Statistical Methods for Campylobacter Outbreak Detection using Genomics and Epidemiological Data

by

Laura Marcela Guzmán Rincón

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Mathematics for Real-World Systems

September 2020

THE UNIVERSITY OF WARWICK

To my brother Juan Pablo,
in loving memory.

Contents

List of Tables

List of Figures

Acknowledgments

I would like to thank my supervisors Noel McCarthy and Simon Spencer, not only for their valuable advice but also for their help, their feedback throughout the thesis, their kindness and their support in this project, which made it enjoyable and worth it. Also, I am very grateful for their help to find funding for this programme.

I would like to thank the EPSRC and Public Health England for funding this PhD, to the University of Warwick and MathSys for the opportunity to course this programme. I am grateful to Colm Connaughton for his help to find funding for my PhD and to develop my professional career. Also, Stefan Grosskinsky and Yulia Timofeeva for their support. I would like to thank the Mathsys department, staff, colleagues and friends for making such an enjoyable journey. Especially I would like to thank Heather Robson for her kind support and her friendship during these years.

Also, I want to thank all my friends on this journey. Especially Charlotte for her amazing friendship and nice discussions about life, Odettis for being such a great companion, Nata for her happiness and kindness, Jack for all the discussions to improve this thesis, Pablo for his great friendship, Paul for his nice discussions about maths, Guillem for his valuable support, the climbing crew Juan, Jack, Odette and Edu for all the amazing time and their awesome friendship. Finally, I would like to thank all my family for their love. My mom for her infinite support in everything I achieve, for her infinite love, and also for proofreading this thesis and providing her wise comments. For my sister Carolina, who I admire and love with my hearth. For my brother Juan Pablo (RIP) who didn't see me finish this journey but he is always inspiring me. To my beloved grandparents (RIP) who could not celebrate with me this time.

Declarations

I declare that that the work contained herein is my own except where explicitly stated otherwise. This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree.

Abstract

Campylobacter infections are the main bacterial cause of gastroenteritis in the UK, causing an estimated 500 thousand cases per year. Health authorities investigate outbreaks to identify the source, control the spread and understand the cause. Outbreak detection mechanisms are potentially improved by the increasing availability of whole-genome sequence alongside other epidemiological data. However, techniques mixing genomics and other epidemiological factors are still underdeveloped. This project aims to develop and apply outbreak detection methods using surveillance data collected from two regions in the UK. The approaches proposed in this thesis are based on an existing spatial-temporal Bayesian hierarchical model, where cases are labelled as potential outbreaks if they comprise an elevated number of cases compared to the expected sporadic count. The model is adjusted to include genetic data using Gaussian random fields, exploiting the capacity of whole-genome sequencing to discriminate closely related isolates. Moreover, a Markov Chain Monte Carlo algorithm is implemented to obtain the posterior distribution of the model parameters. In particular, a sampling strategy is proposed to improve the convergence of the chain for the parameters describing the Gaussian random field. The project dataset is analysed using a spatial-temporal, a spatial-genetic and a temporal-genetic version of the model, where each version explores different types of outbreaks. The proposed approach demonstrates how to organise genetic sequences into a high-dimensional structure and incorporate them into a Bayesian framework. Also, the MCMC sampling algorithm improves the mixing of the chain to estimate the posterior distribution of the model parameters. Finally, all model versions provide the probability that each reported infection is part of a potential outbreak. Comparing the potential outbreaks found by each model provides insights to estimate the real outbreaks. It also identifies cases that are potentially part of a diffuse real outbreak hard to detect by existing approaches. Despite the capability of the model, it requires predefined outbreak sizes and therefore is not flexible at capturing many shapes. Autocorrelated models are a potential improvement to be explored.

Abbreviations

AHC	Agglomerative Hierarchical Clustering
AUC	Area Under the Curve
BHM	Bayesian Hierarchical Model
BIGSdb	Bacterial Isolate Genome Sequence Database
BLAST	Basic Local Alignment Search Tool
BYM	Besag-York-Mollié
CDC	Centers for Disease Control and Prevention
cgMLST	Core Genome Multilocus Sequence Typing
CM	Correlated Model
EARS	Early Aberration Reporting System
ESS	Effective Sample Size
FSA	Food Standards Agency
FPR	False Positive Rate
GAM	Generalized Additive Model
GAMM	Generalized Additive Mixed Models
GMRF	Gaussian Markov Random Field
GRF	Gaussian Random Field
IM	Independent Model
LGCP	Log-Gaussian Cox Process
LSOA	Lower Layer Super Output Area
MCMC	Markov chain Monte Carlo
M-H	Metropolis-Hastings

MLEE	Multilocus Enzyme Electrophoresis
MLST	Multilocus Sequence Typing
MSOA	Middle Layer Super Output Area
NE	North-East Dataset
ONS	Office for National Statistics
OX	Oxfordshire Dataset
PFGE	Pulsed-Field Gel Electrophoresis
PHE	Public Health England
PubMLST	Public databases for molecular typing
RAMMIE	Rising Activity, Multilevel Mixed Effects, Indicator Emphasis
RJMCMC	Reversible-Jump Markov Chain Monte Carlo
rMLST	Ribosomal Multilocus Sequence Typing
ROC	Receiver Operating Characteristic
RW	Random Walk
SE	Squared Exponential kernel
SG	Spatial-genetic model
ST	Spatial-temporal model
TG	Temporal-genetic model
TPR	True Positive Rate
UK	The United Kingdom
wgMLST	Whole-Genome Multi Locus Sequence Typing
WGS	Whole Genome Sequencing

Chapter 1

Introduction

1.1 Topic background

Zoonotic illnesses are a consequence of the complex interactions between humans, animals and their environment. Caused by those interactions, humans are exposed to pathogens present in the products they consume, their contact with animals, and their interaction with the environment. In particular, foodborne infections can be caused by bacteria as *Campylobacter* genus, the main cause of bacterial gastroenteritis worldwide [Kaakoush et al., 2015]. Infections are commonly induced by the ingestion of contaminated poultry, unpasteurised milk and contaminated water [Little et al., 2010; Kaakoush et al., 2015]. Moreover, it causes an estimated 500 thousand infections and 80 thousand general practitioner consultations per year in the UK [Tam et al., 2012]. Public health authorities have a critical role in diminishing the burden of cases and reducing the impact on human health and the global economy [Tauxe et al., 2010].

Understanding the biology of *Campylobacter* and the infections from source to humans has a substantial impact on food-safety policies and the subsequent reduction of the burden [Gerner-Smidt et al., 2017; Tauxe et al., 2010]. Besides the complexity of food-human interaction, the detailed study of outbreaks provides insights into the sources of contamination and supports the strengthening of public control measures [Gormley et al., 2011]. Surveillance systems operate on a routine basis to collect data of reported infections and to investigate potential outbreaks [Gormley et al., 2011]. However, the number of identified outbreaks is low compared to the total burden [Gormley et al., 2011; Frost et al., 2002; Pebody et al., 1997].

Outbreak detection can be improved with the inclusion of quantitative methods into surveillance systems. For instance, Public Health England (PHE) and the Centers

for Disease Control and Prevention (CDC) incorporate outbreak detection algorithms into routine surveillance [Noufaily et al., 2019]. These approaches detect unexpected increases in reported infections, based on daily counts. Similarly, several methodologies have been developed to study the agglomerative nature of spatial reports [Ripley, 1976; Diggle, 2013], or to identify spatial outbreaks [Kulldorff, 1997; Knorr-Held and Raßer, 2000]. Spatiotemporal approaches have also been proposed [Spencer et al., 2011; Kulldorff, 2001], demonstrating the impact of incorporating several data sources into these systems.

The inclusion of genetics into surveillance systems potentially improves the detection of outbreaks [McCarthy, 2017; Cody et al., 2013]. Previous studies have shown how whole-genome sequencing (WGS) of bacteria can distinguish samples that are epidemiologically related [McCarthy, 2017]. For that purpose, a project between the Food Standards Agency (FSA), PHE and the University of Oxford implemented a surveillance system for human *Campylobacter* infections using WGS in a few areas of the UK. The mix of genomic and other epidemiological data provides the means to test combined algorithms for outbreak detection.

The goal of this project is to develop mathematical techniques for outbreak detection, based on the integration of bacterial genomic and epidemiological routine data (Chapter 2). For that purpose, an existing spatiotemporal approach was adapted to incorporate genetic data. First, the adaptation required an exhaustive exploration of the mathematical properties of whole-genome sequences (Section 2.5.1). Then, a general framework was established to provide different methods for different sources of data: spatial, temporal or genetic. As described throughout this project, three Bayesian statistical models were analysed. First, the spatial-temporal model developed by Spencer et al. [2011] was applied to the project dataset (Section 3.1). The genetic sequences linked to the resulting outbreaks were studied to validate the results. Second, a spatial-genetic was designed and implemented, incorporating a Gaussian process to include genetic data into the model structure by Spencer et al. [2011] (Section 4.3). Also, a sampling technique was proposed to improve the model implementation (Section 4.4). Third, a temporal-genetic model was created and implemented (Section 5.2). Finally, outbreaks obtained by each model were contrasted as well as the variations among each model results (Section 6.1).

Methods for outbreak detection, health protocols and communication for the general pubic require a clear definition of outbreaks. Definitions might vary from groups of cases in which at least two infected people might have the same common exposure [Pebody et al., 1997], to infections caused by genetically related pathogens [Robinson

et al., 2013]. In this project, an outbreak is defined as the occurrence of more cases of the infection than expected.

Previous knowledge of several fields is required for the development of this project. First, an introduction to *Campylobacter* species and the epidemiology of the infection is included (Section 1.1.1), as well as the overview of existing *clustering* methods for outbreak detection using genomics (Section 1.2). Moreover, the construction of models and the exploration of genetic sequences requires random fields theory (Section 1.3.1). Since all methodologies are based on Bayesian modelling, the project requires the theory of Bayesian statistics and inference (Section 1.3.2). Spatial analysis using point processes is required to study genetic sequences (Section 1.3.3), as well as hierarchical clustering algorithms (Section 1.3.4). Finally, a review of existing clustering methods using mathematical modelling is included (Section 1.4).

The purpose of this chapter is to establish the goal of the project and provide the background knowledge required to follow the models and arguments discussed throughout the document. The current section, Section 1.1, provides an overview of the project and the introduction to *Campylobacter* species and genetic sequencing. Section 1.2 reviews the current methodologies in outbreak detection in genomics. Section 1.3 summarises the main fields required in the mathematical framework of this project. Finally, Section 1.4 provides an overview of the mathematical models for outbreak detection and disease modelling.

1.1.1 *Campylobacter* epidemiology

Campylobacter is the main bacterial cause of gastroenteritis in many industrialised countries, causing an estimate of 9.3 cases per 1000 person per year in the UK [Tam et al., 2012]. Due to health care and productivity expenditure, legal costs and other expenses, it produced an estimated annual cost of £50 million in the UK in 2008 [Tam and O'Brien, 2016]. The bacterium is found naturally in the intestinal tract of a wide range of warm-blooded animals. Therefore, the slaughter process of poultry and other meat products might be a cause of contamination of carcasses ready for distribution [Silva et al., 2011]. As a consequence, food products derived from poultry, cattle and other animals, unpasteurised milk and contaminated water are usually vehicles of *Campylobacter* poisoning in humans [Silva et al., 2011], although differences in exposure have been found among *Campylobacter* species [Gillespie et al., 2002]. Common symptoms are acute diarrhoea, fever, and abdominal cramping, with a usual incubation period of 24 to 72 hours [Blaser, 1997].

It is estimated that age, season, immunity, and demographic factors influence the prevalence of campylobacteriosis cases [Silva et al., 2011]. Infections are more frequent in children younger than four years old and young adults [Kaakoush et al., 2015; Nichols et al., 2012]. Prevalence is higher in males of all age groups, except for a higher incidence in females between twenty and thirty-six years old [Louis et al., 2005; Gillespie et al., 2008]. Infection reports are lower on weekends, presumably caused by lower access to health services [Nichols et al., 2012]. A 10-years study showed a consistent summer peak of incidences starting in late-May with a maximum between mid-June and mid-July [Louis et al., 2005; Nichols et al., 2012]. The summer peak was also observed during the same weeks in ten different countries with seasonal weather [Nylen et al., 2002; Louis et al., 2005]. In a two decades study, the seasonal trend is correlated with temperature and farming environments [Louis et al., 2005; Nichols et al., 2012]. Incidence among age also exhibits seasonality, with a higher seasonal trend in children younger than four years old [Louis et al., 2005; Nichols et al., 2012]. Although trends are consistent among countries, the number of reported cases between and within countries differ. These differences are possibly due to surveillance methodologies, food practices and environmental exposure [Kaakoush et al., 2015]. The incidence in the UK was positively correlated with the number of rural wards and negatively correlated with population density and deprivation [Louis et al., 2005; Nichols et al., 2012].

1.1.2 Sequence data for bacterial isolates

The sporadic nature of infections, the underreported incidences, the lack of representative samples and the high genetic diversity complicate the categorisation of the bacteria [Dingle et al., 2002]. Therefore, understanding the dynamics and the epidemiology of the disease requires the study of the genetic diversity and evolution of bacteria. *Campylobacter jejuni*, the most common cause of reported campylobacteriosis infections, was firstly sequenced in 2000, revealing 1.641.481 base pairs and 1.654 hypothetical proteins [Parkhill et al., 2000]. Additionally, it has been shown that *C. jejuni* has a weakly clonal population structure and that isolates with diverse origins share same alleles, implying horizontal genetic exchange [Dingle et al., 2001]. Facing these conditions demands a system that collects genetic information of isolates, compares and analyses them to identify the association between reported cases.

Several typing techniques have been developed to classify and trace back isolates of *Campylobacter* species, including phenotypic and genotypic procedures. Phenotypic typing methods are usually based on serotyping. However, the large number of strains

unable to be typed, and the expensive and time-consuming process have restricted the use of these methods [Wassenaar and Newell, 2000]. In contrast, genotypic subtyping such as ribotyping, pulsed-field electrophoresis PFGE, multilocus enzyme electrophoresis MLEE, and *fla* typing have been implemented [Wassenaar and Newell, 2000]. For instance, systems based on the 16s ribosomal RNA gene sequence succeeded in identifying and classifying species, but more resolution was necessary for closely related isolates [Maiden et al., 2013]. Additionally, standardisation of these procedures is still needed, in order to compare large amounts of isolates obtained from diverse locations. To solve these drawbacks, the multilocus sequence typing MLST was proposed [Maiden et al., 1998], providing higher resolution and performing comparisons based on alleles differences rather than the variations of single nucleotides (point mutations as in PFGE). Particularly, a scheme for *C. jejuni* and *Campylobacter coli* was defined based on seven housekeeping genes which provided sufficient diversity to discriminate among isolates [Dingle et al., 2001]. Similarly, seven loci schemes for other *Campylobacter* species have been introduced [Miller et al., 2005].

Isolate discrimination often requires different levels of resolution, depending on the genetic closeness of the samples. This resolution requires the introduction of new typing schemes; particularly, several modifications of the seven-loci MLST described previously have been proposed [Maiden et al., 2013]. The number of loci included in those schemes depends on the taxonomic discrimination required. For instance, whole-genome MLST compares all loci of a set of closed genomes. If not all loci are shared, a core-genome MLST is applied, comparing all available genes. Particularly, a scheme based on the comparison of the 53 ribosomal protein subunit genes, ribosomal MLST, has been proposed, since those genes are commonly founded in all bacteria but have the variability to discriminate among samples [Jolley et al., 2012].

A new generation of sequencing methods has emerged with the development of high-throughput sequencing technologies. Now, it is possible to produce genome reads and obtain Whole Genome Sequencing (WGS) data of a bacteria in a single experiment at a low cost. These technologies produce short or long raw sequencing reads, which are subject to a process for the reconstruction of the whole genome. For instance, the reads can be aligned and compared to a reference genome (read mapping approach) or they can be divided into several pieces to construct an assembly graph and infer the real genome (de novo approach) [Bakker et al., 2017; Loman et al., 2012]. Then, to identify the location of genes in the obtained genome, it is aligned and compared to a set of sequences of similar strains (a set of reference genomes). When aligned genes show similarities, the annotation of the original sequences is transferred to the new one.

Several algorithms are used to perform this task, such as BLAST or FASTA [Richardson and Watson, 2013].

The opportunity to generate WGS data requires a mechanism to store and organise large amounts of data from different sources. The Bacterial Isolate Genome Sequence Database BIGSdb provides a structured system to meet this requirement, extending previous MLST approaches [Jolley and Maiden, 2010]. It consists of two components: the isolate database and the sequence-definition database [Maiden et al., 2013]. When a new sequence is introduced, it is defined as a new allele and scanned in the definition database using the BLAST algorithm. If the allele does not exist, it is created and the sequence is stored in the isolate database. The system also provides flexibility in the inclusion of new schemes (such as rMLST). Additionally, for a set of isolates, BIGSdb calculates a distance matrix accounting the number of allelic differences between each pair of isolates and generates a graph using an algorithm called NeighborNet [Bryant and Moulton, 2004].

1.2 Application of genomics to outbreak detection

In Section 1.1.2, it has been discussed how the multilocus sequence typing of *Campylobacter* species provides an effective means to characterise and discriminate among isolates. In this section, it will be discussed how these approaches offer a new dimension to understand the dynamics of the disease.

Several questions about the epidemiology of campylobacteriosis have been explored since the disease was discovered. However, the inclusion of genomic data provides these studies with novel perspectives. For instance, the distribution of subtypes of *Campylobacter* could be analysed for different locations (i.e. separate countries) [McCarthy et al., 2012]; the incidence peak observed in summer could be explained by a genotypic feature favored by the temperature [McCarthy et al., 2012; Cody et al., 2012]; and the observed differences in incidences as a function of age, gender, or symptomatic reactions of infected patients could be examined [Dingle et al., 2008]. Similarly, it has been demonstrated the contamination of flocks through the slaughter process could be a route of transmission of *Campylobacter* [Colles et al., 2010]. Two additional issues have been especially addressed: source attribution and outbreak detection. This project is focussed fundamentally on outbreak detection.

Outbreak investigations are focussed on the study of abrupt or unanticipated changes in the number of cases of campylobacteriosis. Although definitions vary among studies, they are mainly centered on finding two or more cases with a common exposure,

or observing a higher than expected number of incidences in a period of time, a limited geographical area or a group with similar characteristics[1]. Identifying such outbreaks allow researchers to design intervention strategies and understand the pathways and sources of infections [Little et al., 2010]. Between 2008 and 2015, 143 outbreaks were reported in the UK, with an average size of 26 cases and having poultry as the most common source [Kaakoush et al., 2015; Little et al., 2010]: however, registered outbreaks comprise less than 1% of the reported cases [McCarthy, 2017], a low percentage considering the properties of the *Campylobacter* genus. Slow growing speed and the ability to survive in food products raise the question if apparently sporadic cases are also part of diffuse outbreaks, generated early in the food chain; therefore, being hard to detect [McCarthy, 2017].

Typing methods potentially provide tools to differentiate and compare among strains collected in hospitals, study the diversity of already detected outbreaks, and find possible outbreaks not identified by epidemiological means. In particular, WGS can provide higher resolution in the detection of outbreaks for different types of pathogens; for instance, for bacteria with low levels of recombination, WGS can identify strains and study the evolution of bacteria as in an outbreak of Mycobacterium tuberculosis [Roetzer et al., 2013]. Similarly, WGS and novel schemes of MLST are suitable for understanding the diversity and structure of *Campylobacter*, assessing the relatedness among isolates within a potential outbreak [Llarena et al., 2017]. Nevertheless, the frequent horizontal transference of *Campylobacter* hinders the design of such procedures.

Understanding the nature of campylobacteriosis outbreaks requires the retrospective analysis of already reported ones. Those investigations elucidate the genomic diversity within epidemiologically linked isolates and the existence of possible undetected diffuse outbreaks. Consequently, several studies have retrospectively applied WGS to collected isolates from outbreaks, using, for instance, MLST schemes, the PubMLST databases, and the NeighborNet algorithms. First, a comparison between isolates from a milk-borne outbreak and the background population showed the low genetic variation of isolates contaminated from a common source [Fernandes et al., 2015]. Also, it suggested the existence of diffusely distributed outbreaks and the difficulty of their detection by epidemiological means. Similarly, based on a hierarchical approach as proposed previously ([Maiden et al., 2013; Cody et al., 2013]), the isolates collected on a summer peak in Finland revealed the occurrence of genetically identical isolates spread in different districts of the country [Kovanen et al., 2014]. Subsequently, patient and animal samples of that peak were examined to determine possible sources [Kovanen

[1] http://www.who.int/foodsafety/publications/foodborne_disease/outbreak_guidelines.pdf

et al., 2016]. A similar source analysis was performed for isolates in a milk-borne outbreak, comparing sequences from patients, cattle and contaminated milk [Revez et al., 2014]. Although retrospective investigations improve the understanding of outbreaks, they are limited to the collected epidemiological information [Gerner-Smidt et al., 2017]. Real-time would solve this drawback, guiding the surveillance system and driving the collection of epidemiological data.

Although WGS has been mostly applied to retrospective analysis [Gerner-Smidt et al., 2017], the inclusion of WGS into the surveillance system is required for the early detection of outbreaks. Recently, a hierarchical gene-by-gene approach has been proposed to demonstrate how WGS could be valuable in the real-time characterisation of *C. jejuni* and *C. coli* [Cody et al., 2013]. For the analysis of isolates collected in Oxfordshire, the following steps were applied. First, the BIGSdb autotagger assigned alleles, sequence types and clonal complexes to the collected samples and compared them with the distribution of a control population. Second, phylogenetic trees were generated according to an initial MLST scheme to differentiate within clonal complexes. Finally, the number of loci included in the scheme was increased to improve resolution inside the clusters founded on the phylogenies. For instance, rMLST profiles can be identified and subsequently wgMLST comparisons can be performed. PubMLST repositories and the online tools provided by the BIGSdb genome comparator module facilitate the implementation of this approach in real time. For the Oxfordshire data, locus differences among all isolates were compared against samples taken repeatedly from patients, suggesting a cutoff value to determine if isolates are clustered or not (20 loci or fewer). Also it showed that, despite the complex structure and evolution of *Campylobacter*, it is possible to determine if two isolates are part of the same transmission route. Application of this procedure is gaining acceptance and has been applied in other studies [Kovanen et al., 2014, 2016; Llarena et al., 2017].

1.3 Mathematical background

1.3.1 Random fields

A stochastic process f over an *index space* S is a set of random variables $\{f(s) : s \in S\}$ where f takes values on a *state space*. The process f is a *random field* if the state space is a subset of a Euclidean space. Random fields are commonly used in time series analysis, where S is a subset of $\mathbb{R}$, and in spatial analysis, where S is a subset of $\mathbb{R}^2$. A random field is a *real-valued Gaussian random field* or *Gaussian process* if for every finite

subset $\{s_1, ..., s_n\}$ of S of size $n \in \mathbb{Z}^+$, the vector $(f(s_1), ..., f(s_n))$ follows a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} = [c(s_i, s_j)]_{ij}$, where $c(.,.)$ is a function or *kernel* that maps any pair (s_i, s_j) into the covariance of $f(s_i)$ and $f(s_j)$. For the $\boldsymbol{\Sigma}$ to be a valid covariance matrix, it must be positive semi-definite. The function c is called a *covariance function* if the matrix $[c(.,.)]$ is positive semi-definite.

Covariance matrices encode key properties of the process f, like the continuity or the smoothness. A special type of covariance functions are defined when S is a subset of a d-dimensional Euclidean space. The covariance $c(s,t)$ is *isotropic* if it can be written as a function of $|s-t|$, and it is expressed as a single-valued function $k(r)$ on the non-negative real numbers. Isotropic covariance functions are characterised as Fourier transforms [Rasmussen and Williams, 2005; Stein, 1999], where the power spectrum provides information about the smoothness of the process. The most common type of isotropic functions is the Square Exponential. It is given by:

$$k(r) \propto \exp\left(-\frac{r^2}{2l^2}\right), \tag{1.1}$$

with a real-valued *length-scale* parameter l, $l > 0$. The main property of the Squared Exponential is that processes drawn by this kernel are infinitely differentiable. The unrealistic smoothness property of this function is improved by the Matérn class of covariance functions [Stein, 1999]. Based on the modified Bassel functions B, it is defined as:

$$k_\nu(r) \propto \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}r}{l}\right)^\nu B_\nu \left(\frac{\sqrt{2\nu}r}{l}\right), \tag{1.2}$$

where the positive real-valued parameters ν and l. Two simplified versions of this Matérn classes are given by $\nu = 1/2$ and $\nu = 3/2$. These covariance functions are implemented in Section 4.2.3 for the analysis of genetic sequences.

Other types of Gaussian random fields are defined according to the structure of S. In particular, *Markov Gaussian Random Fields* (GMRFs) are studied when S is finite and has a *neighbouring structure*. Formally, let (S, E) be a graph with vertices S and edges E. Any two elements s, t in S are called *neighbours* if there is an edge connecting them: $(s,t) \in E$. A random field f is a GMRF if the following condition holds for every s, t in S: $f(s)$ and $f(t)$ are conditionally independent given all the other values of f, if and only if s and t are not neighbours. This condition is equivalent to:

$$P(\{f(t)|t \in S \backslash s\}) = P(\{f(t)|t \in N(s)\})$$

for all s in S, where $N(s)$ is the set of neighbours of s.

1.3.2 Bayesian modelling and inference

Bayesian statistics provides a mathematical framework for learning from data. Bayesian modelling employs probability theory to construct and parameterise a model that could explain the data. Also, it can incorporate prior beliefs and expert knowledge and can be used to compare hypotheses given the data. The fundamentals of Bayesian modelling rely on Bayes' theorem. Let $P(\boldsymbol{\mathcal{D}}|\boldsymbol{\theta})$ be the likelihood function describing the data $\boldsymbol{\mathcal{D}}$ given a vector of parameters $\boldsymbol{\theta}$. The uncertainty of the parameters is learned from the data $P(\boldsymbol{\theta}|\boldsymbol{\mathcal{D}})$ subject to prior assumptions $P(\boldsymbol{\theta})$. That is,

$$P(\boldsymbol{\theta}|\boldsymbol{\mathcal{D}}) = \frac{P(\boldsymbol{\mathcal{D}}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}'} P(\boldsymbol{\mathcal{D}}|\boldsymbol{\theta}')P(\boldsymbol{\theta}')d\boldsymbol{\theta}'}, \tag{1.3}$$

where $P(\boldsymbol{\theta}|\boldsymbol{\mathcal{D}})$ is the posterior distribution of $\boldsymbol{\theta}$, $P(\boldsymbol{\mathcal{D}}|\boldsymbol{\theta})$ is the likelihood of the model, $P(\boldsymbol{\theta})$ is the prior of $\boldsymbol{\theta}$, and the integral in the denominator is the marginal probability of the data.

Some models exhibit hierarchical structures organised in sub-models, known as *Bayesian hierarchical models* (BHM). BHMs are organised in conditionally independent layers. The simplest hierarchical structure can be described as follows:

$$P(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{\mathcal{D}}) \propto P(\boldsymbol{\mathcal{D}}|\boldsymbol{\lambda})P(\boldsymbol{\lambda}|\boldsymbol{\theta})P(\boldsymbol{\theta}). \tag{1.4}$$

Therefore, the uncertainties in the model are propagated through the hierarchical structure. The intermediate layer $P(\boldsymbol{\lambda}|\boldsymbol{\theta})$ or latent field is formed of *latent parameters* $\boldsymbol{\lambda}$. This structure explains data produced by complex interactions and captures the uncertainty of those interactions.

GMRFs are frequently used as priors for latent fields [Banerjee et al., 2004]. To incorporate a GMRF into a hierarchical structure as in (1.4), let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_L)$ be a vector of L real-valued latent variables. The vector $\boldsymbol{\lambda}$ can be modelled as the realisation of a GMRF. That is, there is a finite space S such that for each i in $\{1, \ldots, L\}$ there exits an s in S such that $\lambda_i = f(s)$. The prior of the latent parameters can be written as a function of a penalty matrix $\boldsymbol{Q}(\boldsymbol{\theta})$ (or equivalently, $\boldsymbol{Q}$):

$$P(\boldsymbol{\lambda}|\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}\boldsymbol{\lambda}^T\boldsymbol{Q}\boldsymbol{\lambda}\right).$$

In the context of time series and spatial statistics, GMRFs are commonly used as latent

fields in hierarchical models [Rue and Martino, 2007; Blangiardo et al., 2013], as in [Besag et al., 1991; Knorr-Held and Richardson, 2003; Beneš et al., 2005; Spencer et al., 2011].

Bayesian inference

The exact computation of the posterior distribution in (1.3) requires the calculation of the marginal probability, usually involving intractable integrals. Instead, many sampling methods aim to draw from the posterior distribution $P(\boldsymbol{\theta}|\boldsymbol{\mathcal{D}})$. For instance, Markov Chain Monte Carlo methods (MCMC) collect simulated samples from the desired distribution by drawing a Markov Chain $(\boldsymbol{\theta_t})_{t=1}^{\infty}$ in the parameter space Θ. Each $\boldsymbol{\theta_t}$ is sampled from a transition distribution $q(\boldsymbol{\theta}^*; \boldsymbol{\theta_{t-1}})$. If the chain fulfils the detailed balance condition $P(\boldsymbol{\theta'}|\boldsymbol{\mathcal{D}})q(\boldsymbol{\theta'}; \boldsymbol{\theta}) = P(\boldsymbol{\theta}|\boldsymbol{\mathcal{D}})q(\boldsymbol{\theta}; \boldsymbol{\theta'})$, the chain is ergodic; that is, $\lim_{t\to\infty} P(\boldsymbol{\theta_t}|\boldsymbol{\mathcal{D}}) = P(\boldsymbol{\theta}|\boldsymbol{\mathcal{D}})$.

Examples of MCMC methods are Gibbs sampling and the Metropolis-Hastings algorithm. Both methods are described in Algorithm 1 and Algorithm 2, respectively. Let $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$ be a vector of parameters and $P(\boldsymbol{\theta}|\boldsymbol{\mathcal{D}})$ the desired function. The Gibbs sampling requires to sample analytically from the conditional probability distributions $p_i = P(\theta_i|\theta_j, j = 1, ..., i-1, i+1, ..., n)$. At each iteration t, a sample θ_i^* is taken from p_i while the other parameters are fixed. The Metropolis-Hastings algorithm collect samples using the following steps. At iteration t, a proposal value $\boldsymbol{\theta^*}$ for $\boldsymbol{\theta}$ is drawn from a transition distribution $q(\boldsymbol{\theta}|\boldsymbol{\theta_{t-1}})$. The sample is accepted with probability $\alpha(\boldsymbol{\theta^*}|\boldsymbol{\theta_{t-1}}))$ or rejected with probability $1 - \alpha(\boldsymbol{\theta^*}|\boldsymbol{\theta_{t-1}}))$, as in Algorithm 2.

Algorithm 1: Gibbs sampling

```
Result: (θ^(1), ..., θ^(T))
initialise θ^(0) ~ q_0;
for iteration t = 1, ..., T do
    for parameters i = 1, ..., n do
        sample from conditional θ_i^(t) ~ P(θ|θ_1^(t), ..., θ_{i-1}^(t), θ_{i+1}^(t-1), ..., θ_n^(t-1));
    end
end
```

Bayesian inference is typically performed by alternating MCMC techniques, depending on the properties of each parameter. Proposals can be obtained using single-site updates, where parameters are updated one at a time. Conversely, updating can also be implemented using block strategies where sets of parameters are updated simultaneously. This approach is relevant when parameters are highly correlated since single-site

Algorithm 2: Metropolis-Hastings algorithm

Result: $(\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(T)})$
initialise $\boldsymbol{\theta}^{(0)} \sim q_0$;
for *iteration* $t = 1, ..., T$ **do**
  sample from proposal $\boldsymbol{\theta}^* \sim q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t-1)})$;
  compute acceptance probability:
  $\alpha(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)}) = \min\left\{1, \frac{P(\boldsymbol{\theta}^*|\boldsymbol{\mathcal{D}})q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}^*)}{P(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\mathcal{D}})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)})}\right\}$;
  sample $u \sim \text{Unif}(0, 1)$;
  if $\alpha(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)}) > u$ **then**
    $\boldsymbol{\theta}^{(t)} \leftarrow \boldsymbol{\theta}^*$;
  else
    $\boldsymbol{\theta}^{(t)} \leftarrow \boldsymbol{\theta}^{(t-1)}$;
  end
end

updates might have slow convergence. For instance, latent parameters on a GMRF typically are highly correlated [Rue, 2001; Knorr-Held and Rue, 2002; Fahrmeir and Lang, 2001]. Other strategies to improve the efficiency of MCMC samples are the adaptive models, proposed improved jumping rules based on the history of the chain [Gelman et al., 2013].

In particular, Knorr-Held [1999] proposed a block algorithm to overcome the slow convergence for GMRF models and was extended to more general models by Fahrmeir and Lang [2001]. Let f be a GMRF with latent parameters λ_i, $i \in \mathcal{I}$. At each iteration, the latent parameters are partitioned in blocks of a given size m, except for one block. For a given block $\mathcal{J} \subseteq \mathcal{I}$, the proposed jumps for $\boldsymbol{\lambda}_{\mathcal{J}}$ are sampled from a normal distribution with mean μ and covariance matrix Ξ that depend on the penalty matrix $\boldsymbol{Q}$ of the GMRF and the values of the resting parameters $\boldsymbol{\lambda}_{-\mathcal{J}}$. This block strategy is extended to Gaussian processes in the analysis of genetic sequences in Section 4.4.2.

1.3.3 Point Processes

Intuitively, a point process $\mathcal{N}$ on a space S is a mechanism for allocating points randomly on the underlying space S. Given the applicability of point processes, the space S is usually the one, two or three-dimensional Euclidean space. Point processes are the fundamental structure for several spatial and spatial-temporal models [Illian et al., 2007b; Daley and Vere-Jones, 2003] as in disease mapping [Diggle et al., 2005; Brix and Diggle, 2001; Beneš et al., 2005], image analysis [Descombes, 2013] and ecology [Illian

and Burslem, 2017].

Formally, let S be a complete separable metric space equipped with a distance d and a σ-algebra $\mathcal{B}$. Then, $\mathcal{N}$ is a point process on S if $\mathcal{N}$ is a measurable mapping from a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ into the measurable space $(M, \mathcal{M})$. The set M contains all subsets $A \subseteq S$ such that $|A \cap B| < \infty$ for all bounded sets $B \subseteq S$, and $\mathcal{M}$ is the smallest σ-algebra generated by the subsets $A \subseteq S$ such that $|A \cap B| = m$ for a given bounded $B \in \mathcal{B}$ and $m \in \mathbb{Z}^*$ [Moller et al., 2003]. For a subset A of S, $N(A)$ denotes the number of events falling in A.

The Poisson point process is the most common types of point processes, with the assumption that points drawn by the process are not in interaction. Formally, a point process is Poisson if there exists an intensity measure μ such that $\mu(B) < \infty$ for all bounded subsets $B \subseteq S$ and, for every collection $A_1, ..., A_k$ of disjoint Borel sets, the following holds [Daley and Vere-Jones, 2003]:

$$P(N(A_1) = m_1, ..., N(A_k) = m_k) = \prod_{i=1}^{k} \frac{\mu(A_i)^{m_i}}{m_i!} e^{-\mu(A_i)}.$$

1.3.4 Agglomerative hierarchical clustering

In data analysis, clustering is a technique that groups points based on similarity. It covers a wide range of applications including the analysis of genetic sequences and the study of disease locations in a city. Agglomerative hierarchical clustering is one type of clustering algorithm that does not require a predefined number of clusters [Hastie et al., 2009]. The model constructs a hierarchical structure as follows. Suppose points are in a space S equipped with a distance d. Initially, every point is assigned into a single cluster. Iteratively, the two most similar clusters are merged, forming a single cluster. In the last iteration, all points are part of the same cluster. Each iteration results in one level of the hierarchy, characterised by the maximum distance between the resulting clusters, or height h. This strategy is detailed in Algorithm 3, for a finite set $S = \{p_1, ..., p_n\}$. Hierarchical clustering requires a notion of similarity between points and between clusters. For the first case, the distance d measures the similarity between points. For clusters, several measures of dissimilarity l are used. In particular, the *unweighted average linkage* between clusters A, B is:

$$l(A, B) = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a, b), \tag{1.5}$$

and the *complete linkage* between clusters A, B, given by:

$$l(A, B) = \max_{a \in A, b \in B} d(a, b). \tag{1.6}$$

The hierarchy structure obtained is summarised by the heights $h_1, ..., h_T$ and cluster sets $C_1, ..., C_T$ and is represented by *dendrograms*. This clustering approach is applied to genetic sequences in Section 4.4.1.

Algorithm 3: Agglomereative hierarchical clustering

```
Result: Heights h_1, ..., h_T, cluster sets C_1, ..., C_T
initialise h_0 = 0, C_0 = {{p_1}, ..., {p_n}}, i = 1;
while |C_{i-1}| > 1 do
    find A, B ∈ C_{i-1} with minimum distance l(A, B);
    C_i = C_{i-1} \ {A, B} ∪ {A ∪ B} (remove A, B and add A ∪ B);
    update height h_i = l(A, B);
    increase step i = i + 1
end
```

1.4 Spatial and temporal methods for outbreak detection

Spatial analysis literature refers to *cluster detection* to methods studying unusual clusters in disease patterns. In this section, the term is adopted as equivalent to outbreak detection. A range of definitions for clustering has been proposed, grouped into two types [Besag et al., 1991; Lawson and Lawson, 2006]. *General clustering* determines if the disease cases have a clustering tendency. In contrast, *specific clustering* investigates the potential location of clusters. Although this classification was introduced for spatial analysis, it can be extended to temporal and spatial-temporal methods. Some methods currently implemented by public authorities such as PHE are classified as non-Bayesian specific methods and are included in Section 1.4.1. Section 1.4.2 reviews spatial methods aimed to study general clustering. Sections 1.4.3 and 1.4.4 review methods for specific clustering from two perspectives: non-Bayesian and Bayesian modelling, respectively.

1.4.1 Temporal approaches in current surveillance systems

Public health authorities monitor the incidences of diseases to detect outbreaks or trends of potential emerging infections [Lawson, 2005; Noufaily et al., 2019]. Many statistical methods have been developed to monitor surveillance time-series. In particular, PHE

implemented methods like the Rising Activity, Multi-level Mixed Effects - Indicator Emphasis (RAMMIE) [Morbey et al., 2015], Farrington [Farrington et al., 1996] and Farrington Flexible [Noufaily et al., 2013] based on daily reports on the number of cases of notifiable diseases. Similarly, the CDC implemented the Early Aberration Reporting System (EARS) [Hutwagner et al., 2003] based on Shewhart control charts, where an alarm is activated if the counts exceed a given threshold [Shewhart, 1930].

RAMMIE, Farrington and Farrington Flexible apply time-series regression. Each approach model the count of cases y_t on a week t using a log-linear model. The parameters of the model and the expected incidence values are estimated. The observed counts are compared to confidence limits to trigger alarms, where thresholds systems are chosen according to the disease characteristics or other criteria [Lawson, 2005]. In particular, RAMMIE fits the counts y_t into a negative binomial regression where the mean μ_t is given by:

$$\log(\mu_t) = \log(N_t) + \sum_i \beta_i X_{it}.$$

The index i indicates the day of the year, the β_i are regression parameters, and each X_{it} is a 0-1 random variable equal to 1 depending on the day of the year. Comparatively, Farrington assumes that the counts y_t are distributed with mean μ_t and variance $\phi\mu_t$, where ϕ is a dispersion parameter. The parameter μ_t follows a log-linear model as:

$$\log(\mu_t) = \alpha + \beta t, \tag{1.7}$$

where α and β are regression parameters. An alarm is triggered if the observed counts exceed a threshold based on the counts from related weeks in previous years, capturing seasonal effects. Alternatively, a new version of Farrington, denoted as Farrington Flexible, includes an extra term $\delta_{j(t)}$ into the model in (1.7), where $j(t)$ is a seasonal factor level for the week t.

