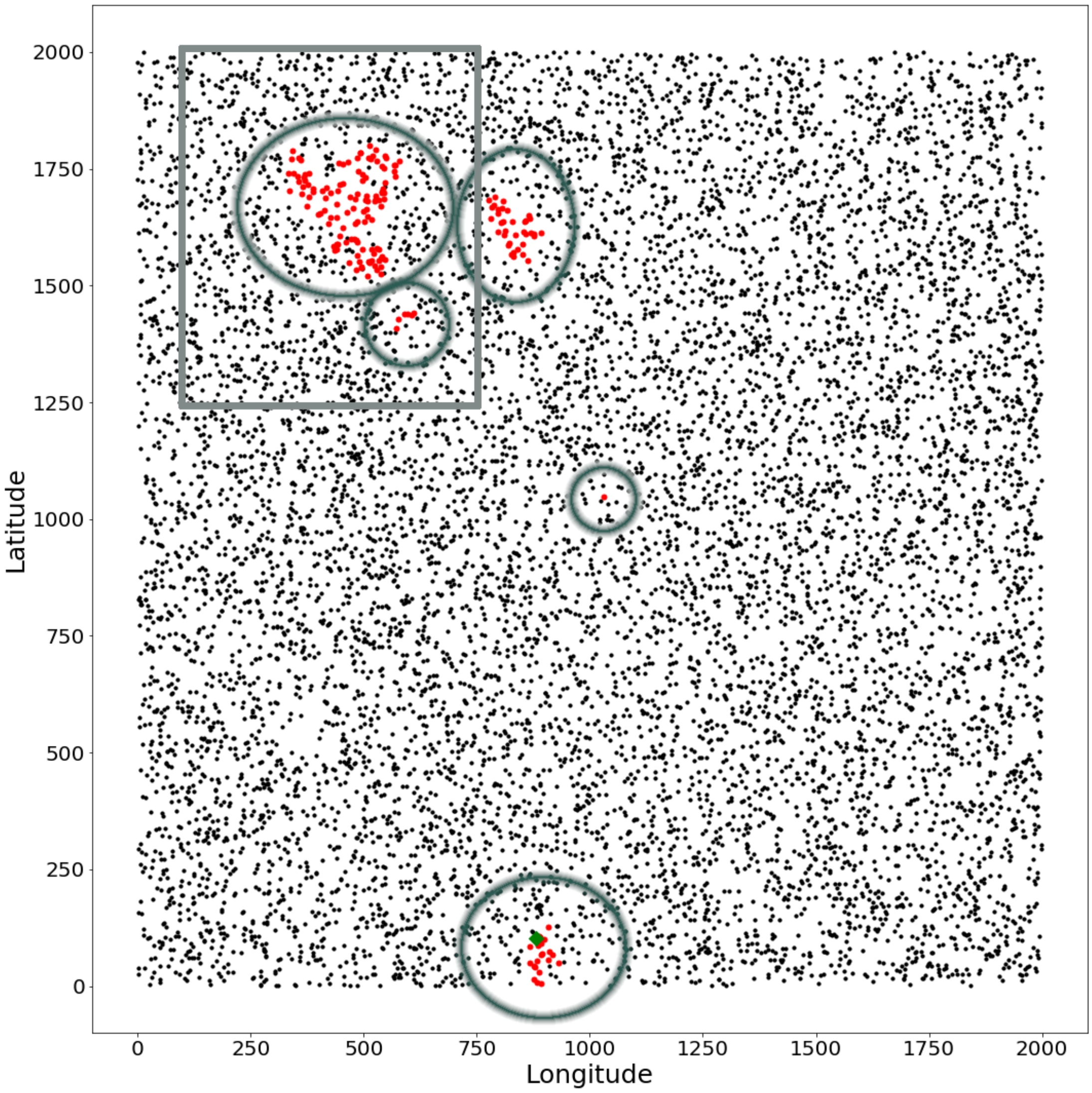


Figure 3.12: The spatial spread for Outbreak k). The green diamond represents the first infected individual. Each circle denotes one outbreak cluster presumably attributable to the background infectious pressure $\epsilon$. The rectangle limits the smaller area for inference.

temporal progression of the outbreak (Figure 3.13), we can make out remarkable rises beginning at around time points 0, 13 and 23, which can be attributed to the three remaining bigger clusters.

From the previous section we learned that the sampler stabilizes when cropping the observation area further towards the infected individuals. Thus, for this out-

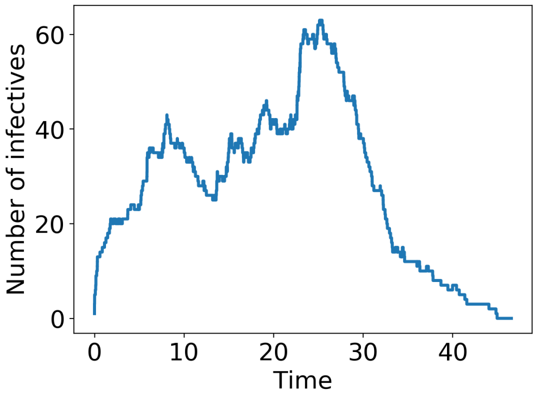


Figure 3.13: The temporal spread of Outbreak k)

| Parameter | True value | Small area<br>Posterior mean | Large area<br>Posterior mean | Small area ESS | Large area ESS |
|---|---|---|---|---|---|
| $\kappa$ | 0.25 | 0.263 | 0.121 | 21 | 19 |
| $\gamma$ | 0.142 | 0.120 | 0.120 | 28 | 43 |
| $\overline{\text{IR}}$ | 6.960 | 8.309 | 8.346 | 13 | 28 |
| $\sigma^2$ | 0.05 | 0.057 | 0.160 | 28 | 18 |
| $\alpha$ | 0.98 | 0.999 | 0.999 | 381 | 56 |

Table 3.4: Mean posterior estimates and effective sample size for simulated Outbreak k)

break we also investigated a smaller area cropped to the two outbreak clusters on the top left, that is $\{(X, Y \mid 100 < X < 770, 1250 < Y < 2000)\}$. This decreased the population size to 1282 out of which 126 were infected.

In Table 3.4 the posterior estimates are displayed. In these scenarios, we fixed the parameter $\epsilon$ to its true value in order to stabilize inference. The two scenarios yield similar reasonable posterior estimates, whereby for the smaller area the posterior estimates are slightly closer to the true values than for the larger area (e.g. $\kappa$ and $\sigma^2$). The larger population size introduces more uncertainty into the model. The small ESS values, in particular for $\alpha$ reflects that the sampler algorithm had difficulties sampling from the joint distribution. A potential approach for the larger area could be to perceive the single outbreak clusters as independent outbreaks with their own dynamics instead of one large outbreak. The likelihood would then be the product of these cluster terms and the parameter estimates would be local. Here, the limitation is that we would have to be able to separate the outbreak clusters clearly.

We furthermore investigated the estimation of the background infectious pressure $\epsilon$. Even an informative Gamma(10, $10^7$) prior slowed down the mixing of the other parameters even further, such that with 50,000 samples there was not a visible convergence of the chain.

The impact of fixing $\epsilon$ to a different value than the true value showed a tendency towards changing the spatial kernel parameter: a too small background infection pressure led to a smaller $\kappa$, thus a less steep decrease of the spatial kernel allowing distant infections to become more likely. On the other hand, a larger background infection pressure led to a higher estimate of $\kappa$, and thus to less infectious pressure in the distance. These results have to be taken with care as the sampling performance again was very slow.

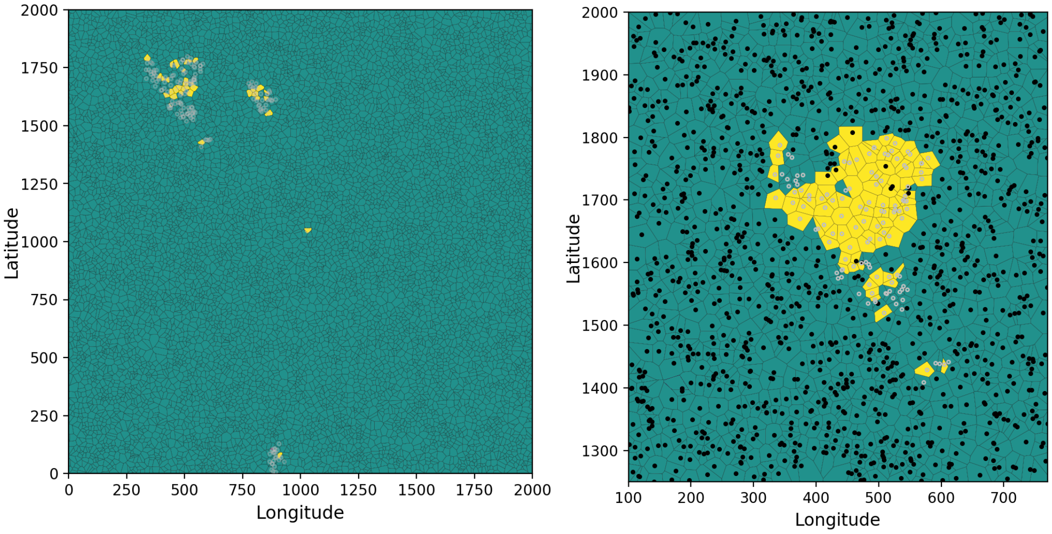


Figure 3.14: Estimated high risk regions $\{\mathbf{s} \mid \mathbb{P}(\mathbf{s}) > 0.8\}$ for Outbreak simulation k). Prediction on the a) larger and b) smaller population.

The prediction in Figure 3.14 displays that in the larger area all five outbreak clusters were captured by the model and high risk regions were identified. The same holds for the smaller area. Comparing both predictions, the larger area shows smaller high risk regions at the same probability level of 80%. This could be attributed to the additional uncertainty introduced when scaling up the population size further.

Considering the dynamics of the outbreak, the background infection pressure can cause local infection foci anywhere in a greater observation area. The exact location, however, determines the further spread and the ongoing spread of the epidemic, e.g. in a more densely populated area, the infection will spread faster than in a area with a lower population density, analogously for high/low spatial risk. The recovery of the spatial and temporal spread cannot be compared well to the original outbreak as they are subject to great variation.

## 3.4 Discussion

In this chapter we explored the integration of Gaussian Markov Random fields (GMRFs) into the epidemiological model framework in order to cope with larger population sizes. Scaling up to a larger population size poses several challenges, such as computationally in terms of memory usage and computation time, in terms of parameter estimation and their identifiability, and in terms of reconstructing the underlying spatial effect (no information for the most area).

We first showed that by means of GMRFs we can represent the main features of Gaussian processes within our epidemiological framework. By the cost of a higher coarseness, we gain computational efficiency through the sparse representation of the precision matrix and the simplified computation of the determinant. One clear limitation of this approach is that we need to find *all* eigenvalues of the matrix $D^{-\frac{1}{2}}AD^{\frac{1}{2}}$. This might run into computational issues for a too large matrix dimension. For example, the calculation of all eigenvalues for a $15,000 \times 15,000$ matrix takes about 15 minutes on an Intel i7 4x2Ghz CPU with 8GB RAM using the python package *NumPy* and its method *linalg.eigvals().* For a matrix of dimension $25,000 \times 25,000$, the computation exceeded resource capacity. In the latter, we would refer the reader to sampling from a GMRF via the sparse Cholesky decomposition from Rue and Held (2005) which is for example provided by the *scikit-sparse* library in Python.

The amount of sparsity of the precision matrix depends on the connectivity of the underlying network. If we choose a larger $k$ in a $k$-th order CAR model, the prediction can be refined, however the methods for sparse matrices might not be fully applicable any more by the cost of the computational advantages.