Other systems are based on control charts, where counts of the disease are a realisation of a stochastic process $Y = (Y_t)_{t=1}^T$. For instance, EARS is a reporting system formed of several variants. In EARS, each count follows a normal distribution as $y_t \sim N(\mu_t, \sigma_t)$. All variants are designed based on the statistic:

$$s_t = \max\{0, s_{t-1} + \left(\frac{y_t - \mu_t}{\sigma_t} - k\right)\},$$

where k is a positive real-valued parameter. Finally, temporal surveillance methods can be modified to include spatial information. For instance, Rogerson and Yamada [2004]

proposed a multiregional surveillance system that implemented CUSUM simultaneously in several regions.

1.4.2 Hypothesis testing for general clustering

To understand the patterns of a disease in a geographical area $S \subseteq \mathbb{R}^2$, many authors have modelled the incidence of cases as a *point process* $\mathcal{N}$ on S [Møller and Waagepetersen, 2007]. As a consequence, analysing the behaviour of such mathematical structures helps to comprehend and predict the dynamics of a disease. Analogous to a real-valued function, a notion of *mean* is defined [Diggle, 2013; Daley and Vere-Jones, 2003]. It is known as the *intensity* λ of the process and is defined as:

$$\lambda(x) := \lim_{|dx| \to 0} \frac{\mathbb{E}(\mathcal{N}(x))}{|dx|},$$

quantifying the expected number of observations in x. Higher-order quantities can be defined as a *second-order intensity* λ_2 or a *covariance density* [Diggle, 2013]. Similarly, a *Ripley function* K, introduced by Ripley [1976], quantifies the expected number of *events* around a randomly selected one as a function of the distance between them. That is,

$$\begin{aligned} K(r) &:= \mathcal{E}(r) \\ &= \frac{1}{\lambda}\mathbb{E}(\#\text{ events at a distance } r \text{ of another event}), \end{aligned} \tag{1.8}$$

definition valid only for *isotropic* and *stationary* processes. Since K counts events within a fixed distance, it evaluates the homogeneity of the process; that is, how much it differs from a homogeneous process. K can be analytically calculated for cases as the homogeneous point process: $K_{\text{Poisson}}(r) = \pi r^2$ [Illian et al., 2007a]. Complementing the information provided by Ripley's function, other second-order quantities have been defined [Diggle, 2013]. For instance, the function $F(r)$ quantifies the distribution of the distance between a random point in S and its nearest event, and $G(r)$ quantifies the distribution of the distance between an arbitrary event and its nearest other event. This characterisation of spatial processes can be applied to clustering detection problems, to determine if there exist agglomerated, regular or random patterns in the studied processes. A hypothesis test can be defined to investigate how the process differs from a random one.

Before making inferences about the patterns in the process, an estimator of the function $K(s)$ is calculated. According to (1.8), $K(s)$ can be estimated if λ and $\mathcal{E}(r)$

are estimated as well. An estimator $\hat{\mathcal{E}}(s)$ is calculated by averaging the number of events within a distance r of any randomly chosen event [Illian et al., 2007c; Diggle, 2013]. That is, let $x_1, \ldots, x_n$ be the events happening in G. For each x_i, the quantity $\sum_{j \neq i} I(|x_i - x_j| \leq s)$ is calculated. Then, the estimator is computed as:

$$\hat{\mathcal{E}}(s) = \sum_{i=1}^{n} \sum_{j \neq i} I(|x_i - x_j| \leq s).$$

The estimator of $\hat{\lambda}$ is calculated using the ratio of the total number of events $\mathcal{N}(G)$ and total area $|S|$.

The estimation of $\mathcal{E}(s)$ requires a boundary correction. Even though the count of events is correct for inner points, the density is potentially underestimated if analysed near the boundaries. Ripley [1976] proposed a correction on the estimation of $\hat{\mathcal{E}}(s)$. For a pair of events in x, y the quantity $I(|x - y| \leq s)$ can be weighted by w_{ij}, representing the proportion of events located at a distance $|x-y|$ of x that lies in the area S. Applying this correction the following estimator is obtained:

$$\hat{\mathcal{E}}(s) = \sum_{i=1}^{n} \sum_{j \neq i} \frac{1}{w_{ij}} I(|x_i - x_j| \leq s).$$

The estimation of the second-order function K can be applied to clustering using hypothesis testing. The test consists of comparing the observed $\hat{K}$ with its value K_0 if the process were random based on the null-hypothesis. For example, if the process is Poisson, K_0 can be computed: $K_0(s) = \pi s^2$. Additionally, the variance of K_0 can also be calculated to test the model. Although some processes have a known $\text{var}(K)$ as the homogeneous Poisson Process, the variance cannot be calculated analytically in some cases. Then the variance is estimated using Monte Carlo simulations at each location in S. Then, the observed $\hat{K}$ is compared to the simulated version.

In most applications, including epidemiological analysis, there is a population at risk underlying the process [Diggle, 2013], as in [Gatrell et al., 1996; Kelsall and Diggle, 1995b, 1998]. The previous hypothesis testing is generalised as follows. Suppose that in S there are two different processes. The first one associated with the observed cases and the second one accounting some control individuals. K is estimated for each process, obtaining $\hat{K}_{\text{cases}}$ and $\hat{K}_{\text{control}}$, respectively. For the null-hypothesis, it is assumed that both processes follow the population at risk distribution, and therefore $D := K_{\text{cases}} - K_{\text{control}}$ will be 0 [Diggle and Chetwynd, 1991]. The variance of $\hat{D}$ under the null-hypothesis can be estimated using Monte Carlo simulations. At each realisation, the

labels between cases and controls are randomly interchanged.

1.4.3 Non-Bayesian modelling for specific clustering

Spatial approaches using kernels

The incidence of a disease in a geographical area S is related to the population at risk underlying the region. Then, any model should consider both the distribution of cases in S as well as the population for each area in S, generating two different point processes. The goal is to quantify the differences between both processes and to test if they follow the same distribution; that is, measuring where the population has a higher risk. A *risk surface* is defined to quantify these differences, assuming that the distribution of cases $\mathcal{P}_1$ and the distribution of population at risk $\mathcal{P}_2$ are Poisson processes with parameters $\lambda_1(x)$ and $\lambda_2(x)$, respectively. Then the risk surface is defined as the ratio $\lambda_1(x)/\lambda_2(x)$ or similarly, the *log risk surface* as $\rho(x) = \log(\lambda_1(x)/\lambda_2(x))$. Based on previous studies Kelsall and Diggle [1995b] developed a methodology for estimating $\rho(x)$ using non-parametrical kernels. Later, they approached the problem using regressions [Kelsall and Diggle, 1998]. The goal of estimating $\rho(x)$ is to test if the risk surface is constant $\rho(x) = \rho_0$, i.e. the relative risk of having a disease is constant in the whole area.

According to Kelsall and Diggle [1995b], estimating $\rho(x)$ is equivalent to finding an estimator for $r(x) = \log(f(x)/g(x))$, where $f(x)$ and $g(x)$ are the distribution functions of cases and the population at risk, respectively. The goal is to estimate $r(x)$, test if $r(x) = 0$, and determine whether there are regions in G with higher risk. Then the estimation of the surface at risk will be expressed in terms of $\hat{f}(x)$ and $\hat{g}(x)$, the estimators of the distributions. The authors proposed to use non-parametric kernels for $\hat{f}(x)$ and $\hat{g}(x)$:

$$\hat{f}(x) = \frac{1}{n_1} \sum_{i=1}^{n} \frac{1}{h_1^2} K\left(\frac{x - x_i}{h_1}\right),$$

where $x_i, i : 1, ..., n_1$ are the locations of observed cases, h_1 is a smoothing parameter or *bandwidth*, and K is a kernel function. Likewise, $\hat{g}(x)$ is defined in terms of h_2 and the location of individuals at risk $y_i, i : 1, ..., n_2$.

Before calculating the kernel for each function $\hat{f}_{h_1}(x)$ and $\hat{g}_{h_2}(x)$, a bandwidth should be carefully selected. Kelsall and Diggle [1995b] performed cross-validation to find the optimal h_1, h_2 that minimises the difference between $\hat{r}(x)$ and $r(x)$. They showed empirically that the optimum scenario is applying the same h for both cases. However, if the region S has boundaries, the kernel will count cases outside S and overestimate $\hat{f}$ near the edges. Therefore, once $\hat{f}_h(x)$ and $\hat{g}_h(x)$ are calculated, a boundary correction

is applied. The authors proposed to divide $\hat{f}(x)$ by a quantity that approximates the proportion of points around x that lies in the region S. Several circles are drawn around x such that they uniformly cover the surroundings. Then for each circle, the proportion of points that lies inside S is calculated. Finally, once $\hat{r}(x)$ is computed, a surface evaluating the significance of the estimation is calculated; that is, a *p-value surface* is obtained through Monte Carlo simulations. For each replicate, the labels among cases and controls are randomly reassigned, and the value of $\hat{r}(x)$ is computed under the null-hypothesis.

The risk surface approach developed by Kelsall and Diggle [1995b] was based on previous studies [Bithell, 1990]. Later, Kelsall and Diggle [1995a] explored the effect of bandwidth choices and included covariates into a regression model [Kelsall and Diggle, 1998]. Kernel smoothing has been applied in point source clustering [Lawson and Williams, 1993] and spatial-temporal analysis in epidemiological applications [Han et al., 2005].

Spatial approaches using hypothesis testing

For cluster analysis on a geographical area S, a hypothesis test can be designed to check if the number of cases in a subregion of S is due to chance. The process can be repeated for several subregions and the relevant areas marked as possible clusters. To this end, a set of subregions or *windows* and a statistical test should be defined. The first of these approaches was introduced by Openshaw et al. [1987] and is known as Geographical Analysis Machine (GAM). Besag and Newell [1991] improved this approach by minimising the number of windows as well as reducing computational time. These methods and other attempts to test possible clusters were generalised by Kulldorff [1997], as described later in this section. For all these approaches, assume that the observations on S are drawn from a point process $\mathcal{N}$.

For GAM, Openshaw et al. [1987] proposed to construct a grid over S such that each intersection of lines in the grid is a potential window location. Each window is a circle of radius r, centred at one intersection of the grid. Additionally, the raduis r is modified to cover a wide range of sizes, ensuring that different window sizes are tested. For each window, a hypothesis test is performed. Openshaw et al. [1987] suggested that, under the null hypothesis H_0, $\mathcal{N}$ can be a Poisson process, but emphasized that any H_0 can be used. Finally, a Monte Carlo simulation is performed to calculate the significance of the test. The GAM approach has been criticised since it performs an exhaustive scan that could be simplified [Besag and Newell, 1991]. Also, GAM performs one test per

circle, which leads to a multiple testing problem. Despite the problems of the GAM machine, it was the initial point of several scan procedures as for [Besag and Newell, 1991; Kulldorff, 1997, 2001].

Besag and Newell [1991] modified the GAM to decrease computational cost by reducing the scan and calculating a test analytically. First, the area S is partitioned into regions. Windows are circles of radius r located at each region's centre. For a fixed circle, the radius is incremented such that the area of the circle intersects more regions. That is, at step i, the circle intersects regions that cover a total area S_i. Additionally, the statistics D_i and P_i store the observed number of cases and the underlying population in S_i, respectively. At iteration N, the studied area S is covered by the circle S_N. Finally, the method defines a hypothesis test that evaluates the statistic $T = \min\{i : D_i \geq k\}$, where k is the minimum cluster expected size. If the process is Poisson under the null hypothesis, then:

$$\mathbb{P}(T \leq t) = 1 - \sum_{s=0}^{k-1} \frac{e^{-\lambda}\lambda^s}{s!},$$

where λ is an estimation of the intensity of the process, approximated as $\lambda \approx P_t D_N / P_N$. Compared to GAM, this procedure reduces the number of circles to be drawn in the scan. However, the model requires aggregated data and needs a value of k.

A general approach known as *scan statistic* aims to generalise previous attempts to detect clusters as well as to solve computational problems. It was introduced by Naus [1965] in a one-dimensional case and adapted to geographical cluster detection by Kulldorff [1997]. Naus [1965] method is interpreted as follows. For the interval $[0, 1]$ and fixed population size of N, the method aims to find a cluster of length p. If the null-hypothesis assumes that points are uniformly random, the probability $P_{N,p}(n)$ of having a cluster with at least $n > N/2$ cases is computed analytically. Then the observed value is compared to obtain the significance of the test. Kulldorff [1997] stated that this analytical result would not be possible for higher-dimensional cases, suggesting a Monte Carlo simulation instead. Moreover, a likelihood ratio test is proposed since the size of the cluster is not fixed as in Naus [1965] model. Then, the scan is performed in three steps. First, determine a set $\mathcal{Z}$ of subsets of S to be tested. Second, calculate the likelihood ratio Λ for the observed cases. Third, calculate Λ under the null-hypothesis using a Monte Carlo simulation. A subset Z in $\mathcal{Z}$ is called *window*.

In Kulldorff [1997] method, individuals are located in a studied area S equipped with a measure μ. The quantity $\mu(A)$ counts the number of individuals in every subset A of S. Each individual can have or not an illness, and individuals having the illness

are drawn by a point process $\mathcal{N}$. For the null-hypothesis, Kulldorff [1997] studied the Bernoulli and Poisson processes as the underlying process $\mathcal{N}$. The model states that for a fixed pair $p, q \in (0, 1)$ there exists a unique window Z such that individuals in Z are ill with probability p and individuals in Z^c are ill with probability q. For the Bernoulli case that is, $\mathcal{N}(A) \sim \text{Bin}(\mu(A), p)$ if $A \subseteq Z$ and $\mathcal{N}(A) \sim \text{Bin}(\mu(A), q)$ if $A \subseteq Z^c$. The alternative hypothesis H_a states that it is more likely to be ill inside Z; that is, $p > q$. Conversely, the null-hypothesis H_0 states that $p = q$. Similarly, for a Poisson process $\mathcal{N}(A) \sim \text{Po}(\mu(A \cap Z)p + \mu(A \cap Z^c)q)$ for all $A \subseteq G$ and $H_0 : p = q$ while $H_a : p > q$.

Finally, the likelihood ratio Λ is computed. First, the likelihood of a given window Z is calculated as

$$L(Z) = \sup_{p>q} L(Z, p, q),$$

where $L(Z, p, q)$ is the likelihood of Z following H_a. Likewise, $L_0(Z)$, the likelihood under H_0, is calculated as

$$L_0(Z) = \sup_{p=q} L(Z, p, q) =: L_0,$$

a result that does not depend on Z. For both Bernoulli and Poisson cases, $L(Z)$ and L_0 can be formulated in terms of μ and $\mathcal{N}$, simplifying the computation. Then, the ratio Λ is defined as:

$$\Lambda = \sum_{Z \in \mathcal{Z}} \frac{L(Z)}{L_0}. \tag{1.9}$$

The set that maximises Λ is $\hat{Z} = \{Z \in \mathcal{Z} | L(\hat{Z}) \geq L(Z), Z \in \mathcal{Z}\}$. A Monte Carlo simulation is implemented to calculate the significance of the test. Each replica samples $\mathcal{N}(S)$ events in S, according to the population at risk $\mu(S)$. Then the ratio Λ is computed for each replica.

The scan statistic is commonly applied cluster detection in epidemiological studies. Subsequent modifications of the statistic have been proposed to cover windows with more flexible shapes. Duczmal and Assunção [2004] proposed a simulated annealing algorithm to find optimal windows. Kulldorff et al. [2006] adapted the windows to include elliptical shapes. Tango and Takahashi [2005] created FlexScan, choosing windows of various shapes but with limited size. Similarly, other algorithms have been proposed covering different shapes, as in [Patil and Taillie, 2004] and [Yao et al., 2011].

Spatial-temporal approaches

The scan statistic has been extensively used in retrospective analysis [Auchincloss et al., 2012]. However, some applications require a routine search for emerging clusters when new data is analysed, refered as prospective analysis. Kulldorff [2001] proposed an extension of the scan statistic to a spatio-temporal space; that is, a subset of $S \times \mathbb{R}$. Le T_c be the time when the analysis is performed. The following modifications were considered:

1. The scanned subsets are cylinders; that is, windows of shape $C_r \times [s, t]$, where C_r is a circle of radius r.

2. The maximum time t within a cylinder has to be the current time T_c; that is, $t = T_c$. This condition ensures that only current potential clusters are considered.

3. For computing the likelihood ratio Λ under H_0, only the cylinders $C_r \times [s, t]$ with $T_s \leq t$ will be included in the Monte Carlo simulation, where T_s is the last time when historical surveillances were analysed using the scan statistics. This condition ensures that clusters tested in the previous historical analysis are included in the significance calculation, avoiding a multiple testing problem.

The approach of Kulldorff [2001] has been extensively used in disease surveillance [Unkel et al., 2012], although further modifications have been proposed to improve the scan. Kleinman et al. [2005] incorporated geographical and temporal trends into the scan, such that seasonality patterns, for instance, are considered. Kulldorff et al. [2005] removed the requirement of using the population-at-risk data for applications where the information is scarce. Sonesson [2007] fitted the scan statistic into the CUSUM framework described in Section 1.4.1. Takahashi et al. [2008] adapted the method to include irregular shapes in the spatial component of the windows, while Costa and Kulldorff [2014] used a graph structure with the same purpose. Several applications in disease cluster detection have been implemented using scan statistics. For instance, it has been implemented in a surveillance analysis in health systems [Kulldorff et al., 2005], the study of daily syndromic surveillance [Takahashi et al., 2008], the study of influenza cases [Costa and Kulldorff, 2014], and the detection of Covid-19 clusters [Desjardins et al., 2020].

1.4.4 Bayesian modelling for specific clustering

Spatial approaches

Several methods have been developed to analyse the incidences of rare diseases in spatial and spatial-temporal frameworks. Bayesian methods have given much attention to this topic, and the computational development of MCMC procedures has let statisticians focus on more accurate formulations instead of searching analytically solvable functions. Some of these initial approaches were proposed in other fields, like ecology or image processing ([Besag et al., 1991] as an example). Initially, these methodologies focussed on *disease mapping*, inferring the risk of having a disease in a geographical area using the registered incidence count. Usually, the obtained risk surface suffers a shrinkage compared to the observed count since the data collected is incomplete, the population is small, or the disease is rare. However, for clustering detection, these smooth maps are not desirable since the risk in areas with potential outbreaks could have been diminished. Some variation to the disease mapping approaches has been developed as in [Knorr-Held and Raßer, 2000; Green and Richardson, 2002; Denison and Holmes, 2001; Gangnon and Clayton, 2000], as described later in this section.

Although there are fundamental differences, the methods explained in this section have some characteristics in common. First, the goal is to find spatial clusters of significantly elevated risk in a geographical area S, as defined by Lawson [2008]. Second, all models are BHM with a GMRF as a prior as in (1.4). For that purpose, S is partitioned into n disjoint regions A_i using post-codes areas or other appropriate divisions. The quantities y_i and E_i represent the observed and the expected number of cases, respectively, estimated using the population of A_i or other covariates. For each $i = 1, ..., n$, the latent parameter λ_i (or $\log(\lambda_i)$) captures the underlying relative (log-) risk of incidence on A_i. Therefore, estimating $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$ is the main purpose of the mapping. For the likelihood of the model, the counts y_i are independent given $\boldsymbol{\lambda}$ and $y_i|\boldsymbol{\lambda} \sim \text{Poi}(\lambda_i E_i)$. Then the likelihood will be given by:

$$P(\boldsymbol{y}|\boldsymbol{\lambda}) = \prod_{i=1}^{n} P(y_i|\lambda_i) \propto \prod_{i=1}^{n} (\lambda_i E_i)^{y_i} \exp\left(-\lambda_i E_i\right). \tag{1.10}$$

Although the likelihood is the same for all models, the latent parameters and hyperpriors vary depending on each model assumptions.

Disease mapping The Besag-York-Mollíe model, known as the BYM model [Besag et al., 1991], was proposed for image restoration and extended to spatial analysis in

epidemiology and disease analysis [Richardson et al., 2004]. The goal of the methodology is to compute the posterior distribution of the log-risk, denoted as $x_i = \log(\lambda_i)$. Then, the likelihood in (1.10) can be rewritten as $P(\boldsymbol{y}|\boldsymbol{x})$. To build a prior for $\boldsymbol{x} = (x_1, \ldots, x_n)$, Besag et al. [1991] assumed that risk values in a region i are similar to the risk of its neighbours (as in a Markov Random Field). To capture this spatial structure, $\boldsymbol{x}$ is decomposed in a *spatially structured field* $\boldsymbol{u} = (u_1, \ldots, u_n)$ and an *spatially unstructured field* $\boldsymbol{v} = (v_1, \ldots, v_n)$ such that $\boldsymbol{x} = \boldsymbol{u} + \boldsymbol{v}$. For the first case, if two regions i, j are contiguous, denoted as $i \sim j$, the model will favour cases where $u_i - u_j$ is small. That is, $P(\boldsymbol{u}) \propto \exp\left[-\sum_{i<j} w_{ij}\phi(u_i - u_j)\right]$, where $\phi(z)$ is an increasing function for $|z|$. For instance, if $\phi(z) = z^2/\kappa$ and $w_{ij} = 1$ if $i \sim j$, the distribution $P(\boldsymbol{u}|\kappa)$, known as *Gaussian intrinsic autoregression*, is:

$$P(\boldsymbol{u}|\kappa) \propto \frac{1}{\kappa^{n/2}} \exp\left\{-\frac{1}{2\kappa}\sum_{i\sim j}(u_i - u_j)^2\right\}.$$

On the contrary, the unstructured field $P(\boldsymbol{v})$ is defined as

$$p(\boldsymbol{v}|\iota) \propto \frac{1}{\iota^{n/2}} \exp\left\{-\frac{1}{\iota}\sum_{i=1}^{n} v_i^2\right\},$$

such that $\boldsymbol{v}$ does not have a spatial structure. Therefore, the posterior distribution of the model is given by

$$P(\boldsymbol{x}, \kappa, \iota|\boldsymbol{y}) \propto P(\boldsymbol{y}|\boldsymbol{x})P(\boldsymbol{u}|\kappa)P(\boldsymbol{v}|\iota)P(\kappa)P(\iota).$$

Gibbs sampling is applied to compute the $P(\boldsymbol{x}, \kappa, \iota|\boldsymbol{y})$, where the conditional probabilities are given by:

$$P(x_i|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n, \kappa) \propto \frac{1}{\kappa^{n/2}} \exp\left\{-\frac{1}{2\kappa}(u_i - u_j)^2\right\}.$$

As the BYM model, several methods have been proposed to estimate the risk surface of a disease within a Bayesian framework. Clayton and Kaldor introduced a preliminary version of the BYM model, estimating the surface based on the spatial structure of the data [Clayton and Kaldor, 1987]. In general, Bayesian hierarchical models have been the natural framework for disease mapping in geographical epidemiology [Best et al., 2005]. In this context, Green and Richardson [2002] and Knorr-Held and Raßer [2000] proposed alternative versions of the BYM model that incorporate spatial heterogeneity.

Best et al. [2005] compared the main types of models in disease mapping, including the BYM model, Green and Richardson [2002] and Knorr-Held and Raßer [2000], showing that all models are flexible in capturing features of the risk surface based on their neighbouring structure. Also, Green and Richardson [2002] and Knorr-Held and Raßer [2000] models are suitable for cluster detection, as explained in the next section.

Cluster detection Although the BYM model has received much attention in epidemiology applications, the underlying risk surface obtained has two main drawbacks for clustering detection. First, the model smoothes discontinuities in the risk surface since it averages the information of the neighbourhood, as in equation (1.4.4). Second, it underestimates high-risk values when the surface is smoothed, losing information about possible clusters. To adapt these approaches to cluster detection, several studies have introduced a partition of S such that every element of the partition, or cluster, has a constant risk. For instance, the models of Knorr-Held and Raßer [2000] and Green and Richardson [2002].

Suppose S is partitioned into n regions that are subsequently merged into k clusters. Knorr-Held and Raßer [2000] constructed a model such that the number of clusters, their location and the underlying risk are unknown. Let Z denotes the partition of clusters and λ_j the risk of the cluster j. One region in each cluster j is chosen as the *centre* of the cluster and is denoted as g_j. The prior of k is proportional to $(1-c)^k$ for a fixed constant parameter $c \in [0,1)$; that is, $p(k) \propto (1-c)^k$. If $c = 1$, k has a flat prior. Also, the prior of all partitions is flat; therefore, $P(Z|k) = \frac{1}{n!/(n-k)!}$. Finally, the risk values $\lambda_1, \ldots, \lambda_k$ are independent and follow a log-normal distribution with hyperparameters μ and σ^2. That is, $\log(\lambda_j) \sim N(\mu, \sigma^2)$ and for $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)$:

$$P(\boldsymbol{\lambda}|\kappa, \mu, \sigma) = \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi}\sigma\lambda_j} \exp\left\{ -\frac{1}{2\sigma^2}(\log(\lambda_j) - \mu)^2 \right\}.$$

Therefore, the prior of the model can be written as:

$$P(k, Z, \boldsymbol{\lambda}, \mu, \sigma) = P(k)P(Z|k)P(\boldsymbol{\lambda}|k, \mu, \sigma)P(\mu)P(\sigma^2).$$

The parameter c and the hyperpriors $p(\mu)$ and $p(\sigma)$ are chosen according to the performance of the simulations.

The core of the model relies on the reversible jump MCMC implementation since it performs a predefined search in the space of all possible parameters $(k, g_j, \lambda_j, \mu, \sigma^2)$ to find the most probable set. Note that the partition is completely determined by the

location of the centres. That is, let A_i be a region of S and let $g_1, \ldots, g_k$ be the centres of the partition. A_i will be part of the cluster whose centre is *closer* to A_i, where closer could be defined in several ways, e.g. the minimum number of borders to be crossed to reach one region from the other. The MCMC is based on the Reversible-Jump MCMC (RJMCMC) introduced by Green [1995], as described below. Suppose that the algorithm starts in the initial state s_t. First, it has to find a new possible new state s_{t+1}. Three types of jumps are defined:

- The parameters k and g_j change, since one cluster center is removed or added. Then, the corresponding λ_j is inserted (drawn from a proposal distribution) or removed.
- One of the λ_j's changes.
- The parameters μ and σ^2 change.

Then, the ratio between the posterior probability of the new state and the previous one, and the acceptance probability are calculated. After several simulations, the posterior distribution of the parameters is estimated. Finally, the output of this model is an underlying risk surface as in the BYM model, but it includes discontinuities and the risk within clusters is constant. The term *cluster* in this model does not coincide with the definition proposed for this project. However, the elements in the obtained partition should be analysed to determine if there are possible clusters (regions with significantly elevated risk).

Green and Richardson [2002] proposed a similar model to the Knorr-Held and Raßer [2000] approach, where the area S is partitioned, and the risk within each element of the partition is constant. However, Green and Richardson [2002] employed a Potts model to quantify spatial correlations favouring partitions where neighbouring regions tend to be aggregated. The likelihood is defined as in previous approaches, following (1.10), and the partition and underlying risks are denoted as Z and λ_j, $j = 1, ..., k$, respectively. Also, the number of clusters ranges from 1 to $k_{\max}$. There are two choices for the prior of k: it is constant for all values in $\{1, ..., k_{\max}\}$ or follows a truncated Poisson distribution, where $k_{\max}$ should be fixed. The prior of each λ_j is independent and follows a Gamma distribution with parameters a, b, chosen such that $a/b = \sum_i y_i / \sum_j E_i$. Finally, the prior of the partition Z is defined, using a Potts model with interaction parameter ψ. That is,

$$P(Z|\psi) \propto \exp\left(\psi \sum_{i \sim i'} \mathbb{I}(z_i = z_{i'})\right),$$

where z_i is the cluster of the region i. The term in the exponential favours partitions such that regions within a cluster are neighbours. Also, the parameter ψ quantifies the level of interaction between neighbours. If the parameter is close to 0, the interaction is weak, and the spatial correlation is low. For the hyperparameter ψ, the prior is constant for all values in $\{0, \Delta_\psi, 2\Delta_\psi, ..., \psi_{\max}\}$, for some $\psi_{\max}$ and Δ_ψ. Finally, the prior distribution of the model is given by:

$$P(k, Z, \boldsymbol{\lambda}, \psi) = P(k)P(Z|k, \psi)P(\boldsymbol{\lambda}|k, a, b)P(\psi).$$

To calculate the posterior distribution of the parameters, a RJMCMC is performed, where the jumping rules are similar to the ones defined in Knorr-Held and Raßer [2000]. Given the current state s_t in the space of all possible parameters (k, ψ, z, λ), a probability distribution of the possible next move s_{t+1} is defined. Two types of jumps are defined:

- Fixed-dimension jump: the value of ψ, the location of clusters in Z, or the risk λ_j change.
- Variable-dimension jump: the value of k changes. In that case, two clusters will be merged, or a cluster will be divided into two parts.

Spatial-temporal approaches

For spatial-temporal analysis, spatial models have been easily adapted to include a temporal dimension. However, the optimal structure depends on the type of data obtained through surveillance. In disease mapping, if the geographical location of cases is known as well as the infection onset time, then point processes provide a suitable framework to formulate disease mappings. Otherwise, statistical analysis is conducted using models based on aggregated data.

When locations are known, incidences are modelled as a realisation of a point process $\mathcal{N}$ in a spatial-temporal space $S \times T$, where the process is characterised by a spatial-temporal intensity $\lambda(s, t)$. If the intensity is not constant, the process is called inhomogeneous. Moreover, if the intensity is drawn from a stochastic process, $\mathcal{N}$ is referred to as a Cox process [Cox, 1955]. A general form for the intensity in $S \times T$ is given by:

$$\lambda(s, t) = \rho g(s, t) f_1(s) f_2(t) f_3(s, t)$$

[Lawson, 2013]. The intensity is determined by a constant rate ρ, a background intensity $g(s, t)$, and functions f_1, f_2 and f_3 describing the spatial, temporal and spatial-

temporal terms, respectively. In particular, if the intensity is written as $\lambda(s,t) = \rho g(s,t) \exp Q(s,t)$, the model is known as a Log-Gaussian Cox Process [Møller et al., 1998], where $Q(s,t)$ is a Gaussian process. Several studies have employed LGCP in spatial-temporal disease modelling and surveillance, as in [Diggle, 2005], [Beneš et al., 2005], and [Diggle et al., 2013]. For instance, Diggle [2005] incorporated a point process methodology developed by Brix and Diggle [2001] into surveillance systems. In their model, the intensity is given by $\lambda(s,t) = \lambda_0(s)\mu_0(t)R(s,t)$, where λ_0 is a smooth spatial surface, μ_0 is the temporal variation, and $R(s,t) = \exp(Q(s,t))$ is a spatial-temporal stochastic process. Diggle [2005] used this framework to define potential outbreaks in spatial-temporal cases. The authors defined an anomaly as a spatial-temporal neighbourhood if every location within the neighbourhood fulfils the following condition $R(s,t) > c$, where c is a threshold. Further applications of LGCP into disease modelling approach are explained in Diggle et al. [2013]. Although there is extensive research on the formulation of spatial-temporal models using Gaussian processes, LGCP is not always employed since the available data format is not always suitable. In many applications, the exact location of the incidences is not always provided, or the disease is rare, and therefore the data is aggregated.

BHM models are a suitable framework for aggregated data, similar to the BYM model presented earlier in this section. First, the space $S \times T$ is partitioned into N regions in S and K intervals in T. Also, the counts on a region i and an interval t are denoted by y_{ti}. In general, counts y_{ti} follow a Poisson distribution where the rate parameter depends on the relative risk λ_{ti}. In general, the model has the following structure:

$$\log \lambda_{ti} = \alpha + R_t + U_i + W_{ti},$$

where R_t is a term associated with time, U_i is related to space, and W_{ti} is an interaction term between space and time. For spatial-temporal clustering, the model is adapted such that the interactive term W_{ti} captures the clustering. For instance, Knorr-Held and Richardson [2003] proposed a model where $W_{ti} = X_{ti}\boldsymbol{z}^T\boldsymbol{\beta}$, where the parameter X_{ti} is a 0-1 random variable indicating when there is a cluster, $\boldsymbol{z}$ is a term that depends on the counts in $t-1$, and $\boldsymbol{\beta}$ is a regression parameter vector. The model also includes a temporal and a seasonal trend such that $R_t = r_t + s_t$, and a spatial component in U_i. Also, priors for each term have a GMRF structure. Spencer et al. [2011] proposed a similar approach, where $W_{ti} = X_{ti}\beta_i$, β_i is an outbreak size parameter, and X_{ti} is a 0-1 outbreak indicator. Moreover, the temporal term and spatial terms are described using GMRF. This model is described in detail in Section 3.1.

1.5 Outline of the thesis

This chapter has established the purpose of this project, described the background knowledge required for the following chapters, and provided an overview of the current methodologies in outbreak detection. The rest of the document is structured as follows. Chapter 2 provides an overview of the dataset studied in this project. The structure of the data is explained in Section 2.1, the properties of its spatial, temporal and genetic variables are described in Sections 2.2, 2.3 and 2.4, respectively. Also, an exploratory data analysis is performed in Section 2.5.

A Bayesian hierarchical model is presented and adapted to this project, using the spatial-temporal outbreak detection model in Spencer et al. [2011]. Chapter 3 describes the original spatial-temporal model and how to apply it to the project dataset. The model and its implementation are explained in Section 3.1 and Section 3.3, respectively. The results are explained in detail in Section 3.4, including a validation strategy using genetic sequences.

In Chapter 4, a spatial-genetic model is proposed, adapted from the spatial-temporal version in Spencer et al. [2011]. The model incorporates a Gaussian Random Field into the Bayesian structure to include genetic sequences in the detector, as explained in Section 4.3. Also, a sampling strategy is proposed to improve the speed of the sampling algorithm, as described in Section 4.4.

Chapter 5 describes a temporal-genetic approach using the model structure in previous chapters. Two versions of the model are proposed such that the seasonality of genetic sequences is taken into account. Finally, outbreaks found by each model are compared in Chapter 6, including a criterium to choose potential outbreaks. The output of this project, its novelty and its limitations are discussed in Section 6.3 and 6.4.

Chapter 2

Reported *Campylobacter* infections dataset

2.1 Overview of the dataset

Campylobacter species are a notifiable organism, as stated by the UK government, meaning that laboratories and medical practitioners are required by law to report suspected cases of infection. This project had access to sentinel surveillance data provided by Public Health England (PHE) and the Food Standards Agency (FSA). As part of a PHE project, the dataset included cases of *Campylobacter* infections reported in the Oxfordshire and Newcastle upon Tyne clinical laboratories between October 2015 and August 2018, inclusive. The database contained a total of 3901 reports and 4207 bacterial samples, where one or more faecal samples are collected per patient. When a patient's sample tested positive for *Campylobacter* species, epidemiological information was collected from the public health data or requested directly from the patient. It might include information about the residence location, time of symptom onset, and other demographic and behavioural questions. As part of the PHE project, samples of the bacteria were taken and the genome sequence data processed, assembled and stored in the PubMLST databases. Whole-genome sequences are publicly available in `pubmlst.org/campylobacter`, stored as the *FS-FS101013* project, where the last update of the dataset was obtained in July 2019.

The main variables related to this project are the patient residence location, the date when the sample was received by the laboratory and the whole-genome sequence of the sampled bacteria, as detailed in Table 2.1. Other temporal variables as the isolation or symptom onset time were incomplete and not considered in the analysis.

Variable	Type of data	Details	Availability
Location	Output Area[1]	Residencial location of the patient	99.2%
Received date	Date format	Submission date on the laboratory (07 Oct 2015 - 30 Aug 2018)	100%
Whole-genome sequence	Allelic profile	List of tagged loci with their corresponding allele(s)	100%

Table 2.1: Variables relevant to this project, including the type of data and the percentage of data available.

The spatial, temporal and genetic data are summarised in this chapter since they are the core variables of the models described in this project. Section 2.2 describes the spatial data, Section 2.3 displays the structure of temporal data, Section 2.4 shows details of the genetic data, and Section 2.5 performs some preliminary data analysis on the genetic sequences data.

2.2 Spatial data

The actual location where a food poisoning occurred is hard to track. For the dataset in this project, the patient residence address provides an approximation of the location of the event. Although this information is always accessible and the availability is high, it provides only an estimate. Moreover, the location of infection and residence might differ due to people's movements. For instance, some cases reported in Newcastle Upon Tyne had a residence address in Wales.

Most patients residences are located in Oxfordshire and Northamptonshire in the south-east and Tyne and Wear and Northumberland in the north-east, as shown in Figure 2.1. Each location is registered based on the *Lower-Layer Output Areas* or *LSOAs*, geography defined by the Office of National Statistics or ONS. LSOAs are constructed to be socially homogeneous and have a population between 1000 and 3000 inhabitants. If further aggregation is required, the ONS merges contiguous LSOAs into *Middle-Layer Output Areas* or *MSOAs*, covering a population between 5000 and 15000. Demographic information for LSOAs and MSOAs can be obtained through the Office for National Statistics at `www.ons.gov.uk`, such as the rural-urban classification of the territory and population estimates.

For this project, the list of reported cases was divided into two datasets. First, the *OX* dataset includes cases in Oxfordshire and six MSOA in Northamptonshire, as

[1]Output Areas are defined by the Office of National Statistics as geospatial units for England.

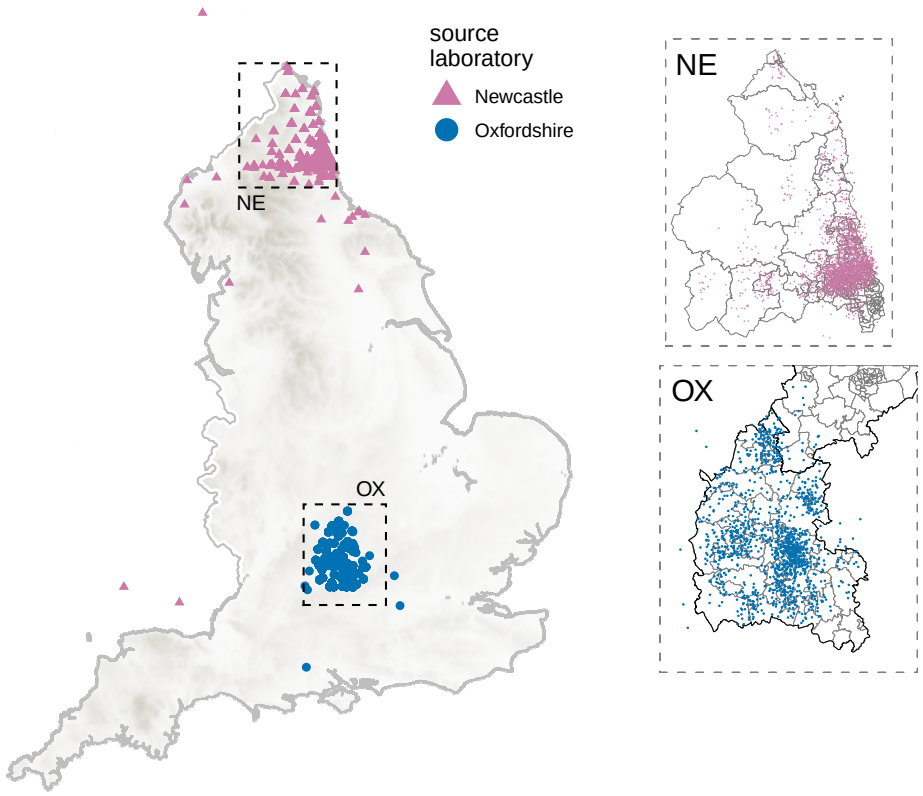


Figure 2.1: Residence location of reported patients in two regions of England, registered as LSOAs. (left) Location of cases registered in the Newcastle upon Tyne (triangles) and Oxfordshire clinical laboratories (circles). (top-right) Location of cases comprising the NE database. (bottom-right) Location of cases comprising the OX database.

shown in Figure 2.2. Second, the *NE* dataset includes cases located in two metropolitan boroughs of Tyne and Wear (Newcastle upon Tyne and North Tyneside) and Northumberland, as shown in Figure 2.3. In summary, 1440 of the 3901 patients were included in OX, 2247 were included in NE, 186 were not located in the areas mentioned, and 30 did not report location, as shown in Figure 2.4. Table 2.2 details the amount of LSOA and MSOA comprising the areas covered by OX and NE. Areas covered by both datasets have a balanced amount of rural and urban areas, as shown in Figure 2.5.

Dataset	Number of patients	Number of LSOA	Number of MSOA
OX	1440	431	92
NE	2247	503	99

Table 2.2: Number of patients, LSOAs and MSOAs comprising the OX and NE datasets.