GMRFs are one possible option to overcome the *Big n problem.* In our scenario it has worked sufficiently well which is why we did not explore further options, such as for example inducing points (Quinonero-Candela and Rasmussen, 2005).

It became clear in the simulation study with a population size of 100, that the parameters were able to be estimated reasonably well, but also the underlying spatial effect was able to be recovered. However, when dealing with a larger population size and a relatively small number of infected individuals (for example less than 2% of the population), we do not have much information about the disease progress or dynamics in most of the area. There, the estimate of the spatial random effect will revert to its mean. As displayed in the simulation studies, it yet provides enough information to capture the outbreak dynamics.

The inference contains a lot of uncertainty which is pronounced in the non-identifiability of the spatial parameters and the slower mixing of the MCMC chain. When scaling up the population size, the sampler has different possibilities of attributing the infectious pressure to individuals far away from the area of infection: one way is to increase the background infection pressure for random appearances of infections within the observation area. Another way would be to scale the spatial kernel in such a way that individuals far away would be also exposed to infectious pressure from currently infectious individuals. To decrease the non-identifiability between the parameters, it is important to have a priori information, such as for example the parametrization and cut off of the spatial kernel. This can furthermore help to stabilize the inference and increase the effective sample size (as discussed in Section 2.8). Assuming a certain level of spatial correlation and fixing $\alpha$ reduces the parameter space and thus can help the sampler to explore the parameter space better. Fixing $\alpha$ to exactly 1 leads to calculating the determinant of a matrix that is not positive definite any more. For this case, a different approach one could consider are intrinsic CAR models (ICAR), though whose joint distribution is improper (Besag et al., 1991).

If possible, it is advisable to scale down the population size as this saves time and computational resources and additionally stabilizes inference. Care has to be given not to crop the observation area too close to the infected herd to avoid edge effects as the model also gets information on individuals nearby not having been infected. The chosen distance could be in line e.g. with the cut off of the spatial kernel to obtain sparsity of the precision matrix for the GMRF.

In summary we can state that including GMRFs into the epidemiological setting is a hard problem, but it can represent Gaussian processes and identify its highest risk regions given that sufficient detailed data on the outbreak and the infected individuals are provided.

# Chapter 4

# Aggregated data collection

This chapter is concerned with the inherent uncertainty in the dataset of the Bluetongue outbreak in Great Britain 2007 (see Section 1.6.3). We observe that at the beginning of the outbreak half of the infected farms were detected simultaneously three days after the first case became known (Figure 1.9). Such a sudden and unusual increase in case numbers suggests that the data does not reflect the underlying truth. One possibility for this phenomenon is that it may result from the increased coverage of testing after the first observed case leading to many more positive farms simultaneously. This however neglects the fact that the virus may have been present on the farm before. An alternative is that the detection date in the data gives the date on which the detection was entered into the database. The time point of detection itself, however, can differ. Therefore, we must assume that the detection date provided in the data set is the latest possible date for detection.

When applying inference methods, such an uncertainty in the data can pose a problem with respect to the reliability of the resulting parameter and outbreak dynamics estimates. In this chapter we will treat the aggregated data as missing data and propose a way to deal with it based on a Bayesian data augmentation approach. Afterwards, we conduct a simulation study to test the limits of how little data is required for good inference on the parameters.

## 4.1 Treating missing removal times

Our epidemiological setting is based on a SIR model. The detection times correspond to the removal times because we assume that as soon as a farm is detected it is removed from the system (Section 2.2). Recall that $\mathbf{R} = \{\mathrm{R}_1, \ldots, \mathrm{R}_n\}$ is the set of the detection times of all $n$ individuals in the population. We denote the reporting time of the cases and thus entering them into the database by $\mathbf{N} = \{\mathrm{N}_1, \ldots, \mathrm{N}_n\}$. The time at which the detection of $i$ is reported, $\mathrm{N}_i$, is left-centered and therefore satisfies $\mathrm{R}_i \leq \mathrm{N}_i$ (detection time before or equal to reporting time).

Assume that for the first $m$ removals $\mathrm{R}_i < \mathrm{N}_i$, that there is a delay between detection and reporting, as displayed in Figure 4.1 scenarios A and B. For the later detections we take $\mathrm{R}_i = \mathrm{N}_i$ (the detection time corresponds to the reporting time, Figure 4.1 scenario C). We do not model a specific distribution for $\mathrm{N}_i - \mathrm{R}_i$ (delay from detection to reporting), but assume reporting kicks in at some time, t say, with detections after time t being reported in a timely manner (ie. when they occur). Until time $t$ we only know that a removal occurred before that time. Thus, for individuals similar to A or B we augment infection and detection time, for individuals similar to C we only augment infection time.

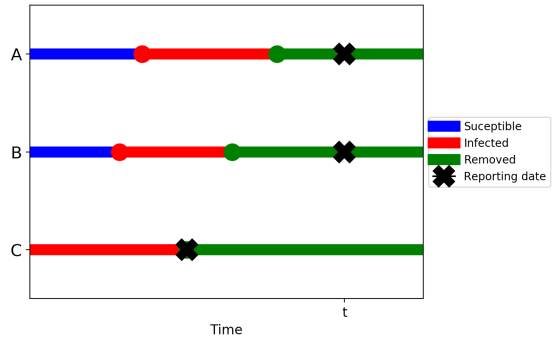


Figure 4.1: Illustration of augmented data. We are given reporting data. Thus, for scenarios A and B we augment both, infection and removal times. For C we only augment infection times, because the reporting date equals the removal date.

Let $\mathbf{R_M} := \{\mathrm{R}_1, \ldots, \mathrm{R}_m \mid \mathrm{R}_1 = \cdots = \mathrm{R}_m\}$ be the removal times of all the $m$ farms that were detected simultaneously on this particular $m$-th day of the outbreak. Then for all $\mathrm{R}_i \in \mathbf{R_M}$ we have $R_i \leq N_i$, while for all $\mathrm{R}_i \in \mathbf{R_T} := \mathbf{R} \backslash \mathbf{R_M}$

we assume $R_i = N_i$. Thus, for the purpose of this thesis we will treat the set $\mathbf{R_M}$ for which the detection time is uncertain as missing data.

Marginalization allows us to deal with missing data while incorporating uncertainty in the removal time. However, in practice this is an infeasible task because it involves summing over all possible removal times of all individuals with removal time in $\mathbf{R_M}$. Therefore we mimic a marginalization by sampling from the joint posterior distribution and eventually neglect the value of the missing data. The joint posterior distribution is given by

$$f(\boldsymbol{\theta}, \mathbf{I}, \mathbf{R_M} \mid \mathbf{R_T}) \propto f(\mathbf{R_T} \mid \boldsymbol{\theta}, \mathbf{I}, \mathbf{R_M}) f(\mathbf{R_M} \mid \boldsymbol{\theta}, \mathbf{I}) f(\mathbf{I} \mid \boldsymbol{\theta}) f(\boldsymbol{\theta}) \tag{4.1}$$

McKinley et al. (2014) used such a Bayesian approach to augment removal times within a day, that is to impute exact removal times given daily data. For our purpose, we assume the unknown detection times lie before the recorded reporting date and we are not interested in imputing the exact date. Instead, we want to use this uncertainty in the detection times in order to gain estimates for the model parameters that govern disease transmission.

## 4.2 Augmenting missing removal times

Our aim is to sample from the joint posterior distribution $f(\boldsymbol{\theta}, \mathbf{I}, \mathbf{R_M} \mid \mathbf{R_T})$ (Equation (4.1)) to marginalize out missing detection times.

In order to do so, we adapt Algorithm 4 for making inference on the parameters using a hybrid NUTS-Metropolis within Gibbs algorithm and include an independence sampler to sample the missing removal times $\mathbf{R_M}$ given the current samples of the parameters and the infection times. This is displayed in Algorithm 7.

For updating the missing removal times, one has to think about how to do this in an efficient manner. As the conditional distribution of the removal times given the parameters and the infection times is too complicated to be able to sample directly from it, we make use of Metropolis within Gibbs. Recall that the infectious period is Gamma$(\alpha, \beta)$ distributed. We use an independence sampler by sampling an infectious period $Q^*$ from Gamma$(\alpha, \beta)$ and proposing a new removal time through $\mathrm{R}_j = \mathrm{I}_j + Q^*$. In case the proposed value lies beyond the original date

**Algorithm 7** Sampling from $f(\boldsymbol{\theta}, \mathbf{I}, \mathbf{R_M} \mid \mathbf{R_T})$ using Hybrid NUTS-Metropolis within Gibbs

```
Input: $s$ number of samples, $n_\mathrm{I}$ number of total infections in the outbreak, $d$ prior belief about length of infectious period, $\mathbf{I}_{n\times 1}$ an $n \times 1$ vector of all ones
Initialize: Set $\mathbf{I}^{(0)} \leftarrow (\mathbf{R} - d\mathbf{I}_{n\times 1})$, $\mathbf{R}_\mathbf{M}^{(0)} \leftarrow \mathbf{R_M}$
for $k = 0, \dots, s-1$ do
    Obtain a sample $\boldsymbol{\theta}^{(k+1)}$ from $(\boldsymbol{\theta} \mid \mathbf{R_T}, \mathbf{R}_\mathbf{M}^{(k)}, \mathbf{I}^{(k)})$    ▷ Update parameters
      using the NUTS sampler
    for i $= 1, \dots, n_\mathrm{I}$ do:    ▷ Update infection times
        Obtain a sample $\mathrm{I}_i^{(k+1)}$ from $(\mathrm{I}_i \mid \mathbf{R}, \mathbf{I}_{-i}^{(k)}, \boldsymbol{\theta}^{(k+1)})$
        using adaptive Metropolis-Hastings
    end for
    for i $= 1, \dots, m$ do:    ▷ Update $m$ removal times
        Choose a removal time $\mathrm{R}_j$ from $\mathbf{R_M}$ randomly
        Obtain a sample $\mathrm{R}_j^{(k+1)}$ from $(\mathrm{R}_j \mid \mathbf{R_T}, \mathbf{R_M}_{-j}^{(k)}, \mathbf{I}^{(k+1)}, \boldsymbol{\theta}^{(k+1)})$
        using Metropolis-Hastings
    end for
    $k \leftarrow k + 1$
end for
```

in the dataset, that is the date in $\mathbf{R}_\mathbf{M}^{(0)}$, we discard it and draw a new sample for the infectious period. For our setting an independence sampler was found to be more effective than a Random Walk Metropolis update. Furthermore, experiments revealed that it is more efficient to update the missing removal times one at a time with the removal time chosen uniformly at random rather than to use a sequential or block updating scheme.

## 4.3 Simulation study: Augmenting detection times

It is of great value to investigate how well the Gaussian process risk surface and the outbreak characteristics can be recovered given only partially observed epidemic outbreak data. Therefore, we performed a simulation study using the Outbreak i) from Table 3.2 and slightly altered the data in similarity to the BTV-8 data by setting a proportion of the infected population to the same calendar date of detection. We investigated the model behaviour for the three different scenarios: augmenting 50%, 60% and 70% of the detection times. This is displayed in Figure 4.2.

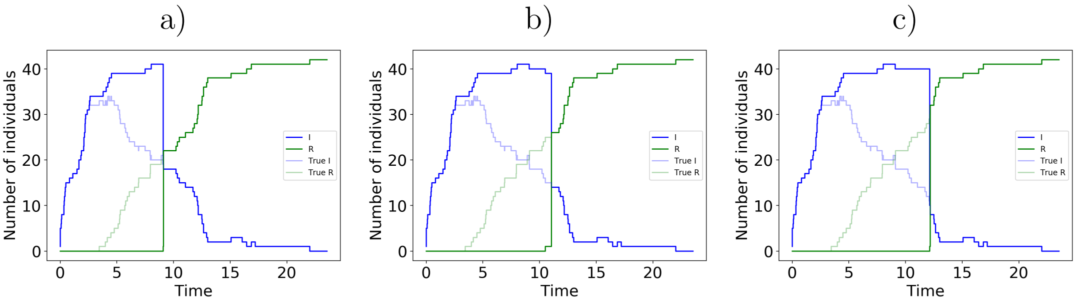


Figure 4.2: Number of infected and removed individuals when altering a) 50%, b) 60% and c) 70% of the data.

In Table 4.1 the mean posterior estimates are displayed. In order to compare the scenarios, we chose the same priors in all of the settings, that is $\kappa, \gamma \sim$ Gamma(0.1, 0.1), $\beta_0, \beta_1 \sim N(0,2), \sigma^2 \sim$ Gamma(2,2) and fixed $\alpha$ and $\epsilon$ to their true values for simplicity reasons. The estimates from the augmented inference scenarios are comparable, in particular the estimate of the spatial kernel and the coefficients of the fixed effects. The mean infectious time becomes shorter as less data on the detection is available. This effect is particularly pronounced with more than 2 days difference in the mean infectious period if we compare the augmented scenarios (4.9 days for 70% augmented data) with the inference on the full data (7 days).

The 95% credible intervals of the parameter estimates provide a measure for uncertainty. While the length of $\kappa$ is about 0.25 without additional augmentation,

| | True value | 50% augmented | 60% augmented | 70% augmented | 0% augmented |
|---|---|---|---|---|---|
| $\kappa$ | 0.3 | 0.323<br>[0.189,0.471] | 0.349<br>[0.198, 0.517] | 0.349<br>[0.201, 0.504] | 0.305<br>[0.184, 0.437] |
| $\gamma$ | 0.143 | 0.170<br>[0.133, 0.212] | 0.201<br>[0.150, 0.254] | 0.207<br>[0.164, 0.254] | 0.143<br>[0.110, 0.177] |
| $\overline{\text{IR}}$ | 7.006 | 5.923<br>[4.867, 6.915] | 5.025<br>[4.003, 6.230] | 4.875<br>[4.098, 5.643] | 7.047<br>[5.695, 8.478] |
| $\beta_0$ | 3 | 2.675<br>[1.721, 3.678] | 2.546<br>[1.536, 3.629] | 2.560<br>[1.695, 3.448] | 2.528<br>[2.086, 2.956] |
| $\beta_1$ | 1 | 0.323<br>[-0.368, 1.010] | 0.222<br>[-0.447, 0.889] | 0.277<br>[-0.410, 0.928] | 0.250<br>[-0.426, 0.911] |
| $\sigma^2$ | 0.1 | 0.622<br>[0.026, 1.756] | 0.719<br>[0.040, 1.895] | 0.503<br>[0.029, 1.299] | 0.959<br>[0.081, 2.283] |

Table 4.1: Mean posterior estimates and 95% credible intervals for augmented detection data of Outbreak i)

it increases to 0.28 for 50% and even 0.3 for 70% augmented data reflecting more uncertainty. On the scale for the spatial kernel function, this can have much effect on the distance of transmission (see Figure 3.1). However, the credible interval of the mean infectious period shortens with more data augmentation. This can be explained by the fact that with more data missing at the start of the epidemic, we can only capture the dynamics of the tail of the epidemic. As displayed in Figure 4.2, there are some new infections after time point 15, which leads to information towards the parameter estimates. This might not reflect the dynamics of the whole outbreak progression, as indicated by the lower mean infectious period estimate (4.9 days compared to 7 days).

Comparing the 0% and the 50% augmented data setting, the credible intervals become wider with less observed data for $\beta_0$, as there is more uncertainty in the estimation. For a higher percentage of augmentation they have approximately at the same length because the number of new infections are the same in all the augmented scenarios. Similarly, $\beta_1$ preserves the same level of uncertainty throughout all scenarios. With a less coarse structure on the fixed effects, the uncertainty in the estimators could be reduced in particular for the 0% augmented case.

$\sigma^2$ had wide credible intervals and differing mean estimates for all of the scenarios. With less information about the reporting time, we also have even less information about the drivers of the infection which is why the sampler had difficulties exploring the state space. In fact, the effective sample sizes for the parameters for the full data were $\kappa : 389, \gamma : 146, \bar{\text{s}} : 1509, \text{IR} : 107, \beta_0 : 1507, \beta_1 : 1794, \sigma^2 : 669$ in comparison to $\kappa : 197, \gamma : 73, \bar{\text{s}} : 928, \text{IR} : 38, \beta_0 : 50, \beta_1 : 372, \sigma^2 : 393$ for the half of the data augmented confirming the statement.

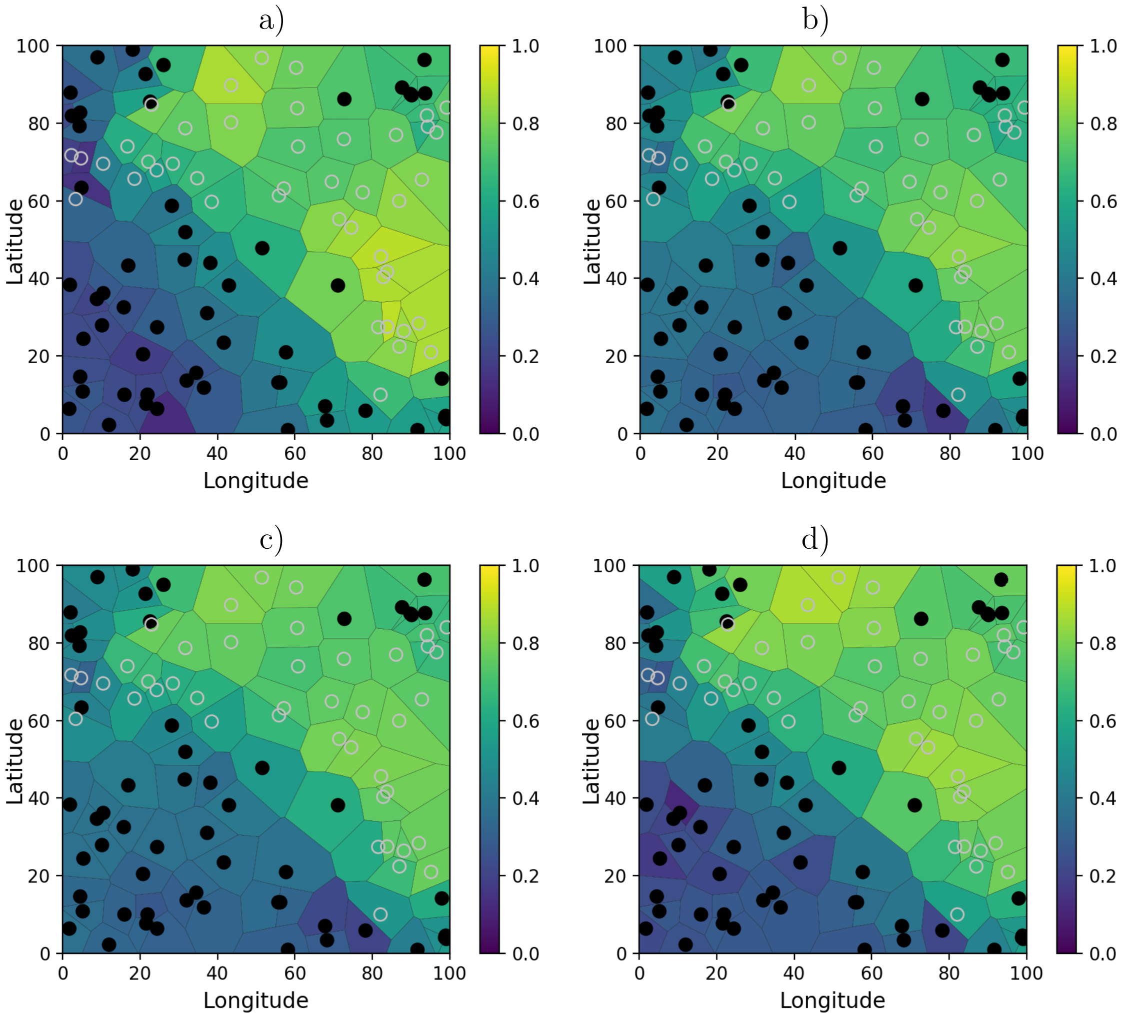


Figure 4.3: Prediction plots of a) 0% b) 50% c) 60% and d) 70% augmented data of Outbreak i)

The prediction plots in Figure 4.3 look very much alike. Here, we do not notice a substantial difference which confirms that even with a high percentage of the reporting times augmented, we achieve similar results in the random risk surface. It is remarkable that the lower risk at the top left within an infected area was identified by all scenarios. There, individuals became infected at a later time of the epidemic and the reporting time corresponds to their detection time. For other locations we have less information in particular with the percentage of reporting times augmented increased.

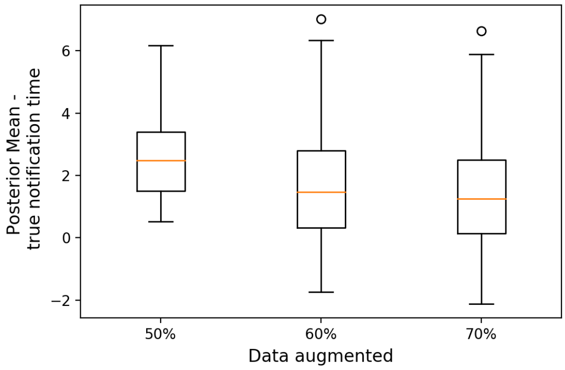


Figure 4.4: Boxplot of the difference between the true removal times and their posterior mean estimate for a) 50% b) 60% and c) 70% augmented data.

Investigating the posterior estimate of the augmented reporting time, the boxplot of the deviations of the posterior mean from the reporting times is displayed in Figure 4.4. In the mean the reporting times are estimated a little bit later. With more data augmented, the variation of the estimates, that is the uncertainty towards the true reporting time, becomes larger.

## 4.4 Discussion

Imputing data can be a very valuable tool when dealing with missing data. For this complex model, we deal with a relatively small amount of data: in our case, we do not only impute the missing infection times but further aim to impute part of the reporting time. Imputing the reporting times leads additionally to a higher dimensional sampler space causing the sampler to slow down and more uncertainty in the deviations in the posterior estimates.

From the simulation study we can infer that this approach can be applied to the data we have, yet it is not an ideal situation. With more data missing at the start of the epidemic, we can only capture the dynamics of the tail of the epidemic. However, these might not reflect the dynamics of the whole outbreak. Less detailed data introduce more uncertainty, which is why stronger priors on the parameters, e.g. the typical infectious period and a good understanding of the effects of the covariates may aid in coping with the difficulties. Thus, in summary care has to be taken when applying this approach to data with other forms of missing data.

# Chapter 5

# Application: BTV-8 outbreak in UK 2007

In this chapter we display the application of our methodology developed in the previous chapters on the BTV-8 data in Great Britain. Details of the disease, the outbreak and the dataset can be found in Section 1.6.

## 5.1 The model

The BTV-8 outbreak in Great Britain emphasizes the probable associated impact of climate change on animal disease. Moreover, it highlights the importance of modelling the introduction of a disease in previously disease free regions with a total susceptible host population. As discussed previously, a stochastic mechanistic vector-borne disease transmission model seems to be suitable in this context (Section 1.5.8). We do not aim to find the best model for BTV transmission, but rather display an application of our methodology in this section.