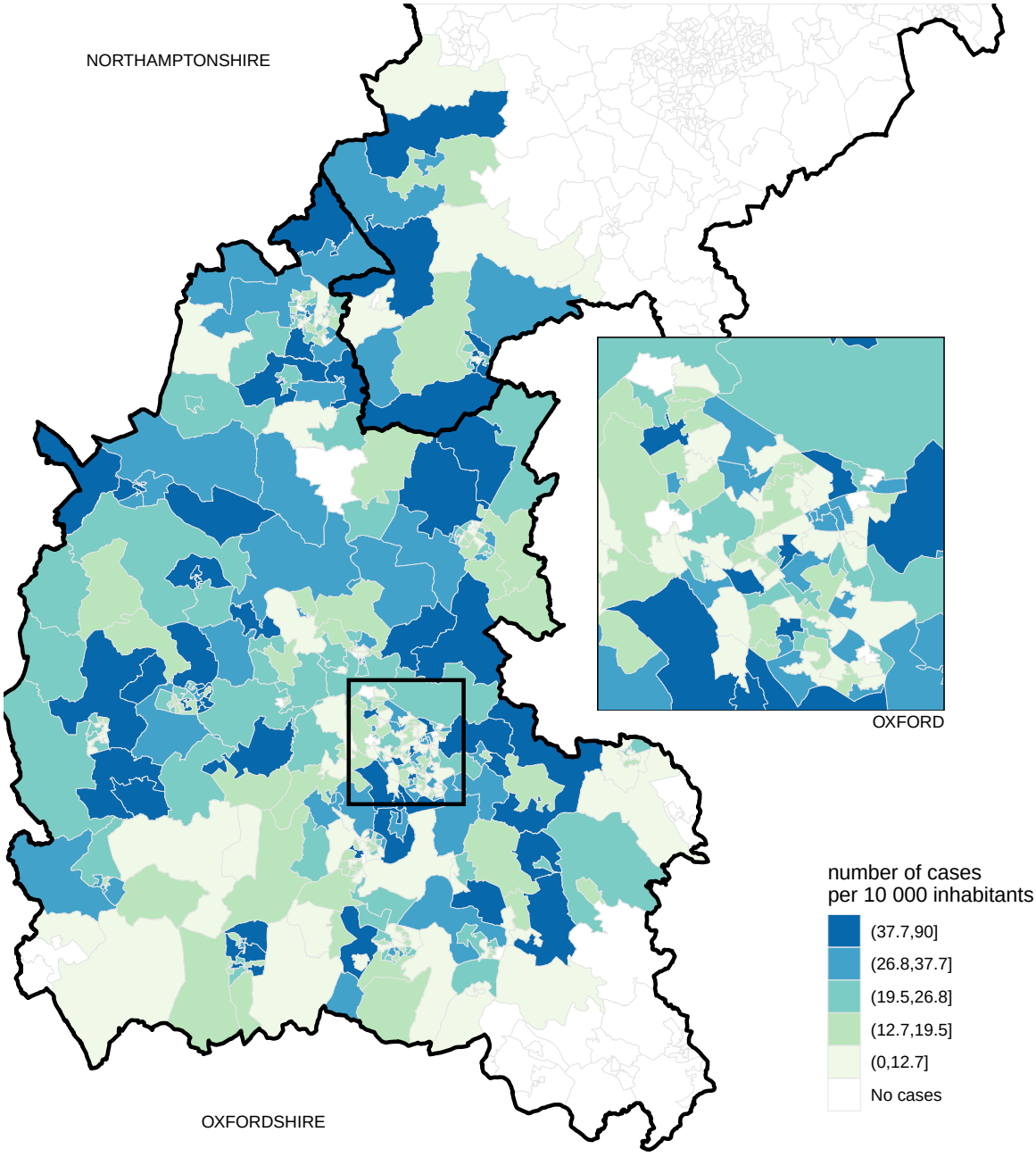


Figure 2.2: Number of cases per 10 000 inhabitants in the Lower Layer Super Output Areas (LSOA) included in the OX dataset, comprising all areas in Oxfordshire and six Middle-Layer Output Areas (MSOA) in Northamptonshire. The augmentation in Oxford is shown on the middle-right side of the figure. The intervals displayed in the colour scheme are the percentiles of the number of cases such that a similar amount of regions correspond to each colour.

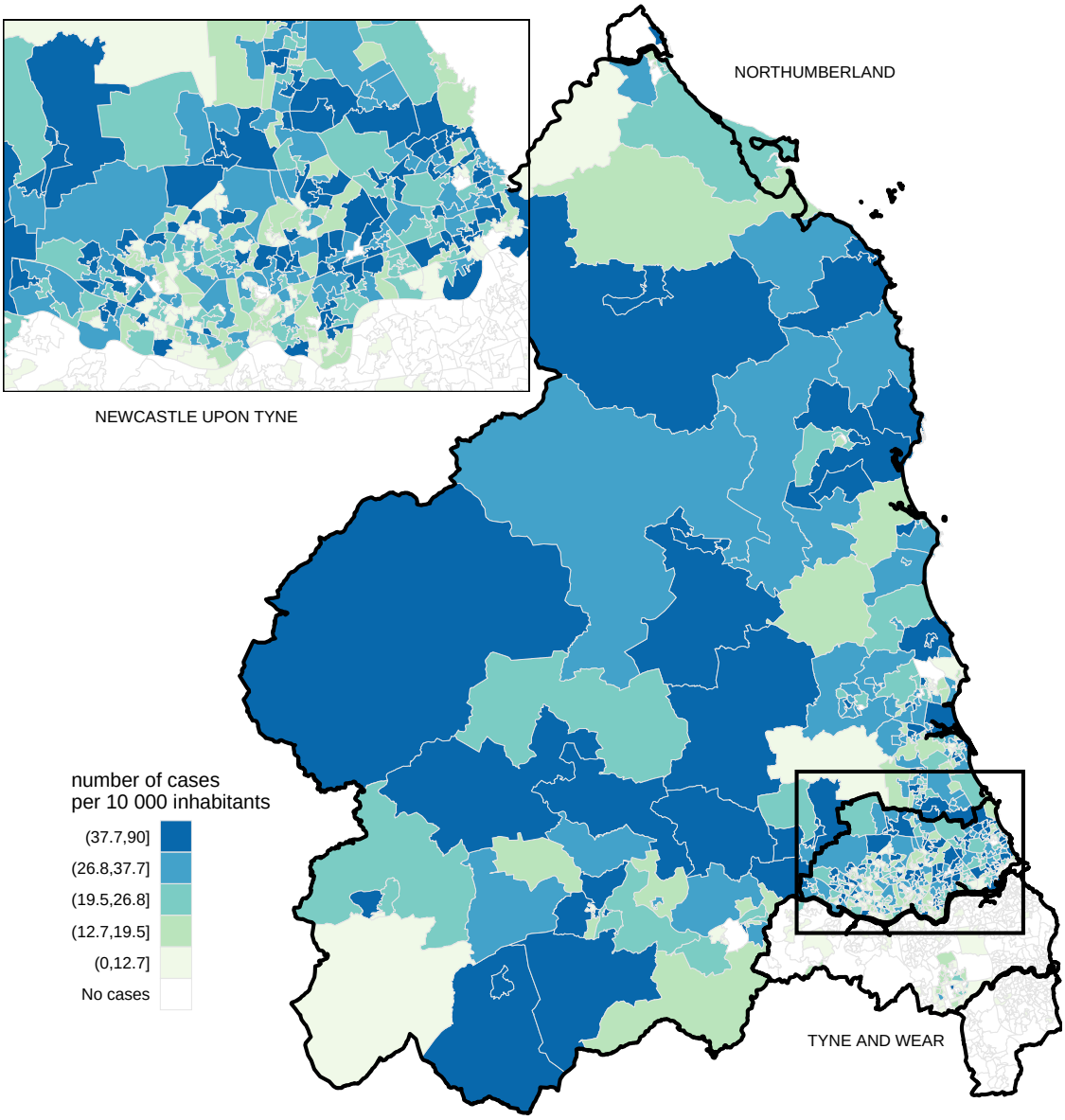


Figure 2.3: Number of cases per 10 000 inhabitants in the Lower-Layer Output Areas (LSOAs) included in the NE dataset, comprising all areas in Newcastle upon Tyne, North Tyneside and Northumberland. The augmentation in Newcastle upon Tyne is shown on the top-left side of the figure The intervals displayed in the colour scheme are the percentiles of the number of cases such that a similar amount of regions correspond to each colour.

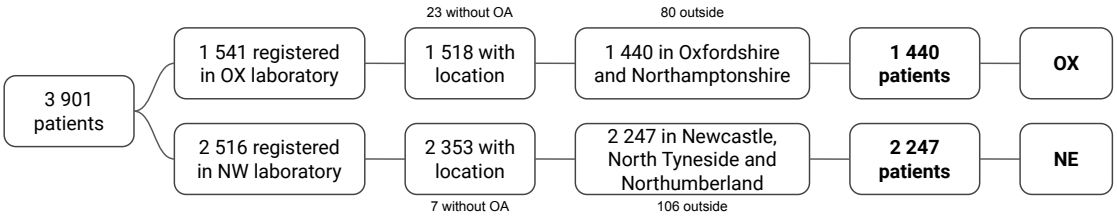


Figure 2.4: Distribution of the 3 901 cases reported in the Oxfordshire and Newcastle laboratories: 1 440 in the OX dataset, 2 247 in the NE dataset, 30 without location, and 186 located outside the chosen MSOA.

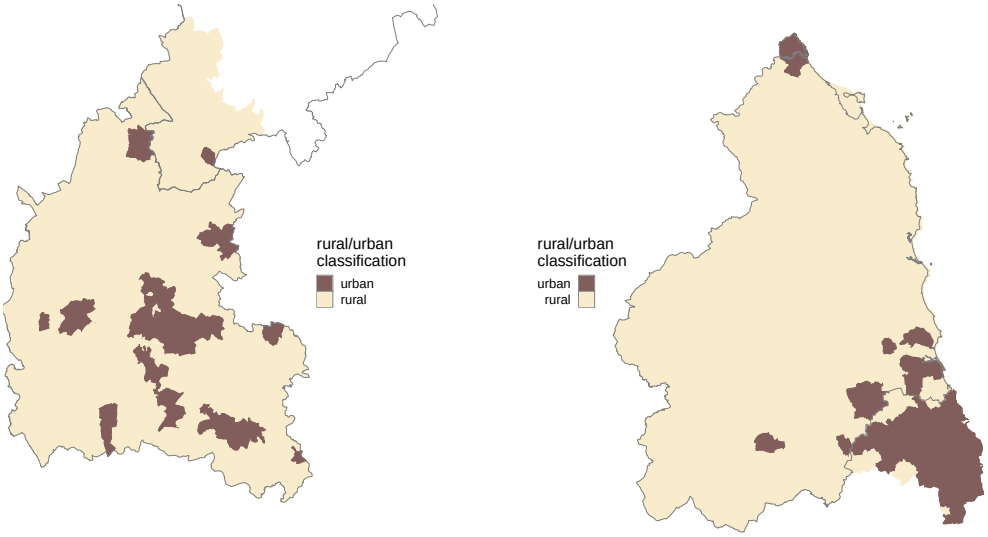


Figure 2.5: The figure shows the rural/urban classification of the Lower-Layer Output Areas of (left) Oxfordshire and selected regions in Northamptonshire, and (right) Newcastle upon Tyne, North Tyneside and Northumberland. The classification is defined by the Office of National Statistics.

2.3 Temporal data

The actual time when the exposure to infection occurred is not available, nor the time when the symptoms started, usually occurring a few days after exposure. Instead, the OX and NE datasets include the date when the isolate was reported to the laboratory, covering cases from the 7th of October 2015 to the 30th of August 2018. Reports occurred only between Monday and Friday with a peak on Tuesday and Wednesday, as shown in Figure 2.6. Therefore, dates were merged into weeks and labelled with the Friday date, covering a total of 152 weeks.

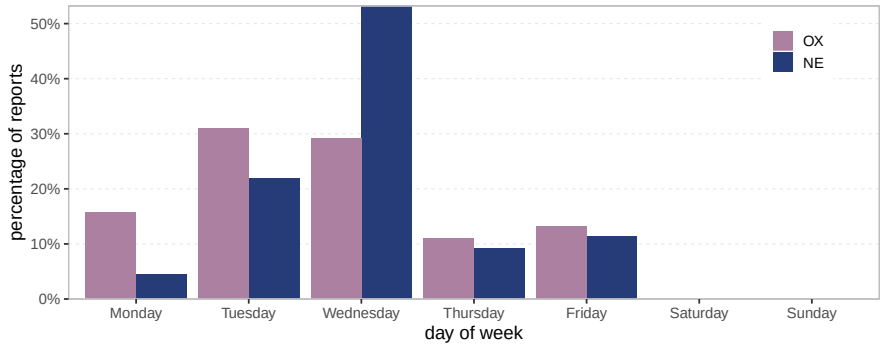


Figure 2.6: Proportion of reported isolates per weekday for the OX and NE datasets.

Cases varied from 0 to 31 counts per week. Peaks of cases occurred between spring and summer each year, for both datasets OX and NE, as shown in Figure 2.7. The background colour of each plot represents the temperature (red scheme) and rainfall (blue scheme) registered in each week and region. The figure provides a visual comparison between meteorological variables and the count of cases, as observed in previous *Campylobacter* studies [Louis et al., 2005]. Weather information was obtained from the Met Office's Weather and Climate records[2].

For both databases, observations at a given week are correlated to the number of cases at contiguous weeks, as shown in the autocorrelation plots in Figure 2.8. Weekly counts are significantly correlated with observations at four and five weeks of lag for the OX and NE datasets, respectively. The autocorrelation plot also exhibits seasonality patterns, clearly marked for the NE case. Periodogram for OX and NE are shown in Figure 2.9 to estimate the length of the seasonality cycles. The periodogram displays the estimated spectral density of the time series for a range of frequencies. Peaks for

[2] `https://www.metoffice.gov.uk/research/climate/maps-and-data/uk-and-regional-series`, visited on January 2019.

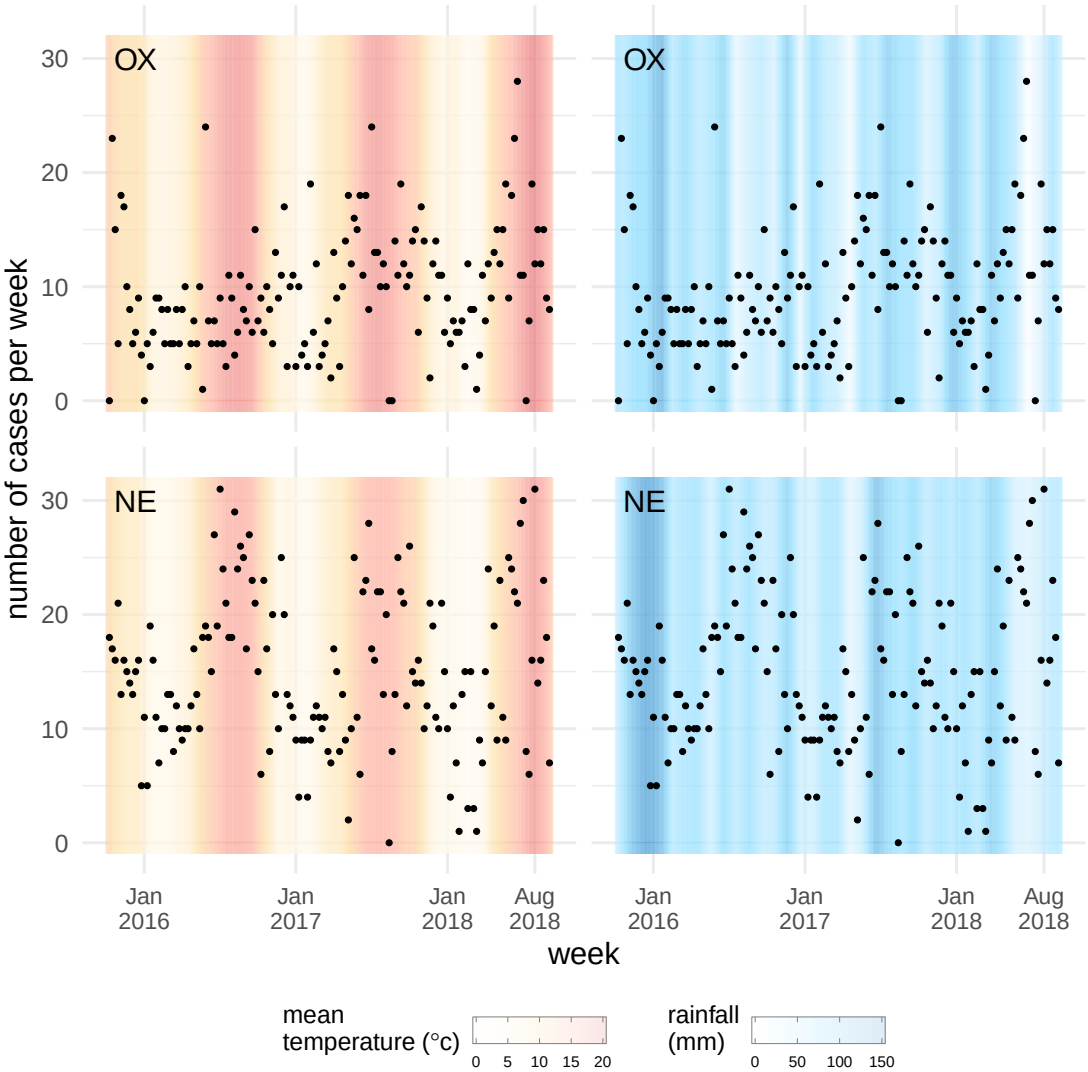


Figure 2.7: The registered date corresponds to the week of submission at the clinical laboratory for the OX database (top) and the NE database (bottom). The plot background colour (left) shows the monthly mean temperature in Celsius registered in the region, interpolated to obtain week values (left). Similarly, the rainfall in each region is represented by the background colour in the plot (right), measured in mm. Meteorological data were obtained from the Met Office's Weather and Climate records.

both datasets occurred at a frequency corresponding to a year cycle. NE showed an additional six months cycle that was not exhibited for OX.

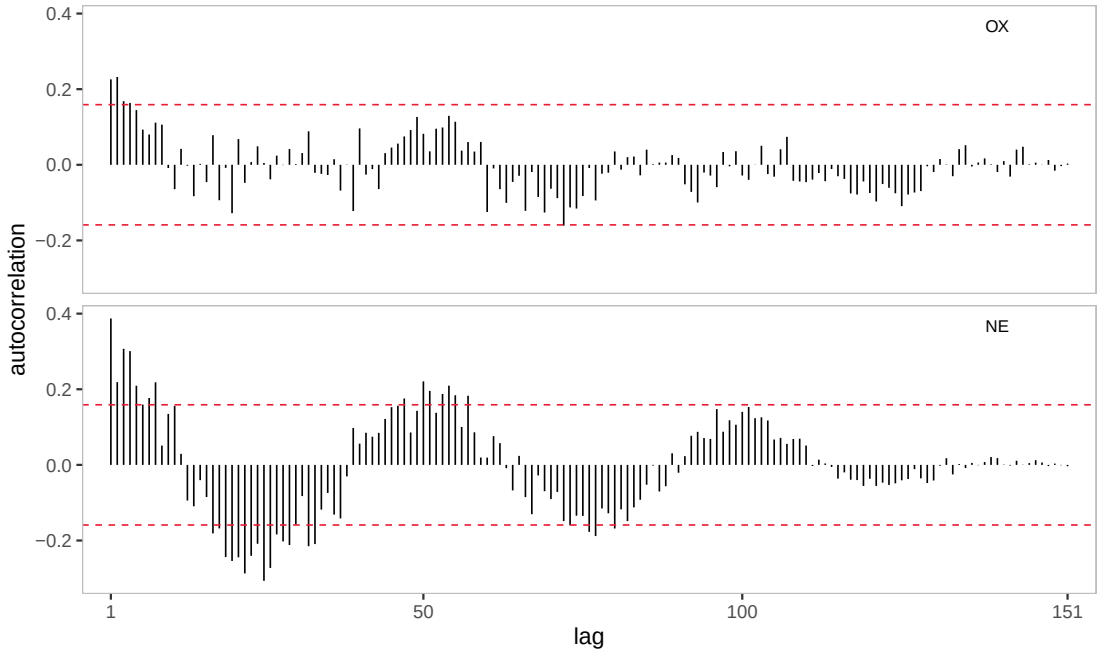


Figure 2.8: Autocorrelation function for OX (top) and NE (bottom). The horizontal axis represents the lag, and the vertical axis indicates the autocorrelation at the given lag. The horizontal dotted lines indicate the 95% bound interval under the hypothesis that the series is not autocorrelated. If the autocorrelation value at lag k is outside the confidence interval, the hypothesis that there is no autocorrelation at any lag greater than k is rejected at a significance level of 95%.

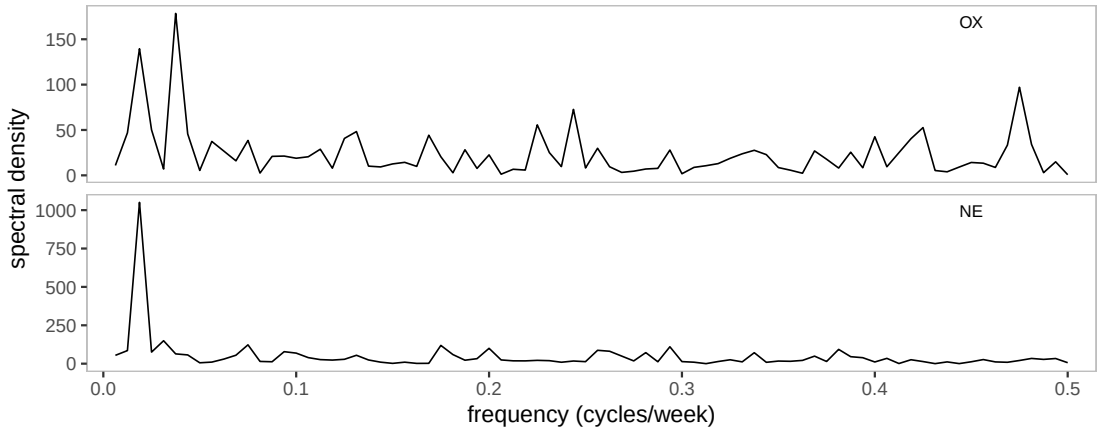


Figure 2.9: Spectral density estimation for OX (top) and NE (bottom). The horizontal axis displays the frequency at which the density is estimated. The periodogram shown corresponds to a non-smooth estimation of the spectral density.

Scheme	Length	Description
MLST	7	Multiple housekeeping gene loci [Maiden et al., 1998].
rMLST	53	53 ribosomal protein loci [Jolley et al., 2012].
cgMLST	1 287	Loci present in at least 95% of the isolates in this project.
wgMLST	1 643	Loci identified and re-annotated in Gundogdu et al. [2007].

Table 2.3: Description of the main schemes defined for the genus *Campylobacter*. The MLST, rMLST and wgMLST are obtained through the PubMLST. The cgMLST scheme is defined for this project based on the loci available for the OX and NE databases.

2.4 Genetic data

To obtain genetic sequences, one or more clinical specimens were sequenced from each patient sample. Whole-Genome Sequences data were assembled *de novo* and annotated by the PubMLST tools. The dataset contains 378 *C. coli* and 3 776 *C. jejuni* sequences with a mean length of 1 694 029 base pairs. A total of 156 patients had more than one sample linked, marked in the dataset as duplicates. The amount of data conveyed by the WGS can be summarised and organised by different typing methods, as described in Section 1.1.2. Several schemes are available through the PubMLST tools, as detailed in Section 1.2. The main schemes for *Campylobacter* species are seven-locus MLST (MLST), ribosomal MLST (rMLST), and whole-genome MLST (wgMLST), as described in Table 2.3. A core-genome MLST scheme (cgMLST) was defined for the joined OX and NE databases, based on the 1 643 loci in wgMLST. A total of 1 287 loci were selected, such that only loci present in at least 95% of isolates were included, as shown in Figure 2.10.

2.5 Exploratory data analysis

2.5.1 Study of the genetic space

Before designing statistical methods, each variable in the data should be embedded in a metric space. For instance, spatial points can be seen as a subset of $\mathbb{R}^2$ with a Euclidean distance, whereas time counts can be seen as subsets of $\mathbb{R}$ or $\mathbb{Z}^+$ with the absolute distance. However, it is unclear how to embed the WGS into a suitable metric space. For a fixed typing scheme of length L, a genetic sequence can be written as a vector $g = (g_1, ..., g_L)$ where each g_i takes values from an arbitrary set $\mathcal{G}_i$. In some cases, g_i can be reported as missing or incomplete. The set $\mathcal{G}$ of all possible sequences can be structured as a metric space with distance d, denoted as *genetic space*. Here, the distance

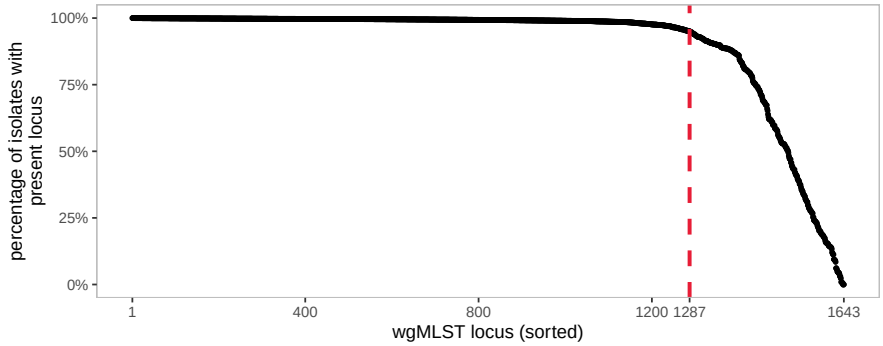


Figure 2.10: Presence of each wgMLST locus in the OX and NE databases, measured as the percentage of isolates in the database with present locus (vertical axis). The horizontal axis corresponds to each of the 1 643 wgMLST loci sorted such that the plot is decreasing. The vertical dotted line indicates the threshold chosen to define the cgMLST.

used in the PubMLST is adopted. It is defined as the number of loci that differ between every pair of isolates. That is, for $g, h \in \mathcal{G}$,

$$d(g,h) = \sum_{i=1}^{L} \mathbb{1}(g_i \neq h_i). \tag{2.1}$$

If both g_i and h_i are missing or incomplete, the term $\mathbb{1}$ is excluded from the sum. Note that this condition does not apply if only one of the alleles is missing, in which case the term is still included. For the subsequent analysis in this chapter, any sequence will be a sample of the metric space $(\mathcal{G}, d)$ with d as the distance measure.

Outbreak detection based on genetic sequences requires a deep understanding of the data and the underlying space $\mathcal{G}$. Any method aimed to detect unusually close sequences must take into account whether points in $\mathcal{G}$ look random or exhibit a clustered pattern. Moreover, if points are not random, some regions in the genetic space will have denser regions than others, and therefore the notion of closeness will differ. For instance, in the project database, the distance between an ST-21 isolate and its closest sequence is 22.6 on average while for an ST-1034 isolate is 320.9 (using the 1643 loci in the wgMLST scheme). In that case, the notion of proximity relies on the abundance of each type. Therefore, it is crucial to understand the pattern of the observed sequences in $\mathcal{G}$. To visualise the structure of the genome data, Figure 2.11 shows the pairwise distance distribution of the 4207 isolates, using the cgMLST loci. Also, Figure 2.12 displays a Minimum Spanning Tree of the sequences, coloured by the dataset.

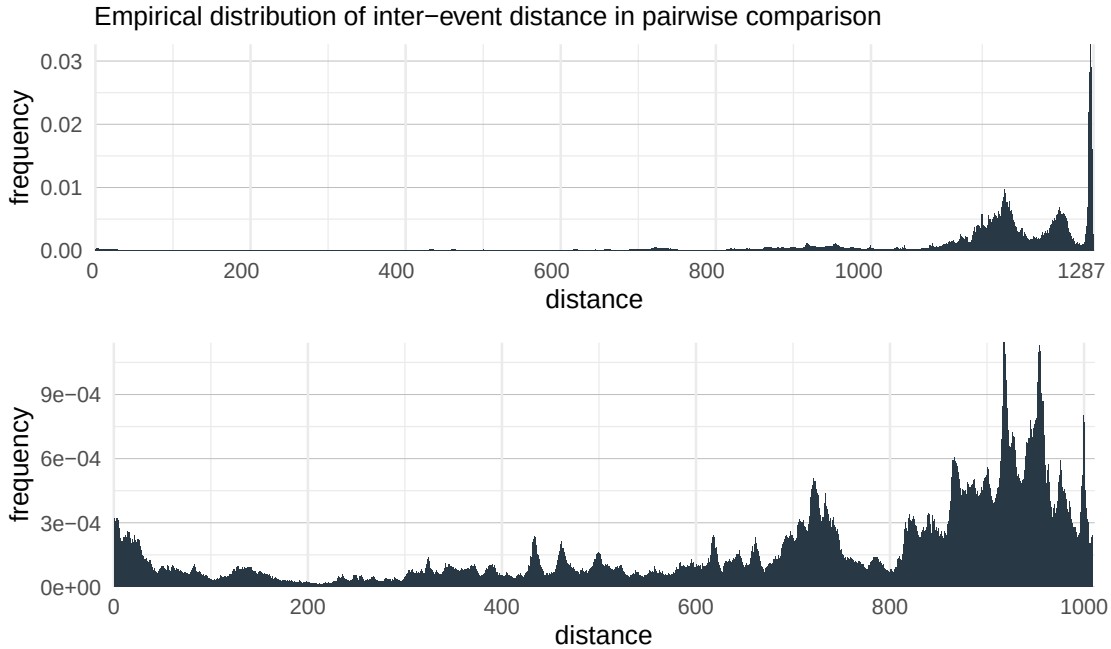


Figure 2.11: Empirical distribution of pairwise distances of all isolates using the cgMLST scheme, where distances range between 0 and 1 287. The top plot shows the histogram for the whole range. The bottom plot displays the histogram of distances below 1 000.

Since a variety of techniques have been used in the analysis of spatial-temporal point patterns, these ideas can be adapted to study the structure of the genetic space. In particular, spatial-temporal data can be modelled as a point process, and its basic features can be analysed through preliminary testing. For instance, several summary statistics have been developed to test randomness, like the distribution of interpoint distances [Diggle, 2013; Møller and Waagepetersen, 2007] (Section 1.4). Most methods in point processes assume that the underlying space has *the orderly property*, property fulfilled by Euclidean spaces, for instance, [Daley and Vere-Jones, 2008]. However, given the discrete and high-dimensional nature of $\mathcal{G}$, these assumptions are not satisfied, and the point processes theory cannot be applied. Instead, it is easy to estimate the distribution of interpoint distances, assuming that every new observed sequence is a random event. That is, for an arbitrary sequence type $g \in \mathcal{G}$, we estimate the distribution of observing s events at a distance r. First, assume that $N(g)$, the number of sequences of type $g \in \mathcal{G}$, follows a Poisson distribution with parameter λ and is independent of $N(h)$ for any other sequence $h \in \mathcal{G}$. Moreover, let $\mathcal{C}_r$ be the number of sequence types around g at a distance r. Then, the probability of having a total of s events at a distance r of

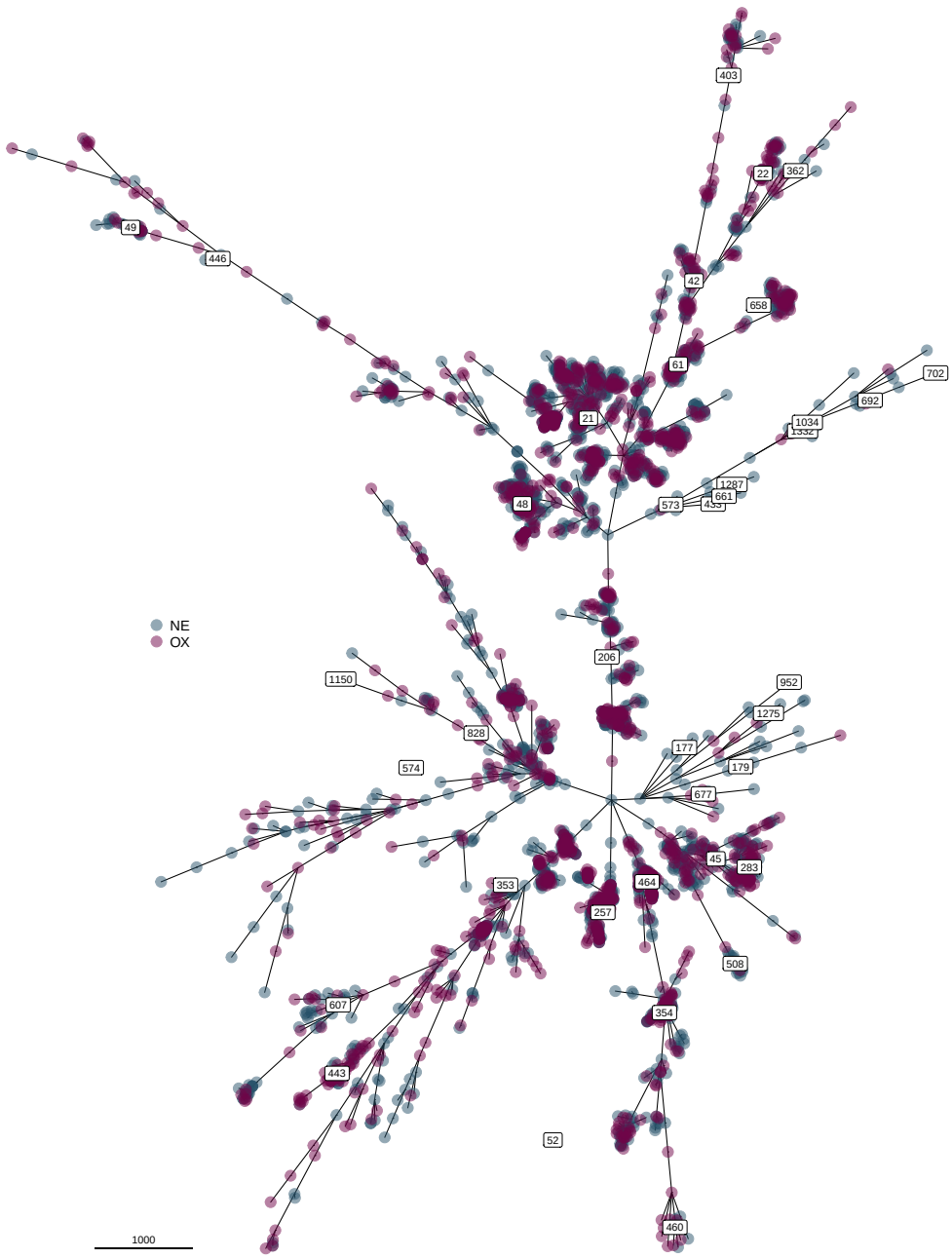


Figure 2.12: Minimum Spanning tree of the genetic sequences in the OX (purple) and the NE dataset (blue), indicated by colour. Labels indicate the clonal complex and are localised in the centre of the clonal complex node locations.

g follows a Poisson distribution with parameter λ times $\mathcal{C}_r$, following that the sum of Poisson-distributed random variables is Poisson. That is,

$$s \text{ events at distance } r \sim \text{Poi}(\mathcal{C}_r \lambda).$$

If we assume that each locus can take values from a set of length X, then $\mathcal{C}_r = \binom{L}{r}(X-1)^r$, where L is the length of the scheme. Analogous to Ripley's function for point processes, a summary function is defined as $K(r) = \lambda^{-1}\mathbb{E}[\text{events at a distance } r]$. Note that $K(r) = \mathcal{C}_r$ under the homogeneous scenario (λ constant). Empirically it can be estimated as

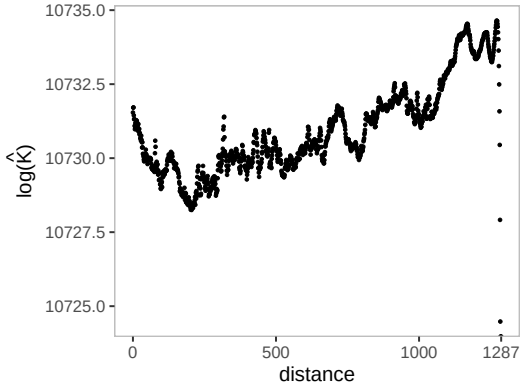
$$\hat{K}(r) = \hat{\lambda}^{-1} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \mathbb{1}[d(g_i, g_j) = r],$$

where n is the total number of observations, and $\mathbb{1}$ is the indicator function. The parameter can be estimated as $\hat{\lambda} = n / \sum_r \mathcal{C}_r$.

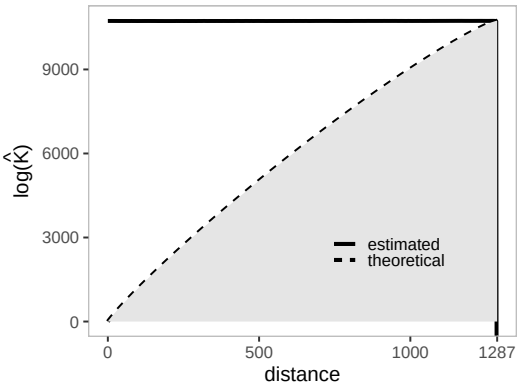
For the project database, $\hat{K}(r)$ is calculated assuming that $X = n$ since n is the maximum number of alleles that can be observed per locus. The estimated $\log(\hat{K}(r))$ is shown in Figure 2.13a, and its compared to the theoretical value $\log(K(r))$ plus or minus one standard deviation in Figure 2.13b. The extreme behaviour exhibited by $\hat{K}(r)$ is a consequence of the high-dimensionality of the space: all neighbours tend to be as far as possible. The non-randomness of the data observed is a consequence of the nature of the genetic evolution since it does not behave randomly.

2.5.2 Analysis of in-patient samples

It is often difficult to construct and validate models for outbreak detection since not all real outbreaks are detected. PubMLST sometimes provides more than one isolate for one patient, providing a set of isolates with a common source of infection. The dataset analysed in this project had 142 patients with repeated samples, a total of 292 isolates marked as duplicates and 158 pairs of isolates coming from the same patient. Figure 2.14 shows the distance distribution of pairs of isolates from the same patient, using the core-genome scheme cgMLST introduced before. Since most of the pairs (79%) had a distance less or equal than 25, Figure 2.15 shows the distribution for pairs with small distances. Also, distant isolates might represent patients simultaneously infected with multiple strains. Understanding the cause of the differences between the remaining pairs of isolates is useful for understanding the features of a real outbreak. For that reason, the loci responsible for each distance greater than 0 are analysed further. In total, 47 loci were identified, and 3 of them were distinct in at least four pairs of isolates from at least three patients, as reviewed in Table 2.4.



(a) Estimation of the statistics $\hat{K}(r)$ for the collected isolates, for distances between 0 and 1287. The values are displayed in the logarithmic scale.



(b) Theoretical value $K(r) = \mathcal{C}_r$ for distances between 0 and 1287, plus and minus one standard deviation coloured in grey. The values are displayed in the logarithmic scale.

Figure 2.13: Comparison between the empirical and theoretical values of the function $K(r)$. For small distances, $\hat{K}(r) \gg K(r)$, proving the non-randomness of the collected data.

Locus	Pairs	Name of locus	Function	Entropy
CAMP1178	7	Major outer membrane protein	Macromolecule metabolism	4.75
CAMP1159	4	Putative periplasmic protein	Cell envelope Unknown	2.85
CAMP0751	4	Hypothetical protein Cj0816	Unknown	3.65

Table 2.4: Description of the loci responsible for in-patient variations of at least three different patients, including the name of the locus, entropy and the functionality, if known.

To examine the variability of a locus, the entropy is introduced as a measure of the average information contained in a locus, and it is calculated using the entire dataset. Let l be a locus in the cgMLST scheme, I_l the number of alleles of l, o_i the number of isolates with allele i, and n_l the number of sequences in the database where locus l is present. Then, the entropy of l is defined as:

$$\hat{H}(l) = -\sum_{a=1}^{A_l} \frac{o_a}{n_l} \log \frac{o_a}{n_l}.$$

If a locus has many types of infrequent alleles the entropy will be high. On the other hand, if a locus has a stable allele found in almost all isolates, the entropy will be low.

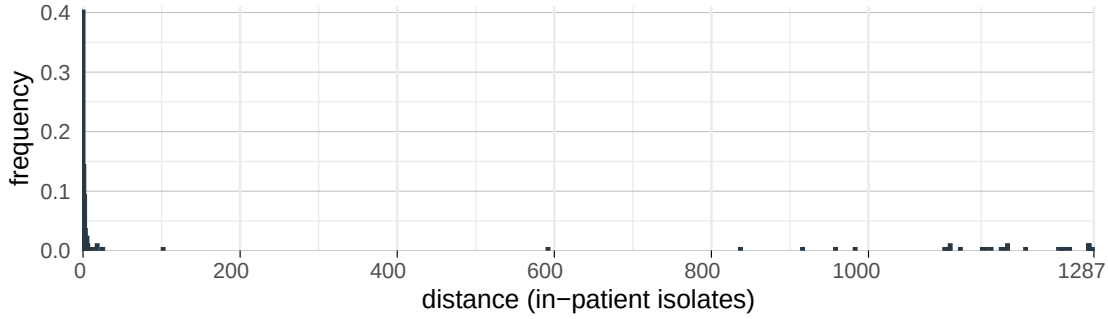


Figure 2.14: Distribution of pairwise distances between isolates taken from the same patient.

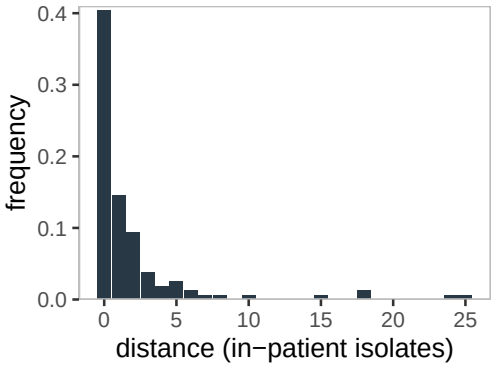


Figure 2.15: Distribution of pairwise distances between isolates taken from the same patient. Pairs with a distance larger than 100 were omitted.

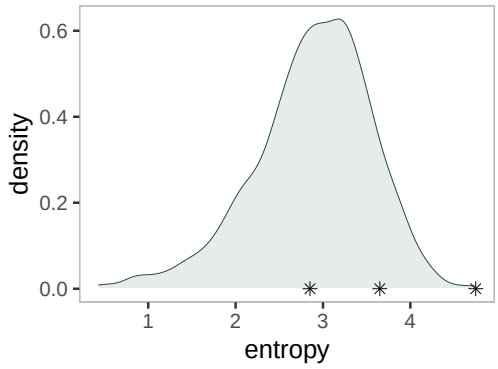


Figure 2.16: Estimated distribution of the entropy of cgMLST loci. Stars indicate the entropy value of the top loci shown in Table 2.4.

Figure 2.16 illustrates the estimated distribution of entropy for the 1287 cgMLST loci. The horizontal line at the bottom shows the entropy of the loci in Table 2.4, responsible for in-patient variations. The cgMLST locus with the highest entropy is CAMP1178, causing variations within seven pairs of in-patient isolates. Entropy potentially is a measure to explain why some loci are responsible for in-patient variation.

2.6 Summary

The purpose of this chapter was to overview the dataset available for this project, review the main variables used throughout the thesis, and perform data analysis to explore some properties of the genetic sequences. Section 2.1 reviewed the content of the data.

The spatial variable in the database was described in Section 2.2. Similarly, Section 2.3 reviewed the temporal variable, and Section 2.4 examined the characteristics of the genetic sequences. Finally, Section 2.5 contained the exploratory data analysis.

Chapter 3

Spatiotemporal model for outbreak detection

The dataset described in Chapter 2 contains three main attributes: a spatial, a temporal and a genetic component. An algorithm to detect outbreaks could combine the information provided by these attributes. For instance, many models suggested in the literature analyse spatial or temporal data, and few of them combine both dimensions, as described in Section 1.1. In this chapter, the spatial-temporal model presented by Spencer et al. [2011] is described and applied to the dataset, where a Bayesian hierarchical model identifies relative peaks of *Campylobacter* incidences and labels them as potential outbreaks. Several versions of the model are studied to capture different types of outbreaks. Also, a mechanism to validate the model is proposed using the genetic component of the data. Finally, the output of the model is compared to the spatial-temporal scan statistics described in Section 1.4.3. The content of this chapter is divided as follows. The structure of the model is described in Section 3.1. In Section 3.2, the whole-genome sequences are used as a measure to validate the model. The model implementation is described in Section 3.3 and the results obtained are analysed in Section 3.4. Finally, the discussion about the model and the results is included in Section 3.5.

3.1 Model for outbreak detection

The spatial-temporal outbreak detection model proposed by Spencer et al. [2011] studies the disease in a particular region and a fixed period of time. It assumes that the count of cases is characterised by a temporal pattern constant in space, a spatial pattern constant in time, and a spatial-temporal term that captures potential outbreaks. To define the

model, the region is divided into I non-overlapping areas labelled as $i = 1, ..., I$ and the time period into T intervals $t = 1, ..., T$. Further, the data is aggregated into the spatial-temporal blocks it, and the counts are stored in y_{it}. Therefore, the incidences can be described by a spatial effect U_i, a temporal effect R_t and a non-negative term W_{it} capturing localised periods of increased risk, or outbreaks. Since outbreaks may have various sizes, the studied areas are combined into larger areas or *outbreak areas* $\sigma(i)$, where σ is a function defined on the set of indices $i = 1, .., I$. Similarly, the intervals t are combined into larger intervals or *outbreak intervals* $\phi(t)$, where ϕ is defined on the set of interval indices $k = 1, ..., K$. This partition gives the model the flexibility to capture potential outbreaks in *blocks* of different sizes, identified with the index $\sigma\phi$. Also, it constrains the model to be identifiable. The term capturing outbreaks is rewritten as $W_{\sigma(i)\phi(t)}$ or equivalently as $W_{\sigma\phi}$.

According to the model, the observed data y_{it} follows a Poisson distribution with mean $n_i\mu_{it}$. The parameter μ_{it} describes the risk of an individual becoming infected in the block it, and the offset n_i is the population in area i. The effects U_i, R_t and $W_{\sigma(i)\phi(t)}$ are included as latent variables as follows,

$$\begin{aligned} y_{it}|\mu_{it}, U_i, R_t, W_{\sigma(i)\phi(t)} &\sim \text{Poisson}(n_i\mu_{it}), \\ \log \mu_{it} &= \alpha + U_i + R_t + W_{\sigma(i)\phi(t)}. \end{aligned} \tag{3.1}$$

Additionally, the intercept α captures the average incidence rate per person per time unit in the studied region.

The parameters U_i quantify the risk at area i, caused by characteristics of the spatial location. The model assumes the value of u_i should be similar to its neighbours, such that the spatial risk is smooth across the studied region. In this case, two areas are neighbouring if they share a border. Therefore, the prior distribution for the terms U_i are formulated as a Gaussian Markov Random Field (GMRF). In particular, each U_i is centred on the mean of its set of neighbours N_i,

$$U_i|\tau_U, U_{-i} \sim \mathcal{N}\left(\frac{1}{|N_i|}\sum_{i'\in N_i} U_{i'}, \frac{\tau_U^{-1}}{|N_i|}\right),$$

where U_{-i} indicates all elements in the vector $U = (U_1, ..., U_I)$ except U_i.

Similarly, the model smoothes the temporal terms $R_1, ..., R_T$ such that the increase in the value from R_t to R_{t+1}, $R_{t+1} - R_t$, behaves similarly to the previous increment $R_t - R_{t-1}$. That is, the prior distribution for the terms R_t is that they follow a

second-order random walk such that:

$$(R_{t+1} - R_t)|\tau_R, R_{1:t} \sim \mathcal{N}(R_t - R_{t-1}, \tau_R^{-1}),$$

for t greater than 1, where $R_{1:t} = (R_1, ..., R_t)$. The priors for the precision hyperparameters τ_U and τ_R follow Gamma distributions such that $\tau_U \sim \text{Gamma}(a_U, b_U)$ and $\tau_R \sim \text{Gamma}(a_R, b_R)$, controlling how smoothly the spatial and temporal risk should change. The parameters R_1 and R_2 follow improper flat priors. To enable the identifiability of the intercept parameter α, an additional sum-to-zero constraint is imposed such that $\sum_i U_i = 0$ and $\sum_t R_t = 0$.

Finally, the outbreak term $W_{\sigma\phi}$ captures any risk not described by the spatial or temporal terms. The model assumes that if there is a potential outbreak in the block $\sigma\phi$, the size of the outbreak will depend only on its location σ. If no outbreak is detected, the term takes the value of 0. Formally, $W_{\sigma\phi}$ is defined as $W_{\sigma\phi} = B_\sigma X_{\sigma\phi}$, where B_σ is the typical size of an outbreak in the area σ and $X_{\sigma\phi}$ is an *outbreak indicator* that can be 0 or 1. The parameter B_σ follows a Gamma distribution such that,

$$B_\sigma|a_B, b_B \sim \text{Gamma}(a_B, b_B).$$

The outbreak indicators $X_{\sigma\phi}$ are 0-1 random variables. If $X_{\sigma\phi}$ is 1, the model captures the existence of an outbreak of size B_σ in the block $\sigma\phi$. If the model does not detect an increased number of cases in the block $\sigma\phi$, the indicator $X_{\sigma\phi}$ takes the value of 0. Furthermore, the posterior distribution of $X_{\sigma\phi}$ provides the probability of observing an outbreak. Real outbreaks could last more than the length of an interval σ or cover larger regions than the size of the area ϕ. To allow the model to approximate realistic outbreaks, the $X_{\sigma\phi}$ could be temporarily independent or dependent on each other. Spencer et al. [2011] proposed two approaches. In the first case, referred as the *independent model* or IM, the $X_{\sigma\phi}$ are independent and identically distributed as follows:

$$X_{\sigma\phi}|p \overset{\text{iid}}{\sim} \text{Bernoulli}(p),$$

where the prior of the hyperparameter p follows a Beta distribution $p \sim \text{Beta}(a_p, b_p)$. Since $\mathbb{E}(X_{\sigma\phi}) = p$ for all σ and ϕ, the hyperparameter p captures the expected number of outbreaks that occurred in each interval and outbreak area $\sigma\phi$. In the second case, referred as the *correlated model* or CM, the $X_{\sigma\phi}$ are correlated in time. That is, the probability of having an outbreak at $\sigma\phi(t)$ depends on the presence of an outbreak at $\sigma\phi(t-1)$. Formally, for a fixed outbreak area σ, the chain $X_{\sigma\phi(1)}, \ldots, X_{\sigma\phi(T)}$ form

a Markov Chain with transition probabilities $\mathbb{P}(X_{\sigma\phi(t)} = 1|X_{\sigma\phi(t-1)} = 0) = p_{01}$ and $\mathbb{P}(X_{\sigma\phi(t)} = 1|X_{\sigma\phi(t-1)} = 1) = p_{11}$, leading to the following prior:

$$X_{\sigma\phi(t)}|p_{01}, p_{11}, X_{\sigma\phi(t-1)} \sim \text{Bernoulli}(\mathbb{1}_{\{X_{\sigma\phi(t-1)}=0\}}p_{01} + \mathbb{1}_{\{X_{\sigma\phi(t-1)}=1\}}p_{11}).$$

Additionally, p_{01}, p_{11} have beta prior distributions with parameters (a_{p01}, b_{p01}) and (a_{p11}, b_{p11}), respectively. Also, the prior distribution of the first position in the chain $X_{\sigma\phi(1)}$ is given by the stationary distribution of the Markov chain. That distribution can be calculated analytically given the transition probabilities p_{00} and p_{01}. That is,

$$X_{\sigma\phi(1)} \sim \text{Bernoulli}\left(\frac{p_{01}}{p_{01} + 1 - p_{11}}\right)$$

This approach using correlated indicator terms allows the model to capture outbreaks with variable duration.

The hierarchical conditional independence structure of the model is displayed in a Directed Acyclic Diagram in Figure 3.1. Circle nodes represent the parameters of the model and square nodes denote constants. Also, the red squares show the two possible priors for the $X_{\sigma\phi}$ terms.

3.2 Validation strategy

The spatial-temporal model described in Section 3.1 provides the probability that a block is a potential outbreak. However, validating the accuracy of the model output is not possible since real outbreaks are unknown. Two validation processes are proposed. First, the validation is performed using genetic sequences as described below. Second, the model is compared to an existing spatial-temporal outbreak detection model, such as the scan statistics.

Genetic sequences validation

An approximate validation can be performed based on genetic sequences, since genetic closeness has been observed between in-patient isolates (Section 2.5), and epidemiologically related isolates [McCarthy, 2017; Cody et al., 2013]. A block $\sigma\phi$ is labelled as a *genetically-linked block* if there exists at least one pair of patients such that the genetic distance between their bacterial samples is less or equal than 20, using the genetic distance in equation (2.1) (Section 2.5). Finally, the potential outbreaks detected by the model are defined. Let $\rho_{\sigma\phi}$ the probability of having an outbreak at block $\sigma\phi$, defined

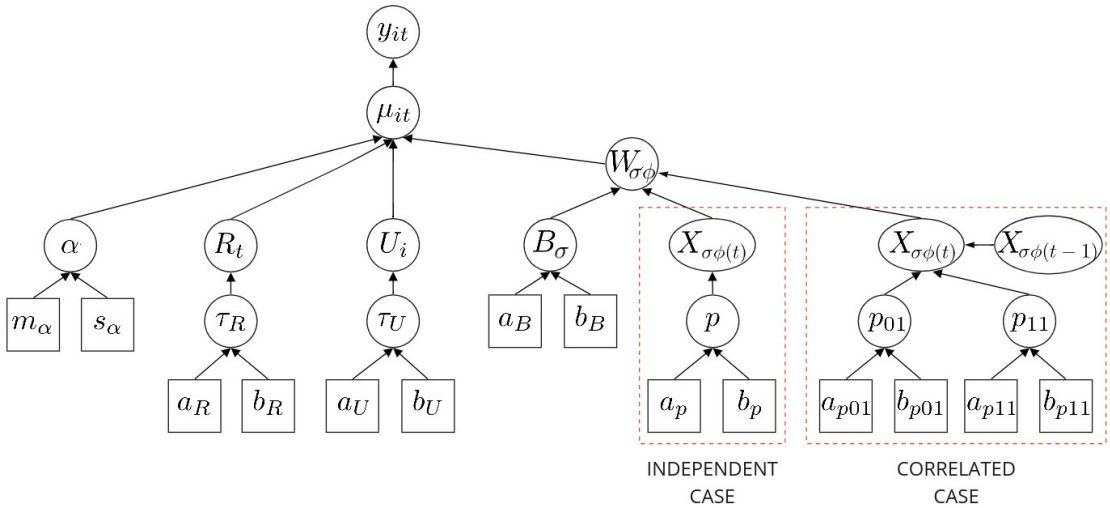


Figure 3.1: Directed Acyclic Graph describing the hierarchical conditional independence structure of the model, the parameters α, R_t, U_i, B_σ, $X_{\sigma\phi}$ and the hyperparameters τ_R, τ_U, p, p_{01}, p_{11}. The red squares are the two possible priors for the $X_{\sigma\phi}$ terms.

	Parameter description	Prior distribution	Sampling algorithm
α	Intercept	Normal	M-H (normal proposal)
R_t	Risk on the interval t	Second order random walk	M-H (normal proposal) M-H (block updates)
τ_R	Temporal term precision	Gamma	Gibbs
U_i	Risk on the spatial region i	GMRF	M-H single updates (normal, conditional prior proposal)
τ_U	Spatial term precision	Gamma	Gibbs
B_σ	Typical size of outbreak on the block σ	Gamma	M-H single update (truncated normal proposal)
$X_{\sigma\phi}$	Outbreak indicator on the block $\sigma\phi$	Bernoulli	Gibbs
p	Probability that a block $\sigma\phi$ is an outbreak (IM)	Beta	Gibbs
p_{01}	Probability that $\sigma\phi(t)$ is an outbreak if $\sigma\phi(t-1)$ is not an outbreak (CM)	Beta	M-H
p_{11}	Probability that $\sigma\phi(t)$ is an outbreak if $\sigma\phi(t-1)$ is an outbreak (CM)	Beta	M-H

Table 3.1: Description of the parameters of the model, including the prior distribution and the sampling algorithm used in the implementation.

as the expected value of $X_{\sigma\phi}$. Then, a block is labelled as a *potential outbreak* if $\rho_{\sigma\phi}$ is greater than the threshold θ.

To assess the accuracy of the model, genetically-linked blocks are compared to the potential outbreaks detected by the model. A Receiver Operating Characteristic (ROC) curve provides a diagnostic of the performance of a classifier for various thresholds. The ROC curve compares the rate of true positives (TPR) in the horizontal axis against the rate of false positives (FPR) in the vertical axis. The area under the ROC curve (AUC) evaluates the accuracy of the classifier. In particular, a perfect classifier has a TPR of 1, an FPR of 0 for every threshold, and an AUC of 1. For the spatial-temporal model, the detected outbreaks are the output of the classifier while the genetically-linked blocks are an approximation of the real outbreaks. Note that the terms *true positives*, *false positives* and *perfect classifier* arise from a comparison between models rather than a validation; therefore, they can be misleading.

Scan statistic comparison

Comparing the output to an existing model does not provide validation, but it highlights the properties of outbreaks detected by both models. Therefore, the model is compared to the retrospective version of the spatial-temporal scan statistics described in Section 1.4.3, using the SatScan^{TM} software[1]. The retrospective version of the spatial-temporal statistic is chosen since the geographical region, the study period is fixed, and the model search for historical clusters.

3.3 Implementation

The implementation of the model is available in the `R` package `epiclustR`[2] for the independent model. The package estimates the posterior distribution of the parameters using a Markov Chain Monte Carlo algorithm. At each iteration, new values for each parameter are proposed. For the spatial component U_i, the new values are suggested alternately by single Gaussian Random Walk proposals and conditional prior proposals [Knorr-Held, 1999]. For the temporal component R_t, block updates are proposed to control the high correlation between the parameters [Knorr-Held, 1999]. Additionally, the new values of B_σ are suggested using single site Gaussian Random Walk proposals.

[1] SaTScan^{TM} is a trademark of Martin Kulldorff. The SaTScan^{TM} software was developed under the joint auspices of (i) Martin Kulldorff, (ii) the National Cancer Institute, and (iii) Farzad Mostashari at the New York City Department of Health and Mental Hygiene.

[2] Available at https://github.com/jmarshallnz/epiclustR. Visited in July 2019.

Finally, for the independent model, the outbreak indicators $X_{\sigma\phi}$ are updated using the Gibbs sampler. Since the package only includes the independent model, it was adapted to include the correlated model. The indicators $X_{\sigma\phi}$ are updated using a Forward Filtering Backward Sampling algorithm (FFBS) [Shephard, 1994].

The package uses parallel computing so that several chains are run at the same time, saving computational time. This method allows the user to assess if different starting locations for the Markov chain result in similar posterior distributions for the parameters.

3.3.1 Model specifications

Before applying the model, the areas i are defined for the covered regions in the OX and NE datasets. Both regions are aggregated using the LSOA since they are the most granular segregation available. To define the intervals t, the reported dates are segregated by week (Monday to Sunday), as described in Section 2.3. Additionally, the model is run separately for the OX and NE datasets since both regions presents different temporal pattern, as shown in Section 2.3.

Outbreaks areas and intervals are also defined. Time intervals are aggregated to form outbreaks intervals. Three possible levels of temporal aggregation are proposed with lengths of 1, 3, and 5 weeks, as shown in Table 3.2. Similarly, the spatial areas i are aggregated to form outbreak areas. The first proposed aggregation uses the MSOA. For the other aggregations, a clustering algorithm is applied to the areas i. A distance matrix is defined using the great-circle distance of the LSOA population centroids[3]. Next, the agglomerative hierarchical clustering (AHC) is applied using the complete linkage algorithm as in (1.6). The resulting dendrogram can then be cut to produce the number of areas required. Table 3.3 shows all outbreak area configurations and their sizes, and Figure 3.2 shows the outbreak areas in the OX and the NE map. The combination of outbreak areas and intervals gives the model the flexibility to capture diverse types of outbreaks: spatial-temporally localised ones and outbreaks covering large areas, which are known to exist for *Campylobacter* [McCarthy, 2017].

[3]LSOA population centroids obtained through the Office of National Statistics for the 2011 UK census.

Configuration number	Number of outbreak intervals	Length of interval
I	152	1 week
II	50	3 weeks
III	30	5 weeks

Table 3.2: Possible configurations for the outbreak intervals.

Configuration number	Number of outbreak areas	Source
A	92 (OX) / 99 (TW)	MSOA
B	60	AHC
C	40	AHC
D	20	AHC

Table 3.3: Possible configurations for the outbreak areas.

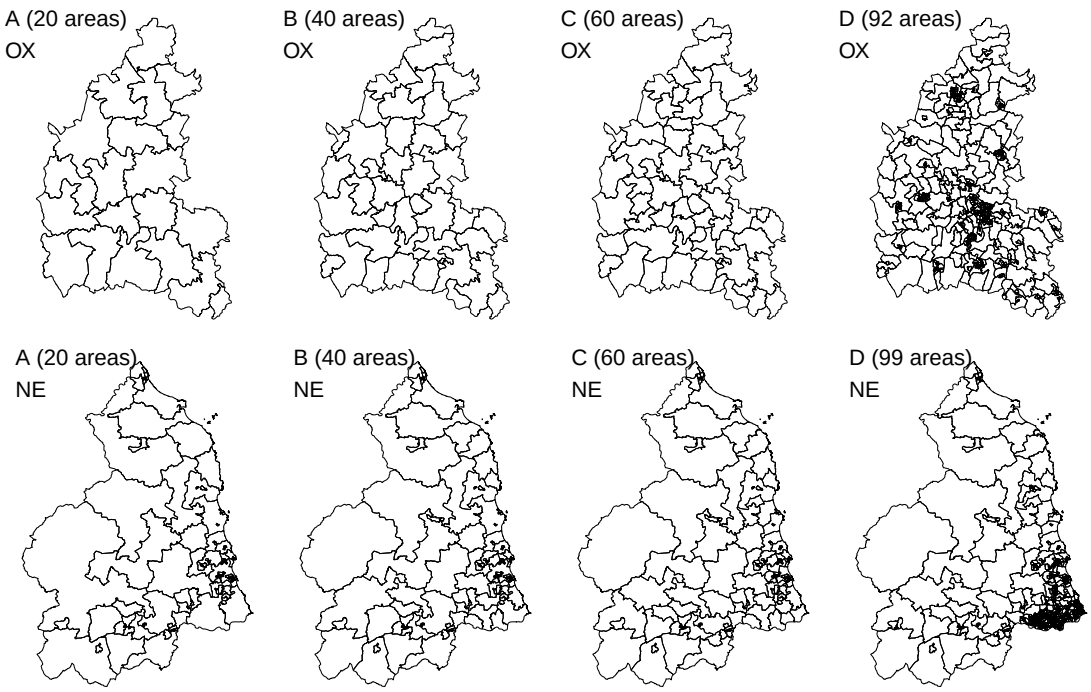


Figure 3.2: Possible configurations for the outbreak areas displayed in the OX and the NE map. The first row corresponds to OX and the second row to NE. Columns from left to right correspond to the configurations A, B, C and D, respectively.

The prior distributions of the hyperparameters are equally defined for both datasets. The spatial parameters U_i are allowed to have higher variance compared to

the temporal parameters R_t, capturing the heterogeneity of the spatial risk surface:

$$\tau_R \sim \text{Gamma}(5, 0.1),$$
$$\tau_U \sim \text{Gamma}(1, 0.5),$$

Also, the outbreak typical size B_σ is chosen as:

$$\beta_\sigma \sim \text{Gamma}(1, 1).$$

For the outbreak indicators $X_{\sigma\phi}$, the independent and the correlated models are implemented. For the independent case, the prior distribution of p is chosen according to the expected number of outbreaks observed in the studied area. As published by Public Health England, there were 9 outbreaks detected in England and Wales in 2017 PHE [2017], although it could under-represent the actual number of outbreaks. As suggested by Spencer et al. [2011], the mean of the parameter p is chosen such that one outbreak is expected per year and per area. Then, p is fixed to 1/52, 3/52, and 5/52 for the configurations I, II and III in Table 3.2, respectively. Also, the prior distribution is chosen to have large variance. Therefore, the distribution is chosen as:

$$p \sim \text{Beta}(1, \tfrac{52}{l} - 1),$$

where l is 1, 3, and 5 for the configurations I, II and III, respectively. For the correlated case, the prior distribution for p_{01} is chosen as in the independent case. That is:

$$p_{01} \sim \text{Beta}(1, \tfrac{52}{l} - 1).$$

For p_{11}, the prior distribution is given by:

$$p_{11} \sim \text{Beta}(2, 2).$$

The MCMC is run for combinations of outbreak area configurations (A, B, C, D), outbreak interval configurations (I, II, III), outbreak indicator cases (independent, correlated model), and dataset (OX, NE). For each combination, results are obtained running three chains in parallel with different starting locations, with 650 iterations each. Trace plots of the parameters are examined to assess the convergence of the chain.

3.3.2 SatScan^{TM} specifications

The retrospective spatial-temporal scan statistic is computed using the SatScan^{TM} software, separately for the OX and the NE database. A Poisson probability model is chosen for the space-time retrospective analysis. Dates are aggregated by week and the length of the clusters is limited to a maximum of five weeks. The studied area is divided into LSOA. The size of the clusters is limited to cover a maximum of 5% of the population. The software provides a list of *clusters* and the resulting p-values.

3.4 Results

In this section, the results of the spatial-temporal model are reviewed, following the implementation described in Section 3.3. In Section 3.4.1, the MCMC output, intercept, and the spatial and temporal terms are analysed using the independent model in OX, with outbreak intervals as in I and outbreak areas as in A. In Section 3.4.2, the independent and correlated models are compared, using all interval and area configurations. Also, the section includes the list of most probable outbreaks using all configurations. Finally, the most probable potential outbreaks are compared to the clusters found by the spatial-temporal scan statistic.

3.4.1 Model general results

Figure 3.3 shows a typical set of traces for the OX dataset, including the hyperparameters τ_U, τ_R, p, and randomly chosen spatial, temporal and outbreak size parameters U_i, R_t and B_σ, respectively. The colours represent the chains starting at different values. Additionally, Table 3.4 compares the Effective Sample Size ESS and the acceptance rate of each parameter for the OX dataset.

Parameter	R_t	U_i	B_σ
ESS	337-650	312-650	286-350
Mean acceptance rate (%)	23.1%	54.5%	47.9%

Table 3.4: Effective Sample Size ESS and mean acceptance rate of samples produced by the MCMC, for the parameters R_t, U_i and B_σ.

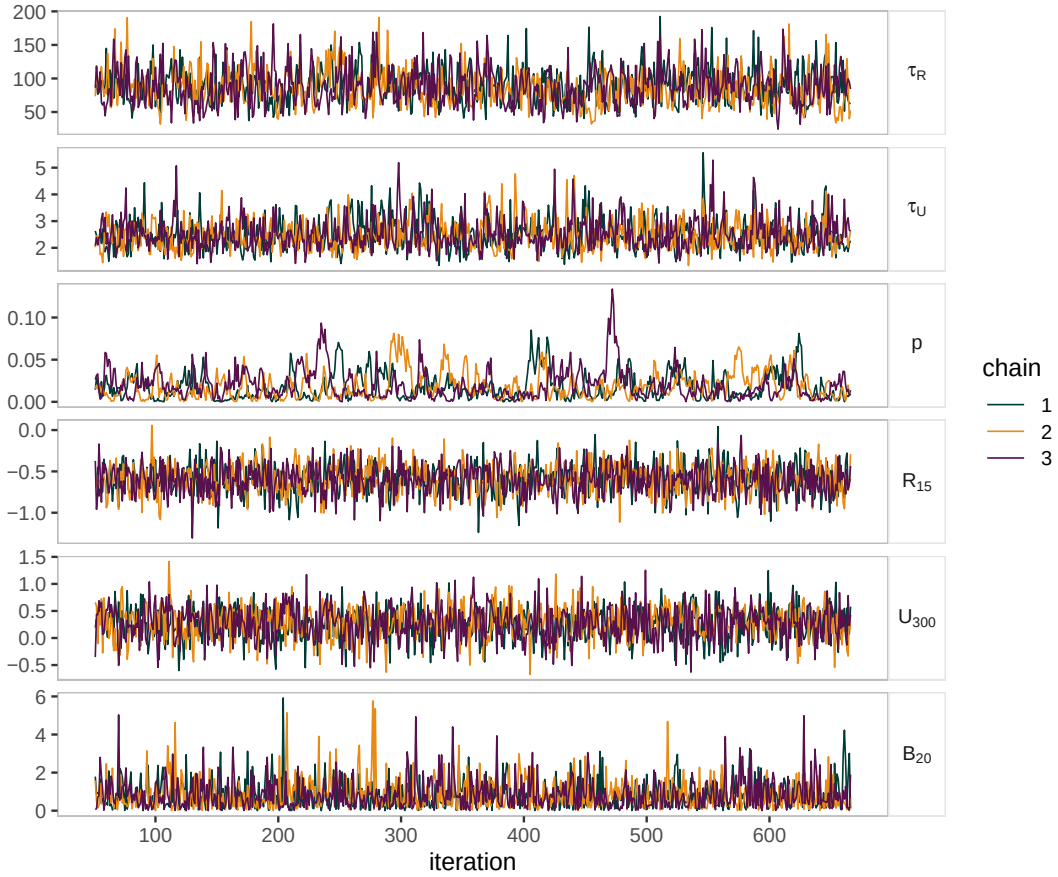


Figure 3.3: A typical set of traces obtained after running the model for the OX dataset, using the independent model, with outbreak intervals in I and outbreak areas in A. It includes traces of the hyperparameters τ_U, τ_R, p, and one sample of the spatial, temporal and the outbreak size parameters U_i, R_t and B_σ, respectively.

The exponential of the intercept α captures the probability of infection of an individual in an area with geometric mean spatial risk during a week with geometric mean temporal risk. For OX, the mean of the posterior distribution of α was -11.47 with a 95% confidence interval of (-11.57,-11.39). Also, for NE, the mean was -11.02 with a 95% confidence interval of (-11.07,-10.97). Figure 3.4 shows the expected number of sporadic and total cases per week, for OX and NE. For both datasets, there was a similar trend per year with the highest peak occurring during summer. However, a different pattern was observed in 2016 in OX, where the peak is not clear. Moreover, some peaks of observed cases were not fully covered by the outbreak indicators, and, therefore, the total and the sporadic number of cases followed a similar trend.

For the OX case, Figure 3.5 shows the spatial relative risk of sporadic cases,

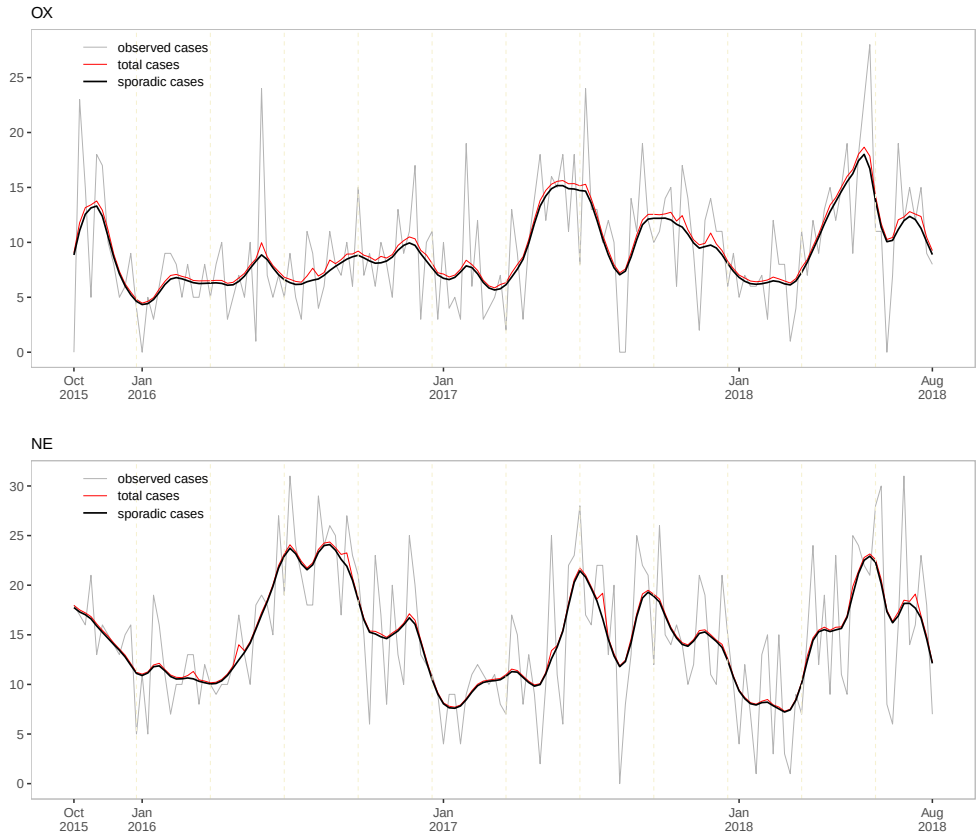


Figure 3.4: Comparison of the observed number of cases with the expected number of sporadic and total cases per week for OX (top) and NE (bottom). Yellow vertical lines denote the week when a new season started.

$\exp U_i$, which values are relative to the geometric mean of $\exp \alpha$. Regions at south-east had the lowest risk in the region. Also, rural areas showed a higher risk than urban regions, on average. That is, the ratio of the geometric mean of $\exp U_i$ for urban areas compared to rural areas was 0.87. Figure 3.7a shows the histogram of relative risk for both urban and rural areas. Similarly, results for the NE dataset are shown in Figure 3.6. Urban areas had a lower risk compared to rural areas, with a ratio of 0.88. The histogram of spatial relative risk is shown in Figure 3.7b.

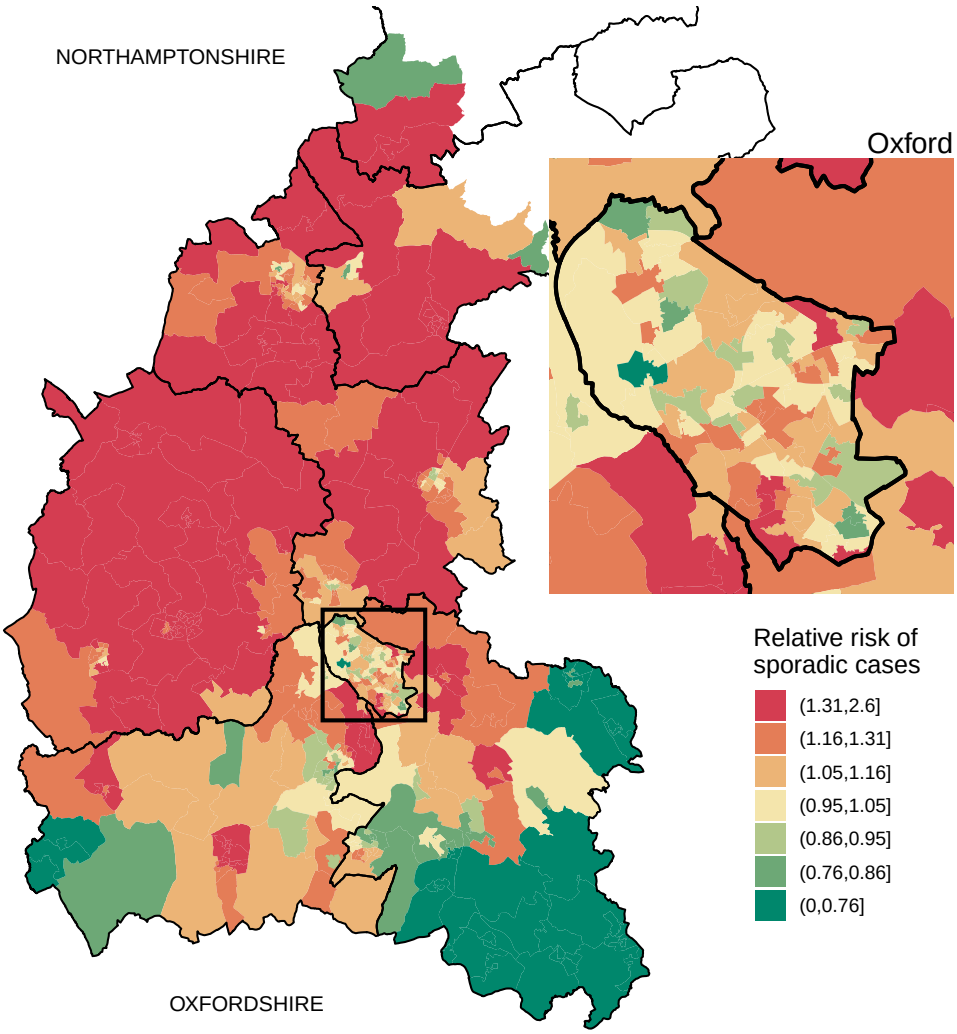


Figure 3.5: Map of the relative risk of sporadic cases per Lower-Layer Output Area in the areas covered by the OX dataset (left) and the augmentation in Oxford (top-right). The intervals displayed in the colour scheme are based on the deciles of the absolute value of the risk, such that a similar amount of regions correspond to each colour.