The final dataset we analyzed contained 14266 livestock farms with information on their location, their date of infection (this is set to the end of the observation period for farms that did not get infected within that time frame) and two meteorological information for each farm's location that resulted from a principal component analysis of the meteorological dataset (Figure 1.8). More detailed in-

| Parameter | Prior choices |
|---|---|
| $\kappa$ | Gamma(0.1, 0.1)<br>Gamma(6.2718, 19.299) |
| $\gamma$ | Gamma(0.1, 0.1) |
| $\beta_0, \beta_1$ | N(0,2) |
| $\sigma^2$ | Gamma(2, 2)<br>Gamma(2, 20) |
| $\alpha$ | Fixed to 0.99 |
| $\epsilon$ | Gamma(0.01, 0.01)<br>Gamma(9, 3) |

Table 5.1: Prior choices for the inference on the BTV-8 dataset

formation on the dataset and its sources can be found in Section 1.6.3.

For the GMRF usage, we first performed a Delaunay triangulation and the resulting network contained 42766 edges. The respective sparse precision matrix of the Gaussian Markov Random field of size $14266 \times 14266$ contained $2 \times 42766 + 14266 = 99798$ non-zero entries (due to symmetry and the diagonal elements), which is only 0.05% of all entries.

For the spatial kernel in the model (Equation (3.1), Figure 3.1), we determined that the cut off distance at 40 is sensible for a midge-borne disease. That implies that infected individuals may exert infectious pressure only within a distance range of 40 units onto susceptibles.

## 5.1.1 Inference and results

### 5.1.1.1 Inference scenarios

As displayed in the chronology of the BTV-8 outbreak in Figure 1.9, 64 out of 129 (50%) infections on farms were reported just shortly after the introductory case. In our model we assume that the disease must have been present on the farms before and augment their date of detection in addition to their infection times. We used different inference scenarios with varying prior distributions in order to investigate the robustness towards varying the prior distributions. Table 5.1 summarizes the different prior combinations we considered.

The parametrization of the informative Gamma(6.2718, 19.299) prior for $\kappa$ was chosen in such a way that the 98% of the distribution mass was contained within the range of [0.1, 0.7], whose values were inferred to be reasonable from the spatial kernel decay in Figure 3.1. A larger background infectious pressure $\epsilon$ led to spatially random infection appearances (Section 3.3.3). Thus, for pushing $\epsilon$ towards greater values, we made use of a Gamma(9,3) prior. To allow the contrary, a small background infection pressure, we also considered the scenario of a Gamma(0.01, 0.01) prior. In order to give more information to $\sigma^2$ (and thus more impact of the spatial random effect towards the infection), we further analyzed the more informative case with a Gamma(2, 20) prior. Estimating $\alpha$ led to a destabilization of the sampling in the other parameters, in particular $\sigma^2$ which is why we fixed it to a value close to 1 accounting for high spatial correlation.

To avoid the conflict of a wrong kernel choice for this example, we also investigated the approach of using a Gaussian Kernel instead of an exponential kernel. This leads to more infection pressure near infected individuals before it decreases with distance.

#### 5.1.1.2 Inference results

In all scenarios the mixing was slow, yet with more information put into the inference the mixing stabilized. In the different inference settings we see a consistency in the estimation. The parameter estimates responded to a change of prior but not drastically (as observed in previous chapters). The length of infection was estimated to be around 5 to 6 days. The coarse meteorological covariates data did not provide much information input towards the inference. $\beta_0$ was estimated close to 0 (posterior mean 0.027 for scenario I) and -0.117 for scenario V) ). $\beta_1$ was more informative with 5 different values on the area instead of 3 (Figure 1.10) and was estimated with -1.903 in scenario I and -4.529 in scenario V. Fixing the covariates to 0 led to a stabilization of the sampler for the other parameters. This can be attributed to the impact of the fixed effects on the outbreak, presumably introducing more non-identifiability between the other spatial variables.

In order to display the spatial prediction for two extremes with more and less

| | Prior choice | $\kappa$ | $\gamma$ | $\overline{\text{IR}}$ | $\beta_0$ | $\beta_1$ | $\sigma^2$ | $\epsilon$ |
|---|---|---|---|---|---|---|---|---|
| I | $\kappa \sim$ Gamma(0.1, 0.1)<br>$\sigma^2 \sim$ Gamma(2,2)<br>$\epsilon \sim$ Gamma(0.01, 0.01) | 0.126 | 0.192 | 5.523 | 0.027 | -1.903 | 0.849 | 0.458 |
| II | $\kappa \sim$ Gamma(6.3, 19.3)<br>$\sigma^2 \sim$ Gamma(2,2)<br>$\epsilon \sim$ Gamma(0.01, 0.01) | 0.158 | 0.210 | 4.973 | 0.069 | -1.857 | 1.014 | 0.303 |
| III | $\kappa \sim$ Gamma(0.1, 0.1)<br>$\sigma^2 \sim$ Gamma(2,2)<br>$\epsilon \sim$ Gamma(9,3) | 0.126 | 0.178 | 5.973 | -0.125 | -4.168 | 0.659 | 2.630 |
| IV | $\kappa \sim$ Gamma(6.3, 19.3)<br>$\sigma^2 \sim$ Gamma(2,2)<br>$\epsilon \sim$ Gamma(9, 3) | 0.151 | 0.193 | 5.449 | -0.117 | -4.530 | 1.014 | 2.601 |
| V | $\kappa \sim$ Gamma(6.3, 19.3)<br>$\sigma^2 \sim$ Gamma(2,20)<br>$\epsilon \sim$ Gamma(9,3) | 0.130 | 0.177 | 5.930 | -0.115 | -3.688 | 0.211 | 2.679 |
| VI | $\kappa \sim$ Gamma(6.3, 19.3)<br>$\sigma^2 \sim$ Gamma(2,2)<br>$\epsilon \sim$ Gamma(0.01,0.01)<br>Fixed $\beta_0 = 0$ | 0.155 | 0.193 | 5.454 | - | -2.135 | 0.841 | 0.353 |
| VII | $\kappa \sim$ Gamma(6.3, 19.3)<br>$\sigma^2 \sim$ Gamma(2,2)<br>$\epsilon \sim$ Gamma(9,3)<br>Fixed $\beta_0 = \beta_1 = 0$ | 0.114 | 0.196 | 5.411 | - | - | 0.050 | 1.890 |
| | Gaussian kernel | | | | | | | |
| VIII | $\kappa \sim$ Gamma(6.3, 19.3)<br>$\sigma^2 \sim$ Gamma(2,2)<br>$\epsilon \sim$ Gamma(9,3) | 0.111 | 0.183 | 5.842 | -0.038 | -3.790 | 1.006 | 2.671 |
| IX | $\kappa \sim$ Gamma(6.3, 19.3)<br>$\sigma^2 \sim$ Gamma(2,2)<br>$\epsilon \sim$ Gamma(0.01,0.01) | 0.113 | 0.179 | 5.921 | 0.077 | -1.895 | 0.742 | 0.866 |

Table 5.2: Mean posterior estimates for BTV-8 data. If not specified else, we chose $\beta_0, \beta_1 \sim N(0, 2)$ and $\gamma \sim$ Gamma(0.1, 0.1) priors and fixed $\alpha = 0.99$.

a priori information, we chose the scenarios of the most difference in priors, that is

I) $\kappa \sim \text{Gamma}(0.1,\ 0.1)$, $\epsilon \sim \text{Gamma}(0.01,\ 0.01)$, $\sigma^2 \sim \text{Gamma}(2,\ 2)$

V) $\kappa \sim \text{Gamma}(6.2718,\ 19.299)$, $\epsilon \sim \text{Gamma}(9,\ 3)$, $\sigma^2 \sim \text{Gamma}(2,\ 20)$.

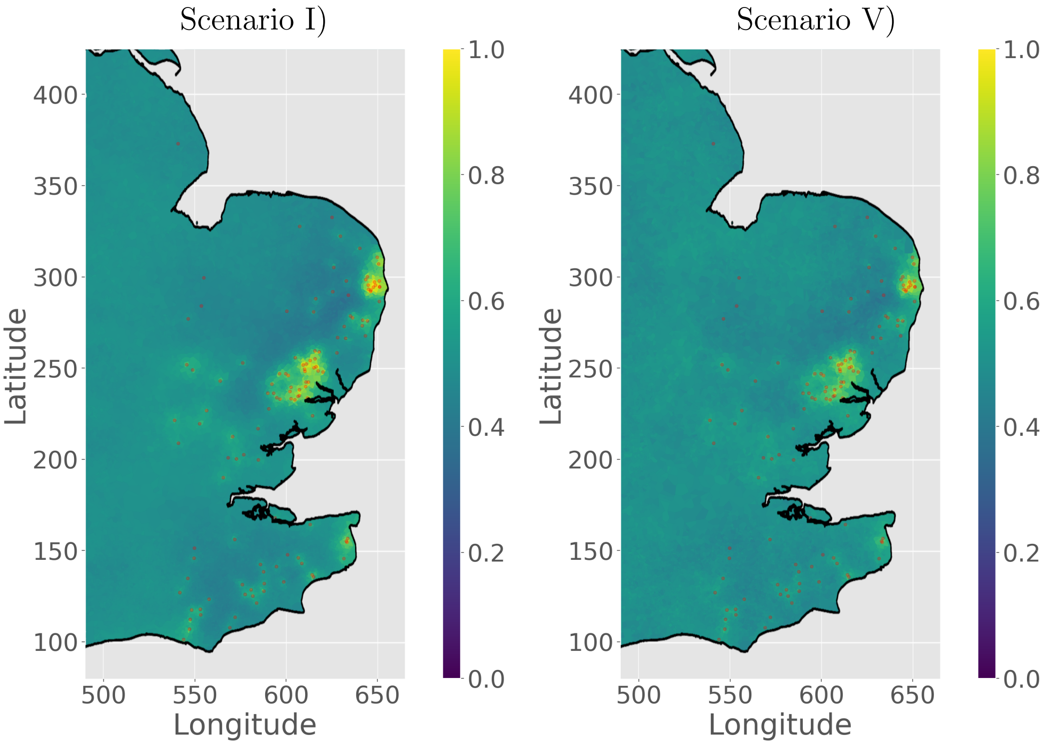


Figure 5.1: Comparison of the estimated random spatial risk surface for the two inference scenarios I) and V).

As displayed in Figure 5.1 we obtained very similar spatial random effect risk surfaces in both cases. This also held true for the other scenarios. Only when leaving out the fixed effect covariates (setting their parameters to 0 in the inference), we obtained a slightly more pronounced spatial random effect taking into account the missing infectious pressure. However, it did not change the overall result.

In their risk prediction, the plots were comparable. We can clearly identify two main hot spots. There, we also have the highest density of infected farms.

In our model the infections that are assumed to arise from the background infections are calculated by $\psi\epsilon \times 14266$ (population size) $\times 75$ (length of outbreak).

Whereas the remaining infections would rather be attributed to the background infection pressure. Between an $\epsilon = 0.458$ and $\psi = 1.3 \times 10^{-5}$ (scenario II) and $\epsilon = 2.679$ and $\psi = 8.302 \times 10^{-8}$ (scenario V), we infer between 6.37 and 2.4 infections due to random background infections.

Using a Gaussian kernel (scenario VIII and IX) in contrast to an exponential kernel had very little effect on the parameter estimates and resulted in a comparable spatial random effect recovery.

#### 5.1.1.3 Model fit

In this section we investigated how well our model fitted the data.

Figure 5.2 displays the outbreak dynamics that were created from posterior samples of scenario V) $\kappa \sim$ Gamma(6.2718, 19.299), $\epsilon \sim$ Gamma(9, 3), $\sigma^2 \sim$ Gamma(2, 20) (see Table 5.2).