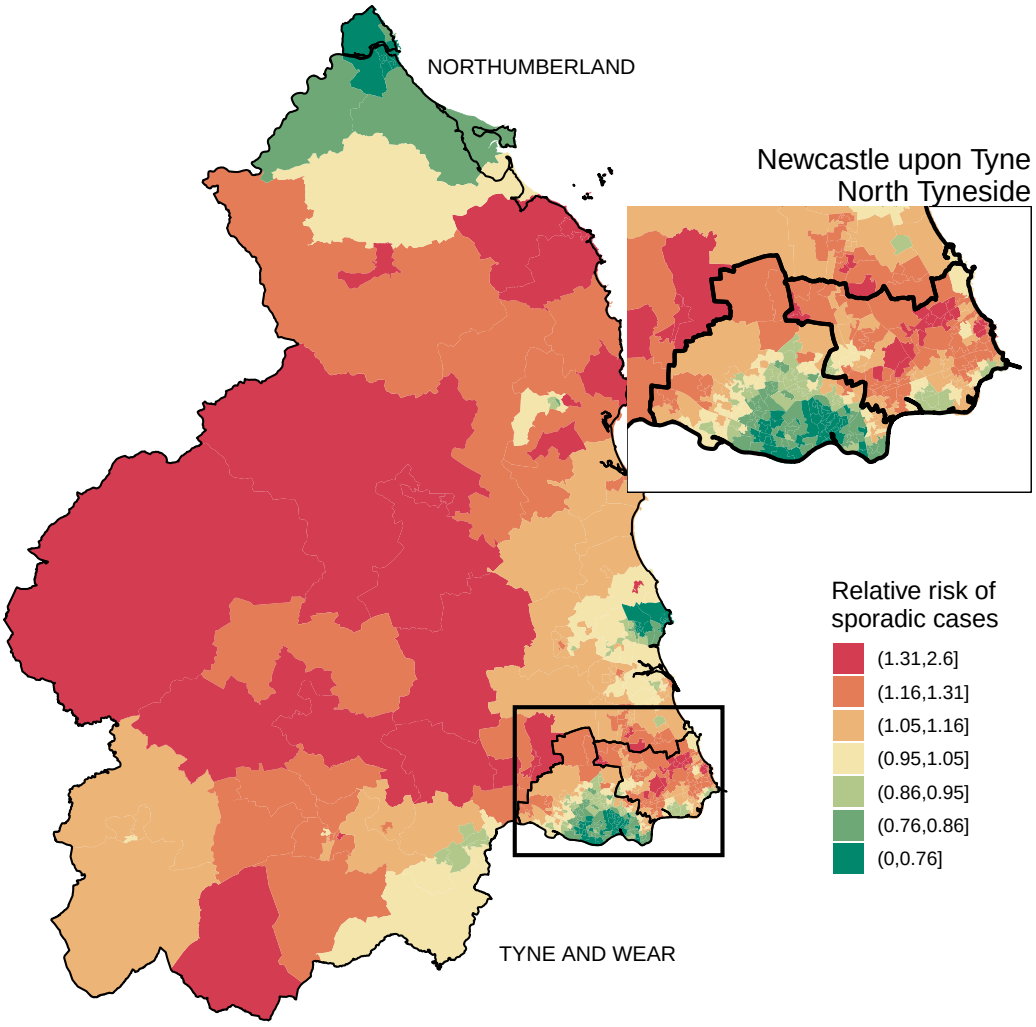
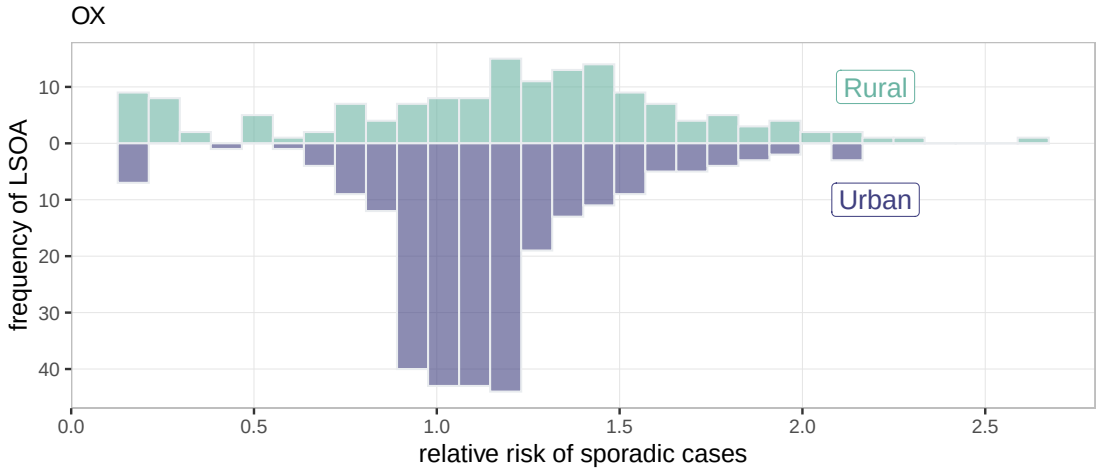
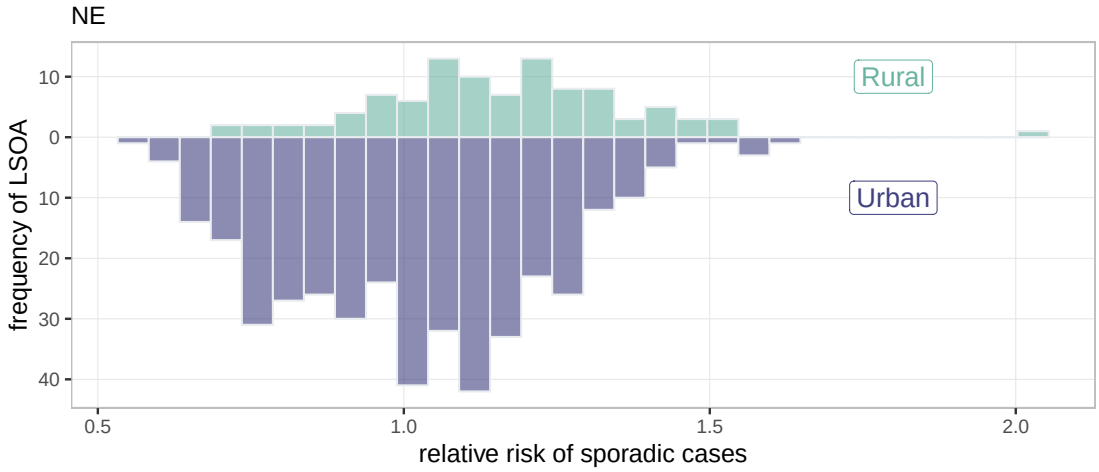


Figure 3.6: Map of the relative risk of sporadic cases per Lower-Layer Output Area in the areas covered by the NE dataset (left) and the augmentation in Newcastle upon Tyne and North Tyneside (top-rigth). The intervals displayed in the colour scheme are based on the deciles of the absolute value of the risk, such that a similar amount of regions correspond to each colour.



(a) Histogram for the OX dataset.



(b) Histogram for the NE dataset.

Figure 3.7: Histogram of the geometric mean of relative risk for urban and rural areas. The horizontal axis shows the geometric mean spatial risk. The height of the bars represents the number of LSOA that falls in each range of spatial risk.

3.4.2 Model validation

The model was run for the different intervals and areas described in Table 3.2 and Table 3.3, respectively. Also, it was run for OX and NE separately, and the independent (IM) and correlated model (CM). Figure 3.8 shows the area under the ROC curve (AUC) using IM for both datasets, whereas Figure 3.9 shows the AUC using CM. Both figures visualise the performance of both models and all configurations. In general, IM and

CM perform similarly. Also, OX registered higher AUC than NE, with better results at medium size areas and short intervals. Maximum performance was obtained for the configuration B-I. The results for NE were lower, where the highest scores occurred for the MSOA and long intervals, with a maximum AUC for the configuration A-III. Finally, *potential outbreaks* in one or more configurations are detailed in Table 3.5 for OX and Table 3.6 for NE. Potential outbreaks are defined as blocks with a probability higher than 60% for OX and 50% for NE. For the OX data, there were eight potential outbreaks. Potential outbreaks with label i., v. and vi. occurred in the same MSOA area. Moreover, isolates from three patients in v. were genetically-linked to the isolate from one patient in vi. Note that potential outbreaks might include isolates that would not be part of the real outbreak.

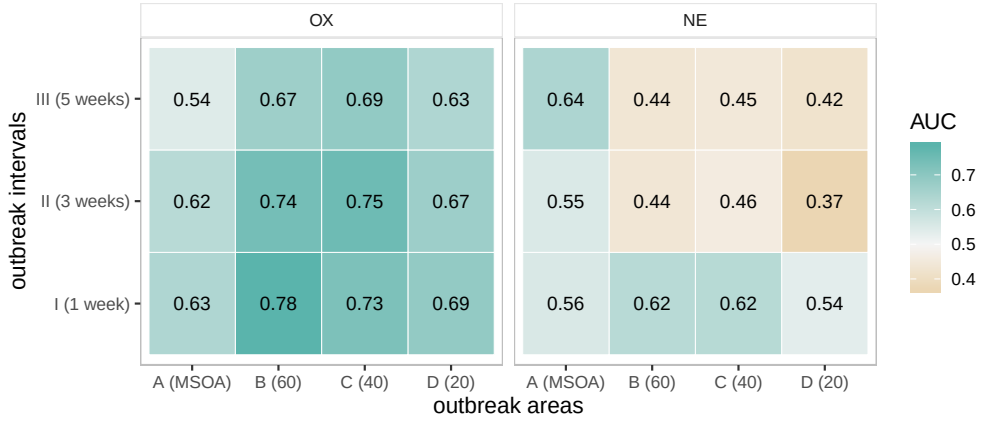


Figure 3.8: Area under the ROC curve for each interval and area configuration for IM.

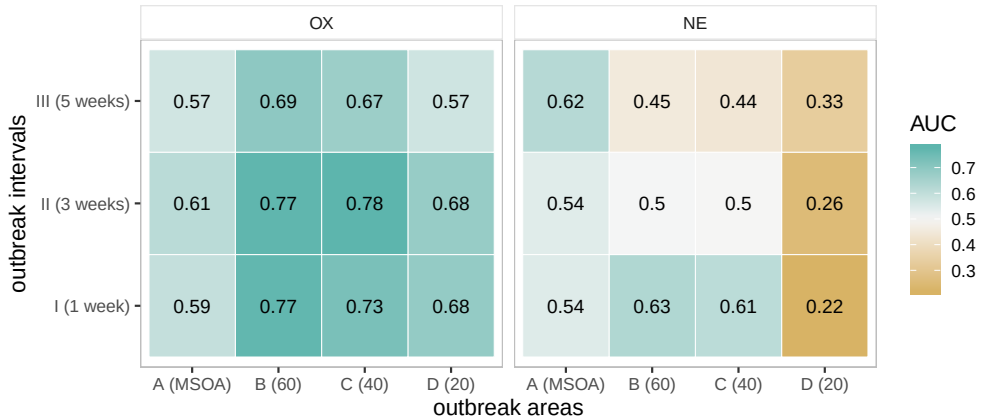


Figure 3.9: Area under the ROC curve for each interval and area configuration for CM.

	Number of cases	Probability of outbreak (%)	(Received) date range	Location type	Genetic distance within sequences
i	11	93.6	16 Oct 2015	Urban	2 (+9)
ii	4	86.0	29 Jul - 5 Aug 2016	Rural	4 (+2)
iii	6	81.4	20 Jul - 17 Aug 2018	Rural	678-1255
iv	4	73.4	17 Nov - 8 Dec 2017	Rural	0 (+2)
v	6	73.2	3 Feb 2017	Urban	697-1192
vi	6	68.9	27 May 2016	Urban	3-5 (+3)
vii	5	66.6	20 Jul 2018	Urban	908-1287

Table 3.5: List of *probable outbreaks* in OX detected by all configurations, using a threshold of 60%. It includes the number of cases, probability of being an outbreak, dates when the isolates were received in the laboratory, rural/urban classification of the spatial block and range of genetic distances within the block. $(+N)$ indicates there are N isolates genetically distant from the other isolates in the group. The list is ordered by probability (in decreasing order).

	Number of cases	Probability of outbreak (%)	(Received) date range	Location type	Genetic distance within sequences
i	4	92.4	21 Jul 2017	Urban	1093-1282
ii	4	59.2	29 Apr 2016	Urban	27 (+2)
iii	4	52.2	10 Aug 2018	Urban	1030-1259

Table 3.6: List of *probable outbreaks* in NE detected by all configurations, using a threshold of 50%. It includes the number of cases, probability of being an outbreak, dates when the isolates were received in the laboratory, rural/urban classification of the spatial block and range of genetic distances within the block. $(+N)$ indicates there are N isolates genetically distant from the other isolates in the group. The list is ordered by probability (in decreasing order)

3.4.3 SatScan™ comparison

SatScan™ provides a list of clusters detected by the model and their assigned p-value. The model was run for both datasets separately, following the specifications in Section 3.3.2. For OX, fourteen clusters were detected. Table 3.7 shows the top clusters with a p-value of less than 0.5. The total number of cases, duration in weeks and genetic distance within the cluster are included in the table. Clusters are compared to the output of the Bayesian independent model. Cases involved in each cluster are compared to the output of the Bayesian model of all configurations. The maximum outbreak probability found among all cases and configuration is listed in the table, as well as the model configuration that provided the maximum probability. For OX, two of five clusters were detected by

	P-value	Number of cases	Duration (weeks)	Genetic distance within sequences	Max. outbreak probability (%)	Model configuration
i	0.042	11	5	84-1258	19.6	C-II
ii	0.051	4	2	4-1251	86.0	A-II
iii	0.082	6	5	776-1260	17.8	C-I
iv	0.286	5	2	955-1285	34.1	A-II
v	0.304	12	5	0-1286	93.6	D-I

Table 3.7: Comparison of the ST model and SatScan^{TM} in OX. The table shows SatScan clusters with a p-value less than 0.50, including the number of cases, duration in weeks, and range of genetic distances within the cluster. Each cluster is compared to the output of all spatial-temporal model configurations. The configuration with the highest probability is shown in the table.

the Bayesian model. That is, clusters with label ii. and v. corresponded to potential outbreaks with label ii. and i., respectively. Also, the two clusters contained cases with a genetic distance of less than 4. Similarly, results for the NE are shown in Table 3.8, where two clusters were identified with a p-value of less than 0.5. Cluster with label ii. corresponded to the potential outbreak with label iii.

	p-value	Number of cases	Duration (weeks)	Genetic distance within sequences	Max. outbreak probability (%)	Model configuration
i	0.050	10	5	7-1263	40.2	A-III
ii	0.115	10	2	478-1285	52.2	A-I

Table 3.8: Comparison of the ST model and SatScan^{TM} in NE. The table shows SatScan clusters with a p-value less than 0.50, including the number of cases, duration in weeks, and range of genetic distances within the cluster. Each cluster is compared to the output of all spatial-temporal model configurations. The configuration with the highest probability is shown in the table.

3.5 Discussion

The chapter focussed on detecting outbreaks based on the spatial and temporal data of reported cases. The Bayesian hierarchical model proposed in Spencer et al. [2011] was described and applied to the Oxfordshire (OX) and Tyne and Wear (NE) datasets presented in Chapter 2. The model divides the spatial-temporal space into blocks, estimates the risk of sporadic cases, and labels any localised increase in risk as a potential outbreak. Therefore, the model provides a list of potential outbreaks with an associated probability as well as the posterior distribution of the parameters describing the risk of

sporadic infections.

The risk of sporadic cases was described by three terms: the average infection rate, the temporal effect and the spatial effect. The posterior distribution of each term displayed the spatial and temporal patterns of sporadic cases and served as a reference point to compare the epidemiology of the disease in both datasets. For instance, the exponential of the intercept quantifies the average risk of one person becoming infected as a sporadic case, and it was higher for NE compared to OX. Similarly, the spatial terms for OX showed higher fluctuations than the terms for NE, as shown in Figure 3.5 and Figure 3.6. These differences can be caused by variations in the geography: travel distances to food suppliers and water sources that might be highly exposed to contamination. Also, the distribution of the land could affect the risk factors. For instance, 12% of the area of OX is urban, where 65% of the population lives, whereas the urban area is 7% for NE containing 80% of the population. Additionally, rural regions exhibit a larger risk on average than urban areas for both datasets. These variations might be caused by higher exposure to farm animals sources or preferences in the consumption of local sources of food. However, some bordering regions exhibited low values in risk, in particular, the south-east and south-west of Oxfordshire as well as the south of Newcastle-upon-Tyne. This effect is possibly caused by a decrease in reports since patients' samples could be sent to hospitals outside the study region. For instance, cities as Reading and Swindon are closer to the south of Oxfordshire, as well as Gateshead to Newcastle-upon-Tyne.

Figure 3.4 shows the reported number of cases (grey) and the expected number of sporadic cases (black) per week. The variations of the former were smoothed by the trend of sporadic cases. Although both datasets have been analysed separately, they exhibited similar patterns. Major peaks occurred in July, August and September, similar to summer peaks found in England and Wales in a previous study [Louis et al., 2005]. Figure 3.4 also shows the expected number of cases including outbreaks (red). High peaks in the reported cases were not completely captured by the outbreak detection since the increase in incidences was not localised in a single outbreak region.

The accuracy of the outbreak detection was studied for different configurations, changing the duration and the size of the outbreak regions. Moreover, two versions of the model were applied, according to the correlation between the outbreak indicators: the independent model (IM) and the correlated model (CM). Detected outbreaks were *validated* using a genetically-linked measure, although highlighting that this is a comparison rather than a validation . Then, the accuracy of the detection was analysed using a ROC curve. Figures 3.8 and 3.9 show the area under the ROC curve for both models using all configuration proposed. Results were similar for both approaches, also showing that the

model performed better in detecting genetically-linked outbreaks for the OX data. For the NE data, the AUC was close to 0.5, suggesting that the model was unable to detect genetically-linked outbreaks. The poor performance of the model might be caused by a low incidence of localised outbreaks in NE. For instance, if a widely distributed food is infected early on the food chain, it could result in a spatially or temporally dispersed outbreak McCarthy [2017], difficult to be detected by the model. Also, the number of real outbreaks is unknown, and this measure relies on the genetic proximity of isolates.

A threshold of 60% for OX and 50% for NE were chosen to analyse the detected outbreaks in detail. Three of the eight outbreaks in OX occurred in the same area in Oxfordshire. Further inspection showed that two of these outbreaks were genetically linked even when they were separated by nine months. This event might be a persistent outbreak spread in time that was not captured by the sporadic risk. These results suggest the benefits of mixing genetic data with epidemiological data. The spatial-temporal model presented in this chapter could incorporate genetic distances to detect outbreaks that are not localised in space and time.

Finally, the spatial-temporal model discussed in this chapter was compared to a common existing outbreak detection mechanism known as the scan statistics. The model scans spatial-temporal cylinders of different sizes and compares the number of the observed and expected number of cases in the cylinder. If there are more cases than expected, the cylinder is labelled as a cluster. Table 3.7 and Table 3.8 displays the clusters found by the scan. Both approaches captured three clusters in common out of seven found by the scan. The spatial-temporal Bayesian model had several advantages over the scan. First, it provided the probability of being a potential outbreak, a flexible measure to evaluate the detection mechanism. Second, the model considers the seasonality pattern. However, the scan does not require to run several configurations to evaluate several regions and outbreak lengths. This comparison between the Bayesian model and the scan statistic is not a validation since real outbreaks are unknown.

Chapter 4

Spatial-genetic model for outbreak detection

The model in Chapter 3 aimed to identify localised spatial-temporal outbreaks, where the dynamics of the sporadic cases were captured by a spatial and a temporal *smooth* surface. The inclusion of genetic data to this type of model provides information to detect other types of outbreaks, such as localised ones in space but dispersed in time [McCarthy, 2017]. In this chapter, the spatial-temporal model is adapted to detect spatial-genetic outbreaks that may not necessarily be localised in time. To incorporate genetic information, the model includes a smooth surface describing the risk of sporadic cases for each genetic type using a Gaussian process. Any unexpected increase in cases is labelled as a possible outbreak. The model is applied to the OX dataset described in Chapter 2. The spatial and genetic risk distributions are analysed, and the potential outbreaks are compared to the results in the spatial-temporal model in Chapter 3. Although a spatial-temporal-genetic model could be studied, outbreaks localised in time, space and genetics would not be more informative than the results in Chapter 3. Therefore, the main goal of this chapter is to describe the mathematical aspects of the spatial-genetic model, as well as to present the results for the OX dataset. The model applied to the NE dataset is studied in Chapter 6.3.

The chapter is organised as follows. Section 4.1 explains the motivation to propose a spatial-genetic model and Section 4.2 formulates a model for sporadic cases, using a Bayesian hierarchical model. The model for spatial-genetic outbreak detection is presented in Section 4.3 and the details of the implementation are explained in Section 4.4. Section 4.5 presents the results of applying the model on the data available for this project. Finally, the analysis of the results, limitations and further work are discussed

in Section 4.6.

4.1 Motivation

The goal of this chapter is to produce a method for outbreak detection using epidemiological and genetic information of reported cases in some regions in the UK. The data available for this purpose were presented in Chapter 2. It contains three main sources of information: space, time, and the whole genome sequences of sampled bacteria. The method described in this chapter analyses only spatial and genetic data, intending to find outbreaks localised in space and with similar genetic sequences. However, this approach requires a notion of proximity between genome sequences. With this notion, it is possible to understand which genetic types are causing sporadic cases, and allows the model to determine when there is an outbreak. That is, if the sporadic cases are studied, it is possible to label outbreaks when there is a striking increase in the number of cases. This approach was used in the spatial-temporal model presented in Chapter 3, using a Bayesian hierarchical model. In this chapter, the spatial-temporal model is adjusted to use the spatial data and whole-genome sequences.

The spatial-temporal model decomposed the log-risk of sporadic cases into purely spatial and temporal components, where the priors of each component produced a smooth spatial and temporal surface, respectively. To include a genetic term into a similar model, a smooth surface of the log-risk of genetic types should be proposed. Before presenting a formulation for the problem, the space of genetic sequences is explored to understand what a sporadic surface will capture and how to smooth it.

In Section 2.5.1, a space of all possible sequences was defined and denoted by $\mathcal{G}$ with a *distance measure* d given by (4.1):

$$d(g,h) = \sum_{i=1}^{L} \mathbb{1}(g_i \neq h_i).$$

for $g, h \in \mathcal{G}$, and where the pair $(\mathcal{G}, d)$ forms a metric space. L indicates the length of the typing scheme, and it is chosen to be cgMLST for the following chapters. In Section 2.5.1 it was also shown that the sequences collected in the dataset are not uniformly distributed in $\mathcal{G}$. For instance, if a ball of radius 20 is drawn around an observed ST-21 sequence, the ball will contain an average of 82 other sequences from the dataset. That is, on average there exist 82 sequences that differ in less than or equal to 20 alleles to an ST-21 sequence. Conversely, a ball around an ST-464 sequence will contain an average

of 33.

There are certain genetic types such as ST-21 that have a higher risk of being observed, while types as the ST-464 are less common. Observing a high number of *uncommon* sequences could be suggestive of an outbreak. To build a model to understand the behaviour of the sequences of sporadic cases, a smooth surface describing the risk of genetic sequences should be constructed.

4.2 Sporadic cases in the genetic space

4.2.1 Formulation of a smooth risk surface

A smooth surface should capture the non-uniformity of the observed sequences on the space $\mathcal{G}$, similar to the formulation for the spatial-temporal model (Section 3.1). For the spatial case, a Gaussian Markov Random Field (GMRF) was used to compute the log-risk associated with the space, where a Gaussian field was defined over the discrete domain of regions with a notion of a neighbourhood. The idea of the GMRF could be extended but instead of using a precision matrix encoding the notion of neighbourhood, a kernel can be used. The kernel is described by a covariance function, as detailed in Section 1.3.2. In that sense, the risk of every type of sequence is described as a Gaussian process over the space $(\mathcal{G}, d)$. That is, the log-relative risk of the vector of observed sequences $\boldsymbol{G} = (g_s)_{s \in S}$, with indexes in S, follows a multivariate normal distribution:

$$\boldsymbol{G}|\boldsymbol{\Sigma} \sim \text{MVN}(0, \boldsymbol{\Sigma}), \tag{4.1}$$

with mean $\mathbf{0}$, and covariance matrix $\boldsymbol{\Sigma}$ given by the covariance function: $\mathcal{C}_w : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$:

$$[\boldsymbol{\Sigma}]_{ss'} = \mathcal{C}_{\boldsymbol{w}}(g_s, g_{s'}),$$

where $s, s' \in S$ and with parameters $\boldsymbol{w}$. The formulation of the prior in (4.1) will be updated in Section 4.3.1, to resolve some identifiability issues with the outbreak detection model.

4.2.2 Model for sporadic cases

The prior defined in (4.1) is used to extend the spatial-temporal model to include genetic data. In particular, a spatial-genetic model is proposed to study possible localised outbreaks in space that are persistent for a long period. Similarly to the spatial-temporal case, the count of cases is defined as y_{is}, for a region i and a genetic sequence s, and

follows a Poisson distribution $y_{is} \sim \text{Poisson}(n_i \mu_{is})$, where n_i is the population in the region i. Before studying an outbreak detection mechanism, a model estimating the risk of sporadic cases is described. Therefore, the log-risk of sporadic cases is given by:

$$\log \mu_{is} = \alpha + U_i + G_s, \tag{4.2}$$

where U_i is the logarithm of the spatial risk of a region i and G_s is the logarithm of the genetic risk of a sequence s. However, the model in (4.2) might fail at capturing the heterogeneity of the risk in the genetic space. That is, in the current dataset, 82% of the observed sequences are unique. The respective terms G_s for these sequences will behave similarly, regardless of the density of cases in the neighbourhood. To overcome this issue, the genetic space is partitioned into regions or *clusters* of similar size. Then the model can estimate the risk per cluster instead of the risk per sequence. That is, the set of observed sequences S is partitioned into clusters c_k, $k \in \mathcal{K}$, and the genetic risk is labeled by the cluster index k instead of the sequence index s. This also reduces the computational cost of calculating the inverse and the determinant of $\mathbf{\Sigma}$, both of which are required for the estimation of parameters, as described in Section 4.4. To extend the notion of cluster proximity, a distance between clusters is defined as follows:

$$d_c(k, k') = \frac{1}{|c_k||c_{k'}|} \sum_{s \in c_k} \sum_{s' \in c_{k'}} d(s, s'),$$

equivalent to the unweighted average linkage dissimilarity in (1.5). For notation, K is introduced as the number of clusters: $K = |\mathcal{K}|$. Now, the terms G_s are rewritten as G_k and are given by a multivariate Normal distribution as in equation (4.1). The notion of similarity between neighbouring clusters and the smoothness of the surface are given by a covariance matrix, as explained in Section 4.2.3. Additionally, details on the partition into the clusters c_k is described in Section 4.4.1.

4.2.3 Covariance functions

Two types of covariance functions are used. First, the Squared Exponential covariance or SE is given by

$$[\mathbf{\Sigma}]_{kk'} = \tau_G^{-1} \exp\left(\frac{-d_c(k, k')^2}{2\rho^2}\right),$$

with two parameters: a precision parameter τ_G, and a length-scale ρ controlling the notion of closeness. Second, the Matérn kernel with parameters ν, ρ and τ_G, as in (1.2). Note that the Matérn function is equivalent to the Squared Exponential when $\nu \to \infty$.

Two Matérn functions with parameters $\nu = 1/2$ and $\nu = 3/2$ are used, given by:

$$[\mathbf{\Sigma}]_{kk'} = \tau_G^{-1} \exp\left(\frac{-d_c(k,k')}{\rho}\right),$$
$$[\mathbf{\Sigma}]_{kk'} = \tau_G^{-1}\left(1 + \frac{\sqrt{3}d_c(k,k')}{\rho}\right)\exp\left(-\frac{\sqrt{3}d_c(k,k')}{\rho}\right),$$

respectively, also written as M1/2 and M3/2. The shape of each covariance function is shown in Figure 4.1 for distances between 0 and 50 and a length-scale of 10. For each case, τ_G is a non-negative parameter controlling the precision of the values of G_k, and it is the maximum covariance that can be obtained. The length-scale ρ controls the distance at which two sequences are close enough to have a similar risk. In the SE and the M1/2 case, ρ is the distance where the covariance between two sequences is $1/e \approx 0.37$ times the maximum possible variance τ_G. The three covariance functions applied here share the same property: $[\mathbf{\Sigma}]_{kk'}$ strictly decreases when $d_c(k,k')$ increases. However, the SE is infinitely differentiable and therefore it is strongly smooth compared to the Matérn cases.

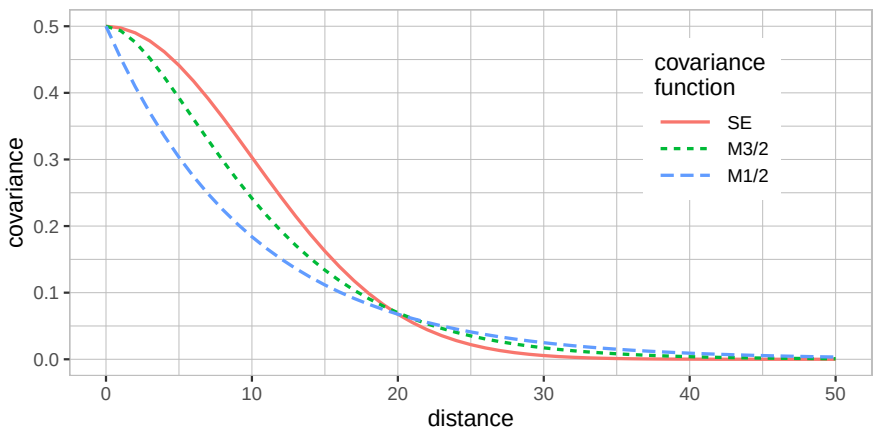


Figure 4.1: Covariance functions studied for the model: Squared Exponential (SE), Matérn function with $\nu = 3/2$ (M3/2), and Matérn function with $\nu = 1/2$ (M1/2), with $\rho = 10$. The functions are strictly decreasing.

4.3 Model for outbreak detection

In this section, the model in 4.2 is modified to include a term to study spatial-genetic outbreaks. Similarly, the counts of cases in a region i and a genetic cluster k are given

by:

$$y_{ik}|n_i, \mu_{ik} \sim \text{Poisson}(n_i \mu_{ik}),$$

where the risk μ_{ik} is given by

$$\log \mu_{ik} = \alpha + U_i + G_k + X_{\sigma(i)\xi(k)} B_{\xi(k)},$$

with an extra term $X_{\sigma(i)\xi(k)} B_{\xi(k)}$ capturing outbreaks in the block $\sigma(i)\xi(k)$. As in the previous model, the regions i are grouped into larger *spatial blocks* $\sigma(i)$, where the function σ gives the index of the block containing the region i. Similarly, the genetic clusters k are grouped into *genetic blocks* $\xi(k)$ in $\mathcal{G}$. The function ξ gives the index of the block containing the cluster k. The existence of blocks gives flexibility to the outbreak coverage and also avoids identifiability problems.

The term $X_{\sigma(i)\xi(k)}$ is a 0-1 random variable capturing outbreaks, where 1 means the block $\sigma(i)\xi(k)$ is an outbreak, and 0 otherwise. X follows a Bernoulli distribution as before, with parameter $p \in [0, 1]$. The term $B_{\xi(k)}$ captures the typical size of an outbreak having values in the genetic cluster $\xi(k)$, and it is given by,

$$B_{\xi(k)}|a_B, b_B \sim \text{Gamma}(a_B, b_B), \tag{4.3}$$

Finally, the terms α and U_i are given by:

$$\begin{aligned} \alpha|m_a, s_a &\sim \mathcal{N}(m_a, s_a) \\ U_i|\tau_U, U_{-i} &\sim \mathcal{N}\left(\frac{1}{|N_i|}\sum_{i' \in N_i} U_{i'}, \frac{\tau_U^{-1}}{|N_i|}\right), \end{aligned} \tag{4.4}$$

A review of the model structure and the detail of each parameters is shown in Figure 4.2 and in Table 4.1. Details on the algorithm tuning are included in Section 4.4.

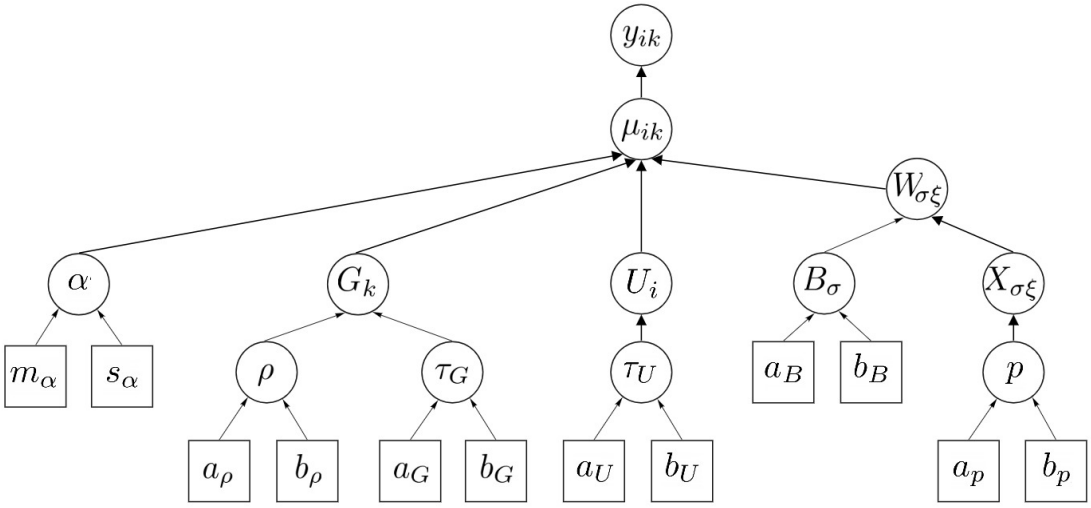


Figure 4.2: Directed Acyclic Graph describing the hierarchical conditional independence structure of the model, including the parameters α, U_i, G_k, B_ξ, $X_{\sigma(i)\xi(k)}$ and the hyper-parameters τ_U, τ_G and p.

	Parameter description	Prior distribution	Sampling algorithm
α	Intercept	Normal	M-H (normal proposal)
G_k	Risk of the genetic type in cluster k	Multivariate normal	M-H single updates (normal proposal)
ρ	Length-scale of covariance function	Gamma	M-H (truncated normal proposal)
τ_G	Precision of covariance function	Gamma	Gibbs (in block with ρ)
U_i	Risk on the spatial region i	GMRF	M-H single updates (normal, conditional prior proposal)
τ_U	Spatial term precision	Gamma	Gibbs
$X_{\sigma\xi}$	Outbreak indicator on the block $\sigma\xi$	Binomial	Gibbs
B_ξ	Typical size of outbreak on the cluster σ	Gamma	M-H single update (truncated normal proposal)
p	Probability that a block $\sigma\xi$ is an outbreak	Beta	Gibbs

Table 4.1: Description of the parameters of the model, including the prior distribution and the sampling algorithm using in the implementation.

4.3.1 Constraints

Additional identifiability constraints are imposed on the model. Since the terms α, U_i and G_k can take values in $\mathbb{R}$ and thus, there are several assignments to the parameter values that result in the same likelihood. For instance, $\alpha + U_i + G_k = 0$ can be obtained with infinitely many configurations of each term. Therefore, a constraint can be applied to the spatial and genetic terms to solve this restriction. In the first case, the spatial terms are subject to $\sum_i U_i = 0$. In the second case, a similar linear constraint can be applied to the genetic terms. However, imposing $\sum_k G_k = 0$ alters the original prior shown in (4.1). Therefore, the prior is modified such that the vector $\boldsymbol{G} = (G_k)_{k \in \mathcal{K}}$ is centred in the mean $\bar{\boldsymbol{G}}$. That is,

$$\boldsymbol{G} - \bar{\boldsymbol{G}} | \boldsymbol{\Sigma} \sim \text{MVN}(\boldsymbol{0}, \boldsymbol{\Sigma}).$$

This distribution can be rewritten as $\boldsymbol{AG} \sim \text{MVN}(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{A}$ is a $K \times K$ matrix such that:

$$\boldsymbol{A} = \mathbb{I}_K - \frac{1}{K}\mathbb{J}_K, \tag{4.5}$$

where $\mathbb{I}_K$ is the $K \times K$ identity matrix and $\mathbb{J}_K$ is the $K \times K$ matrix of ones. Following the linear properties of the multivariate Normal distribution, the prior can be rewritten as:

$$\mathbb{P}(\boldsymbol{G}|\boldsymbol{P}) \propto \exp\left(-\frac{1}{2}\boldsymbol{G}^T\boldsymbol{PG}\right), \tag{4.6}$$

where $\boldsymbol{P} = \boldsymbol{A\Sigma}^{-1}\boldsymbol{A}^T = \boldsymbol{A\Sigma}^{-1}\boldsymbol{A}$ is the precision matrix. Then, the constrain $\sum_k G_k = 0$ is applied to the genetic terms.

4.4 Implementation

The posterior distribution of the parameters of the model is estimated using a Markov Chain Monte Carlo algorithm. At each iteration, a new value of each parameter is proposed using different sampling algorithms, as described in Table 4.1. The parameter α is sampled using univariate Gaussian Random Walk proposals. As in the spatial-temporal model, the update of U_i is alternated between single Gaussian Random Walk proposals and conditional proposals [Knorr-Held, 1999], explained in Section 1.3.2. The proposed ρ and $B_{\sigma\xi}$ are sampled from a truncated Gaussian distribution on $\mathbb{R}^+$. The remaining parameters τ_G, τ_U, $X_{\sigma\xi}$ and p are updated using Gibbs sampling. Values of τ_G and ρ are proposed and accepted jointly to allow a good mixing between both parameters. Finally, the sampling algorithm for updating G_k is described in Section

4.4.2.

The model was implemented in `R` and is available on a public GitHub repository [Guzmán-Rincón, 2021]. The results in Section 4.5 were obtained using the dataset from Oxfordshire. Three parallel MCMC were run with different starting points, with 5000 iterations each. Additionally, the prior distribution of the hyperparameters was set as follows:

$$\begin{aligned} \tau_U &\sim \text{Gamma}(1, 0.1), \\ \tau_G &\sim \text{Gamma}(1, 0.01), \\ B_{\sigma\xi} &\sim \text{Gamma}(2, 1), \\ p &\sim \text{Beta}(1, 611). \end{aligned}$$

For the parameter p, the distribution was chosen such that the mean is $a_p/(a_p + b_p) = 1/612$ and the variance is large, where 612 is the number of genetic blocks $\xi(k)$. That is, the expected number of outbreaks per block cluster during the studied period is expected to be 1. The priors for the precision parameters τ_U and τ_G were chosen such that more variation is expected in the spatial component than in the genetic component. Also, two changes were made compared to the prior choices for the ST model. First, the prior for τ_U was modified to allow higher precision values. Second, the prior of the parameters $B_{\sigma\xi}$ was modified such that the mean size of a typical outbreak was 2. These changes enabled the spatial component to be smoother and imposed the model to capture larger outbreaks. Otherwise, the model did not capture potential outbreaks. Finally, the construction of the spatial and genetic *blocks* $\sigma(i)$ and $\xi(k)$ is explained in Section 4.4.1.

4.4.1 Blocks construction

For the implementation of the model, the following specifications are considered:

- The spatial area was divided into the *regions* i, which in turn were grouped into spatial blocks $\sigma(i)$. The choice for spatial regions and blocks are the LSOA (431 areas) and the MSOA (92 areas) in Oxfordshire, respectively (Section 2.2).

- The set of observed sequences $\{g_s, s \in S\}$ was grouped into the *clusters* c_k, $k \in \mathcal{K}$, which in turn were grouped into the *genetic blocks* $\xi(k)$. Consequently, an agglomerative hierarchical clustering algorithm was applied to $\{g_s, s \in S\}$ using the genetic distance d and the unweighted average linkage (Section 1.3.4). The

obtained dendrogram, denoted as $\mathcal{D}$, was cut twice to obtain clusters and genetic blocks. For the former case, the dendrogram was cut such that the average linkage distance within clusters is less than or equal to 10; the distances between the resulting clusters are shown in the histogram in Figure 4.3. For the genetic blocks, the dendrogram is cut such that the average linkage distance within genetic blocks is 50 or less. Figure 4.4 shows the dendrogram $\mathcal{D}$, where the two blue horizontal lines represent the two cutting heights: 10 and 50.

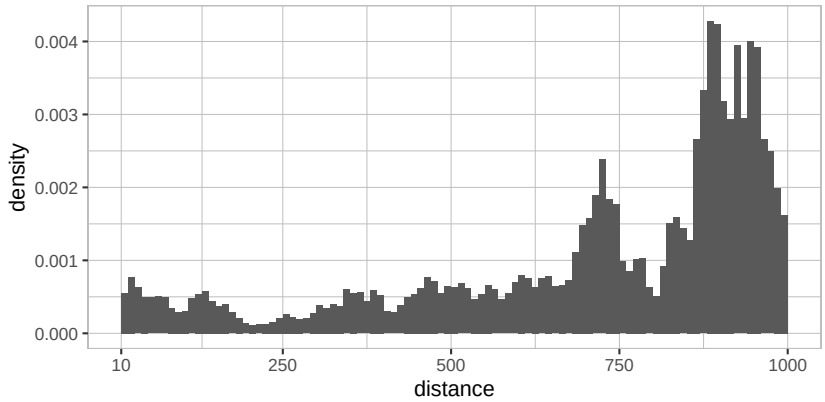


Figure 4.3: Histogram of distances between the clusters c_k, applying a hierarchical clustering with unweighted average linkage (only distances smaller than 1000 are shown).

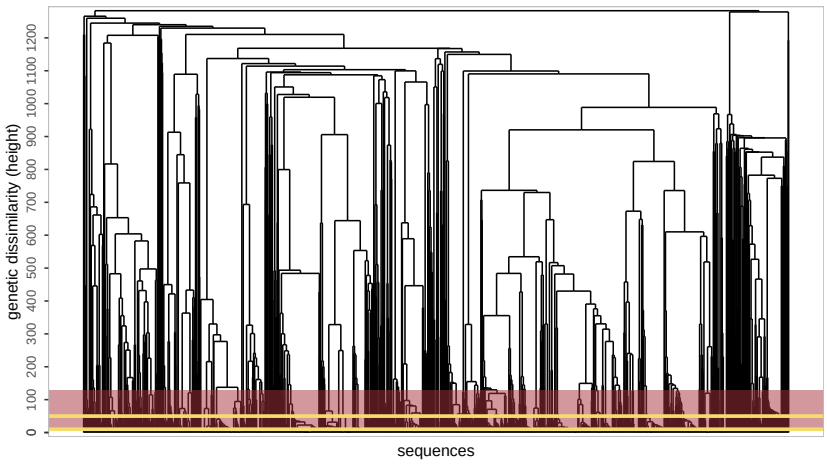


Figure 4.4: Dendrogram of a hierarchical clustering applied to the set of observed genome sequences g_s, based on the distance d in (4.1). The yellow lines are the cut heights to define the clusters c_k and the genetic blocks $\xi(k)$. The red band represents the range of heights chosen to update the parameters G_k for the MCMC implementation.