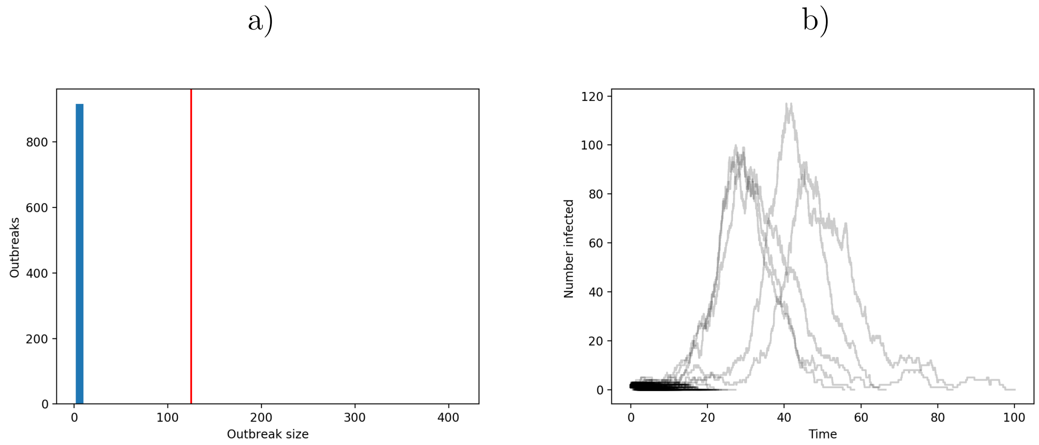


Figure 5.2: Posterior outbreak dynamics of BTV-8 outbreak data with a) histogram of the outbreak size; the red lines refer to the original outbreak size. b) temporal progression of posterior outbreaks.

It became clear that the outbreaks created from the posterior samples differed from the original BTV-8 outbreak data. Most of the simulated outbreaks from the posterior samples led to outbreak sizes that were far below the original outbreak. Only a very little number of the simulations led to more than 300 infected individuals, which is more than double the number of infections in the real data. Posterior predictive plots of other inference scenarios (Table 5.2) led to similar results. This suggests that there are features in the real data generating process that we did not capture. Possible reasons for this lack of fit might be for example the spatially coarse structure of the meteorological covariates as introduced earlier. Additionally, we only considered the mean values of these variables over a period of three months. Furthermore, BTV-8 can also be asymptomatic which we did not account for in the model. Further work is needed to address those issues.

In summary, from the posterior predictive plots we can infer that the data was not generated by the model we are proposing. However, we did not mean to provide a detailed analysis of the 2007 UK BTV-8 outbreak. Instead, our goal was to demonstrate our proposed joint modelling framework.

## 5.2 Discussion

Using the BTV-8 outbreak as an example, we showed that our methodology is effective at identifying spatial hotspots and differentiated between inter-farm transmission and background infections.

For a more precise modelling of the BTV-8 disease dynamics other features should be included. For example the incubation period, the time from the individual being infected to when they become infectious, or a less perfect detection probability can be added, as BTV-8 can be asymptomatic. The background infectious pressure took into account spontaneous infections in previously disease free regions which could not be attributed to the transmission from a proximate infected farm. This also comprised strong winds that carried infected midges over long distances leading to cases in distant locations. Although we included wind direction and speed within the fixed effects, the information does not contain enough detail, including time-dependence, to allow for detailed inference.

This chapter illustrates the importance of the data quality for modelling and parameter estimation. For the case data we had information on who is infected. However, the main limitation of the dataset was that we did not know when the infection happened and, additionally, for a considerable proportion of the farms the detection date was imprecise. Proper record keeping to obtain better detection time estimates enhances the model's estimates and predictions as shown.

Besides, with a more heterogeneous structure in the fixed effects, we would have been able to obtain a better prediction of the risk surface in particular in areas of no infection. There, the Gaussian process reverts to its mean, neglecting the strength of the model. Higher resolution earth observational data can provide an alternative to missing or highly coarse freely available meteorological data.

Additionally, better data quality leads to less imputation and thus a shorter runtime of the model. The BTV-8 model from this chapter took about 2 weeks for 50,000 samples, which does not qualify as a modeling tool for emergency response. Thus, apart from more informative data, further software engineering is needed to be able to inform epidemiologists in a short time.

We can conclude that this modelling approach needs detailed data for a rather complicated inference setting. The low level of infections (less than 1% of the individuals) increased the complexity to differentiate between the importation of cases and the farm-to-farm transmission. Furthermore, we imputed half of the infection times and dealt with a population size of almost 15,000 individuals introducing much more uncertainty into the model. With infection data that is very limited by its location and date, more detail on the fixed covariates can inform the model better. Furthermore, to stabilize inference prior distributions should be informed by epidemiologists and experts.

# Chapter 6

# Conclusions and Discussion

This thesis aimed at developing an approach for emergency response to vector-borne disease outbreaks when there is not enough data on the vector population yet available. In contrast to other models that deal with vector density or biting rate, we inform the model indirectly about vector presence by introducing covariates that influence the vectors' habitat and behaviour. This allowed us to estimate key disease dynamic parameters, such as transmission and infection-to-detection rates, as well as to recover the spatial random effect of the vector risk across the country.

We showed that epidemiological models can be combined with methods from spatial statistics. In the simulation studies, we were able to capture the dynamics of the outbreak and recover the risk surface of the spatial random effect where possible. We estimated the parameters that govern the transmission process and investigated non-identifiabilities between parameters accounting for spatial variation. Through various simulation studies we could infer that the mixing in particular of the augmented infection times was slow, yet reasonable. In order to decrease the sampling space, we analytically integrated out the infection rate parameter $\psi$, which stabilized the mixing of the remaining parameters. Different attempts did not improve the sampling performance of the infection times and remains subject to further research.

Depending on the population size the computational complexity must be taken into account. While Gaussian processes are a common tool in geostatistics, it comes with a cubic complexity. Thus, for an increased population suitable alternatives need to be considered. In this thesis we investigated the representation of Gaussian processes by Gaussian Markov random fields within the epidemiological framework and made use of the computational benefits through its sparse precision matrix. The estimates of both approaches were reasonable with less precision in the risk surface through the GMRF but with gains in computational efficiency. We could conclude that estimating the hyperparameters is difficult and destabilizes the sampling in both scenarios. In fact, fixing one of the parameters, e.g. the marginal variance of the Gaussian process or the spatial correlation parameter for the GMRF, can help with the inference.

Furthermore, scaling up the population size leads to large areas of no infection, and thus only very little information. Using meteorological data with a sufficiently fine grid adds heterogeneity and yields more information outside the infected herds when fitting the model. For computational benefits, the spatial kernel can be defined in such a way that no infectious pressure is able to be transmitted given a certain distance. This way, efficient matrix multiplication can speed up the runtime of the MCMC sampling. In all cases, informative prior distributions should be preferred where possible to stabilize inference. The estimates do respond to a change of priors but not in a drastic way still capturing the dynamics of the outbreak.

We applied this approach to a real disease outbreak which is the UK 2007 BTV-8 outbreak, where half of the population had accumulated detection data. We acknowledge that a small proportion of infected individuals(less than 1%) provides only limited data for a complex model to fit. Though, accumulated data in addition includes more uncertainty into the model as the dimension of the sampling space and thus the computational time to sample from the posterior distribution increases. However, we showed that this provides a useful tool for dealing with accumulated or missing data given sufficient computational resources.

We illustrated that our methodology is working well as we could identify high risk regions and obtain reasonable parameter estimates. However, we also concluded that there are certain constraints on the data: for this rather complex model we need sufficient information for good predictions and inference. The larger the population size and the more data that needs to be augmented, the greater the runtime of the model and more uncertainty is included in the estimation. Discussions with experts and epidemiologists can help specify the prior distributions and thus stabilize inference. A crucial aspect to exploit the strength of this approach, is that the spatial heterogeneity in the data is provided. The density of observation points has to be high enough in relation to the variation in the surface. This is important to give the model a chance to succeed in capturing the characteristics of the random surface. Furthermore, the remote sensing data needs to be sufficiently detailed in order to predict the risk for areas that have not yet experienced the outbreak.

Thus, the availability and quality of data are practical limitations to the use of the model in the real world. For emergency response to disease outbreaks, we recommend that fine-scale remote sensing data should be made freely - and therefore quickly - available. The runtime of the model benefits from the above mentioned data availability. Furthermore, for the policy-makers to be able to act upon the model predictions, more software engineering is needed to shorten the runtime from the range of a couple weeks to a few days or even overnight.

The outcome of this thesis is a generic vector-borne disease forecasting system, capable of assimilating case, demographic, and remotely sensed environmental data to rapidly predict high risk areas of transmission should a new incursion occur even if the new incursion requires a hitherto understudied vector population. As such, it yields rigorous disease control policy evidence for a future outbreak, even in the presence of considerable uncertainty concerning the disease transmission process.

## 6.1 How to take this thesis forward?

To generalize this method, the type and features of the disease outbreak we wish to model have to be considered and adapted to the purpose. For example, we did not account for the movement of the individuals as we assume a movement ban was set in place shortly after the outbreak started and thus farms did not trade during the observation period. However, additional features such as partial trading restrictions or moving individuals, can be included by making use of contact networks and graph theory.

This method can be applied as well to other forms of disease transmission.

For instance, for bovine tuberculosis Woodroffe et al. (2016) suggest, to consider disease transmission through contamination of the host's and vector's shared environment. In this case, the time independent spatial risk would highlight areas of contaminated cattle pasture and hence areas of high disease risk.

In cases where for example the environment and hence the location of the vector with the pathogen changes rather quickly, the method can be extended by a time dependent spatial random effect and further by temporarily changing covariates. A piecewise constant intensity function for the Poisson process is not applicable here, as the value of the covariates are only able to change when an event (infection or removal) happens and is not independent of the event. For this, the nonhomogeneous Poisson process is suitable (Appendix A.1.2) for which new ways for efficient and fast computation have to be considered.

Besides, in a future study it might be useful to investigate how many infected individuals are needed before predictions become practically useful, given some policy-derived objective function (e.g. predictive variance of the epidemic size).

For simplicity reasons we used a standard SIR epidemic model, though, our approach leaves room for extensions to other epidemic models, e.g. with additional states such as exposed (SEIR) or vaccinated (SVIR), or the chance of recovering from the infection (SIS), as well as lifelong infection (SI).

In this thesis, we only considered completed outbreaks. However, when working with online data, that is, data of an ongoing outbreak, we would want to predict the likelihood of the disease arriving in currently uninfected areas of the country and predict the vector population's propensity to spread the disease once it has arrived. Besides, the framework is an individual-based model and thus can stand only if the data on the location and status of *all* farms (infected and, in particular, not infected) are available. Further research is needed on how to relax this requirement.

Chapter 4 stands alone. This approach can be used in different areas, such as for example in diagnostics: the development of a far more sensitive test would lead to a sudden and steep increase of the number of infected individuals, which is difficult for epidemiological models to capture. These detection times can be augmented as shown to fit the model better and predict the ongoing of the outbreak dynamics.

This approach can presumably be applied to a variety of diseases, in particular the whole class of vector-borne diseases. In can be highly beneficial in situations with a lack of or obsolete information about the vector, with environmental data in the public domain particular provided. We highly recommend and anticipate the use of the self implemented epidemiological extension to the PyMC3 library as it facilitates the use of statistical computing and its applications. The programming code is available under https://fhm-chicas-code.lancs.ac.uk/koeppell/PyMC3_Epi_Extension.

# Appendix A

# Appendix

## A.1 Poisson Process

In the following, we consider a stochastic process which enables us to model the occurrence of infection events. In particular, we focus on the so-called Poisson point process or Poisson process. We utilize the concept of Poisson process to model the time points of infections instead of their location. A reference to an article where Poisson processes were used in a one dimensional, temporal setting to model epidemics would be for example Andersson and Britton (2012), and a selection of articles in an applied setting is Jewell et al. (2009); Jewell and Brown (2015); Xiang and Neal (2014).

In the following, we introduce the Poisson process first in its simplest form before we gradually extend it for our use. Being part of our modelling framework, we derive its likelihood, which enables us to do inference on its parameters.

### A.1.1 Homogeneous Poisson process

A Poisson process can be defined in several ways (Ross, 2013). In a one dimensional setting, the most intuitive approach is to model a sequence of waiting times each ending with an event. Hence, these waiting times are also referred to as *inter-event times*. The simplest continuous distribution for modelling the inter-event times is the exponential distribution as it has the memorylessness property: the waiting time until the next event in the future is independent of how much time we have already waited. In other words, for a random waiting time $T \in [0, \infty)$ and any given time $t'$ with $0 \leq t' \leq t$ we have

$$\mathbb{P}\left(T < t \mid T > t'\right) = \mathbb{P}\left(T < t - t'\right) \tag{A.1}$$

The probability that we at most wait $t$ given that we have already waited $t'$ is the same as the probability that we at most wait $t - t'$. The only continuous probability distribution that fulfils this property is the exponential distribution (Forbes et al., 2010). Hence, for any random variable $T \in [0, \infty)$ that fulfils (A.1), the probability density is given by

$$\mathbb{P}\left(T \leq t\right) = \int_0^t \lambda e^{-\lambda x} dx$$

for all $t \in [0, \infty)$ and a fixed $\lambda \in (0, \infty)$. The parameter $\lambda$ can be interpreted as a rate parameter and governs the length of the waiting time:

$$\begin{aligned}
\mathbb{E}(T) &= \int_0^\infty x \cdot \lambda e^{-\lambda x} dx \\
&= \big[ - x e^{-\lambda x} \big]_0^\infty + \int_0^\infty e^{-\lambda x} \\
&= (0 + 0) + \big[ - \frac{1}{\lambda} e^{-\lambda x} \big]_0^\infty \\
&= \frac{1}{\lambda}
\end{aligned}$$

The mean of the inter-event time states how long we have to wait on average for an event to occur. So if we wait $1/\lambda$ for one event, we expect $\lambda$ events to take place within one time unit. Hence, the rate parameter $\lambda$ controls the intensity of occurring events over time; a higher $\lambda$ leads to an increasing number of events happening, a lower $\lambda$ leads to less events.

We are now able to introduce the Poisson process. We will start with the definition of a simple Poisson process, which models inter-event times at a constant rate. Then, we extend this concept to the heterogeneous Poisson process to model short and long inter-event times. Finally, we further condition the occurrence of new events on the information of all events in the past by introducing a conditional intensity function for the Poisson process. We close this chapter with deriving the formula for the likelihood.

**Definition 8:** Let $T_1, T_2, \ldots$ be i.i.d. random variables that follow an exponential distribution with parameter $\lambda$. The counting process

$$N_t := \max\{n \mid \mathrm{T}_0 + T_1 + \cdots + T_n \leq t\}, \quad \text{with } \mathrm{T}_0 := 0$$

is called *homogeneous Poisson process* with intensity $\lambda$.

$N_t$ counts the number of events that happen up to time $t$. We now show that this number of occurred events until time $t$ is Poisson distributed with parameter $t\lambda$.

**Theorem 3:** $\mathbb{P}\left(N_t = n\right) = \frac{(t\lambda)^n}{n!} e^{-t\lambda}$

*Proof.* First it is to note that the sum of independent identical exponential distributions is Erlang distributed (Forbes et al., 2010). In other words, let $T_1, T_2, \ldots$ be independent identical exponential distributed random variables with a common rate $\lambda$, then

$$\mathbb{P}\left(\mathrm{T}_0 + T_1 + \cdots + T_{n+1} \leq t\right) = 1 - \sum_{i=0}^{n} \frac{1}{i!} e^{-\lambda t} (t\lambda)^i$$

From Definition (8) we have $\{N_t \geq n\} = \{\mathrm{T}_0 + T_1 + \cdots + T_n \leq t\}$ for $n \in \mathbb{N} \cup \{0\}$ and thus $\{N_t > n\} = \{\mathrm{T}_0 + T_1 + \cdots + T_{n+1} \leq t\}$. Then

$$\begin{aligned}
\mathbb{P}\left(N_t \leq n\right) &= 1 - \mathbb{P}\left(N_t > n\right) \\
&= 1 - \mathbb{P}\left(\mathrm{T}_0 + T_1 + \cdots + T_{n+1} \leq t\right) \\
&= \sum_{i=0}^{n} \frac{1}{i!} e^{-\lambda t} (t\lambda)^i
\end{aligned}$$

Since this holds for each $n$, we get

$$\mathbb{P}\left(N_t = n\right) = \frac{1}{n!} e^{-\lambda t}(t\lambda)^n$$

□

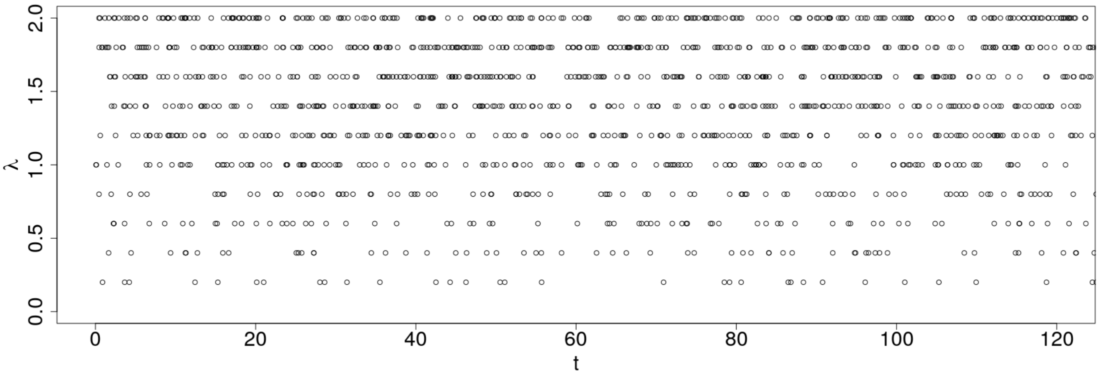


Figure A.1: Simulation of 10 Poisson processes (horizontal) for $\lambda = 0.2, 0.4, \ldots, 2.0$. Every point marks one event.

Figure A.1 displays the simulation of 10 homogeneous Poisson processes (Definition (8)) each with different rate parameter $\lambda$. It becomes apparent how the rate $\lambda$ is influencing the event count: a high $\lambda$ leads to short inter-event times and thus many events over time, whereas a low $\lambda$ entails longer waiting times and thus a smaller number of events over time.

Although this process enables us to model short and long inter-event times and thus low and high intensities, it is often too simple for practical matters as we will see in the next section.

### A.1.2 Nonhomogeneous Poisson process

In the previous section we introduced a Poisson process with a constant intensity $\lambda$. Hence, this process is called *homogeneous Poisson process*. In real word applications, though, we are often interested in intensities that change over time for example due to seasonal variations. To include such a feature we should be able to alter $\lambda$ over time as opposed to leaving it constant. This can be done by considering the rate parameter $\lambda$, meaning the intensity, as function of time $t \mapsto \lambda(t)$ with $\lambda(t) \in [0, \infty)$ for all $t \in [0, \infty)$. The *intensity function* $\lambda(t)$ governs how the occurrence of events of a Poisson process is changing depending on time. To do so, we need the integral $\Lambda(t)$ of $\lambda$ on the interval $[0, t]$

$$\Lambda(t) := \int_0^t \lambda(s)ds$$

$\Lambda$ is monotonic increasing, though not strictly monotonic increasing: $\lambda(s)$ can be 0 over a certain interval leaving the inverse function not uniquely defined. Therefore, $\Lambda^{-1}(s) = \inf\{t \mid \Lambda(t) = s\}$.

**Definition 9:** Let $T_1, T_2, ... \in [0, \infty)$ be identically, independent random variables that follow an exponential distribution with mean 1. The counting process $N_t$ with

$$N_t := \max\{n \in \mathbb{N} \cup \{0\} \mid \mathrm{T}_0 + \Lambda^{-1}(T_1 + ... + T_n) \leq t\}, \quad \text{whereby } \mathrm{T}_0 := 0$$

is called *nonhomogeneous Poisson process.*

Here $\Lambda^{-1}$ transforms the sum of the inter-event times and stretches or compresses it according to the intensity function $\lambda$. Please note that if no event happens until time $t$, we get $N_t = 0$.

**Theorem 4:** $\mathbb{P}(N_t = n) = \frac{\Lambda(t)^n}{n!} e^{-\Lambda(t)}$

*Proof.* This can be verified similarly to the homogeneous Poisson process case (proof of Theorem (3)).

□

Comparing the homogeneous Poisson process (Definition (8)) with the nonhomogeneous Poisson process (Definition (9)), we notice that the latter is a generalization of the former one: we obtain the homogeneous Poisson process as a special case of the heterogeneous Poisson process by taking a constant intensity $\lambda$.

Figure A.2 illustrates a nonhomogeneous Poisson process with intensity function $\lambda(t) = 1 + \cos(t)$. It becomes clear that the number of events increases and

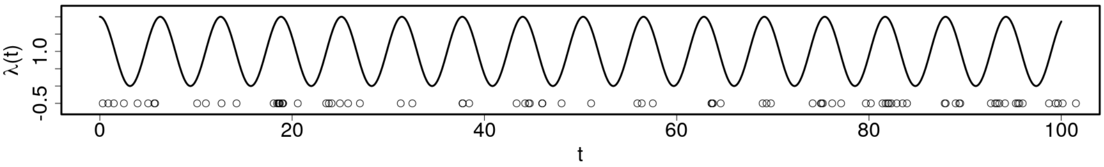


Figure A.2: Nonhomogeneous Poisson process with intensity function $\lambda(t) = 1 + \cos(t)$. Every point marks one event.

decreases with increasing and decreasing values of the periodic function $\lambda(t)$.

### A.1.3 The likelihood

In the previous chapter we introduced a Poisson process which enables us to model a process with varying event frequencies over time. The amount of variation may be determined by the parameters of its intensity function. We are interested in drawing inference on these parameters given only the observed event times in reality. Hence, we rely on the likelihood function, which is the interface between the observed event times and the unknown parameters: it is a function of the parameters of an empirical model given the data.

Therefore, we now derive the likelihood of a nonhomogeneous and thus homogeneous Poisson process given the event times; that is, the value of the density function evaluated at the observation. First we define random variables $X_1, X_2, \ldots$ by

$$X_k := \Lambda^{-1}(T_1 + ... + T_k) \tag{A.2}$$

In the following we will denote the density of $(X_k \mid X_{k-1} = x_{k-1})$ evaluated at $x_k$ by $p_{X_k|X_{k-1}}(x_k \mid x_{k-1})$. This conditional distribution is necessary to derive the density of a Poisson process. The following Lemma helps us to understand its functional form.

**Lemma 1:**

$$p_{X_k|X_{k-1}}(t \mid x_{k-1}) = \mathbb{1}_{\{t>x_{k-1}\}}\lambda(t)e^{-\left(\Lambda(t)-\Lambda(x_{k-1})\right)} \tag{A.3}$$

*Proof.* First we show that the sequence $X_1, X_2, ...$ forms a Markov chain, that is $X_k$ is independent of $X_1, \dots, X_{k-2}$ given $X_{k-1}$, and then utilize it to prove the functional form of the conditional distribution $p_{X_k|X_{k-1}}(x_k \mid x_{k-1})$.