4.4.2 Sampling algorithm for the genetic risk parameters

This section provides a detailed explanation of the sampling algorithm used to update the genetic risk parameters G_k. Initially, single random walk jumps (RW) are proposed, such that for each k the proposed G'_k is obtained from a normal distribution around the current value G_k: $G'_k \sim \mathcal{N}(G_k, .)$. The RW jumps are accepted or rejected using an adaptive Metropolis-Hastings algorithm, where the variance of the jumps is automatically scaled to achieve an acceptance rate of approximately 44% [Garthwaite et al., 2016]. Samples of the posterior distributions of the model are obtained using an MCMC and the RW proposals for G_k to obtain preliminary results. Traces of three chains for one randomly chosen G_k are shown in Figure 4.6. Each chain is exploring different regions of the space of possible values of the G_k, showing low convergence of the RW sampling algorithm.

The slow convergence is due to the high correlations between the parameters G_k. Inspection of the covariance matrix shows that parameters are highly correlated for any value of ρ. Similar difficulties with sampling algorithms occur with other type of models (as the Generalised Additive Mixed Models or GAMM), where the parameters are highly correlated, and therefore the univariate RW sampling has slow convergence [Knorr-Held, 1999; Fahrmeir and Lang, 2001]. To overcome the issue, Fahrmeir and Lang [2001] proposed a generalised block update algorithm for problems similar to GAMM.

Fahrmeir and Lang [2001] describe an updating mechanism for a vector $\boldsymbol{G}$ that follows a distribution as in (4.6). Let $\boldsymbol{G} = (G_k)_{k\in\mathcal{K}}$. Instead of updating each G_k individually, the updates are done in blocks. At each MCMC iteration, $\mathcal{K}$ is partitioned into groups of size m. For a group $S \subset \mathcal{K}$, $\boldsymbol{G}_S$ denotes the subvector $(G_k)_{k\in S}$, S^c is the set of indices of $\mathcal{K}$ not in S, and $\boldsymbol{\Sigma}_{SS^c}$ denotes the submatrix of $\boldsymbol{\Sigma}$ with rows in S and columns in S^c. The proposal value for $\boldsymbol{G}_S$ is sampled from the conditional Normal distribution of $\boldsymbol{G}_S$ given $\boldsymbol{G}_{S^c}$ and $\boldsymbol{\Sigma}$. That is, $\boldsymbol{G}^*_S|\boldsymbol{G}_{S^c}, \boldsymbol{\Sigma} \sim \mathcal{N}(\boldsymbol{\mu}_S, \boldsymbol{\Xi}_S)$ with mean and covariance matrix given by:

$$\boldsymbol{\mu}_S = -\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\Sigma}_{SS^c}\boldsymbol{G}_{S^c}, \tag{4.7}$$

$$\boldsymbol{\Xi}_S = \boldsymbol{\Sigma}_{SS}^{-1}, \tag{4.8}$$

respectively. Since the proposal distribution of $\boldsymbol{G}_S$ is its conditional prior, the proposal is accepted with probability

$$\min\left\{1, \frac{\mathcal{L}(\boldsymbol{G}^*_S|\cdot)}{\mathcal{L}(\boldsymbol{G}_S|\cdot)}\right\}, \tag{4.9}$$

where $\mathcal{L}$ is the likelihood of $\boldsymbol{G}_S$. Finally, the set of indices $\mathcal{K}$ is partitioned into groups. At each iteration, a random integer m is drawn such that the indices are partitioned

into groups of size m. Large values of m improve convergence while random values of m improve the mixing of parameters. This updating scheme was applied in Section 3.3 to improve the mixing of the temporal parameters R_t.

The method proposed by Fahrmeir is adapted to update the genetic risk parameters G_k. For a subset S of $\mathcal{K}$, the proposal for $\boldsymbol{G}_S$ follows a Normal distribution with mean and covariance as in (4.7) and (4.8) respectively. Additionally, a strategy for partitioning the set of indices $\mathcal{K}$ into groups is proposed. Groups are constructed such that parameters within each group are highly correlated. Therefore, the partition is performed by cutting the dendrogram $\mathcal{D}$ introduced in Section 4.4.1 for different cutting heights. To vary the group sizes and avoid correlation problems between groups, random cuts of the dendrogram are proposed at each MCMC iteration, as detailed in Algorithm 4. This updating strategy is equivalent to the blocking strategy of Knorr-Held and Rue [2002], proved to be a valid MCMC technique, and then it produces an ergodic chain. Therefore this algorithm converges to a unique stationary distribution.

Algorithm 4: Updating strategy for $\boldsymbol{G}$

```
Result: (G^(1), ..., G^(T))
initialise G^(0);
for iteration t = 1, ..., T do
    choose a random height h from [10, 130] to cut the dendrogram;
    partition K into n_h groups a_1, ..., a_{n_h} by cutting the dendrogram;
    for l = 1, ..., n_h do
        compute μ_{a_l}, Ξ_{a_l} as in (4.7) and (4.8);
        sample from proposal G*_{a_l} | G^(t-1), Σ ~ N(μ_{a_l}, Ξ_{a_l});
        compute acceptance probability: α(G*_{a_l}; G^(t-1)_{a_l}) = min{1, L(G*_{a_l}|·) / L(G_{a_l}|·)};
        sample u ~ Unif(0, 1);
        if α(G*_{a_l}; G^(t-1)_{a_l}) > u then
            G^(t)_{a_l} ← G*_{a_l};
        else
            G^(t)_{a_l} ← G^(t-1)_{a_l};
        end
    end
end
```

The cutting height at each iteration should target an optimal acceptance rate to improve mixing. Fahrmeir suggests varying the size of groups between 1 and 40 to keep acceptance rates between 30% and 80%. Figure 4.5 shows the range of group sizes obtained for different dendrogram heights. Also, preliminary results were run to

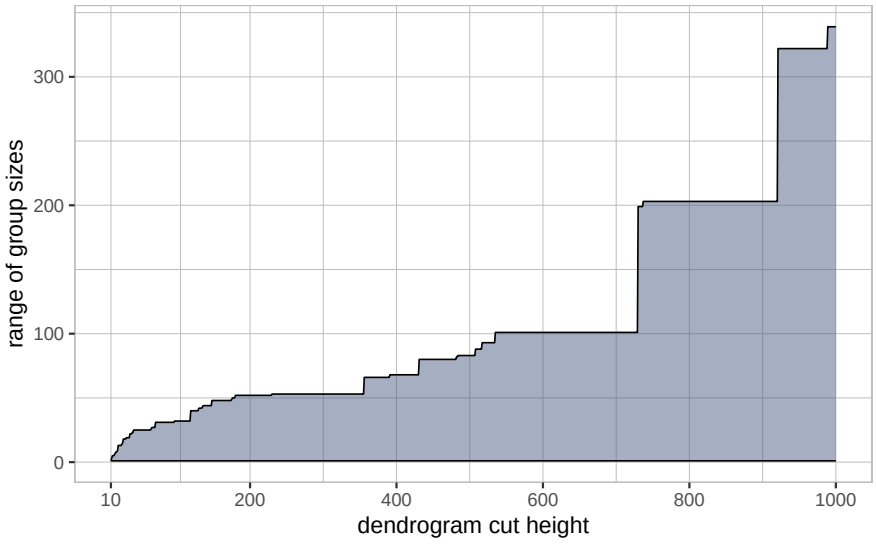


Figure 4.5: Range of group sizes of clusters obtained for each dendrogram cut-off height.

determine the optimal cuts for obtaining acceptance rates between 30% and 80%, as shown in Figure 4.8 (Section 4.5). Cuts from 1 to 1000 were considered to verify the behaviour between acceptance rates and group sizes. Cutting heights are chosen from a range between 10 and 130: 10 is chosen to ensure that groups are not smaller than the clusters c_k, and 130 is chosen such that group sizes do not exceed 40. In Figure 4.4, the red band corresponds to the range of heights chosen for cutting the dendrogram.

4.5 Results

In this section, the results of the model are presented, following the implementation details in Section 4.4. First, the convergence and correct mixing of the MCMC was assessed by examining the traces of each parameter. Initially, the parameters G_k were updated using single RW updates. Figure 4.6 shows the traces of a randomly chosen G_k. The slow convergence in the RW case is compared to the conditional updates proposed in Section 4.4.2. For the second case, Figure 4.7 shows the traces of τ_U, τ_g and p and random choices of G_k, U_i, B_ξ. Additionally, the Effective Sample Size ESS was computed to determine the number of effective samples the MCMC has produced, as shown in Table 4.2. The acceptance rate of each parameter is also shown in Table 4.2. The results displayed correspond to the kernel M1/2 since this function was chosen for the analysis, as explained later in Section 4.6.

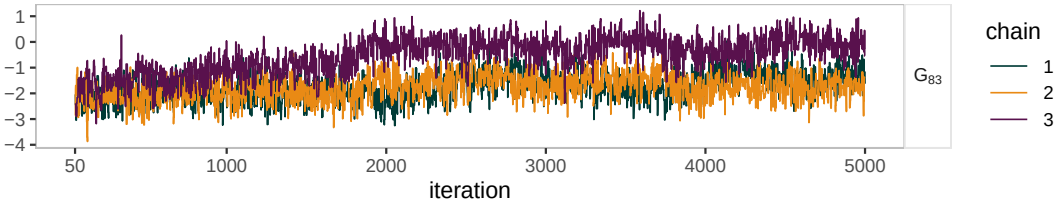


Figure 4.6: Traces of a randomly chosen G_k obtained after running the model, using single Random Walk (RW) updates on three chains with different starting points. Chains display poor mixing and convergence.

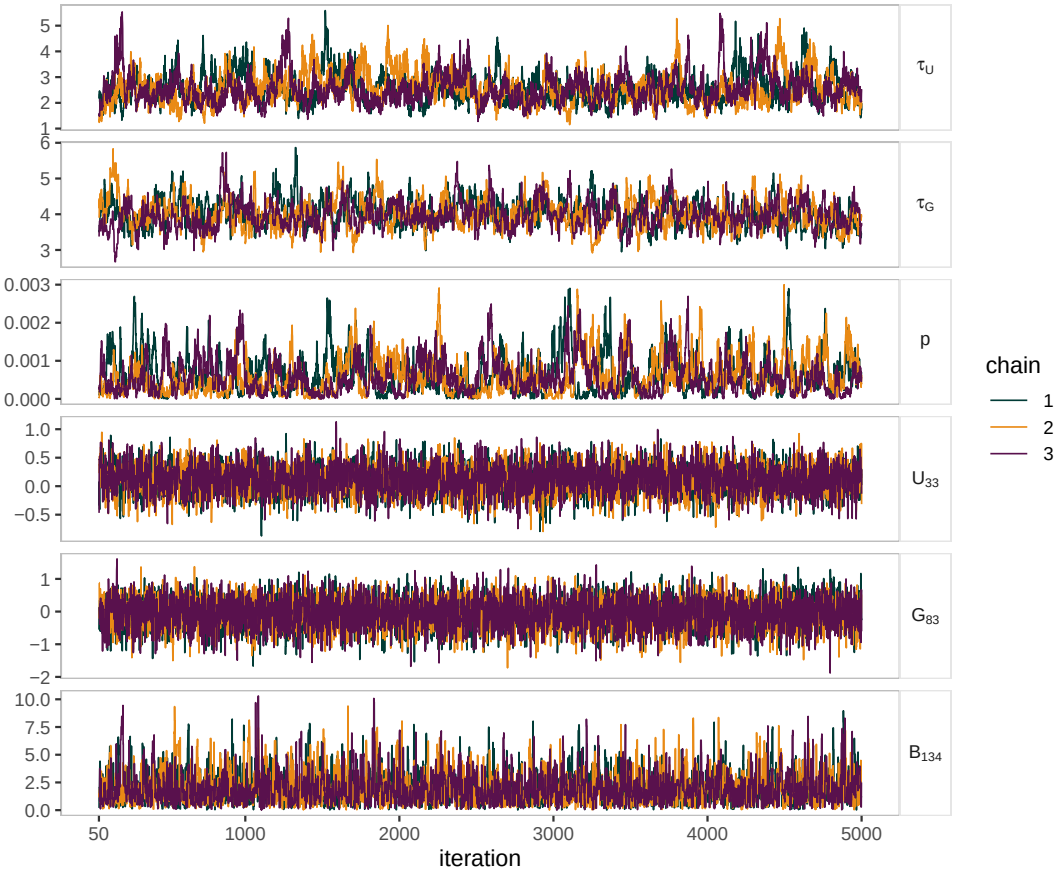


Figure 4.7: A typical set of traces obtained after running the model, including the traces of the hyperparameters τ_U, τ_G, p, and one sample of the spatial, genetic and outbreak size parameters U_i, G_k and B_ξ, respectively. G_k traces are obtained using the block updating strategy. Mixing is greatly improved in comparison to single RW updates, as shown in Figure 4.6 using the same value of k. The model required approximately 90 minutes of CPU time to run 5000 iterations.

Parameter	ρ	τ_G	G_k	U_i	B_ξ
ESS	182	313	648-8700	486-6289	825-2972
Acceptance rate (%)	44.0	44.0	43.8-44.1	43.8-44.0	44.0-44.1

Table 4.2: Effective Sample Size ESS and acceptance rate of samples produced by the MCMC, for the parameters ρ, τ_G, G_k, U_i and B_ξ.

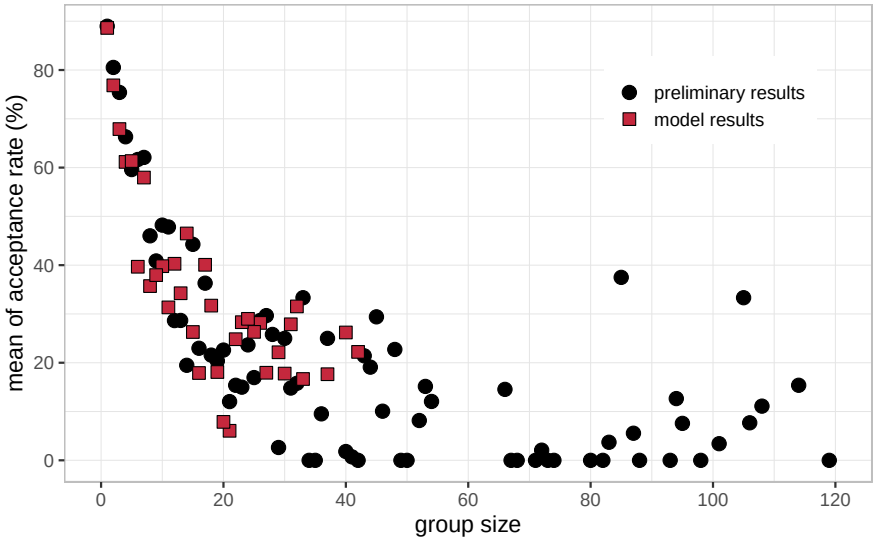


Figure 4.8: Mean of the acceptance rate in groups of different sizes. For the preliminary results, groups were obtained using random dendrogram cuts with heights between 10 and 1000. For the model results, the dendrogram cuts were obtained for heights between 10 and 130.

The mean of the posterior distribution of α was -12.96. Figure 4.9 displays the mean of log-risk per LSOA in Oxfordshire and selected areas in Northamptonshire. Some of the areas are classified as urban and some as rural. Figure 4.10 shows the histogram of mean log-risk for urban and rural areas.

For the genetic terms, the posterior distribution of each kernel parameter ρ and τ_G is shown in Figure 4.11 (bottom), for three types of the kernel. In the same figure (top), kernel functions are displayed as a function of distance. To explore the mean log-risk of the genetic terms in the genetic space, a Minimum Spanning Tree for the observed sequences is displayed in Figure 4.12. The colour of nodes represent the mean value of the log-risk, the black and white labels display the MLST of the nodes, and the blue and white labels indicate the number of the outbreak, as defined later in Table 4.3.

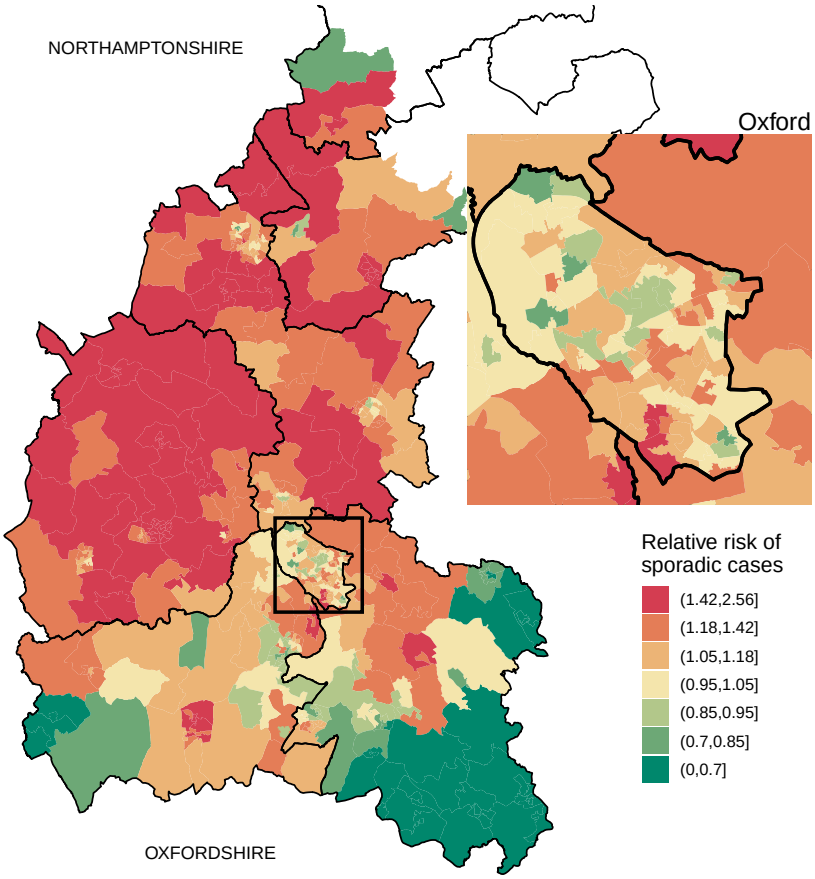


Figure 4.9: Map of the relative risk of sporadic cases per Lower-Layer Output Area in Oxfordshire and some areas in Northamptonshire (left) and the augmentation in Oxford (right). The intervals displayed in the colour scheme are based on the quantiles of the absolute value of the risk, such that a similar amount of regions correspond to each colour.

Additionally, the model output contains the list of blocks $\sigma\xi$ and the corresponding posterior distribution of $X_{\sigma\xi}$ yielding a posterior outbreak probability. Only blocks containing at least 2 cases are considered in the model analysis. The probability $p_{\sigma\xi}$ that a block $\sigma(i)\xi(k)$ is an outbreak is defined as the mean of the posterior distribution of $X_{\sigma\xi}$. The list of more probable outbreaks is listed in Table 4.3; that is, blocks with probability greater than $\theta = 0.15$. It includes the number of cases involved, the registered date intervals, the clonal complex to which the bacteria belong (if any), and the location area. The temporal span of each of block $\sigma\xi$ is shown in Figure 4.13, compared

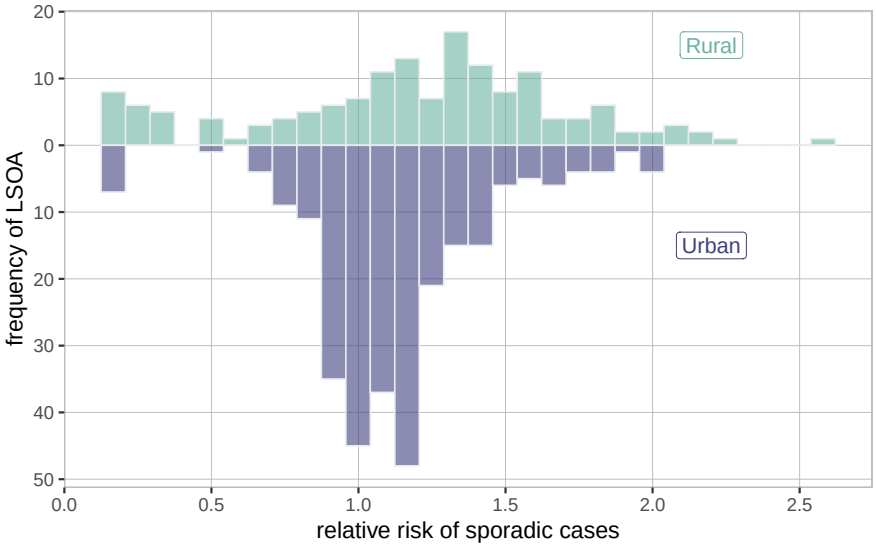


Figure 4.10: Histogram of the geometric mean of relative risk for urban and rural areas. The horizontal axis shows the geometric mean spatial risk. The height of the bars represents the number of LSOA that falls in each range of spatial risk.

	Number of cases	Probability of outbreak	(Received) date range	Clonal complex	Location
I	3	0.49	05-12 Sep 2017	ST-21	West Oxfordshire
II	2	0.25	07 Sep 2016	ST-48	Oxford
III	2	0.19	21 Mar 2018	ST-464	South Oxfordshire

Table 4.3: List of outbreaks for a threshold of $\theta = 0.15$, including the number of cases, probability of being an outbreak, dates when the isolates were received in the laboratory, clonal complex of the isolate and location reported by the patient.

to the posterior probability of being an outbreak. Similarly, a comparison between the spatial-genetic model and the spatial-temporal model in Chapter 3 is shown in Figure 4.14. In the figure, each point corresponds to an isolate, the horizontal axis displays the probability that the isolate is labelled as an outbreak by the spatial-temporal model, and the vertical axis shows the probability for the spatial-genetic model. Note that some isolates were part of a block in one model and not in the other. For instance, the top left points in Figure 4.14 were not part of the same spatial-genetic block since they occurred in different weeks.

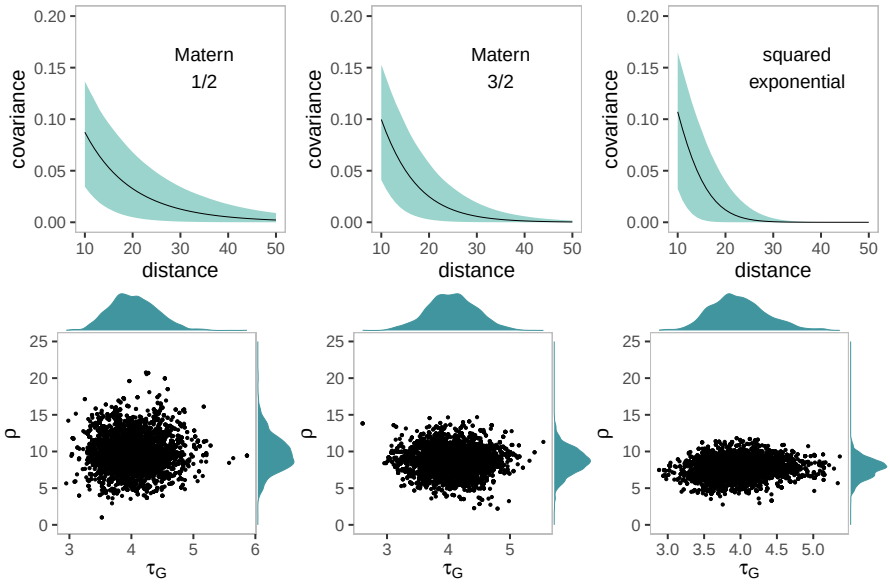


Figure 4.11: (Top) Covariance function according to the posterior distribution of the kernel parameters ρ and τ_G. The distances in the horizontal axis are greater than 10, since the distance between any pair of clusters is greater than or equal to 10, by definition. (Bottom) Posterior distribution of the kernel parameters ρ and τ_G, for three types of kernel. Left to right: Matérn covariance with $\nu = 1/2$, Matérn covariance with $\nu = 3/2$, squared exponential.

4.6 Discussion

The main goal of this project is to find outbreaks of campylobacteriosis, analysing the reported cases in some regions of the UK for a time frame of 3 years. In this Chapter, a method mixing the spatial and the genetic data is proposed, as an extension of the spatial-temporal model in Chapter 3 [Spencer et al., 2011]. The model aims to find spatial and genetic localised outbreaks that might be spread in time. For this purpose, the spatial and genetic parameters are obtained, describing the sporadic cases. Consequently, any unexpected increase in the observed number of sporadic cases is labelled as an outbreak. The spatial region is divided into small areas and aggregates all genetic sequences into clusters. Then, the risk of observing the disease in each area and genetic cluster is obtained. With these results, the probability that a localised set of cases is an outbreak is calculated.

This model has a similar spatial structure than the model presented in Chapter 3, and the log-risk per spatial region is similar in both cases (Figure 4.9). Some regions in the south of Oxfordshire had the lowest risk in the studied area. This border effect

could be explained by the location of hospitals in the area. Patients could receive clinical attention in other counties and report the infection elsewhere. Based on the Rural/Urban classification of the area, the relative risk of having the disease is compared for both categories in Figure 4.10. The areas with the highest risk were found in rural areas and, in general, urban areas had a lower risk than rural ones. This pattern shows a slightly higher trend of registering an infection in rural areas.

In comparison to the spatial-temporal case, the model in this chapter constructs a surface describing the risk of observing sporadic cases as a function of the genetic type of the bacteria. First, the set of genetic sequences is clustered in small balls. Second, the risk surface defined over the set of clusters follows a multivariate normal distribution. The kernel function associated with the covariance matrix determines the shape and smooth quality of the surface. Three kernels were chosen as well as the parameters, such as the length-scale ρ and the precision τ_G. The inspection of the covariance function explains how smooth is the surface, and it is calculated based on the posterior of the parameters ρ and τ_G, as shown in Figure 4.11. The three kernels obtained similar results. The mean of the posterior of the length-scale is approximately 13.2, 10.1 and 8.7 for M1/2, M3/2 and SE respectively. Although the values differ, the shape of the covariance functions is similar and has most of their weight for values lower than 50. These length-scale values agree with the notion of similarity observed in the isolates extracted from same patients (Figure 2.15), and also agrees with the assumptions made about 'close isolates' in previous work with *Campylobacter* [Cody et al., 2013]. However, the covariance function for the squared exponential (Figure 4.11) is small for distances greater than 30 compared to the Matérn covariances. Functions generated by the squared exponential kernel are characterised by a strong smoothness and its ineffectiveness to describe natural phenomena Stein [1999]. In this case, the length-scale decreased to allow the risk surface to change quickly and overcome the strong smoothness.

The minimum spanning tree in Figure 4.12 shows the genetic smooth surface of the relative risk of genetic sequences. Denser regions have a higher risk than the less concentrated regions, as expected since the surface describes the risk of observing a sequence. The outbreaks listed in Table 4.3 are also displayed in the tree. Although the model is not restricting the outbreaks to occur in a short period, the cases obtained appeared in a lapse of less than a week. Moreover, Figure 4.13 compares the span of each block with the probability of being an outbreak. This result confirms that the model did not find extended outbreaks that could stay hidden from detection even if they occurred in a localised area. Also, although all detected outbreaks were spatially and temporally localised, the spatial-temporal model did not capture them, as shown in Figure 4.14.

This suggests that outbreaks might remain hidden if analysed from the spatial-temporal perspective. Also, peaks on the number of cases in certain times of the year could be caused by a single strain.

Comparison between spatial-temporal and spatial-genetic models suggests the potential of mixing sources of information to detect outbreaks. Detailed analysis of the results obtained by both methods can also elucidate the existence of hidden outbreaks and suggests the creation of models that handle all sources of information at once. However, the capacity of the model to detect that an isolate is part of an outbreak is dependant on the choice of blocks (for instance, the top left points in Figure 4.14, captured by the spatial-temporal but not by the spatial-genetic model). Moreover, the model is dependant on the distance metric used in the genetic space and its ability to describe proximity. The model can be adapted to overcome those shortcomings by incorporating a comprehensive study of the genetic evolution of sequences.

Models incorporating spatial and genetic data have not previously been studied and therefore the model presented here is not compared to alternative approaches.

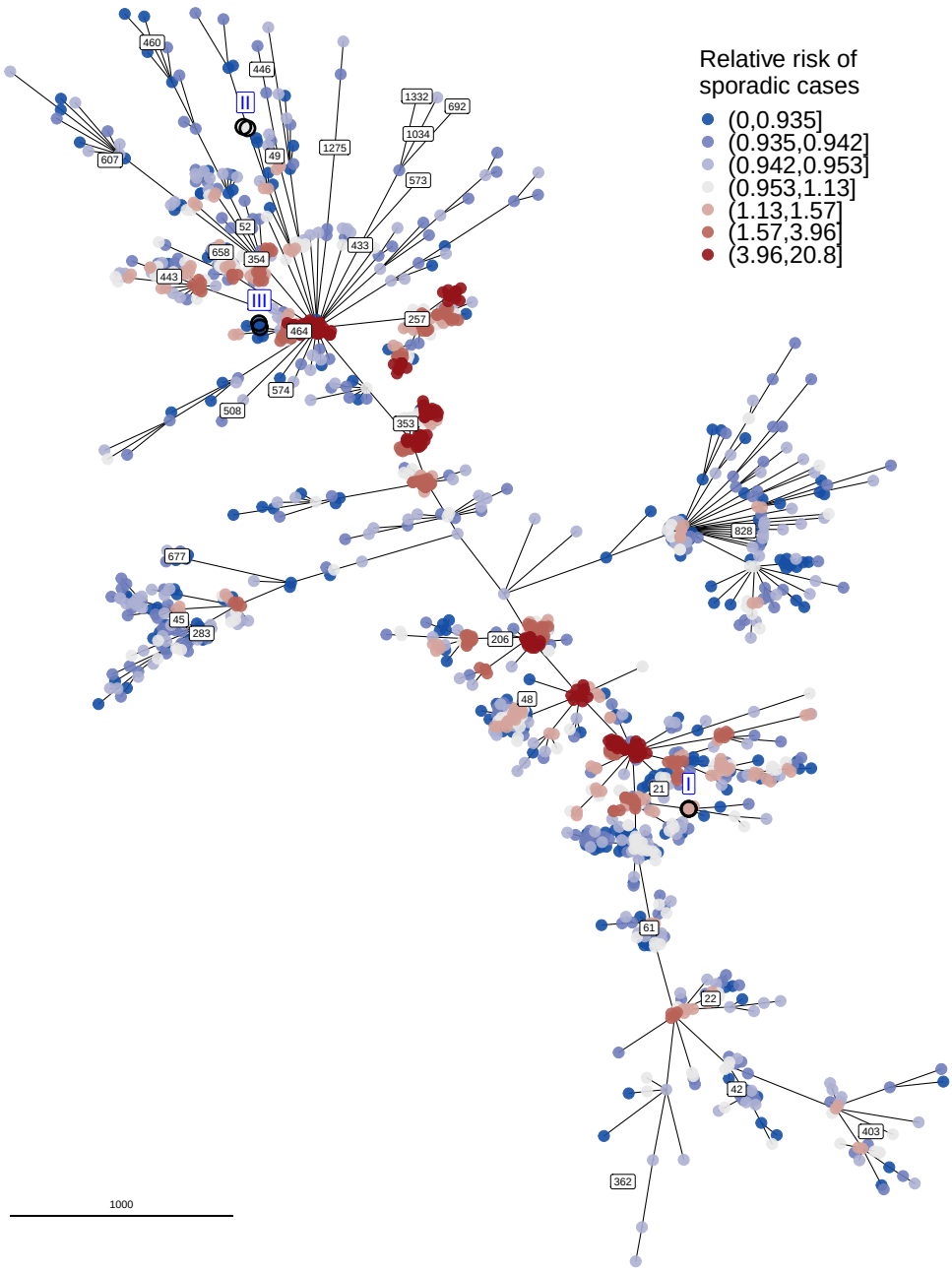


Figure 4.12: Minimum Spanning tree of the genetic sequences used in the model for the OX dataset. The colour of the nodes indicate the relative risk of sporadic cases. The intervals displayed in the colour scheme are based on the quantiles of the value of the risk, such that a similar amount of regions correspond to each colour. Also, potential outbreaks in Table 4.3 are labelled as I, II, III in blue squares.

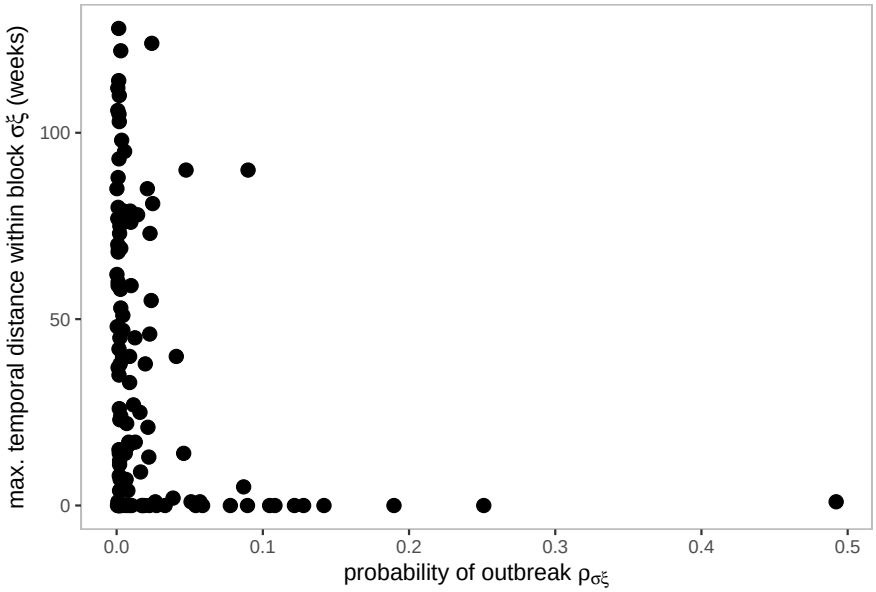


Figure 4.13: Probability that a block is an outbreak, compared to the maximum temporal distance within cases in the block (in weeks).

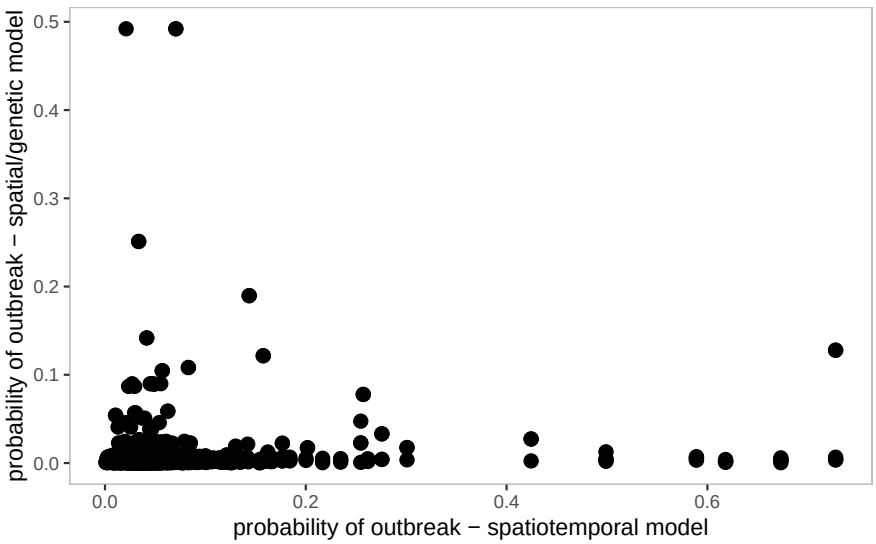


Figure 4.14: Comparison between the outbreak probability of each isolate for the spatial-temporal model in Chapter 3 and the spatial-genetic model in this chapter.

Chapter 5

Temporal-genetic model for outbreak detection

In previous chapters, a spatiotemporal and a spatial-genetic model were developed to find potential outbreaks of campylobacteriosis in some regions in the UK. Both models have a similar structure based on Bayesian hierarchical models, and each model has been adapted to process different combinations of data: spatial, temporal and genetic. In this chapter, an alternative version of the model is proposed, processing temporal and genetic data. It is referred to as the temporal-genetic model. The main goal is to detect potential outbreaks localised in time and caused by bacteria with similar genetic sequences, disregarding the spatial location where the infection occurred.

The explanation and motives to construct a temporal-genetic model are described in Section 5.1. Details of the model constructions are presented in Section 5.2.1. A modified version of this approach aims to analyse the temporal seasonality of different genotypes. The modified version is described in Section 5.2.2. Implementation aspects are described in Section 5.3 and employed to produce the results in Section 5.4. Finally, the discussion about the model and results is presented in Section 5.5.

5.1 Motivation

Although *Campylobacter* is the main pathogen causing foodborne illnesses in the UK, few outbreaks are detected, comprising only 0.1% of reported cases [Pebody et al., 1997]. Several studies have suggested that sporadic cases might be part of outbreaks dispersed in large regions and that these diffused outbreaks are not detected by epidemiological means [Fernandes et al., 2015]. For instance, outbreaks could be caused by contamination

in the early stages of the food chain [McCarthy, 2017]. However, it has been suggested that spatially diffused outbreaks are potentially observed when analysing genetic data [Besser et al., 2018; McCarthy, 2017], as seen in other pathogens outbreaks [Fittipaldi et al., 2013; Butcher et al., 2016].

The spatiotemporal and the spatial-genetic model proposed in this project aimed to detect two different types of potential outbreaks: spatiotemporal localised outbreaks, and genetically related outbreaks closely localised in space. However, both approaches are not able to capture diffuse outbreaks. In this chapter, a temporal-genetic model is proposed based on the hierarchical structure used for previous models, and it is denoted by *global model*. When applied to the UK dataset, genetically related outbreaks could be identified, with cases close in time. Moreover, the model can be adapted to detect potential outbreaks that last several weeks, since diffuse outbreaks could be longlasting.

Although the temporal-genetic approach estimates the pattern caused by sporadic cases, it assumes that all sporadic cases follow a similar temporal trend, regardless of the genotype of the bacteria. Previous studies have shown how different types of *Campylobacter* genotypes produce different temporal trends in the reported cases [Cody et al., 2012]. The authors showed how some strains as the ST-45 has some peaks during summer compared to other strains, like the ST-353, that have a peak during winter. Based on this observation, the global model is adapted to identify temporal patterns per genotype and it is denoted by *genotype-based model*, as described in Section 5.2.2.

Both proposed models, the global and the genotype-based model, are aimed to label outbreaks if certain selected sets of cases are unexpectedly large compared to the sporadic trend. However, the potential outbreaks captured by both models might differ. For instance, an outbreak produced by a genotype with peak cases in summer could remain hidden within the global summer pattern. Similarly, some cases could mistakenly be labelled as outbreaks if they appear in an off-peak part of the year, produced by a genotype with a typical increase in that period. The genotypes to be included in the alternative model should be large enough such that the sporadic trend of the genotype does not confuse potential outbreaks with seasonal trends.

5.2 Model for outbreak detection

5.2.1 Global model

The count of cases is modelled using a Bayesian hierarchical model. The span of the data is divided into T intervals of a fixed number of weeks. The week index is denoted with

the subscript $t = 1, .., T$. The genetic space $\mathcal{G}$ is partitioned into clusters as described in Section 4.4.1, for the spatial-genetic model. Subindices $k = 1, ..., K$ denotes the cluster index, where K is the number of clusters in the partition. Therefore, the count of cases within the interval t and with genetic type in k is denoted by y_{tk} and follows a Poisson distribution with rate μ_{tk}. The logarithm of the rate μ_{tk} is written as:

$$\log \mu_{tk} = \alpha + R_t + G_k + X_{\phi(t)\xi(k)} B_{\xi(k)}, \tag{5.1}$$

where α is the intercept. The term R_t describes the logarithm of the temporal risk of sporadic cases for the interval t. The term G_k denotes the logarithm of the genetic risk of sporadic cases associated with the cluster k. The function $\phi(t)$ denotes the partition of intervals into larger *temporal blocks* and $\xi(k)$ denotes the partition of genetic clusters into the larger *genetic blocks*, similar to the approach in previous models. In the last term, the term $B_{\phi(t)\xi(k)}$ denotes the typical size of an outbreak with a genetic sequence in the genetic block $\phi(t)\xi(k)$. Finally, the term $X_{\phi(t)\xi(k)}$ is a 0-1 random variable capturing if there is an outbreak in the genetic interval and genetic block $\phi(t)\xi(k)$.