Taking into account Equation (A.2), the distribution of $X_k$ given all previous random variables is

$$\begin{aligned}(X_k \mid X_{k-1} = x_{k-1}, X_{k-2} = x_{k-2}, \dots, X_1 = x_1) &= \Lambda^{-1}(t_1 + \cdots + t_{k-1} + T_k)\\ &= \Lambda^{-1}\Big(\Lambda(\Lambda^{-1}(t_1 + \cdots + t_{k-1})) + T_k\Big)\\ &= \Lambda^{-1}(\Lambda(X_{k-1} = x_{k-1}) + T_k)\\ &= (X_k \mid X_{k-1} = x_{k-1})\end{aligned}$$

Thus, the sequence $X_1, X_2, ...$ forms a Markov chain. We infer its functional form by applying a transformation on $T_k$ with the function $f(s) = \Lambda^{-1}(\Lambda(x_{k-1}) + s)$. We know that $T_k$ is exponentially distributed with mean 1, that is $\mathbb{P}(T_k \leq t) = 1 - e^{-t}$.

$$\begin{aligned}\Rightarrow \quad \mathbb{P}(X_k \leq t \mid X_{k-1} = x_{k-1}) &= \mathbb{P}\left(\Lambda^{-1}(\Lambda(x_{k-1}) + T_k) \leq t\right)\\ &= \mathbb{P}(T_k \leq \Lambda(t) - \Lambda(x_{k-1}))\\ &= 1 - e^{-(\Lambda(t) - \Lambda(x_{k-1}))}\end{aligned}$$

Taking the derivative, we obtain the density for $X_k$

$$\lambda(t) e^{-\Big(\Lambda(t) - \Lambda(x_{k-1})\Big)} \tag{A.4}$$

In (A.4), the next event time $t$ is not depending on any previous events and thus can be any positive value. However, by definition events happen consecutively and thus the event can only occur only after $x_{k-1}$, the time of the $(k-1)^{st}$ event. Hence, we have to restrict the density to be 0 unless $t > x_{k-1}$ holds and this completes the proof.

□

We are now able to derive the likelihood function of the nonhomogeneous Poisson process with observed event times $t_1 < t_2 < ... < t_n$ within the observation period $[0, t_{end}]$. Because the chronology of the event times in the observation is known a priori, the indicator function in Equation (A.3) of Lemma 1 is always 1 given the (ordered) observation. Thus, we will neglect this term in the following. Let $p_{X_1}(t_1)$ constitute the density of $X_1$ evaluated at $t_1$, then the likelihood $L(t_1, ..., t_n, t_{end})$ of this observation reads

$$\begin{aligned}
L(t_1, ..., t_n, t_{end}) &= \Big(p_{X_1}(t_1) \prod_{k=2}^{n} p_{X_k|X_{k-1}}(t_k \mid t_{k-1})\Big)\mathbb{P}\left(X_{n+1} > t_{end} \mid X_n = t_n\right) \\
&= \Big(\lambda(t_1)e^{-\Lambda(t_1)} \prod_{k=2}^{n} \lambda(t_k)e^{-(\Lambda(t_k)-\Lambda(t_{k-1}))}\Big)\Big(1-(1-e^{-(\Lambda(t_{end})-\Lambda(t_n))})\Big) \\
&\overset{*}{=} e^{-\Lambda(t_1)+\Lambda(t_n)-\Lambda(t_{end})+\sum\limits_{j=2}^{n}\Lambda(t_{j-1})-\Lambda(t_j)} \prod_{k=1}^{n} \lambda(t_k) \\
&= e^{\,-\Lambda(t_{end})} \prod_{k=1}^{n} \lambda(t_k)
\end{aligned}$$

The asterisk indicates that the product of the exponential distributions can be simplified to a telescoping sum in the exponent, in which the terms cancel each other out, leaving one summand, namely $e^{\,-\Lambda(t_{end})}$.

### A.1.4 Conditional Poisson process

The inhomogeneous Poisson process seems to be quite flexible because its intensity function can be adjusted to one's needs to model short as well as long inter-event times. However, for practical matters we face a significant drawback: dependencies cannot be modelled. For example in a disease outbreak, infected individuals are more likely to infect individuals that are located nearby than the ones that are further away. Hence, for a given individual, the probability of infection depends on the location of the other infected individuals. When incorporating such dependencies into our model, the stochastic process has to be able to include the past events up to any given time $t$. Generally speaking, the process has to condition on the past! With respect to the Poisson process, this can be realized by working with a *conditional intensity function*. This function conditions on all events that occurred up until any time point $t$ and changes its values accordingly. It is even possible to consider several Poisson processes in parallel, whose intensity functions at a given time $t$ condition on all events of *all* processes up until $t$. This becomes important especially when we consider the infection process of an individual as a Poisson process that conditions on all infection events of all other Poisson processes until a certain time point $t$.

In the following we refer to a Poisson process with a conditional intensity function as *conditional Poisson process*. Let $\mathcal{H}_t$ denote all information until time point $t$ (excluding). Then, $\lambda^*(t \mid \mathcal{H}_t)$ is the intensity function that conditions on all past events that happened before $t$, that is on all events at time points $t_1 < t_2 < ... < t_n < t$.

**Example 1:** (Hawkes process) Let $t_1 < ... < t_n$ be the time points of events until time $t$. The so-called *Hawkes process* (Hawkes, 1971) is defined by its intensity function

$$\lambda^*(t \mid \mathcal{H}_t) = \mu + \alpha \sum_{i=1}^{n} e^{-(t-t_i)} \mathbb{1}_{\{t \geq t_i\}}$$

This process behaves as follows: whenever an event occurs, the intensity increases by $\alpha$ and thereafter decreases exponentially over time. The minimal intensity is determined by $\mu$. Figure A.3 displays a simulation of a variety of Hawkes processes with $\mu = 0.2$ and $\alpha = 1, \ldots, 8$. As expected the events tend to occur in clusters and are not uniformly distributed (homogeneous Poisson process) or do not occur with respect to the value of a function that is independent of the past (heterogeneous Poisson process).

In Example 1 it becomes clear that conditioning on past events is an additional feature of a Poisson process. When extending the nonhomogeneous and hence the homogeneous Poisson process by this feature, it is to note that the increments of the conditional Poisson process are not independent anymore.

We are now able to express the likelihood of the conditional Poisson process with respect to the previous Section A.1.3. With $\Lambda^*(t \mid \mathcal{H}_t) = \int_0^t \lambda^*(t \mid \mathcal{H}_t)dt$, the

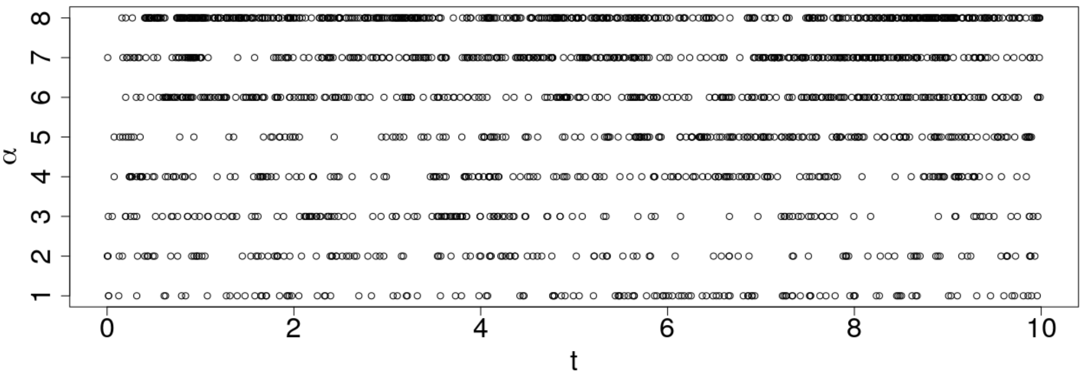


Figure A.3: Simulation of a Hawkes process with $\mu = 0.2$ and $\alpha = 1, ..., 8$.

likelihood reads

$$L(t_1, ..., t_n, t_{end}) = e^{-\Lambda^*(t_{end}|\mathcal{H}_{t_{end}})} \prod_{k=1}^{n} \lambda^*(t_k \mid \mathcal{H}_{t_k})$$

We have now expanded the Poisson process to such an extent that we are able to model many complex processes of the real world including changes in the intensity functions at arbitrary time points.

# Bibliography

Abramowitz, M. and Stegun, I. *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables.* Applied mathematics series. Dover Publications, 1965.

Anderson, R. and May, R. *Infectious Diseases of Humans: Dynamics and Control.* Oxford University Press, 1991.

Andersson, H. and Britton, T. *Stochastic epidemic models and their statistical analysis*, volume 151. Springer Science & Business Media, 2012.

Aristotelous, G. *Topics in Bayesian inference and model assessment for partially observed stochastic epidemic models.* PhD thesis, University of Nottingham, 2020.

Bailey, N. T. *The mathematical theory of infectious diseases and its applications.* Griffin, London, 2nd ed. edition, 1975.

Bartlett, M. Some evolutionary stochastic processes. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(2):211–229, 1949.

Bayes, T. 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, 1763.

Beer, M., Conraths, F., and Van der Poel, W. Schmallenberg virus – a novel orthobunyavirus emerging in europe. *Epidemiology & Infection*, 141(1):1–8, 2013.

Besag, J. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225, 1974.

Besag, J., York, J., and Mollié, A. Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43 (1):1–20, 1991.

Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

Box, G. E. P. and Muller, M. E. A note on the generation of random normal deviates. *Ann. Math. Statist.*, 29(2):610–611, 06 1958.

Braks, M., van der Giessen, J., Kretzschmar, M., van Pelt, W., Scholte, E.-J., Reusken, C., Zeller, H., van Bortel, W., and Sprong, H. Towards an integrated approach in surveillance of vector-borne diseases in europe. *Parasites & vectors*, 4(1):192, 2011.

Brauer, F. Mathematical epidemiology: Past, present, and future. *Infectious Disease Modelling*, 2(2):113–127, 2017.

Brauer, F., Castillo-Chavez, C., Mubayi, A., and Towers, S. Some models for epidemics of vector-transmitted diseases. *Infectious Disease Modelling*, 1(1): 79–87, 2016.

Britton, T. and O'Neill, P. D. Bayesian inference for stochastic epidemics in populations with random social structure. *Scandinavian Journal of Statistics*, 29(3):375–390, 2002.

Burgin, L., Gloster, J., and Mellor, P. Why were there no outbreaks of bluetongue in the uk during 2008? *Veterinary Record*, 164(13):384–387, 2009.

Campbell-Lendrum, D., Manga, L., Bagayoko, M., and Sommerfeld, J. Climate change and vector-borne diseases: what are the implications for public health research and policy? *Phil. Trans. R. Soc. B*, 370(1665), 2015.

Coutinho, F. A. B., Massad, E., Lopez, L. F., Burattini, M. N., Struchiner, C. J., and Azevedo-Neto, R. S. d. Modelling heterogeneities in individual frailties in epidemic models. *Mathematical and computer modelling*, 30(1-2):97–115, 1999.

Daley, D. J. and Gani, J. *Epidemic modelling: an introduction*, volume 15. Cambridge University Press, 2001.

Dean, K. R., Krauer, F., Walløe, L., Lingjærde, O. C., Bramanti, B., Stenseth, N. C., and Schmid, B. V. Human ectoparasites and the spread of plague in europe during the second pandemic. *Proceedings of the National Academy of Sciences*, 115(6):1304–1309, 2018.

Diekmann, O. and Heesterbeek, J. *Mathematical epidemiology of infectious diseases : model building, analysis, and interpretation.* Wiley series in mathematical and computational biology. John Wiley, Chichester, 2000.

Dietz, K. and Heesterbeek, J. Bernoulli was ahead of modern epidemiology. *Nature*, 408(6812):513, 2000.

Diggle, P. and Giorgi, E. *Model-based geostatistics for global public health : methods and applications.* Taylor & Francis, Boca Raton, 2019.

Diggle, P. J. and Ribeiro, P. J. *Model-based Geostatistics.* Springer Series in Statistics. Springer, 2007.

Diggle, P. J., Thomson, M. C., Christensen, O., Rowlingson, B., Obsomer, V., Gardon, J., Wanji, S., Takougang, I., Enyong, P., Kamgno, J., et al. Spatial modelling and the prediction of loa loa risk: decision making under uncertainty. *Annals of Tropical Medicine & Parasitology*, 101(6):499–509, 2007.

Fine, P. E. A commentary on the mechanical analogue to the reed-frost epidemic model. *American journal of epidemiology*, 106(2):87–100, 1977.

Fischer, D., Thomas, S. M., Niemitz, F., Reineking, B., and Beierkuhnlein, C. Projection of climatic suitability for aedes albopictus skuse (culicidae) in europe under climate change conditions. *Global and Planetary Change*, 78(1-2):54–64, 2011.

for Disease Control, C. and Prevention. Lesson 1: Introduction to Epidemiology, Section 11: Epidemic Disease Occurrence, Level of disease. `https://www.cdc.gov/csels/dsepd/ss1978/lesson1/section11.html/`, 2012. Accessed: 2022-10-23.

Forbes, C., Evans, M., Hastings, N., and Peacock, B. *Statistical distributions.* John Wiley & Sons, New York, 4 edition, 2010.

Friedrich-Loeffler-Institut. Profile Bluetongue Disease. `https://www.fli.de/en/news/animal-disease-situation/bluetongue-disease/`, 2014. Accessed: 2019-06-23.

Ganter, M. Bluetongue disease—global overview and future risks. *Small Ruminant Research*, 118(1-3):79–85, 2014.

Gibson, G. J. Markov chain monte carlo methods for fitting spatiotemporal stochastic models in plant epidemiology. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(2):215–233, 1997.

Gloster, J., Mellor, P., Burgin, L., Sanders, C., and Carpenter, S. Will bluetongue come on the wind to the united kingdom in 2007? *Veterinary Record*, 160(13): 422–426, 2007.

Gloster, J., Burgin, L., Witham, C., Athanassiadou, M., and Mellor, Y. Bluetongue in the united kingdom and northern europe in 2007 and key issues for 2008. *Veterinary Record*, 162(10):298–302, 2008.

Green, P. J. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

Greenwood, P. and Gordillo, L. Stochastic epidemic modeling. In Chowell, G., Hyman, J., Bettencourt, L., and Castillo-Chavez, C., editors, *Mathematical and Statistical Estimation Approaches in Epidemiology*, pages 31–52. Springer, Dordrechtr, 2009.

Hamer, W. Epidemic disease in england—the evidence of variability and persistence. *The Lancet*, 167:733–738, 1906.

Hanski, I. A., Gaggiotti, O. E., and Gaggiotti, O. F. *Ecology, genetics and evolution of metapopulations*. Academic Press, 2004.

Hastings, W. K. *Monte Carlo sampling methods using Markov chains and their applications*. Oxford University Press, 1970.

Hawkes, A. G. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

Hoffman, M. D. and Gelman, A. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15 (1):1593–1623, 2014.

Jewell, C. and Brown, R. Bayesian data assimilation provides rapid decision support for vector-borne diseases. *Journal of The Royal Society Interface*, 2015.

Jewell, C., Kypraios, T., Neal, P., and Roberts, G. Bayesian analysis for emerging infectious diseases. *Bayesian Analysis*, 2009.

Jin, X., Carlin, B. P., and Banerjee, S. Generalized hierarchical multivariate car models for areal data. *Biometrics*, 61(4):950–961, 2005.

Jones, K. E., Patel, N. G., Levy, M. A., Storeygard, A., Balk, D., Gittleman, J. L., and Daszak, P. Global trends in emerging infectious diseases. *Nature*, 451 (7181):990, 2008.

Joseph, M. Exact sparse CAR models in Stan. `https://mc-stan.org/users/documentation/case-studies/mbjoseph-CARStan.html`, 2016. Accessed: 2020-05-09.

Kao, R. R. The impact of local heterogeneity on alternative control strategies for foot-and-mouth disease. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1533):2557–2564, 2003.

Keeling, M. and Rohani, P. Modeling infectious diseases in humans and animals. *Princeton University Press*, 2007.

Kermack, W. and McKendrick, A. A contribution to the mathematical theory of epidemics. *Proceedings of the Royals Society A*, 1927.

Koeppel, L., Siems, T., Fischer, M., and Lentz, H. H. Automatic classification of farms and traders in the pig production chain. *Preventive Veterinary Medicine*, 150:86 – 92, 2018.

Krige, D. G. A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6):119–139, 1951.

Kypraios, T. Efficient bayesian inference for partially observed stochastic epidemics and a new class of semi-parametric time series models. Master's thesis, Lancaster UNiversity, 2007.

Lekone, P. E. and Finkenstädt, B. F. Statistical inference in a stochastic epidemic seir model with control intervention: Ebola as a case study. *Biometrics*, 62(4): 1170–1177, 2006.

Lentz, H. H. K., Koher, A., Hövel, P., Gethmann, J., Sauter-Louis, C., Selhorst, T., and Conraths, F. J. Disease spread through animal movements: A static and temporal network analysis of pig trade in germany. *PLOS ONE*, 11:1–32, 05 2016.

Lindgren, F., Rue, H., and Lindström, J. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.

Maclachlan, N. J. Bluetongue: history, global epidemiology, and pathogenesis. *Preventive veterinary medicine*, 102(2):107–111, 2011.

McCann, R. S., Messina, J. P., MacFarlane, D. W., Bayoh, M. N., Gimnig, J. E., Giorgi, E., and Walker, E. D. Explaining variation in adult anopheles indoor resting abundance: the relative effects of larval habitat proximity and insecticide-treated bed net use. *Malaria journal*, 16(1):288, 2017.

McElreath, R. Markov Chains: Why Walk When You Can Flow? `https://elevanth.org/blog/2017/11/28/build-a-better-markov-chain/`, 2017. Accessed: 2020-10-29.

McKendrick, A. Applications of mathematics to medical problems. *Proceedings of Edinburgh Mathematical Society*, 44:98–130, 1926.

McKinley, T. J., Ross, J. V., Deardon, R., and Cook, A. R. Simulation-based bayesian inference for epidemic models. *Computational Statistics & Data Analysis*, 71:434–447, 2014.

Medlock, J. M. and Leach, S. A. Effect of climate change on vector-borne disease risk in the uk. *The Lancet Infectious Diseases*, 15(6):721–730, 2015.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

Mugenyi, L., Abrams, S., and Hens, N. Estimating age-time-dependent malaria force of infection accounting for unobserved heterogeneity. *Epidemiology & Infection*, 145(12):2545–2562, 2017.

Murray, N. E. A., Quam, M. B., and Wilder-Smith, A. Epidemiology of dengue: past, present and future prospects. *Clinical epidemiology*, 5:299, 2013.

Neal, P. and Roberts, G. A case study in non-centering for data augmentation: stochastic epidemics. *Statistics and Computing*, 15(4):315–327, 2005.

Neal, R. M. et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

Nelder, J. A. and Wedderburn, R. W. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

Nishio, M. and Arakawa, A. Performance of hamiltonian monte carlo and no-u-turn sampler for estimating genetic parameters and breeding values. *Genetics Selection Evolution*, 51(1):73, 2019.

Okabe, A., Boots, B., Sugihara, K., and Chiu, S. N. *Spatial tessellations: concepts and applications of Voronoi diagrams.* John Wiley & Sons, 2009.

O'Neill, P. and Roberts, G. Bayesian inference for partially observed stochastic epidemics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 162(1):121–129, 1999.

Ostfeld, R. S., Glass, G. E., and Keesing, F. Spatial epidemiology: an emerging (or re-emerging) discipline. *Trends in Ecology and Evolution*, 20(6):328 – 336, 2005.

Papaspiliopoulos, O. and O Roberts, G. Non-centered parameterisations for hierarchical models and data augmentation. 7:307–326, 01 2003.

Parham, P. E., Waldock, J., Christophides, G. K., Hemming, D., Agusto, F., Evans, K. J., Fefferman, N., Gaff, H., Gumel, A., LaDeau, S., et al. Climate, environmental and socio-economic change: weighing up the balance in vector-borne disease transmission. *Phil. Trans. R. Soc. B*, 370(1665), 2015.

Penrose, M. et al. *Random geometric graphs*, volume 5. Oxford university press, 2003.

Preciado, V. M. and Jadbabaie, A. Spectral analysis of virus spreading in random geometric networks. In *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*, pages 4802–4807. IEEE, 2009.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, 2 edition, 1992.

Purse, B. V., Mellor, P. S., Rogers, D. J., Samuel, A. R., Mertens, P. P., and Baylis, M. Climate change and the recent emergence of bluetongue in europe. *Nature Reviews Microbiology*, 3(2):171, 2005.

Quinonero-Candela, J. and Rasmussen, C. E. A unifying view of sparse approximate gaussian process regression. *The Journal of Machine Learning Research*, 6:1939–1959, 2005.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

Riley, S., Eames, K., Isham, V., Mollison, D., and Trapman, P. Five challenges for spatial epidemic models. *Epidemics*, 10:68–71, 2015.

Roberts, G. O., Gelman, A., Gilks, W. R., et al. Weak convergence and optimal scaling of random walk metropolis algorithms. *The annals of applied probability*, 7(1):110–120, 1997.

Ross. University of Münster, Mathematics, Lecture Notes Stochastik: Chapter I The Poisson Process, 2013. URL: `https://www.uni-muenster.de/Stochastik/lehre/WS1314/BachelorWT/Daten/StPro_Ross1.pdf`. Accessed 2019/02/12.

Ross, R. The prevention of malaria, with addendum on the theory of happenings. *Murray, London*, 1911.

Rubin, D. B. Bayesian data analysis. 2013.

Rue, H. and Held, L. *Gaussian Markov random fields: theory and applications*. Chapman and Hall/CRC, 2005.

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2:e55, 2016.

Samat, N. and Percy, D. Vector-borne infectious disease mapping with stochastic difference equations: an analysis of dengue disease in malaysia. *Journal of Applied Statistics*, 39(9):2029–2046, 2012.

Science Media Centre. bluetongue disease. https://www.sciencemediacentre.org/blue-tongue-disease/, 2007.

Serfling, R. Historical review of epidemic theory. *Human Biology*, 24(3):145–166, 1952.

Sumner, T., Orton, R. J., Green, D. M., Kao, R. R., and Gubbins, S. Quantifying the roles of host movement and vector dispersal in the transmission of vector-borne diseases of livestock. *PLoS computational biology*, 13(4):e1005470, 2017.

Szmaragd, C., Wilson, A. J., Carpenter, S., Wood, J. L., Mellor, P. S., and Gubbins, S. A modeling framework to describe the transmission of bluetongue virus within and between farms in great britain. *PLOS ONE*, 4(11):e7741, 2009.

Tierney, L. Markov chains for exploring posterior distributions. *the Annals of Statistics*, pages 1701–1728, 1994.

White, M. T., Griffin, J. T., Drakeley, C. J., and Ghani, A. C. Heterogeneity in malaria exposure and vaccine response: implications for the interpretation of vaccine efficacy trials. *Malaria journal*, 9(1):82, 2010.

Woodroffe, R., Donnelly, C., Ham, C., Jackson, S., Moyes, K., Chapman, K., Stratton, N., and Cartwright, S. Badgers prefer cattle pasture but avoid cattle: implications for bovine tuberculosis control. *Ecology Letters*, 2016.

World Climate Research Programme (WCRP) (2016): CMIP5 monthly data on single levels, IPSL-CM5A-MR (IPSL, France), Copernicus Climate Change Service (C3S) Climate Data Store (CDS). `https://cds.climate.copernicus.eu/cdsapp#!/dataset/projections-cmip5-monthly-single-levels?tab=overview`. Accessed 2019/07/02.

World Health Organization. Vector-borne diseases. Technical report, WHO Regional Office for South-East Asia, 2014.

Xiang, F. and Neal, P. Efficient mcmc for temporal epidemics via parameter reduction. *Computational Statistics & Data Analysis*, 80:240–250, 2014.