The prior for the temporal terms R_t is copied from the spatiotemporal model and is given by:

$$R_{t+1} - R_t | \tau_R, R_{1:t} \sim \mathcal{N}(R_t - R_{t-1}, \tau_R^{-1}),$$

where R_1 and R_2 have flat priors [Spencer et al., 2011]. Similarly, the prior of the genetic terms G_k is given by

$$\mathbb{P}(\boldsymbol{G}|\boldsymbol{P}) \propto \exp\left(-\frac{1}{2}\boldsymbol{G}^T \boldsymbol{P} \boldsymbol{G}\right),$$

as described in Section 4.3. $\boldsymbol{P} = \boldsymbol{A}\boldsymbol{\Sigma}^{-1}\boldsymbol{A}$ is the precision matrix, $\boldsymbol{\Sigma}$ is the covariance matrix and $\boldsymbol{A}$ is given in (4.5).

The remaining parameters follow the same priors as in previous models. That is, the α follows a Normal distribution with mean m_a and variance s_a. The $B_{\phi(t)\xi(k)}$ terms follow a Gamma distribution with shape a_B and rate b_B. Finally, the indicators $X_{\phi(t)\xi(k)}$ follow a Bernoulli distribution with parameter p. The structure of the genotype-based model is shown in Figure 5.1, and the parameters are described in Table 5.1.

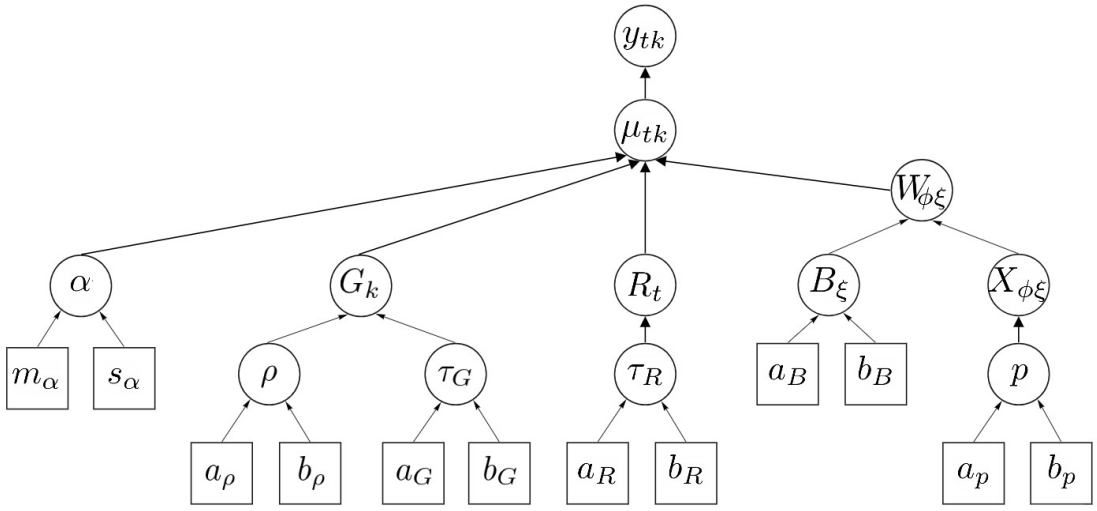


Figure 5.1: Directed Acyclic Graph describing the hierarchical conditional independence structure of the global model, including the parameters α, G_k, R_t, B_ξ, $X_{\phi(t)\xi(k)}$ and the hyperparameters ρ, τ_G, τ_R, τ_G and p.

	Parameter description	Prior distribution	Sampling algorithm
α	Intercept	Normal	M-H (normal proposal)
G_k	Risk of the genetic type k	Multivariate normal	M-H (normal proposal) M-H (block updates)
ρ	Lengthscale of covariance function	Gamma	Gibbs (in block with τ_G)
τ_G	Precision of covariance function	Gamma	M-H (truncated normal proposal)
R_t	Risk on the interval t	Second order random walk	M-H (normal proposal) M-H (block updates)
τ_R	Temporal term precision	Gamma	Gibbs
$X_{\phi\xi}$	Outbreak indicator on the block $\phi\xi$	Binomial	Gibbs
B_ξ	Typical size of outbreak on the block $\sigma\xi$	Gamma	M-H single update (truncated normal proposal)
p	Probability that a block $\phi\xi$ is an outbreak	Beta	Gibbs

Table 5.1: Description of the parameters of the global model, including the prior distribution and the sampling algorithm using in the implementation.

5.2.2 Genotype-based model

The model proposed in (5.1) is adapted to capture temporal patterns based on genotype. The parameters R_t are rewritten as $R_t^{\omega(k)}$ or equivalently as R_t^{ω}, where $\omega(k)$ is a function that assigns a cluster k to a genotype $\omega = \omega(k)$. The set of parameters $R_1^{\omega}, ..., R_t^{\omega}$ denotes the temporal risk linked to the genotype ω. Therefore, the Poisson parameters μ_{tk} can be rewritten as:

$$\log \mu_{tk} = \alpha + R_t^{\omega} + G_k + X_{\phi(t)\xi(k)} B_{\xi(k)}.$$

For each ω, the R_t^{ω} parameters follow a second-order Random Walk as follows:

$$R_{t+1}^{\omega} - R_t^{\omega} | \tau_R, R_{1:t}^{\omega} \sim \mathcal{N}(R_t^{\omega} - R_{t-1}^{\omega}, \tau_R^{-1}).$$

The parameters R_1^{ω} and R_2^{ω} are given flat priors. To avoid identifiability problems, a constraint is applied such that $\sum_t \sum_{\omega} R_t^{\omega} = 0$.

5.3 Implementation

The posterior sampling of the Bayesian hierarchical models proposed in Section 5.2 is estimated using MCMC techniques. The intercept α, the length-scale of the covariance function ρ, and the typical size of outbreaks $B_{\phi(t)\xi(k)}$ are updated using Metropolis-Hastings sampling, as described in Table 5.1. The precision of the covariance matrix τ_G, the temporal precision τ_R, the parameter p and the indicators $X_{\phi(t)\xi(k)}$ are updated using Gibbs sampling. The genetic terms G_k are sampled alternating single site Gaussian random walk proposals and block updates with conditional prior proposals, as described in Section 4.4.2. Similarly, the temporal terms R_t are updated by alternating two different methods: single site Gaussian random walk proposals and block updates, as described by Knorr-Held [1999]. For the genotype-based model, the proposed value of R_t^{ω} is obtained using block updates. The updating procedure is run independently for each ω.

The model is run with two different settings or *scenarios* to study and compare the global and the genotype-based model. The scenarios are described as follows:

i. The first scenario applies the global model to the combination of the OX and NE datasets. Therefore, a unique temporal trend is produced for both regions. The length of the temporal blocks is one week. Clusters are obtained as in Section 4.4 with cutting heights of 10 for the clusters and 50 for the genetic blocks. Thus the number of clusters is $K = 2228$, and the number of genetic blocks is 1143. The number of clusters is larger than in Section 4.4 since both datasets are merged.

ii. The second scenario studies the genotype-based model applied to the combination of the OX and NE datasets. The length of the intervals, the number of clusters and the number of genetic blocks are set as in scenario i. The K clusters are grouped into four sets $\omega_1, ..., \omega_4$ based on their clonal complex designation based on MLST. Three clonal complexes are chosen to define the ω-groups: ST-21, ST-353 and ST-45. The ST-21 type is chosen since it is the most abundant clonal complex in the dataset. The ST-353 and ST-45 are also abundant clonal complexes that have a strong seasonal pattern [Cody et al., 2012]. The last group ω_4 consists of the remaining sequences not included in the selected clonal complexes.

In an alternative third scenario, the model could be adapted to have different temporal trends for OX and NE (as it does for different genetic sets). However, this scenario is not included in the analysis. For both the global and the genotype-based model, the prior of the hyperparameters are chosen as:

$$\begin{aligned} \tau_R &\sim \text{Gamma}(5, 0.1), \\ \tau_G &\sim \text{Gamma}(1, 0.01), \\ \beta_{\sigma\xi} &\sim \text{Gamma}(2, 1), \\ p &\sim \text{Beta}(1, 1142). \end{aligned}$$

The prior of p has a mean of $1/(1 + 1142)$ such that the expected number of outbreaks per genetic block per week is 1. The kernel employed for the covariance matrix is the Matérn kernel with parameter $\nu = 1/2$. The algorithm is run for 5 000 iterations with different starting points. A burn-in period of 500 iterations is implemented. The results obtained are detailed in Section 5.4.

5.4 Results

The output of the model is described in two sections. Section 5.4.1 displays the results of the global and the genotype-based model. Section 5.4.2 shows the potential outbreaks found by both models.

5.4.1 Model results

For the global model, a set of traces of the MCMC output is shown in Figure 5.2. Visual inspection confirms the convergence of the MCMC chain. Also, Table 5.2 shows the Effective Sample Size and the acceptance rate of each parameter. The time-series of

cases is shown at the top of Figure 5.3. The grey line represents the number of observed cases, the black line shows the expected number of sporadic cases, and the red line shows the expected number of total cases including outbreaks. Equivalent plots are shown at the top of Figure 5.4 and Figure 5.5, showing cases associated with genotype ST-353 and ST-45 respectively.

Parameter	ρ	τ_G	τ_R	G_k	R_t	B_ξ
ESS	190	173	89	109-2463	57-396	188-953
Acceptance rate (%)	44.0	44.0	-	22.0-66.1	28.6-48.1	43.9-44.1

Table 5.2: Effective Sample Size ESS and acceptance rate of samples produced by the MCMC, for the parameters ρ, τ_G, τ_R, G_k, R_t and B_ξ, and 4500 iterations.

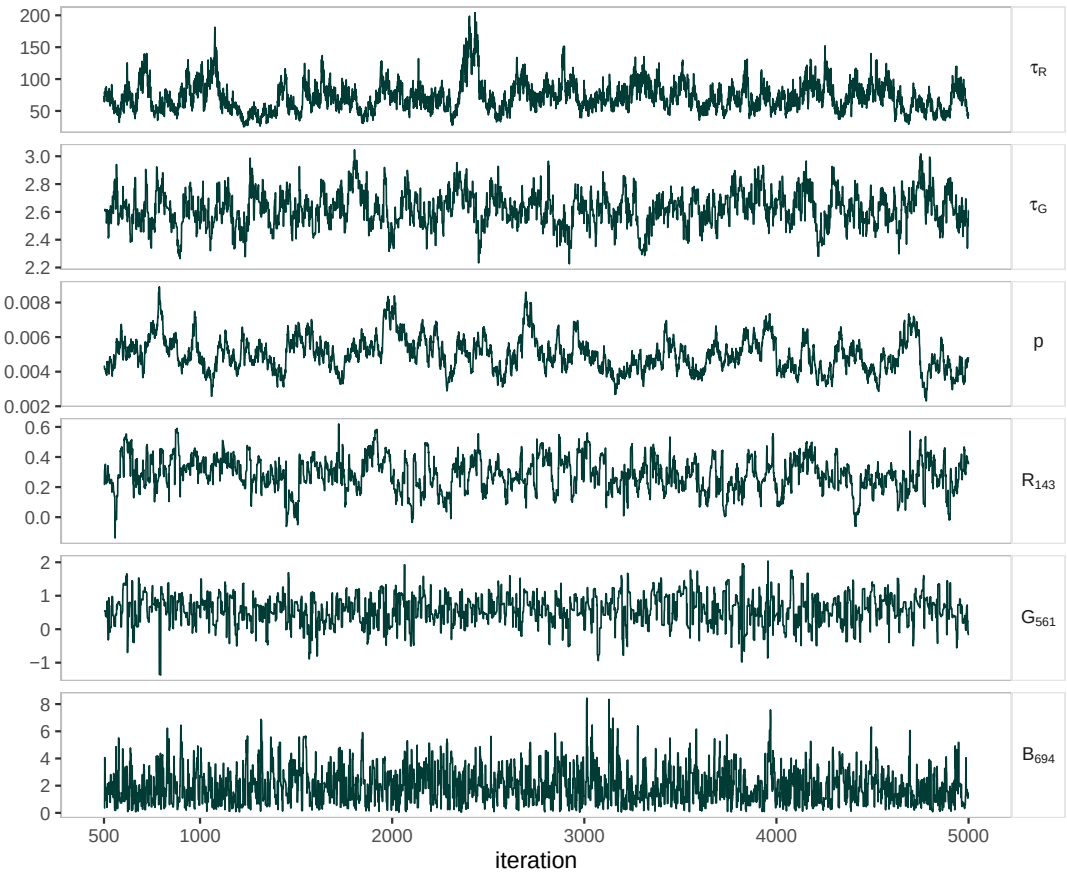


Figure 5.2: A typical set of traces obtained after running the global model applied to OX and NE datasets, including the traces of the hyperparameters τ_G, τ_R, p, and one sample of the genetic, temporal and outbreak size parameters G_k, R_t and B_ξ, respectively.

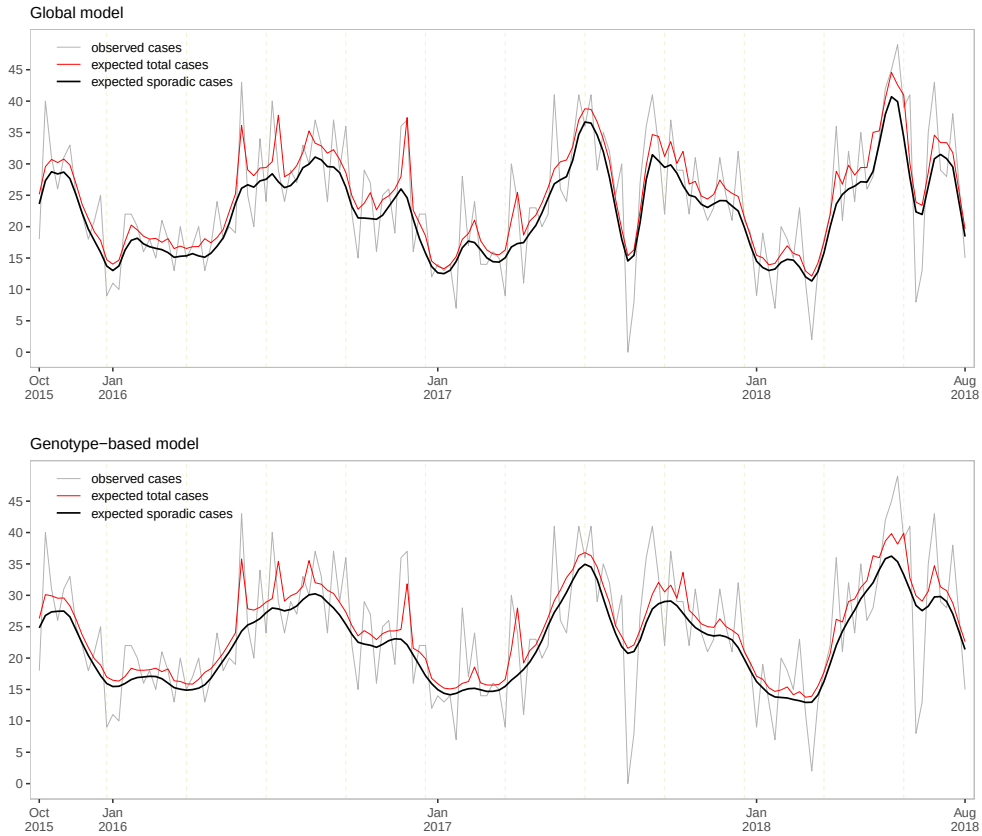


Figure 5.3: Comparison of the observed number of cases with the expected number of sporadic and total cases per week for Scenario i. with the global model (top) and ii. with the genotype-based model (bottom). Yellow vertical lines denote the week when a new season started.

For the genotype-based model, the time-series of cases is shown at the bottom of Figure 5.3. Similarly, time-series of selected genotypes are shown at the bottom of Figure 5.4 for ST-353, and at the bottom of Figure 5.4 for ST-45. Each plot displays the number of observed cases, the expected number of sporadic and total cases per genotype.

5.4.2 Potential outbreaks

Each model provides a list of blocks and the probability that they are part of an outbreak. A block is labelled as *probable outbreak* if it fulfils the following criterium: the probability that it is an outbreak is greater than 90%. Also, if the probability is greater than 50% but less than 90%, it is labelled as a probable outbreak if occurred immediately before or after a block with probability greater than 90%. Figure 5.6 shows a comparison of

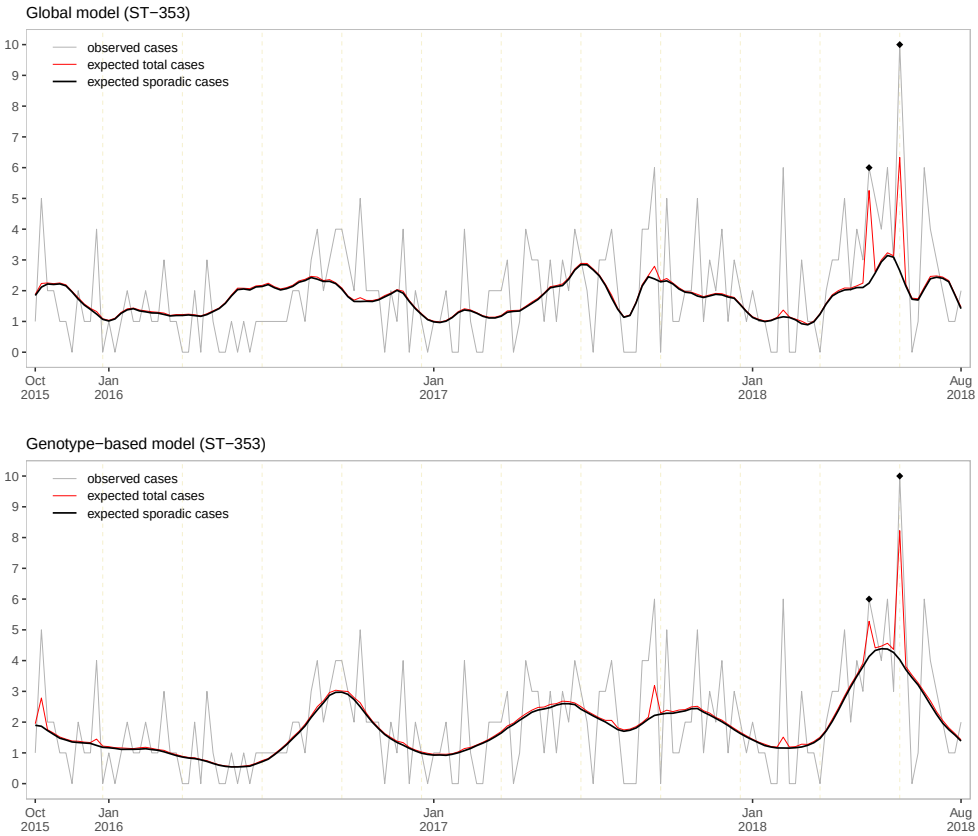


Figure 5.4: Comparison of the observed number of cases with the expected number of sporadic and total cases per week for the genotype ST-353: i. global model (top) and ii. genotype-based model. Yellow vertical lines denote the week when a new season started.

the probabilities provided by both models. The blocks that had the largest difference between both models are marked in red: two correspond to cases belonging to the ST-353 clonal complex, while one corresponds to the ST-45. Black diamonds in Figure 5.4 and Figure 5.5 indicate the time when these cases happened and are marked with a star. Table 5.3 details the most probable outbreaks obtained by the global model, including the number of cases involved, the probability of being an outbreak, the dates when it occurred and the clonal complex of the sequences involved. Similarly, Table 5.4 shows the most probable outbreaks obtained by the genotype-based model, not shown in Table 5.3.

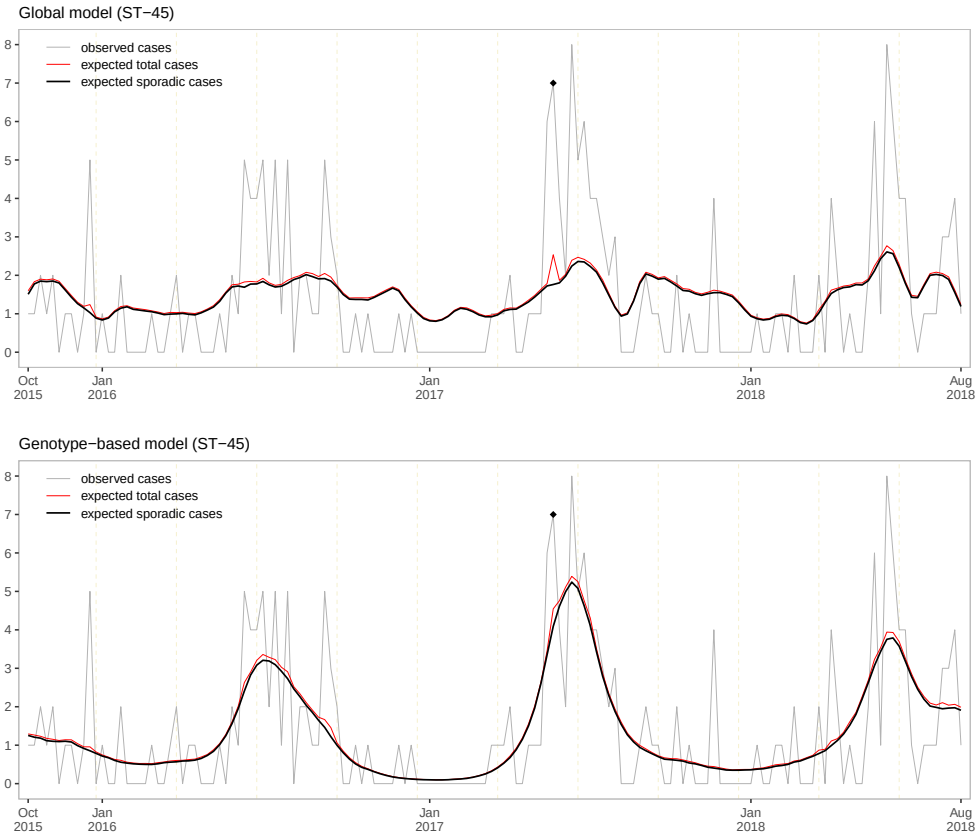


Figure 5.5: Comparison of the observed number of cases with the expected number of sporadic and total cases per week for the genotype ST-45: i. global model (top) and ii. genotype-based model. Yellow vertical lines denote the week when a new season started.

5.5 Discussion

In this chapter, a temporal-genetic model for the detection of outbreaks was presented, following a similar structure as the models in previous chapters. The proposed Bayesian model studied the sporadic trend of cases per time and also analysed the incidences of sporadic cases per genotype. The cases not explained by the sporadic trend are labelled as outbreaks. Finally, the construction of the model was based on MCMC techniques similar to the previous models, using the same mixing strategies. Since the spatial component is not part of the analysis, the model was applied to the combination of the OX and the NE dataset. It allowed the model to study cases that might be part of outbreaks spread in Oxfordshire and Tyne and Wear or subsets of large national outbreaks.

	Number of cases	Probability of outbreak (%)	(Received) date range	Clonal complex	Genetic distance within sequences
I	3	94.0	16-23 Oct 2015	ST-21	0-40
II	5 (NE)	52.3-93.7	8-22 Apr 2016	ST-2274	0-14
III	6	97.0	25 Nov - 2 Dec 2016	ST-257	0-5
IV	6	99.8	31 Mar - 7 Apr 2017	ST-257	0-48
V	3	98.3	8-15 Sep 2017	ST-21	1-7
VI	3 (NE)	100	16-23 Feb 2018	ST-206	2-19
VII	5	91.2	11-18 May 2018	ST-353	1-9
VIII	6	98.0	15-22 Jun 2018	ST-353	0-35

Table 5.3: List of *probable outbreaks*, defined as blocks with probability greater than 90% using the global model. It includes the number of cases (and database of origin), probability of being an outbreak (a range if it consists of more than one block), dates when the isolates were received in the laboratory, clonal complex of the isolate and range of genetic distances within the block.

	Number of cases	Probability of outbreak (%) (genotype)	Probability of outbreak (%) (global)	(Received) date range	Clonal complex	Genetic distance
I	3	91.4	84.7	04 Nov 2016	ST-206	1-18
II	3	90.1	74.9	17 Feb 2017	ST-206	2-25

Table 5.4: List of *probable outbreaks* detected by the genotype-based model and not by the global model, defined as blocks with probability greater than 90% using the genotype-based model not included in Table 5.3. It includes the number of cases (and database of origin), probability of being an outbreak by both models (a range if it consists of more than one block), dates when the isolates were received in the laboratory, clonal complex of the isolate and range of genetic distances within the block.

For the global model, the output estimated the temporal trend of sporadic cases, showing a seasonal pattern during the years covered in the study. In each year, there were two large peaks of reported cases at the beginning and the end of each summer. Conversely, winter peaks were only observed in 2016 and 2017 and not seen in 2018. For the genotype-based model, the temporal pattern had similar characteristics to the global model, as shown in Figure 5.3. Additionally, the genotype model provided the temporal pattern of the clonal complexes ST-21, ST-353 and ST-45. It showed that the ST-45 had distinct peaks in every summer. For the ST-353, it showed peaks in every winter period and also a large summer peak in 2018. The ST-21, the largest clonal complex in the database, showed a similar general trend as the global pattern with a larger peak at the beginning of winter 2017.

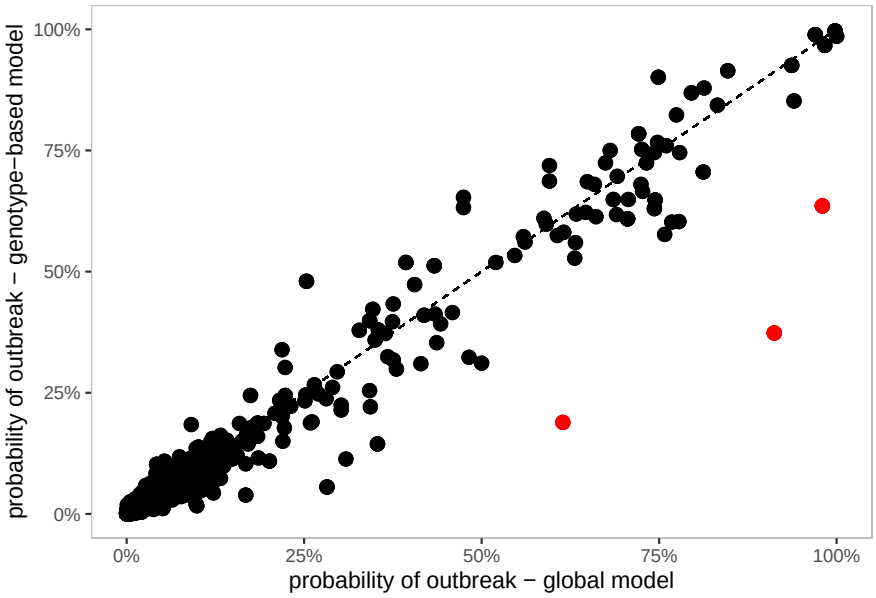


Figure 5.6: Comparison between the probability that a block $\phi(t)\xi(k)$ is an outbreak according to each model. The horizontal axis corresponds to the probabilities for the global model. The vertical axis shows the probabilities for the genotype-based model. The red dots represent the blocks that differed the most when comparing both models.

Both models provided a list of cases and the probability that they were part of an outbreak. In general, the probabilities are consistent between the two models, except for three potential outbreaks which probability was reduced when using the genotype-based model (Figure 5.6). Each of these cases was part of a seasonal peak of the genotypes ST-353 and ST-45, a peak not detected by the global model. For instance, the clonal complex ST-45 showed a summer seasonality not captured by the global model. Therefore, it detected a potential outbreak that could be part of the seasonal trend. These contrasting outputs raise the question of which model is providing the most accurate results.

The main difference between both models relies on the mechanism they use to define the sporadic trend. That is, each model proposed an alternative version of how the sporadic cases are understood. For instance, the global model based the temporal trend on the total number of cases. By contrast, the genotype-based model used seasonality to understand the sporadic pattern. The genotype-based model can be applied when the genotypes used as input are abundant and have a strong seasonal pattern. Otherwise, it would serve as a complication of the model that will not improve the results. That is, the global model should be favoured unless there are abundant genotypes that have strong seasonal patterns.

Based on the global model, the top potential outbreaks were explored in Table 5.3, including cases with probabilities larger than 90%. Most of the potential outbreaks persisted one week, while six of them were found jointly in the datasets OX and NE. Further analysis of these outbreaks is presented in Chapter 6.

Chapter 6

Summary and comparison of previous results

The main goal of this project is to develop mathematical techniques for the detection of outbreaks of *Campylobacter* infections. In previous chapters, three models were introduced to achieve this goal, based on Bayesian models. The structure and assumptions of each model have been described previously, including an output summary. The analysis in this chapter aims to compare the potential outbreaks reported by each model, guiding conclusions about the actual outbreaks that might have happened in the regions studied.

Each model provided a list of cases and their associated probability of being part of an outbreak. If a threshold is chosen, a final set of the potential outbreaks is listed. However, it produces independent lists per model. Here a new criterion is designed to choose potential outbreaks, by combining the output of more than one model. The resulting potential outbreaks are described in detail; in particular, the one with the largest amount of cases. The spatial-temporal model is abbreviated throughout the chapter as ST. Similarly, the spatial-genetic model is denoted as SG and the global temporal-genetic model as TG.

The chapter is organised as follows. First, a brief review of the models' assumptions is listed in Section 6.1, describing the basic configuration used to run the model. Also, the top outbreaks produced by each model is shown. A new criterion to simultaneously analyse all models is described, including the list of the potential outbreaks. Section 6.2 discusses the results obtained in the chapter and provides a final review comprising the similarities and differences between each model results. A summary of the thesis is described in Section 6.3 and potential areas of further research in Section 6.4.

6.1 Outbreak selection

Each model was run independently using the data described in Section 2.1, covering cases reported in the period October 2015-August 2018. The spatial-temporal model (Section 3.1) was run independently for the OX and NE databases, respectively. The spatial-genetic model (Section 4.3) was run independently for the OX and NE databases. The global version of the temporal-genetic model (Section 5.2) was run for the whole dataset since it does not have a spatial distinction. The details of the priors used for each model is in Table 6.1. For all of the models, the region was partitioned in MSOA and time was partitioned into intervals of 1-week length (if applies). The genetic space was partitioned with a cut-off of 10 using a hierarchical clustering method, as described in Section 3.2. The spatial blocks were defined using LSOA. Temporal blocks were set to have one-week length, and the genetic blocks were defined with a cut of 50, as described in Section 3.2. Since the SG model did not capture many outbreaks, we considered alternative parameters for the prior distribution on the spatial terms U_i compared to the ST model.

A list of blocks and their probability of being an outbreak is provided by each model. The posterior distribution of p, the probability that a block is an outbreak, is shown in Figure 6.1. A block is named as *top probable outbreak* if it is part of the blocks with the highest probabilities. To define which blocks are top, a threshold must be defined. Figure 6.2 shows the percentage of cases that will be labelled as an outbreak if a given threshold were chosen. For the ST model and the SG model, a threshold of 90% is chosen. Since the highest probability in the SG model is 29% for the OX database and 49% for the NE database, 90% is not adequate. Therefore, a threshold of 25% is set for this model. The list of all potential outbreaks obtained by these five models is shown in Table 6.2.

Although all potential outbreaks in Table 6.2 had a probability greater than 90%, none of them was chosen by two or more models. That is, the blocks with high probability for one model are marked with low probabilities by the other two models. In Figure 6.3, the ST and TG global model probabilities are contrasted (left) as well as the SG against the TG model (right). In both figures, each dot corresponds to a reported case. This comparison might seem to suggest that the models are not consistent. However, these differences are a consequence of the diverse data and different assumptions handled by each model, as will be discussed in Section 6.2. To study these variations, a new criterion can be developed to label outbreaks using the output of two or more models. Although models are not consistent, a new criterion can be developed to label outbreaks using the

output of two or more models. In particular, four sets of cases are scored at least 50% of probability by the ST and TG models. Note that some potential outbreaks might include isolates that are not part of the outbreak.

Model-Region	Spatial partition	Prior						
		τ_R	τ_U	τ_G	B_*	p	p_{01}	p_{10}
		$\Gamma(\ ,\ )$	$\Gamma(\ ,\ )$	$\Gamma(\ ,\ )$	$\Gamma(\ ,\ )$	$B(\ ,\ )$	$B(\ ,\ )$	$B(\ ,\ )$
ST-OX	MSOA	(5, 0.1)	(1, 0.5)	-	(1,1)	-	(1, 51)	(2,2)
ST-NE	MSOA	(5, 0.1)	(1, 0.5)	-	(1,1)	-	(1, 51)	(2,2)
SG-OX	MSOA	-	(1, 0.1)	(1, 0.01)	(2,1)	(1, 611)	-	-
SG-NE	MSOA	-	(1, 0.1)	(1, 0.01)	(2,1)	(1, 758)	-	-
TG	-	(5, 0.1)	-	(1, 0.01)	(2,1)	(1, 1142)	-	-

Table 6.1: List of each model configuration used to reproduce the results explained in this chapter, including parameter priors.

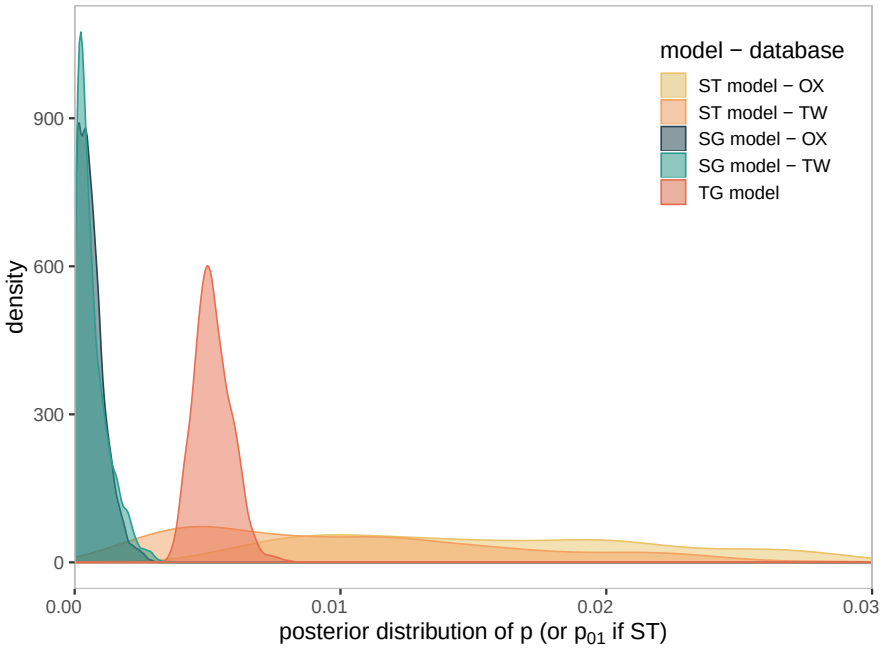


Figure 6.1: For SG and TG model, the figure shows the posterior distribution of the probability that a block is a potential outbreak p. For the ST model, the figure shows the posterior distribution of the probability that an outbreak starts at each block p_{01}.

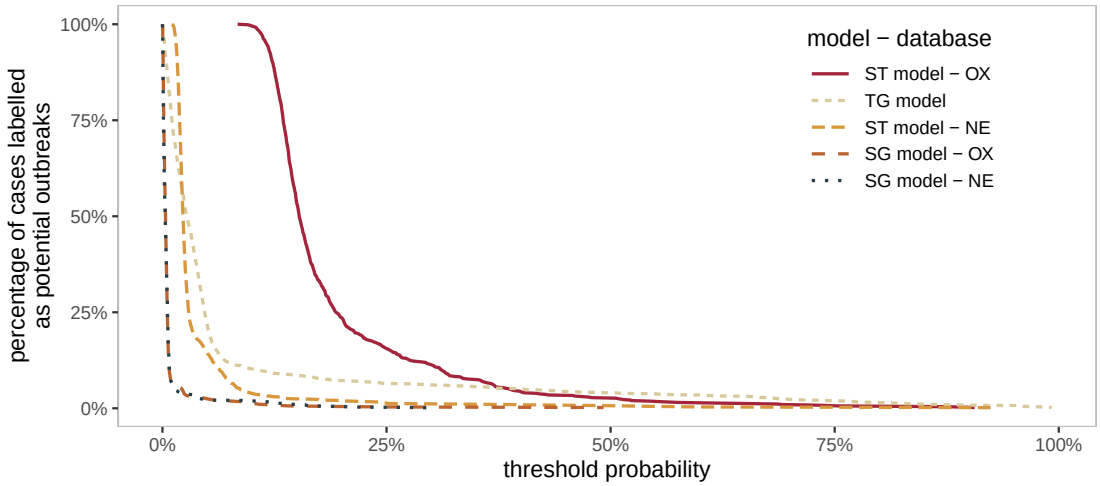


Figure 6.2: Percentage of cases labelled as potential outbreaks, as a function of the threshold chosen to define potential outbreaks. Each curve corresponds to different models and databases.

Model	Number of cases	Database	Probability of outbreak (%)	Clonal complex	Genetic distance within sequences
ST	4	NE	92.4	-	1093-1282
ST	3	OX	90.6	ST-45 (+1)	4 (+1)
SG	3	OX	49.2	ST-21	0
SG	3	NE	29.7	ST-45	7-12
SG	2	NE	27.6	ST-828	0
SG	2	OX	25.1	ST-48	2
TG	3	NE	100	ST-206	2-19
TG	6	OX-NE	99.8	ST-257	0-48
TG	3	OX-NE	98.3	ST-21	1-7
TG	6	OX-NE	98.0	ST-353	0-8 (+1)
TG	6	OX-NE	97.0	ST-257	0-5
TG	3	OX-NE	94.0	ST-21	0 (+1)
TG	3	NE	93.7	ST-2274	0
TG	5	OX-NE	91.2	ST-353	0-9

Table 6.2: List of most probable outbreaks per model, using a threshold of 90% for ST and TG, and 25% for SG. It includes the number of cases, database of origin, probability of being an outbreak, clonal complexes of the isolates and range of genetic distances within the block. (+1) indicates there is an isolate genetically distant from the other isolates in the group. The list is ordered by model and probability (in decreasing order).

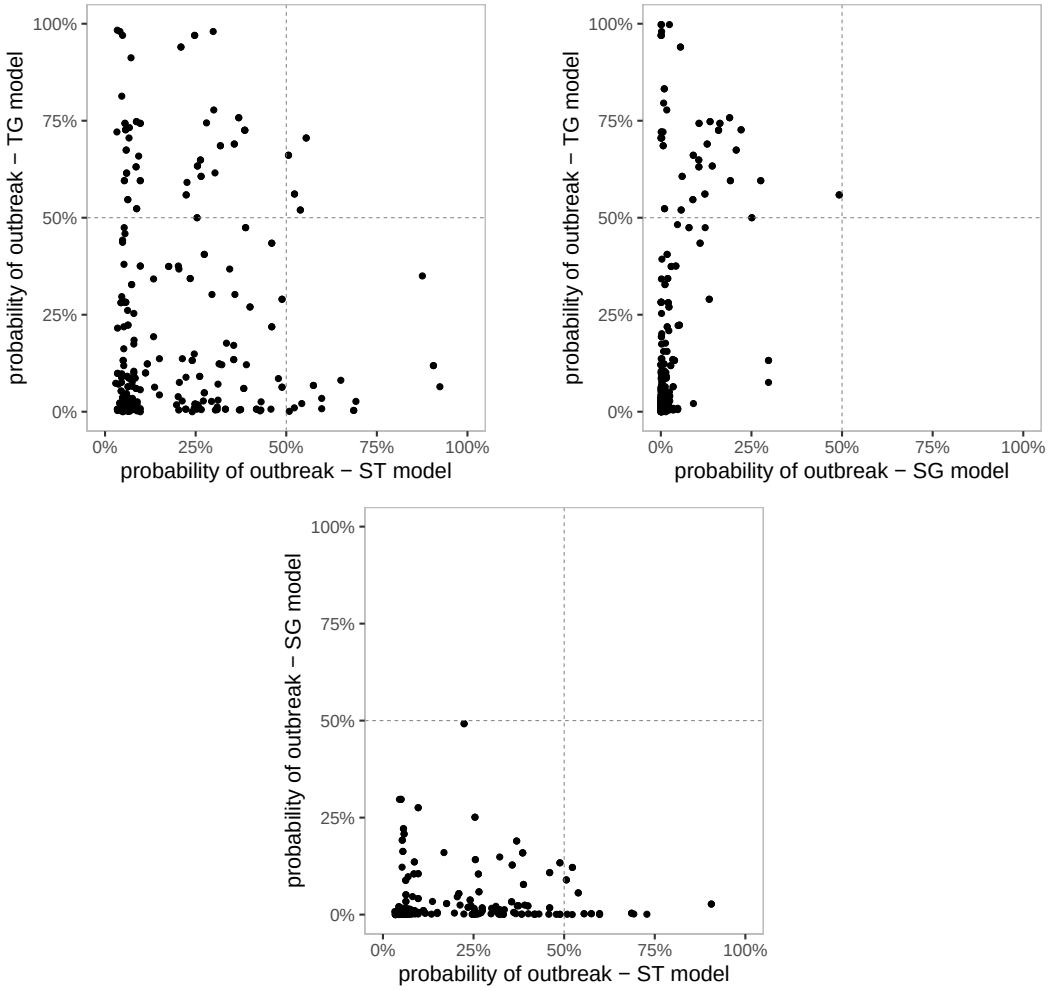


Figure 6.3: Comparison between the outbreak probability of each isolate for the ST model and the TG global model (top-left), comparison between the SG model and the TG model (top-right), and comparison between the ST model and the SG model (bottom).

Based on Figure 6.3, a new criterion for choosing top outbreaks is proposed. Any block with probability greater than 50% in the TG model is labelled as a top outbreak if it contains cases being part of a top ST outbreak; that is, a block with probability greater than 50% in the ST model. This new criterion aims to capture potential outbreaks that are not highly scored by one model but have considerably high scores in at least two models. It might capture outbreaks that are diffuse in space but with spatial-temporal localised sub-outbreaks. For instance, the largest potential outbreak in Table 6.3 consists of 7 cases, 2 of them contained in an ST outbreak. Figure 6.4 shows the location of these cases, where the red dot indicates the location of cases involved in the ST outbreak.

TG global model					ST model	
Number of cases	Database	Probability of outbreak (%)	Clonal complex	Genetic distances	Number of cases	Probability of outbreak (%)
7	OX-NE	70.6	ST-21	0-7	2	55.5
2	OX	66.1	ST-2274	2	2	50.6
2	OX	56.1	ST-42	2	2	52.3
2	NE	52.0	ST-48	27	2	53.9

Table 6.3: Temporal-genetic potential outbreaks with probability greater than 50%, comprising a spatial-temporal outbreak with probability greater than 50 %. It includes the number of cases in the TG-outbreak, the database of origin, probability of being a TG-outbreak, clonal complexes of the isolates and range of genetic distances within the block, number of cases in the ST-outbreak, probability of being an ST-outbreak.

6.2 Discussion

In previous chapters, outbreak detection models have been introduced and implemented using three data sources: temporal, spatial and genetics. In this chapter, potential outbreaks found by each model were compared, aiming to determine if there were outbreaks detected simultaneously by the three models and why there are differences among them. First, the most probable outbreaks were listed and compared, showing that all of them were detected by only one model. Second, a smoother criterion was designed to label potential outbreaks, where a set of cases was chosen if had at least 50% of probability in two models.

Different model outputs are a consequence of the different assumptions in each model design. First, each model examines different data dimensions and aims to capture outbreaks potentially caused by different types of contamination sources. Contamination in a local event is more likely to be captured by a spatial-temporal model since it will be localised in space and time. Conversely, contamination occurred high up in the food chain would be more likely to be detected by a temporal genetic model since the spatial localisation is unlikely. Furthermore, each model has different assumptions about how sporadic cases are explained. Spatial-temporal model, for instance, assumes that the sporadic trend is caused by a temporal trend independently combined with a spatial trend. As a consequence of these different assumptions, the posterior distribution of a case being part of an outbreak behaves differently for each model (Figure 6.1). For instance, under the temporal-genetic case, p is almost surely away from 0, confirming the existence of outbreaks in this scenario. In contrast, under the spatial-temporal case, there is high posterior uncertainty, and therefore potential outbreaks, if any, would have

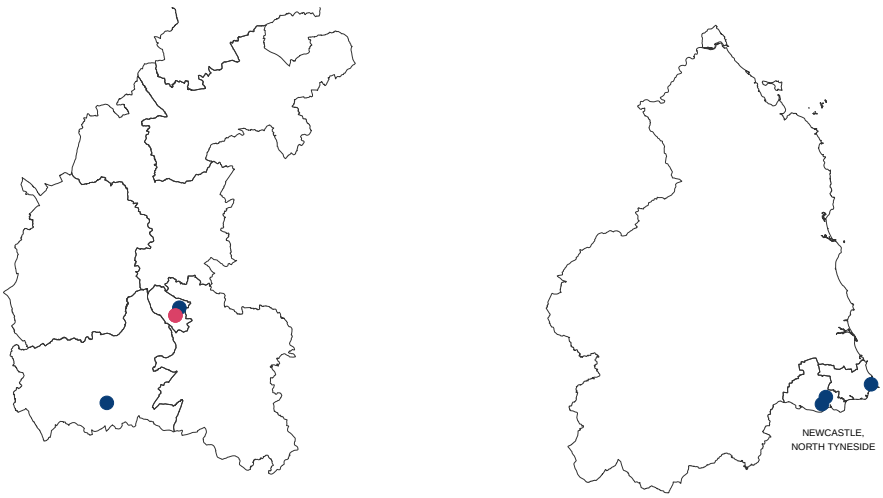


Figure 6.4: Map of the areas covered in the study, showing the location of the cases involved in the largest outbreak in Table 6.3. All cases shown were captured by the TG model. The cases captured by the ST model are shown in red.

larger probabilities (as for the ST model in OX).

Besides the differences between the model assumptions, the number of labelled potential outbreaks also differ. The threshold chosen to mark potential outbreaks was compared to the percentage of cases involved in outbreaks if the threshold was chosen (Figure 6.2). Almost every case in the spatial-genetic model had a probability lower than 25%, showing that potential outbreaks are unlikely under the spatial-genetic assumptions. On the contrary, the temporal-genetic model assigned higher probabilities in most cases. Those differences might rely on how certain types of outbreaks are harder to be seen or are uncommon. For instance, in the spatial-temporal case, each block contains all cases that occurred in a given region and period, including cases that are potentially not part of a real outbreak. It might reduce the specificity of the model and obscures real spatial-temporal outbreaks.

Comparing the probabilities shown by the spatial-genetic model and the temporal-genetic outbreak opens the discussion of the importance of including time into these models. For a fixed genetic type, the time dimension was able to capture unexpected peaks marked as potential outbreaks. On the contrary, understanding the distribution of cases using the spatial dimension did not provide a clear distinction between outbreak cases and sporadic cases.

Although there are many differences between the findings of each model, 94.7% of cases were consistently labelled as non-outbreaks. This result substantially limits the

number of cases that must be investigated to clarify the sources of outbreaks. Moreover, it shows the potential of integrating the output of the three models.

Top outbreaks per model were chosen using thresholds based on the probabilities in Figure 6.2. However, an alternative criterion for labelling potential outbreaks can be based on the results of the three model outputs simultaneously. Figure 6.3 (left) shows a comparison between the probabilities provided by the spatio-temporal and the temporal-genetic model. The new criterion chose cases with probabilities greater than 50% in both models (Table 6.3). This new criterion reveals the existence of potential outbreaks in the spatio-temporal model that might be a subset of a larger outbreak in the temporal genetic model. For instance, there is a potentially diffuse outbreak shown in Figure 6.4. Two of the seven cases involved were located in the same region and were simultaneously detected by the spatial-temporal outbreak.

In conclusion, potential outbreaks occurring in two regions of the UK were analysed, based on three models. Each model was based on diverse assumptions and provided different sets of potential outbreaks. Moreover, the potential of integrating the analysis of the three models may improve the detection of diffuse and other types of outbreaks.

6.3 Summary of the thesis

The main goal of this project is to create mathematical models for *Campylobacter* outbreak detection using a variety of data sources. Mathematical frameworks facing this problem should consider the complexity of the transmission of the disease, the apparent randomness of sporadic cases, and the infection spread dynamics. Therefore, models should identify potential outbreaks out of a set of apparently sporadic cases. This project studied data of reported infections collected in two regions of the UK for three years, covering Oxfordshire, Northhamptonshire, Newcastle upon Tyne, North Tyneside and Northumberland.

Since *Campylobacter* infections are a notifiable disease, data collection is controlled by health authorities. Records of cases provide rich datasets, including information of the patient, residence location, approximate infection time and the whole-genome sequences of a bacteria sampled from the patient. This variety of data sources requires models that handle the structures provided by each data type. Outbreak detection mechanisms have been developed in multiple studies, mainly focussed on temporal and spatial analysis. Moreover, some researches have also incorporated spatial-temporal models. Most of the spatial and temporal approaches treat the problem as point processes or aggregated data depending on the nature of the reports. Bayesian statistics

has been a common framework to face aggregated data problems, like the outbreak detection approach proposed in Spencer et al. [2011], where a Bayesian hierarchical model (BHM) detects potential outbreaks of *Campylobacter* infections using spatial-temporal data. In addition to space and time approaches, the recent whole-genome sequencing techniques have shown the potential of genomics data to discriminate closely related isolates. Comparison of genetic sequences is a potentially powerful tool to understand when two cases are epidemiologically related. Answering the open question in outbreak detection relies on how to mix these varieties of data structures, as the model proposed in this thesis.

In the model studied in this project, outbreaks are defined as unexpected peaks in the number of cases compared to a sporadic trend. The model incorporates whole-genome sequencing data into the spatial-temporal approach in Spencer et al. [2011]. In this spatial-temporal model, sporadic cases are described as an independent combination of a temporal and a spatial trend. First, the spatial trend is described using a Gaussian Markov Random Field (GMRF) while the temporal trend is described using a Random Walk (RW). Genetic sequences can be incorporated in this model if they are embedded in a mathematical structure. A metric space for the genetic sequences is proposed, where the distance between two sequences depends on the similarity of the core genes present in each pair of sequences. This metric space also referred to as the genetic space, provides the notion of neighbourhood required to generalise the GMRF structure. Intuitively, the risk of observing a genetic sequence is similar to the risk of the neighbouring sequences. Therefore, a Gaussian Random Field (GRF) is applied using different kernels to capture the neighbouring structures. That is, the BHM can have an extra latent space linked to the GRF, where the kernel and the distance control the structure of the genetic space and the notion of proximity.

The BHM proposed has the flexibility to incorporate spatial, temporal and genetic data. However, the structure can be adjusted depending on the types of outbreaks investigated. For instance, if only spatial and temporal sporadic surfaces are included in the model, it would search for localised spatial-temporal outbreaks. Therefore, three types of structures are studied further: a spatial-temporal, spatial-genetic and temporal-genetic model. For each of the proposed models, a Monte Carlo Markov Chain (MCMC) is run to estimate the posterior distribution of the parameters involved. However, there are approximately a thousand latent parameters involved in the genetic surface, and they are highly correlated. That causes the MCMC to have a low convergence and therefore requires an update strategy to improve the convergence. In this project, a new MCMC strategy is proposed, where latent parameters in the GRF are updated using blocks, as

a generalisation of the block strategy described in Fahrmeir and Lang [2001].

The models provide the structure of the sporadic trend and the probability that each case is part of a potential outbreak. If the model includes time, the temporal trend of sporadic cases can be examined, including the seasonal patterns of different regions. Similarly, if the model includes spatial data, it provides the risk of observing sporadic cases per region. It allows the comparison of risk for different areas and its comparison to other covariates as the rural-urban classification. Finally, if the model includes genomic data, the genetic trend of sporadic cases provides the risk of observing a sequence as part of the genetic space. This output offers an overview of which genetic types are more common to be observed and quantifies the risk associated with each genetic type.

The output for the spatial-temporal, spatial-genetic and temporal-genetic models can be compared to investigate potential outbreaks. It provides a list of the cases that are potentially involved in outbreaks in each of these models. Therefore, other strategies can be proposed to understand the properties of potential outbreaks. For instance, a national outbreak partially detected by the spatial-temporal model and fully detected by the temporal-genetic model would be hard to observe by local authorities.

In summary, the proposed model gives a flexible approach to study different types of outbreaks. It provides a list of potential outbreaks, the probability associated with each reported case, and the trend of the sporadic cases, including the risk of genetic sequences. It successfully generalises the current spatial-temporal approach, defined in a Bayesian framework that is common at solving outbreak detection. Also, it shows how it can incorporate complex structure data using GRFs. However, the proposed approach has some limitations. First, in an outbreak investigation, several models have to be run with different configurations, such that enough information is obtained. Then, the computational time and effort increases. Also, the time period, region and genetic space should be partitioned to capture different outbreak structures, and the choices might affect the model output. Finally, one-dimensional outbreaks can be harder to detect. For instance, long-lived genetic clusters spread by a farm producing contaminated chickens persistently.

6.4 Further work

The model proposed in this thesis aims to detect outbreaks using spatial-temporal and genetic data. However, this approach has several limitations that can be addressed in future research.

First, the model requires a partition of the studied area, period and genetic space

such that outbreak blocks can be defined. Indicators are linked to each block, specifying if they are potential outbreaks. Therefore, the model is sensitive to the partition of each space, while real outbreaks have unknown shapes. Spencer et al. [2011] provided a correlated version of the model to cover outbreaks of different duration, as described in Section 3.1. Alternative models could be designed, applying the correlated version to spatial and genetic parameters.

Second, some parameters of the model were defined using generic structures. For instance, the kernels used for the GRF were the squared exponential and the Matérn kernel, functions that are originally defined for continuous Euclidean spaces. Alternative kernels have been proposed for discrete spaces with similar properties than the genetic space [Rasmussen and Williams, 2005]. These functions could improve the performance of the model.

Also, the model has been structured to analyse outbreaks retrospectively. However, the Bayesian structure of the model and the properties of the GRF provide flexibility to detect outbreaks on a routine basis. For instance, initially, the posterior distribution of the model parameters can be obtained as described in this project. Then, the probability of every new case can be computed based on the posterior distributions obtained.

Besides these limitations, this thesis has started to explore a new and important area making initial progress that is only going to develop through future developments.

Bibliography

Auchincloss, A. H., Gebreab, S. Y., Mair, C., and Roux, A. V. D. A Review of Spatial Methods in Epidemiology, 2000–2010. *Annual review of public health*, 33:107, April 2012.

Bakker, H. C. d., Strawn, L. K., and Deng, X. Bioinformatics Aspects of Foodborne Pathogen Research. In *Applied Genomics of Foodborne Pathogens*, pages 51–64. Springer, Cham, 2017.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press, June 2004.

Beneš, V., Bodlák, K., Møller, J., and Waagepetersen, R. A Case Study on Point Process Modelling in Disease Mapping. *Image Analysis & Stereology*, 24(3):159–168, September 2005.

Besag, J. and Newell, J. The Detection of Clusters in Rare Diseases. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 154(1):143–155, 1991.

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, March 1991.

Besser, J., Carleton, H. A., Gerner-Smidt, P., Lindsey, R. L., and Trees, E. Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clinical Microbiology and Infection*, 24(4):335–341, April 2018.

Best, N., Richardson, S., and Thomson, A. A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, 14(1):35–59, February 2005.

Bithell, J. F. An application of density estimation to geographical epidemiology. *Statistics in Medicine*, 9(6):691–701, 1990.

Blangiardo, M., Cameletti, M., Baio, G., and Rue, H. Spatial and spatio-temporal models with R-INLA. *Spatial and Spatio-temporal Epidemiology*, 4(Supplement C): 33–49, March 2013.

Blaser, M. J. Epidemiologic and Clinical Features of Campylobacter jejuni Infections. *The Journal of Infectious Diseases*, 176:S103–S105, 1997.

Brix, A. and Diggle, P. J. Spatiotemporal Prediction for Log-Gaussian Cox Processes. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 63(4):823–841, 2001.

Bryant, D. and Moulton, V. Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks. *Molecular Biology and Evolution*, 21(2):255–265, February 2004.

Butcher, H., Elson, R., Chattaway, M. A., Featherstone, C. A., Willis, C., Jorgensen, F., Dallman, T. J., Jenkins, C., McLauchlin, J., Beck, C. R., and Harrison, S. Whole genome sequencing improved case ascertainment in an outbreak of Shiga toxin-producing Escherichia coli O157 associated with raw drinking milk. *Epidemiology and Infection*, 144(13):2812–2823, 2016.

Clayton, D. and Kaldor, J. Empirical Bayes Estimates of Age-Standardized Relative Risks for Use in Disease Mapping. *Biometrics*, 43(3):671–681, 1987.

Cody, A. J., McCarthy, N. M., Wimalarathna, H. L., Colles, F. M., Clark, L., Bowler, I. C. J. W., Maiden, M. C. J., and Dingle, K. E. A Longitudinal 6-Year Study of the Molecular Epidemiology of Clinical Campylobacter Isolates in Oxfordshire, United Kingdom. *Journal of Clinical Microbiology*, 50(10):3193–3201, October 2012.

Cody, A. J., McCarthy, N. D., Jansen van Rensburg, M., Isinkaye, T., Bentley, S. D., Parkhill, J., Dingle, K. E., Bowler, I. C. J. W., Jolley, K. A., and Maiden, M. C. J. Real-Time Genomic Epidemiological Evaluation of Human Campylobacter Isolates by Use of Whole-Genome Multilocus Sequence Typing. *Journal of Clinical Microbiology*, 51(8):2526–2534, August 2013.

Colles, F. M., McCarthy, N. D., Sheppard, S. K., Layton, R., and Maiden, M. C. Comparison of Campylobacter populations isolated from a free-range broiler flock before and after slaughter. *International journal of food microbiology*, 137(0):259–264, February 2010.

Costa, M. A. and Kulldorff, M. Maximum linkage space-time permutation scan statistics for disease outbreak detection. *International Journal of Health Geographics*, 13(1):1–14, December 2014.

Cox, D. R. Some Statistical Methods Connected with Series of Events. *Journal of the Royal Statistical Society. Series B (Methodological)*, 17(2):129–164, 1955.

Daley, D. J. and Vere-Jones, D. *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. Probability and Its Applications, An Introduction to the Theory of Point Processes. Springer-Verlag, New York, 2 edition, 2003.

Daley, D. J. and Vere-Jones, D. *An introduction to the theory of point processes*. Springer, New York, 2nd ed edition, 2008.

Denison, D. G. T. and Holmes, C. C. Bayesian Partitioning for Estimating Disease Risk. *Biometrics*, 57(1):143–149, 2001.

Descombes, X. Marked Point Processes for Object Detection. In Descombes, X., editor, *Stochastic Geometry for Image Analysis*, pages 11–27. John Wiley & Sons, Inc., 2013.

Desjardins, M. R., Hohl, A., and Delmelle, E. M. Rapid surveillance of COVID-19 in the United States using a prospective space-time scan statistic: Detecting and evaluating emerging clusters. *Applied Geography*, 118:102202, May 2020.

Diggle, P. J. and Chetwynd, A. G. Second-order analysis of spatial clustering for inhomogeneous populations. *Biometrics*, 47(3):1155–1163, September 1991.

Diggle, P. Spatio-temporal Point Processes: Methods and Applications. *Johns Hopkins University, Dept. of Biostatistics Working Papers*, June 2005.

Diggle, P., Zheng, P., and Durr, P. Nonparametric estimation of spatial segregation in a multivariate point process: bovine tuberculosis in Cornwall, UK. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):645–658, April 2005.

Diggle, P. J. *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns, Third Edition*. CRC Press, July 2013.

Diggle, P. J., Moraga, P., Rowlingson, B., and Taylor, B. M. Spatial and Spatio-Temporal Log-Gaussian Cox Processes: Extending the Geostatistical Paradigm. *Statistical Science*, 28(4):542–563, 2013.

Dingle, K. E., Colles, F. M., Wareing, D. R. A., Ure, R., Fox, A. J., Bolton, F. E., Bootsma, H. J., Willems, R. J. L., Urwin, R., and Maiden, M. C. J. Multilocus Sequence Typing System for Campylobacter jejuni. *Journal of Clinical Microbiology*, 39(1):14–23, January 2001.

Dingle, K. E., Colles, F. M., Ure, R., Wagenaar, J. A., Duim, B., Bolton, F. J., Fox, A. J., Wareing, D. R., and Maiden, M. C. Molecular Characterization of Campylobacter jejuni Clones: A Basis for Epidemiologic Investigation. *Emerging Infectious Diseases*, 8(9):949–955, September 2002.

Dingle, K. E., McCarthy, N. D., Cody, A. J., Peto, T. E., and Maiden, M. C. J. Extended Sequence Typing of Campylobacter spp., United Kingdom. *Emerging Infectious Diseases*, 14(10):1620–1622, October 2008.

Duczmal, L. and Assunção, R. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics & Data Analysis*, 45(2): 269–286, March 2004.

Fahrmeir, L. and Lang, S. Bayesian inference for generalized additive mixed models based on Markov random field priors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(2):201–220, January 2001.

Farrington, C. P., Andrews, N. J., Beale, A. D., and Catchpole, M. A. A Statistical Algorithm for the Early Detection of Outbreaks of Infectious Disease. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 159(3):547–563, May 1996.

Fernandes, A. M., Balasegaram, S., Willis, C., Wimalarathna, H. M. L., Maiden, M. C., and McCarthy, N. D. Partial Failure of Milk Pasteurization as a Risk for the Transmission of Campylobacter From Cattle to Humans. *Clinical Infectious Diseases*, 61 (6):903–909, September 2015.

Fittipaldi, N., Tyrrell, G. J., Low, D. E., Martin, I., Lin, D., Hari, K. L., and Musser, J. M. Integrated whole-genome sequencing and temporospatial analysis of a continuing Group A Streptococcus epidemic. *Emerging Microbes & Infections*, 2(3):e13, March 2013.

Frost, J. A., Gillespie, I. A., and O'Brien, S. J. Public health implications of campylobacter outbreaks in England and Wales, 1995-9: epidemiological and microbiological investigations. *Epidemiology and Infection*, 128(2):111–118, April 2002.

Gangnon, R. E. and Clayton, M. K. Bayesian Detection and Modeling of Spatial Disease Clustering. *Biometrics*, 56(3):922–935, September 2000.

Garthwaite, P. H., Fan, Y., and Sisson, S. A. Adaptive optimal scaling of Metropolis–Hastings algorithms using the Robbins–Monro process. *Communications in Statistics - Theory and Methods*, 45(17):5098–5111, September 2016.

Gatrell, A. C., Bailey, T. C., Diggle, P. J., and Rowlingson, B. S. Spatial Point Pattern Analysis and Its Application in Geographical Epidemiology. *Transactions of the Institute of British Geographers*, 21(1):256–274, 1996.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. *Bayesian Data Analysis*. Chapman and Hall/CRC, Boca Raton, 3 edition edition, November 2013.

Gerner-Smidt, P., Carleton, H., and Trees, E. Role of Whole Genome Sequencing in the Public Health Surveillance of Foodborne Pathogens. In *Applied Genomics of Foodborne Pathogens*, pages 1–11. Springer, Cham, 2017.

Gillespie, I. A., O'Brien, S. J., Frost, J. A., Adak, G. K., Horby, P., Swan, A. V., Painter, M. J., and Neal, K. R. A Case-Case Comparison of Campylobacter coli and Campylobacter jejuni Infection: A Tool for Generating Hypotheses. *Emerging Infectious Diseases*, 8(9):937–942, September 2002.

Gillespie, I., O'Brien, S., Penman, C., Tompkins, D., Cowden, J., and Humphrey, T. Demographic determinants for Campylobacter infection in England and Wales: implications for future epidemiological studies. *Epidemiology and Infection*, 136(12): 1717–1725, December 2008.

Gormley, F. J., Little, C. L., Rawal, N., Gillespie, I. A., Lebaigue, S., and Adak, G. K. A 17-year review of foodborne outbreaks: describing the continuing decline in England and Wales (1992–2008), May 2011.

Green, P. J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, December 1995.

Green, P. J. and Richardson, S. Hidden Markov Models and Disease Mapping. *Journal of the American Statistical Association*, 97(460):1055–1070, 2002.

Gundogdu, O., Bentley, S. D., Holden, M. T., Parkhill, J., Dorrell, N., and Wren, B. W. Re-annotation and re-analysis of the Campylobacter jejuni NCTC11168 genome sequence. *BMC Genomics*, 8:162, June 2007.

Guzmán-Rincón, L. Bayesian Hierarchical Model for Outbreak Detection, February 2021. URL `https://doi.org/10.5281/zenodo.4537911`.

Han, D., Rogerson, P. A., Bonner, M. R., Nie, J., Vena, J. E., Muti, P., Trevisan, M., and Freudenheim, J. L. Assessing spatio-temporal variability of risk surfaces using residential history data in a case control study of breast cancer. *International Journal of Health Geographics*, 4:9, April 2005.

Hastie, T., Tibshirani, R., and Friedman, J. Unsupervised Learning. In Hastie, T., Tibshirani, R., and Friedman, J., editors, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Series in Statistics, pages 485–585. Springer, New York, NY, 2009.

Hutwagner, L., Thompson, W., Seeman, G. M., and Treadwell, T. The bioterrorism preparedness and response Early Aberration Reporting System (EARS). *Journal of Urban Health : Bulletin of the New York Academy of Medicine*, 80(Suppl 1):i89–i96, March 2003.

Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. The Homogeneous Poisson Point Process. In *Statistical Analysis and Modelling of Spatial Point Patterns*, pages 57–98. John Wiley & Sons, Ltd, 2007a.

Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. Introduction. In *Statistical Analysis and Modelling of Spatial Point Patterns*, pages 1–56. John Wiley & Sons, Ltd, 2007b.

Illian, J., Penttinen, A., Stoyan, H., and Stoyan, D. Stationary Marked Point Processes. In *Statistical Analysis and Modelling of Spatial Point Patterns*, pages 293–361. John Wiley & Sons, Ltd, 2007c.

Illian, J. B. and Burslem, D. F. R. P. Improving the usability of spatial point process methodology: an interdisciplinary dialogue between statistics and ecology. *AStA Advances in Statistical Analysis*, 101(4):495–520, October 2017.

Jolley, K. A. and Maiden, M. C. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics*, 11:595, December 2010.

Jolley, K. A., Bliss, C. M., Bennett, J. S., Bratcher, H. B., Brehony, C., Colles, F. M., Wimalarathna, H., Harrison, O. B., Sheppard, S. K., Cody, A. J., and Maiden, M. C. J. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology*, 158(Pt 4):1005–1015, April 2012.

Kaakoush, N. O., Castaño-Rodríguez, N., Mitchell, H. M., and Man, S. M. Global Epidemiology of Campylobacter Infection. *Clinical Microbiology Reviews*, 28(3):687–720, January 2015.

Kelsall, J. E. and Diggle, P. J. Spatial variation in risk of disease: a nonparametric binary regression approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(4):559–573, April 1998.

Kelsall, J. E. and Diggle, P. J. Kernel Estimation of Relative Risk. *Bernoulli*, 1(1/2): 3–16, 1995a.

Kelsall, J. E. and Diggle, P. J. Non-parametric estimation of spatial variation in relative risk. *Statistics in Medicine*, 14(21-22):2335–2342, November 1995b.

Kleinman, K. P., Abrams, A. M., Kulldorff, M., and Platt, R. A model-adjusted space-time scan statistic with an application to syndromic surveillance. *Epidemiology and Infection*, 133(3):409–419, June 2005.

Knorr-Held, L. and Raßer, G. Bayesian Detection of Clusters and Discontinuities in Disease Maps. *Biometrics*, 56(1):13–21, March 2000.

Knorr-Held, L. and Richardson, S. A hierarchical model for space–time surveillance data on meningococcal disease incidence. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52(2):169–183, May 2003.

Knorr-Held, L. and Rue, H. On Block Updating in Markov Random Field Models for Disease Mapping. *Scandinavian Journal of Statistics*, 29(4):597–614, December 2002.

Knorr-Held, L. Conditional Prior Proposals in Dynamic Models. *Scandinavian Journal of Statistics*, 26(1):129–144, March 1999.

Kovanen, S., Kivistö, R., Llarena, A.-K., Zhang, J., Kärkkäinen, U.-M., Tuuminen, T., Uksila, J., Hakkinen, M., Rossi, M., and Hänninen, M.-L. Tracing isolates from domestic human Campylobacter jejuni infections to chicken slaughter batches and swimming water using whole-genome multilocus sequence typing. *International Journal of Food Microbiology*, 226:53–60, June 2016.

Kovanen, S. M., Kivistö, R. I., Rossi, M., Schott, T., Kärkkäinen, U.-M., Tuuminen, T., Uksila, J., Rautelin, H., and Hänninen, M.-L. Multilocus Sequence Typing (MLST) and Whole-Genome MLST of Campylobacter jejuni Isolates from Human Infections in Three Districts during a Seasonal Peak in Finland. *Journal of Clinical Microbiology*, 52(12):4147–4154, January 2014.

Kulldorff, M. A spatial scan statistic. *Communications in Statistics - Theory and Methods*, 26(6):1481–1496, January 1997.

Kulldorff, M. Prospective time periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 164 (1):61–72, January 2001.

Kulldorff, M., Heffernan, R., Hartman, J., Assunção, R., and Mostashari, F. A Space–Time Permutation Scan Statistic for Disease Outbreak Detection. *PLoS Medicine*, 2(3), March 2005.

Kulldorff, M., Huang, L., Pickle, L., and Duczmal, L. An elliptic spatial scan statistic. *Statistics in Medicine*, 25(22):3929–3943, November 2006.

Lawson, A. B. and Williams, F. L. R. Applications of extraction mapping in environmental epidemiology. *Statistics in Medicine*, 12(13):1249–1258, January 1993.

Lawson, A. B. Spatial and Spatio-Temporal Disease Analysis. In *Spatial and Syndromic Surveillance for Public Health*, pages 53–76. John Wiley & Sons, Ltd, 2005.

Lawson, A. B. Disease Cluster Detection. In *Bayesian Disease Mapping*, Chapman & Hall/CRC Interdisciplinary Statistics Series, pages 119–150. Chapman and Hall/CRC, August 2008.

Lawson, A. B. *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology, Second Edition*. CRC Press, March 2013.

Lawson, A. B. and Lawson, A. B. Small Scale: Disease Clustering. In *Statistical Methods in Spatial Epidemiology*, pages 109–141. John Wiley & Sons, Ltd., 2006.

Little, C. L., Gormley, F. J., Rawal, N., and Richardson, J. F. A recipe for disaster: outbreaks of campylobacteriosis associated with poultry liver pâté in England and Wales. *Epidemiology & Infection*, 138(12):1691–1694, August 2010.

Llarena, A.-K., Taboada, E., and Rossi, M. Whole-Genome Sequencing in Epidemiology of Campylobacter jejuni Infections. *Journal of Clinical Microbiology*, 55(5):1269–1275, January 2017.

Loman, N. J., Constantinidou, C., Chan, J. Z. M., Halachev, M., Sergeant, M., Penn, C. W., Robinson, E. R., and Pallen, M. J. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. *Nature Reviews Microbiology*, 10(9):599, September 2012.

Louis, V. R., Gillespie, I. A., O'Brien, S. J., Russek-Cohen, E., Pearson, A. D., and Colwell, R. R. Temperature-Driven Campylobacter Seasonality in England and Wales. *Applied and Environmental Microbiology*, 71(1):85–92, January 2005.

Maiden, M. C. J., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D. A., Feavers, I. M., Achtman, M., and Spratt, B. G. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences*, 95(6):3140–3145, March 1998.

Maiden, M. C. J., Jansen van Rensburg, M. J., Bray, J. E., Earle, S. G., Ford, S. A., Jolley, K. A., and McCarthy, N. D. MLST revisited: the gene-by-gene approach to bacterial genomics. *Nature Reviews. Microbiology*, 11(10):728–736, 2013.

McCarthy, N. D., Gillespie, I. A., Lawson, A. J., Richardson, J., Neal, K. R., Hawtin, P. R., Maiden, M. C. J., and O'Brien, S. J. Molecular epidemiology of human Campylobacter jejuni shows association between seasonal and international patterns of disease. *Epidemiology and Infection*, 140(12):2247–2255, December 2012.

McCarthy, N. Campylobacter. In *Applied Genomics of Foodborne Pathogens*, pages 127–143. Springer, Cham, 2017.

Miller, W. G., On, S. L. W., Wang, G., Fontanoz, S., Lastovica, A. J., and Mandrell, R. E. Extended Multilocus Sequence Typing System for Campylobacter coli, C. lari, C. upsaliensis, and C. helveticus. *Journal of Clinical Microbiology*, 43(5):2315–2329, May 2005.

Moller, J., Waagepetersen, R. P., and Waagepetersen, R. P. *Statistical Inference and Simulation for Spatial Point Processes*. Chapman and Hall/CRC, September 2003.

Morbey, R. A., Elliot, A. J., Charlett, A., Verlander, N. Q., Andrews, N., and Smith, G. E. The application of a novel 'rising activity, multi-level mixed effects, indicator

emphasis' (RAMMIE) method for syndromic surveillance in England. *Bioinformatics*, 31(22):3660–3665, November 2015.

Møller, J. and Waagepetersen, R. P. Modern Statistics for Spatial Point Processes. *Scandinavian Journal of Statistics*, 34(4):643–684, September 2007.

Møller, J., Syversveen, A. R., and Waagepetersen, R. P. Log Gaussian Cox Processes. *Scandinavian Journal of Statistics*, 25(3):451–482, September 1998.

Naus, J. I. The Distribution of the Size of the Maximum Cluster of Points on a Line. *Journal of the American Statistical Association*, 60(310):532–538, 1965.

Nichols, G. L., Richardson, J. F., Sheppard, S. K., Lane, C., and Sarran, C. Campylobacter epidemiology: a descriptive study reviewing 1 million cases in England and Wales between 1989 and 2011. *BMJ Open*, 2(4), July 2012.

Noufaily, A., Enki, D. G., Farrington, P., Garthwaite, P., Andrews, N., and Charlett, A. An improved algorithm for outbreak detection in multiple surveillance systems. *Statistics in Medicine*, 32(7):1206–1222, March 2013.

Noufaily, A., Morbey, R. A., Colón-González, F. J., Elliot, A. J., Smith, G. E., Lake, I. R., and McCarthy, N. Comparison of statistical algorithms for daily syndromic surveillance aberration detection. *Bioinformatics*, 35(17):3110–3118, September 2019.

Nylen, G., Dunstan, F., Palmer, S. R., Andersson, Y., Bager, F., Cowden, J., Feierl, G., Galloway, Y., Kapperud, G., Megraud, F., Molbak, K., Petersen, L. R., and Ruutu, P. The seasonal distribution of campylobacter infection in nine European countries and New Zealand. *Epidemiology and Infection*, 128(3):383–390, June 2002.

Openshaw, S., Charlton, M., Wymer, C., and Craft, A. A Mark 1 Geographical Analysis Machine for the automated analysis of point data sets. *International Journal of Geographical Information Systems*, 1(4):335–358, January 1987.

Parkhill, J., Wren, B. W., Mungall, K., Ketley, J. M., Churcher, C., Basham, D., Chillingworth, T., Davies, R. M., Feltwell, T., Holroyd, S., Jagels, K., Karlyshev, A. V., Moule, S., Pallen, M. J., Penn, C. W., Quail, M. A., Rajandream, M.-A., Rutherford, K. M., Vliet, A. H. M. v., Whitehead, S., and Barrell, B. G. The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature*, 403(6770):665, February 2000.

Patil, G. P. and Taillie, C. Upper level set scan statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics*, 11(2):183–197, June 2004.

Pebody, R. G., Ryan, M. J., and Wall, P. G. Outbreaks of campylobacter infection: rare events for a common pathogen. *Communicable disease report. CDR review*, 7(3): R33–7, March 1997.

PHE. Campylobacter data 2008 to 2017, 2017.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

Revez, J., Zhang, J., Schott, T., Kivistö, R., Rossi, M., and Hänninen, M.-L. Genomic Variation between Campylobacter jejuni Isolates Associated with Milk-Borne-Disease Outbreaks. *Journal of Clinical Microbiology*, 52(8):2782–2786, August 2014.

Richardson, E. J. and Watson, M. The automatic annotation of bacterial genomes. *Briefings in Bioinformatics*, 14(1):1–12, January 2013.

Richardson, S., Thomson, A., Best, N., and Elliott, P. Interpreting Posterior Relative Risk Estimates in Disease-Mapping Studies. *Environmental Health Perspectives*, 112 (9):1016–1025, 2004.

Ripley, B. D. The Second-Order Analysis of Stationary Point Processes. *Journal of Applied Probability*, 13(2):255–266, 1976.

Robinson, E. R., Walker, T. M., and Pallen, M. J. Genomics and outbreak investigation: from sequence to consequence. *Genome Medicine*, 5:36, 2013.

Roetzer, A., Diel, R., Kohl, T. A., Rückert, C., Nübel, U., Blom, J., Wirth, T., Jaenicke, S., Schuback, S., Rüsch-Gerdes, S., Supply, P., Kalinowski, J., and Niemann, S. Whole Genome Sequencing versus Traditional Genotyping for Investigation of a Mycobacterium tuberculosis Outbreak: A Longitudinal Molecular Epidemiological Study. *PLOS Medicine*, 10(2):e1001387, February 2013.

Rogerson, P. A. and Yamada, I. Approaches to syndromic surveillance when data consist of small regional counts. *MMWR supplements*, 53:79–85, September 2004.

Rue, H. Fast Sampling of Gaussian Markov Random Fields. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 63(2):325–338, 2001.

Rue, H. and Martino, S. Approximate Bayesian inference for hierarchical Gaussian Markov random field models. *Journal of Statistical Planning and Inference*, 137(10): 3177–3192, October 2007.

Shephard, N. Partial Non-Gaussian State Space. *Biometrika*, 81(1):115–131, 1994.

Shewhart, W. A. Economic Quality Control of Manufactured Product1. *Bell System Technical Journal*, 9(2):364–389, 1930.

Silva, J., Leite, D., Fernandes, M., Mena, C., Gibbs, P. A., and Teixeira, P. Campylobacter spp. as a Foodborne Pathogen: A Review. *Frontiers in Microbiology*, 2, September 2011.

Sonesson, C. A CUSUM framework for detection of space-time disease clusters using scan statistics. *Statistics in Medicine*, 26(26):4770–4789, November 2007.

Spencer, S. E. F., Marshall, J., Pirie, R., Campbell, D., and French, N. P. The detection of spatially localised outbreaks in campylobacteriosis notification data. *Spatial and Spatio-temporal Epidemiology*, 2(3):173–183, September 2011.

Stein, M. L. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Science & Business Media, June 1999.

Takahashi, K., Kulldorff, M., Tango, T., and Yih, K. A flexibly shaped space-time scan statistic for disease outbreak detection and monitoring. *International Journal of Health Geographics*, 7:14, April 2008.

Tam, C. C. and O'Brien, S. J. Economic Cost of Campylobacter, Norovirus and Rotavirus Disease in the United Kingdom. *PLoS ONE*, 11(2), February 2016.

Tam, C. C., Rodrigues, L. C., Viviani, L., Dodds, J. P., Evans, M. R., Hunter, P. R., Gray, J. J., Letley, L. H., Rait, G., Tompkins, D. S., and O'Brien, S. J. Longitudinal study of infectious intestinal disease in the UK (IID2 study): incidence in the community and presenting to general practice. *Gut*, 61(1):69–77, January 2012.

Tango, T. and Takahashi, K. A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics*, 4(1):11, May 2005.

Tauxe, R. V., Doyle, M. P., Kuchenmüller, T., Schlundt, J., and Stein, C. E. Evolving public health approaches to the global challenge of foodborne infections. *International Journal of Food Microbiology*, 139:S16–S28, May 2010.

Unkel, S., Farrington, P., Garthwaite, P., Robertson, C., and Andrews, N. Statistical methods for the prospective detection of infectious disease outbreaks: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 175(1):49–82, January 2012.

Wassenaar, T. M. and Newell, D. G. Genotyping of Campylobacter spp. *Applied and Environmental Microbiology*, 66(1):1–9, January 2000.

Yao, Z., Tang, J., and Zhan, F. Detection of arbitrarily-shaped clusters using a neighbor-expanding approach: A case study on murine typhus in South Texas. *International Journal of Health Geographics*, 10(1):23, March 2011.